

应用概率统计基础及算法

The Elements of Applied Probability Theory and Mathematical
Statistics, and Some Related Algorithms

于江生

前 言

概率论 (probability theory) 源于十七世纪几位大数学家对赌博的研究，统计实践则可追溯到几千年以前的人口普查^{*}。时至今日，概率论已经发展成为公理化了的纯粹数学分支，用于探索随机现象的数量规律，大大提高了人类的思考能力 [133]。而数理统计学 (mathematical statistics)，亦称“统计学”，则是在概率论基础上发展起来的一门应用数学的学问。在自然科学、工程学、社会学、人文学、军事学等诸多应用领域，凡是涉及数据的收集、处理、分析、可视化和解释等方面的问题，都是统计学大显身手的舞台。由此可见概率统计的重要性，它已成为理工学科高等教育中的必修课程，也是很多研究领域的理论基础和应用工具。

随着计算机科学的发展，概率统计的实用价值也越来越得以凸显 [43]。例如，在信息科学领域，出现了一些与数据处理和分析有关的新学科，如模式识别 (Pattern Recognition)、机器学习 (Machine Learning) [16]、贝叶斯数据分析 (Bayesian Data Analysis) [55]、数据挖掘 (Data Mining) [69]、大数据分析 (Big Data Analysis)、模式理论 (Pattern Theory) [65, 112] 等，它们都与概率统计有着千丝万缕的联系，归为数据科学 (Data Science) 一类。

既然概率统计有这么广泛的应用背景，学会利用概率统计的方法来做数学建模，进而设计合理的算法加以实现就变得尤为重要。本着学以致用的想法，作者强调计算机科学与概率统计的紧密结合，为此推荐三类软件来完成概率统计实践，旨在以一种直观的方式提供概率统计最基本的一些实用方法和技巧，并向读者展示从建模到算法实现的过程。

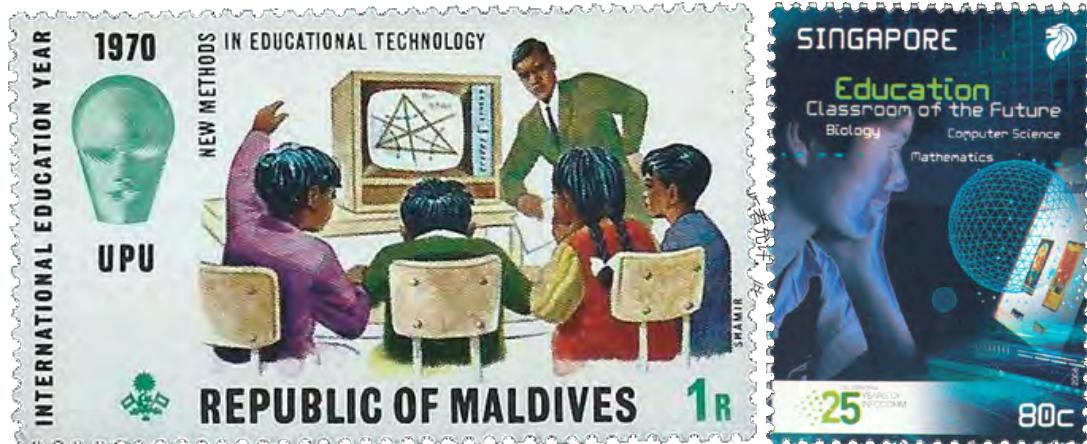
- 符号计算：计算机代数系统 Maxima、Maple、Mathematica 等。
- 数值和统计计算：Fortran、C/C++、R/S-Plus [154]、Octave/MatLab、Python 等，其中 Python 也擅长作符号计算。
- 科学绘图：GnuPlot、MathGL 等。

每个时代都有流行的语言，那些专注于某个语言的数学书籍，若干年后，总会因为过时的语言给读者带来无尽的痛苦。另外，具体的应用决定合适的语言，例如像 R、MatLab、Python 等解释性语言适合对小规模的原型 (prototype) 进行快速的

^{*}公元前二千年，我国的夏朝就出现了为统计人口而设立的国家部门“筹司”。

实验，像 Fortran、C/C++ 等编译性语言适合高性能计算，而集群计算多采用 Java、Scala 等。所以，作者一方面强调概率统计的计算机实践，一方面又极力避免陷入某些具体的语言。无疑，这样既留给读者更多的自由空间，又无损作者的初衷。

为什么学习概率统计需要计算机的辅助或实践？一方面因为计算机科学是成功地运用了数学的典范，理论计算机科学的核心——算法理论离不开概率统计。计算机实践突出了能用于计算机科学的那部分概率统计知识，也就是被计算机科学大师 Donald Ervin Knuth (1938-) 称之为“具体数学”的东西 [63]。另一方面，计算机辅助使得抽象的数学概念和理论变得比较容易理解，甚至有助于更加深入的研究。此外，计算机实践还能从“纸上谈兵”的数学模型切实做出结果，使理论在充分显示其强大威力的同时展现出它极富趣味的一面。



数学家 Richard Courant (1888-1972) 说过，“不顾及应用和直观，将导致数学的孤立和衰退。”在信息时代，概率统计与计算机实践难分难舍，谁都无法忽略计算机的作用，它是帮助人们走向应用和直观的工具，本书正想阐明这一观点。

更宽泛地说，计算机影响了数学研究的方式，为数学增添了些“实验科学”的色彩。除了有助于实现合情推理 (plausible reasoning)*，计算机之于数学家，如同射电望远镜之于天文学家，粒子加速器之于物理学家，它们都是研究工具，都是为了使研究对象更加直观。只不过计算机常面对抽象的东西，如定理的机器证明。虽然 A. Wiles 已经证明了 Fermat 大定理，如何利用计算机证明它或验证人类的证明，对计算机科学依然是一个挑战。

*美籍匈牙利裔数学家、教育家 George Pólya (1887-1985) 在两卷本《数学与合情推理》中提出的一种启发式的推理模式。例如“不完全归纳”：Riemann 猜想至今未被证明或证伪，无论计算机提供多少正例都不算证明，但要证伪它，一个反例就足够了。目前找到的正例越来越多，所以人们在心理上更倾向于认为 Riemann 猜想是对的。另外一个例子是 Fermat 大定理 (Fermat's last theorem)，在 1995 年英国数学家 Andrew Wiles (1953-) 证明它之前，人们已经利用计算机验证了对于不超过四百万的奇素数 n 皆有 “ $x^n + y^n = z^n$ 无非零整数解”。



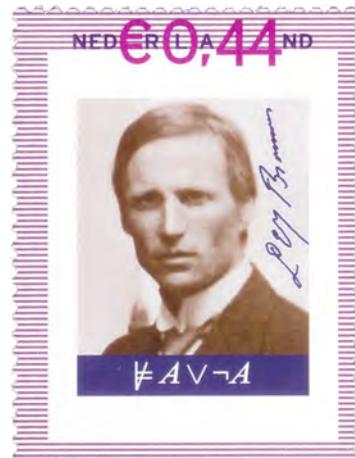
荷兰数学家、哲学家 Luitzen Egbertus Jan Brouwer (1881-1966) 在他的博士论文《论数学的基础》(1907) 中明确提出了直觉主义哲学，从而成为直觉主义数学的代表人物之一。直觉主义深刻影响了二十世纪数学的发展，特别是构造性数学的崛起*。

直觉主义需要构造性数学，但反之不然。对机器而言，Brouwer 直觉主义 (intuitionism) 的信条“存在即被构造”更具吸引力，不论是数值计算还是符号计算，“被构造”是至高无上的标准。构造性数学是强调“做”的数学，非构造性数学是强调“在”的数学，二者相得益彰。构造性数学是计算机科学的基础，也是机器证明和数学机械化的理论支持。



构造性数学 (constructive mathematics) 虽然无法做到把整个数学机械化，仍有相当可观的一部分能够剥离出来，用计算机解放人类的脑力劳动，图论中“四色定理”的机器证明就是一个很好的例子。另外，构造性数学扩展了数学问题的求解，尤其当解没有显式表达的时候，一个可行的求解算法可被视为问题的答案。

美籍华人数理逻辑学家、计算机科学家、哲学家王浩 (1921-1995) 是定理机器证明的先驱之一。1959 年，王浩在 IBM 704 计算机上用了几分钟证明了逻辑主义代表人物、英国数学家、哲学家 A. N. Whitehead (1861-1947) 和其学生英国哲学家、数学家、逻辑学家、文学家 Bertrand Russell (1872-1970) 倾注多年心血合著的三卷《数学原理》(Principia Mathematica) 中数百条数理逻辑定理，该工作于 1983 年荣获人工智能国际联合会的第一个自动定理证明里程碑奖。我国著名数学家吴文俊 (1919-2017) 在上世纪七十年代后期倡导的机械化数学，在几何定理的机器证明方面取得了国际领先的成果。没人知道计算机在这条路上到底能走多远，有笑话说



*感兴趣的读者可参阅荷兰数学家 Arend Heyting (1898-1980) 的著作《直觉主义导论》[73]。1967 年，美国数学家 Errett Albert Bishop (1928-1983) 出版了专著《构造性分析基础》[17]，提供了二十世纪分析学大部分结果的构造性实现。以 Andrey Andreevich Markov (1903-1979) 为首的苏联学派对构造性数学亦做出过杰出的贡献。

Hilbert 千年后复活，睁眼便问 Riemann 猜想解决了吗？答曰，解决了，请看代码和演示。



在数学里，少有像概率论这样的分支，既蕴藏着自然而朴素的真理，又距离应用如此之近：一边植根于测度论，一边面对各种随机现象。而统计学则是推断的艺术，它以概率论为坚实的基础，透过有限的观察，探知其间隐藏着的总体信息，或用于预测或帮助决策。

除了实用性，数学体现出的人类理性认知的水平和数学本身的和谐之美也是值得追求的。德国数学家 Carl Gustav Jacob Jacobi (1804-1851) 在给友人的一封信中说道，“Fourier 确实有过这样的看法，认为数学的主要目的是公共事业和对自然现象的解释；但像他这样的哲学家应当知道，科学的唯一目的是人类心智的荣耀……。”把数学单纯视为意志的产物并沉醉于它的美妙，是很多数学家乐此不疲的原动力。英国数学家 Godfrey Harold Hardy (1877-1947) 曾说，“数学家的模式正像画家或诗人的模式一样，必须是充满美感的；数学的概念就像画家的颜色或诗人的文字一样，也必须和谐一致。美感是首要的试金石，丑陋的数学在世上是站不住脚的。”概率论和数理统计里到处充满美妙的结果，等待着有心人的欣赏；还有各种方法论的思辩，等待着更加深邃的理解。

爱美之心，人皆有之。我们把数学的美当作艺术，若不能作美的创造者，就作美的传播者；若不能作美的传播者，就作美的欣赏者；若不能作美的欣赏者，就作美的追随者。

我们有一个普遍的共识：算法理论是计算机科学的核心，而数学则是算法的灵魂。应用概率统计方法最终都要落实到可行的算法上，数学里没有一无是处的美。无论是为了“心智的荣耀”还是解决实际问题，皆离不开基础理论之可靠和算法设计之精巧。



概率统计的理论和应用中蕴涵的一些朴素的思想，经过历史的沉淀，成为人类智慧宝库中璀璨的明珠。法国数学大师 Henri Poincaré (1854-1912) 说过，“如果我们想要预见数学的将来，适当的途径是研究这门学科的历史和现状。”为此，本书另一个具有革新意义的地方是增加了对概率统计历史和现状的简介，包括近些年本领域取得的一些成果，以及相关数学家的学术功绩和思想等。因为数学的历史是这些数学英雄创造的，他们的思想最能揭示理论的本质和发展脉络，也是数学文化不可缺少的组成部分，值得传承和永世的纪念^{*}。尤其是那些引人深思的哲学思想，它们更应该被津津乐道和传颂。



图 1：法国数学大师 Henri Poincaré (1854-1912)、奥地利数学家、伟大的逻辑学家和哲学家 Kurt Gödel (1906-1978) 和苏联数学大师 Andrey Kolmogorov (1903-1987) 是二十世纪现代数学不同领域的开拓者。

对那些注意事项、关键概念、引申思考、美妙的经典结果、习题难度、初次阅读可选择跳过的例子、证明或者章节等，书中都给出了特殊的标记，其含义说明如下。

表 1：书中用到的一些特殊标记及其含义。

	特别注意的事项		想得再远一点
	关键概念的定义		选读的例子、证明等
	令人怦然心动的结果	*	选读的补充章节
	证明完毕		条目、特款
§1.2.3	第一章第二节的第三小节		第六列举或步骤
☆, ★, ★★	标注在习题前，分别表示“有点难”、“难”、“很难”		

本书在每一章的开始都有一个脉络清晰的“导游图”，提纲挈领地描述本章的基本概念、主要结果及其之间的关系，有助于梳理思路和宏观掌控。而在每一节的开

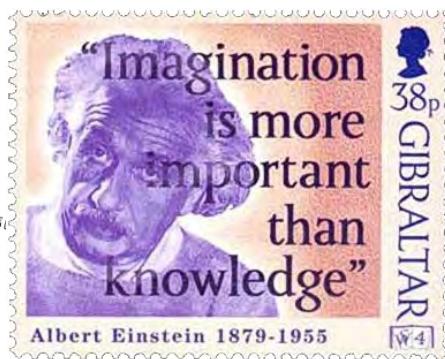
^{*}数学是所有自然科学的基础，数学强则科技强。为教化民众，伟大数学家和思想家的肖像常常出现在货币或邮票上。数学应该得到尊崇，因为这世上没有什么比真理更令人敬畏。

始，作者还对本节内容和关键知识做出概括性的提示，以便于读者抓住要点，理解和掌握即将学习的内容。

正文中的“练习”比课后习题简单，多是为了引发思考或强调某种方法或让读者“照葫芦画瓢”，一般都有答案或提示。课后习题都标有难度级别，为便于自学，附录 H 提供了“习题答案或提示”。

有一些较难的例子和证明，标注有剪刀符号 ，如果读不过去可以先暂时放下来，在掌握了正文主要内容后再来阅读比较好些。另外，本书还为读者提供了一些课外阅读的建议，有助于拓展知识面和增加趣味性，读者可依自己的学习目标和实际情况而定取舍。只有反复地钻研和实践、不断地提高认识才有可能窥见真理，而寻求真理、理解真理的过程本身也能为我们带来快乐。

Einstein 说过，“想象力比知识更重要”。首先，对概率统计的深刻理解往往不能单纯借助数学公式，而是需要直观想象。想象在先，数学在后，数学只不过是为了让理论显得严谨而已。其次，概率统计毕竟是数学工具，要漂亮地解决实际问题还得靠工具使用者的想象力和驾驭知识的能力。现实世界里的问题比教科书里的案例要复杂得多，真实数据往往很脏。通过应用培养对概率统计的“感觉”，遇到问题时就会有大局观去抓住本质并且知道用什么方法最有希望。另外，计算机在数值计算、科学绘图、符号推理、知识检索、过程模拟等方面有助于拓广我们的想象力，所以必须用好这一工具。



数学和禅有点像，需要用心去悟。数学的真理初见它时看山是山；知道了来龙去脉后触类旁通、举一反三，看山不是山；最后真理变成理所当然的东西，当我们能用质朴的话语自由地表达它，看山还是山。这是宋代禅宗大师青原行思 (671-740) 感悟的参禅的三重境界，和数学的认知过程不谋而合。我国著名数学家华罗庚 (1910-1985) 也曾说过，“要真正打好基础，有两个必经的过程，即‘由薄到厚’和‘由厚到薄’的过程。”华先生的话和行思的三重境界是同样的道理，我们最终必须“把那些学到的东西，经过咀嚼、消化，融会贯通，提炼出关键性的问题来。”

教科书里的知识都是最基本的，它是前人工作的精华和沉淀，可以用作建模的工具。博观而约取，厚积而薄发。然而，如何取和发？在多数情况下还需要有“应用”的经验和灵感，有那么多工具可供选择，用得好可不是一件容易的事情。所以，有两个坎要过，一个是继承前人的知识，一个是学会如何使用这些知识。我们不能闭门造车，学习概率统计的过程要特别注意过这两个坎，缺一不可。

本书中的术语在第一次出现时一般都给出了对应的英文，多采用国内既定的或流行的译法。外国人名基本采用英文，但对一些带有词根变化、用作形容词的外国人名，还是保留了其中文译名，如 Bayesian data analysis 译作“贝叶斯数据分析”、 Jacobian determinant 译作“雅可比行列式”等。对一些新术语，作者参考《英汉数学词汇》[2] 和《现代数学手册》[3] 给出适当的命名。读者可通过术语的索引表在正文中找到这些术语。

书中试验涉及的真实数据都标明了出处，模拟数据则给出相应的产生算法。本书利用 X_ET_EX 系统进行排版^{*}，所有科学计算和绘图都是通过开源的 GCC (GNU Compiler Collection)、R、Maxima、GnuPlot、MetaPost 完成的。人物肖像、漫画的图片取自互联网（如 Wikipedia [78] 等），恕不一一标明其出处。

正文几乎不涉及抽象测度论 (Measure Theory)[†]，仅假定读者已经学过集合论、数学分析或高等数学、线性代数等课程。该书可作概率统计课程的教材或参考书，主要面向高等院校非数学专业理工类的本科生和低年级研究生，以及在各自领域需要了解概率统计基础知识的科技人员。

由于概率论与数理统计学都已得到充分的发展，理论分支庞大，结果星罗棋布。要写一本面面俱到涵盖所有重要结果的基础教科书几乎是不可能的事情，我们只能有选择地把重点放在一些基本概念和经典成果上。因为本书的目标是为统计机器学习、模式识别、数据挖掘、大数据分析、人工智能等学科提供概率统计基础，我们既要保证一定的严谨性，又要在知识的组织架构上更侧重应用一些。有的时候，为了严谨需要交代很多概念和结果，而应用中又极少用到它们，作者一般会牺牲掉一点严谨避免读者陷于细节不能自拔。虽然作者尽力去把握严谨和实用的平衡，依旧有众口难调的情形需请读者们谅解。本书的内容共分四个部分，大致如下：



概率论基础: 随机事件，随机变量及其数字特征，特征函数，一些常见的分布，大数律与中心极限定理，随机过程简介等。

数理统计学初步: 一些基本概念，参数估计，假设检验，线性模型的回归分析与方差分析，时间序列分析，统计决策与贝叶斯分析概要等。

^{*}感谢 D. Knuth 大师对史上最优排版系统 T_EX 的杰出贡献 [90]。读者可从 TeXLive 获取不同平台之下的 T_EX 支持，包括各种宏包、字体和自由软件。

[†]测度论是现代分析数学的基础，研究的是一般集合上的测度和积分理论 [68]。二十世纪初建立的 Lebesgue 测度和 Lebesgue 积分理论以及随后的抽象测度和积分理论为概率的公理化奠定了基础。

一些实用算法: 概率图模型（隐 Markov 模型、条件随机场、贝叶斯网络）的几个算法，期望最大化算法，随机模拟算法等。

附录: 一些重要的但不适宜在正文中介绍的补充内容放在了附录里，如正态分布的由来，卷积的物理意义，Riemann-Stieltjes 积分，可测函数与 Lebesgue 积分，矩阵计算，凸性与 Jensen 不等式，软件 R、Maxima 和 GnuPlot 简介等。

第一、二章是基础，概念很多并且篇幅也较长，所用的教学时间也应相对多些。书中某些较复杂的内容带有标记，读者可有选择地略过而不会影响后续的阅读。对于某些重要的结论，虽然正文未给出严格的证明，但结论本身还是值得了解的，我们视其为“边缘”知识。这些证明被省略掉，并不是因为它们不重要，而是因受篇幅和主题所限，想了解细节的读者仍可根据作者提供的参考文献按图索骥找到详解。

感谢蔡延亮、李德珠、李霄翔、张力等几位研究生助教，他们帮助作者收集整理了与正文配套的大部分课后习题，并标注了难度。感谢北京大学信息科学技术学院的屈婉玲教授、王捍贫教授，他们对初稿提出了宝贵的意见。

借此书深切缅怀恩师程民德先生，他引导作者由数学转入信息科学领域。二十世纪七十年代，程先生最早在国内领导开展了模式识别与图像处理的研究，建立了北大信息数学专业并培养了许多优秀的人才。我有幸成为程先生的学生得到他的指导，先生的谆谆教诲令我终生铭记，时常提醒自己不要满足于肤浅狭隘的认知。

最后要说的是，家人多年的无私关爱与支持是作者完成此书的动力。特别地，谢谢女儿经常及时地打断我的工作让我跳出写作的困局全神贯注地陪她玩耍。

本书的大多数章节，曾作为北京大学信息科学技术学院的本科生主干基础课《概率统计 A》的教学内容多次使用，其余部分在研究生课程《统计机器学习》和《贝叶斯数据分析》中讲授过。感谢听过这些课程的学生们，他们容忍了讲义不断更新带来的不便。虽几经易稿，由于作者能力所限，书中仍难免有错讹或不妥之处，诚恳地欢迎读者指出，以便在后续的版本中予以修正，不断提高它的质量。希望本书能对读者有所裨益，并带来阅读的快乐。



于江生
邮件地址: yujiangsheng@gmail.com

目 录

概率论基础	1
1 随机事件与概率论的公理化	8
1.1 古典概率模型	14
1.1.1 计数概率	21
1.1.2 几何概率	28
1.1.3 Monte Carlo 方法	36
1.1.4 对随机性的思考	40
1.2 概率论的公理化	45
1.2.1 σ 域与样本空间	50
1.2.2 Kolmogorov 公理体系	56
1.2.3 概率的一些基本性质	70
1.3 条件概率与随机事件的独立性	75
1.3.1 条件概率及其性质	77
1.3.2 全概率公式与 Bayes 公式	82
1.3.3 随机事件的独立性	87
1.3.4 条件独立性及其性质	98
1.4 习题	101
2 随机变量及其数字特征	106
2.1 随机变量及其基本性质	112
2.1.1 随机变量的分布与分布函数	117
2.1.2 离散型与连续型随机变量	122
2.1.3 随机变量的函数	130
2.2 随机向量及其基本性质	133
2.2.1 边缘分布和条件分布	140
2.2.2 随机变量间的独立性	146
2.2.3 条件独立性	149
2.2.4 随机向量的函数	154
2.3 随机变量的数字特征	164

目 录	11
-----	----

2.3.1 数学期望	168
2.3.2 条件期望与双期望定理	174
2.3.3 方差与条件方差	177
2.3.4 熵、互信息和 Kullback-Leibler 散度	182
2.3.5 原点矩、中心矩和绝对矩	188
2.3.6 概率不等式	191
2.4 随机变量之间的关系	199
2.4.1 相关系数	204
2.4.2 最小二乘法和回归	206
2.4.3 随机向量的主成分	210
2.5 习题	212
3 特征函数	217
3.1 特征函数的基本性质	225
3.1.1 独立随机变量之和的特征函数	228
3.1.2 特征函数与矩	231
3.2 特征函数与分布函数的关系	234
3.2.1 Lévy 反演公式	240
3.2.2 Lévy 连续性定理	246
3.3 习题	249
4 一些常见的分布	251
4.1 离散型随机变量的分布	258
4.1.1 单点分布和两点分布	261
4.1.2 二项分布	263
4.1.3 Pólya 分布及其特例（超几何分布）	267
4.1.4 几何分布和负二项分布	269
4.1.5 Poisson 分布	273
4.2 连续型随机变量的分布	278
4.2.1 均匀分布	280
4.2.2 三角形分布	284
4.2.3 正态分布、对数正态分布和偏正态分布	285
4.2.4 Laplace 分布	290
4.2.5 Cauchy 分布	292
4.2.6 Gamma 分布及其特例（ χ^2 分布和指数分布）	293
4.2.7 Beta 分布	301
4.2.8 t 分布和 F 分布	305
4.2.9 Pareto 分布	308

4.2.10 以物理学家命名的分布	311
4.3 随机向量的分布	317
4.3.1 高维均匀分布	320
4.3.2 多项分布	324
4.3.3 Dirichlet 分布	328
4.3.4 多元正态分布与多元 t 分布	334
4.3.5 随机矩阵与 Wishart 分布	340
4.4 习题	342
5 大数律与中心极限定理	345
5.1 大数律	350
5.1.1 弱大数律	355
5.1.2 强大数律与重对数律	360
5.2 中心极限定理	366
5.2.1 Lindeberg-Feller 中心极限定理	370
5.2.2 中心极限定理的应用	376
5.3 习题	380
6 随机过程简介	382
6.1 离散时间 Markov 链	393
6.1.1 状态的分类	398
6.1.2 Markov 链的遍历性与平稳分布	405
6.1.3 分支过程	410
6.2 连续时间过程	415
6.2.1 Poisson 过程与更新过程	419
6.2.2 生灭过程	426
6.2.3 布朗运动	430
6.3 随机分析	437
6.3.1 伊藤积分	438
6.3.2 随机微分方程	440
6.4 习题	441
数理统计学初步	443
7 数理统计学的一些基本概念	449
7.1 样本的特征	454
7.1.1 次序统计量	460

目 录	13
7.1.2 经验分布及其性质	463
7.1.3 样本矩及其极限分布	471
7.2 样本统计量及其性质	473
7.2.1 统计量的抽样分布	475
7.2.2 重抽样和自助法	480
7.2.3 统计量的充分性	484
7.3 习题	489
8 参数估计理论	491
8.1 点估计及其优良性	494
8.1.1 Fisher 信息量与信息矩阵	495
8.1.2 相合性与渐近正态性	499
8.1.3 无偏性和有效性	503
8.1.4 刀切法	510
8.1.5 点估计之矩方法和最大似然法	512
8.2 Neyman 置信区间估计	522
8.2.1 基于 Markov 不等式的区间估计	524
8.2.2 枢轴量法	526
8.2.3 大样本区间估计	531
8.2.4 Fisher 的信任估计	535
8.3 习题	537
9 假设检验	539
9.1 Neyman-Pearson 假设检验理论	546
9.1.1 功效函数与两类错误的概率	548
9.1.2 Neyman-Pearson 引理与似然比检验	553
9.1.3 广义似然比检验	559
9.1.4 假设检验与置信区间估计的关系	566
9.2 大样本检验	571
9.2.1 拟合优度检验	575
9.2.2 独立性的列联表检验	581
9.3 习题	583
10 回归分析与方差分析	585
10.1 线性回归模型	591
10.1.1 最小二乘估计	593
10.1.2 线性回归的若干性质	598
10.1.3 回归模型的假设检验	602

10.1.4 正交多项式回归	605
10.2 方差分析模型	610
10.2.1 单因素方差分析	613
10.2.2 两因素方差分析	616
10.3 习题	620
11 时间序列分析	621
11.1 ARMA 模型	626
11.1.1 趋势性和季节性	628
11.2 预测	629
11.2.1 ARMA 模型的预测	630
11.3 参数估计	631
11.3.1 谱密度的估计	632
11.3.2 ARMA 模型的估计	633
11.4 习题	634
12 统计决策理论与贝叶斯分析概要	635
12.1 统计决策理论中的基本概念	638
12.1.1 贝叶斯学派的期望损失原则	639
12.1.2 频率派的决策方法	641
12.2 贝叶斯分析	644
12.2.1 后验 \propto 似然 \times 先验	648
12.2.2 参数的先验分布	653
12.2.3 后验分布及其期望的计算	661
12.2.4 贝叶斯模型选择	665
12.2.5 层级贝叶斯模型	666
12.3 习题	672
概率统计中的一些实用算法	673
13 概率图模型	677
13.1 隐 Markov 模型及其算法	678
13.1.1 观察序列的概率: 向前算法与向后算法	681
13.1.2 状态序列的概率: Viterbi 算法	684
13.1.3 模型参数的训练: Baum-Welch 算法	686
13.2 无向图模型与 Bayes 网络	688
13.2.1 Markov 网络与条件随机场	689

目 录	15
13.2.2 Bayes 网络	691
13.3 习题	692
14 期望最大化算法	693
14.1 完全数据与最大似然估计	698
14.1.1 EM 算法及其收敛速度	699
14.1.2 EM 算法的若干变种	702
14.2 期望最大化算法的应用	703
14.2.1 分支个数已知的高斯混合模型	704
14.2.2 针对删失数据的 EM 算法	706
14.2.3 指数族的 EM 算法	707
15 随机模拟技术	708
15.1 产生随机数的传统方法	712
15.1.1 von Neumann 舍选法	714
15.1.2 复合抽样和重要度抽样	720
15.2 Markov 链 Monte Carlo 方法	723
15.2.1 Metropolis 算法	724
15.2.2 Metropolis-Hastings 算法	729
15.2.3 Gibbs 抽样与切片抽样	732
15.3 随机模拟技术的应用	736
15.3.1 模拟退火算法	737
15.3.2 缺失数据的多重填补算法	744
15.3.3 数据增扩算法	746
15.4 习题	750
附录	751
A 正态分布的由来	752
B 卷积的物理意义	754
C Riemann-Stieltjes 积分	755
D 可测函数与 Lebesgue 积分	759
E 矩阵计算的一些结果	764
F 凸性与 Jensen 不等式	775

G 软件 R、Maxima 和 GnuPlot 简介	778
G.1 R: 最好的统计软件	778
G.2 Maxima: 符号计算的未来之路	780
G.3 GnuPlot: 强大的函数绘图工具	784
H 习题答案或提示	786
I 参考文献	813
J 符号表	826
K 术语索引	830

未经作者允许，
本书不得转售和批发。
声明：本书不得用于任何商业活动。

第一部分

概率论基础

概率论简史

大江东去，浪淘尽，千古风流人物。

苏轼《赤壁怀古》

概率论起源于十七世纪中叶。当时，法国流行一种赌博游戏：连续掷一个均匀的骰子 4 次，赌是否出现 1 点。热衷于赌博的法国显贵 Chevalier de Méré 对赌博机理深感兴趣，他发现在这个游戏中选“是”赢的机会更大，并在实战中屡屡得手。而对这个游戏的升级版：连续掷一对骰子 24 次，赌是否出现一对 1 点，de Méré 觉得两个骰子同时掷出 1 点的机会显然是单个骰子掷出 1 点的 $1/6$ ，所以掷 24 次这对骰子出现一对 1 点的机会等同于掷 4 次单个骰子出现 1 点的机会。于是，他想当然地认为同样选“是”赢的机会更大，但事与愿违。

de Méré 百思不得其解，只好求助于法国大哲学家兼数学家 Blaise Pascal (1623-1662)。Pascal 与他的好友、著名的非职业数学家 Pierre de Fermat (1601-1665) 经过多次通信讨论，最终解决了 de Méré 问题*，并第一次系统地阐述了概率的加法与乘法。另外，通过讨论赌资分配问题（见第 168 页的例 2.51），二人还提出了概率论中的一个重要概念——数学期望，简称“期望”。作为一门研究机会的学问，概率论诞生了。



*现在人们通过简单的计算，就能给出 de Méré 问题的解：两个赌博游戏中选“是”的胜率分别是 $1 - (5/6)^4 \approx 0.5177$ 和 $1 - (35/36)^{24} \approx 0.4914$ 。Pascal 提到如果把游戏升级版的规则改为掷 25 次，选“是”的胜率又将超过 $1/2$ ，约为 0.5055。

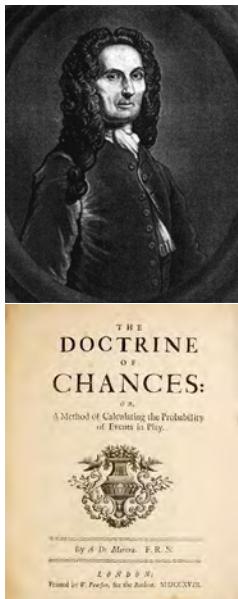
荷兰数学家 Christiaan Huygens (1629-1695) 在巴黎访学期间受到了 Pascal 和 Fermat 工作的启发，在给友人的信中他说到，“我坚信我们对凡事都不确定，而是或然地了解万物。”1657 年，Huygens 发表了概率论的第一篇正式论文《论赌博中的推断》，正式命名了“数学期望”，并探讨了概率的原理和计算方法。

瑞士数学家 Jacob Bernoulli (1654-1705) 研读了 Huygens 的著作并与 Huygens 保持通信交流，从而继承了 Pascal、Fermat 等前人对概率论开创性的工作。



十八世纪

社会需求在很大程度上刺激了概率论的研究，例如当时欧洲的保险、精算等商业实践需要分析大量偶然现象以找出其背后的规律。1713 年，Jacob Bernoulli 的遗著《猜度术》(也译作《推测术》)出版，其内容涉及了排列组合的一般理论和当时概率论的经典结果及应用，包括著名的 Bernoulli 弱大数律 (weak law of large numbers)。Bernoulli 弱大数律的意义在于首次以严格的方式给出概率的频率解释，我们称之为客观概率，即概率作为随机事件的本质属性，它的存在是不以个人的意志而改变的。在 Jacob Bernoulli 的推动下，概率论在欧洲得到了更深入的发展。



此后，旅居英国的法国数学家 Abraham de Moivre (1667-1754) 发现连续抛一枚均匀的硬币 n 次出现正面的次数服从二项分布 $B(n, 1/2)$ ，相继在他的论著《机遇论》(1718) 和《级数和求积的分析杂论》(1730) 中给出了二项分布的严格定义并做了系统的研究。De Moivre 于 1733 年发表了一个重大的研究成果：当 n 足够大时可用正态分布来逼近二项分布 $B(n, 1/2)$ 。可惜的是，这一成果的思想超越了那个时代，当时并未引起数学家们的重视。直到法国数学家 Pierre-Simon Laplace (1749-1827) 在其著作《概率的分析理论》(1812, 1814) 中再度提起，并将之推广到二项分布的一般形式。此结果就是著名的 de Moivre-Laplace 中心极限定理 (central limit theorem)，它堪称“古典概率论的无冕之王”。

1763 年，英国数学家 Reverend Thomas Bayes (1701?-1761) 的遗作《论有关机遇问题的求解》发表，概率被 Bayes 理解为主观的信念度 (belief degree) 用作统计推断。论文中的命题 9 便是著名的 Bayes 公式：后验信念度正比于先验信念度与似然的乘积。Bayes 开创了主观概率的先河，是贝叶斯学派的鼻祖。从此，主观概率与客观概率之争纷纷扰扰几个世纪，至今尚未统一。

概率论的研究引发了人们对知识、推理等哲学问题的思考。Laplace 在《概率的分析理论》第二版的长篇绪论《概率的哲学探讨》中表达了与 Huygens 类似的思想，

他说“严格地讲，我们几乎所有的知识都是或然性的，只有很少的事物我们能确定无疑地了解。在数学科学里，归纳和类比这些发现真理的基本方法也是建基于概率的，以至于人类知识的整个系统都和概率论息息相关。”

十九世纪

概率论开始迅猛发展，逐渐成为自然科学的工具。如天文学、大地测量学急需随机误差分析的数学理论，生物统计学、物理学开始意识到自然科学的随机法则等。

初叶，概率论先驱 Laplace 的《概率的分析理论》一书奠定了古典概率论的基础，是该领域早期的集大成之作。在概率论的发展史中，Laplace 是个承前启后的关键人物，他的思想对后世影响巨大。譬如，作为决定论者，Laplace 认识到概率论对很多现象的研究是必需的。他首次严格地阐述了 Bayes 公式，并将之用于贝叶斯推断。尽管有时缺乏严谨性，Laplace 在概率论上的洞察力仍令人折服。例如，他在误差理论中使用特征函数和反演公式；在应用极限定理时将偏微分方程引入概率论；提出观察误差是一些独立的小误差之和，在一般条件下，观察误差服从正态分布等。



十九世纪概率论的重大成果还包括德国数学天才 Carl Friedrich Gauss (1777-1855) 和法国数学家 Adrien-Marie Legendre (1752-1833) 独立提出的最小二乘法；法国数学家 Siméon Deni Poisson (1781-1840) 发现的 Poisson 弱大数律；俄国圣彼得堡学派数学家 Pafnuty Lvovich Chebyshev (1821-1894) 及其学生 Andrey Andreyevich Markov (1856-1922)、Aleksandr Mikhailovich Lyapunov (1857-1918) 发现的概率不等式、弱大数律、中心极限定理等。

十九世纪末至二十世纪初，Lyapunov 开辟了利用特征函数的方法证明中心极限定理之路。从此，极限定理本身的重要性和特征函数方法的重要性都逐渐被世人所知，由它激发起来的研究热情一直持续到二十世纪中叶。如今，极限定理已是概率论的重要内容，也是统计学的基石之一。

十九世纪的概率论积累了很多漂亮的结果，也有了更广泛的应用。遗憾的是，当时的概率论仍缺乏一些基本概念（如概率、随机事件、随机变量等）的清晰定义。由于没有严格的逻辑基础，一些悖论应运而生，其中比较著名的是法国数学家 Joseph Bertrand (1822-1900) 在其著作《概率计算》(1889) 中给出的几何概率的



悖论（见第 31 页的例 1.24）。Bertrand 悖论敲响了警钟，人们不得不重新审视概率论的数学基础。

二十世纪以来

1900 年，德国数学大师 David Hilbert (1862-1943) 在巴黎第二届国际数学家大会上作了题为《数学问题》的讲演，提出了 23 个指引二十世纪数学发展的关键问题，其中的第六问题涉及概率论的公理化。

1909 年，法国数学家 Émile Borel (1871-1956) 首次把概率论与测度论结合起来，定义了可数事件集的概率。相对古典概率而言，这一工作拓展了对概率的认识。另外，他还证明了 Borel 强大数律。比 Borel 强大数律更精细的结果是由苏联数学家 Aleksandr Yakovlevich Khinchin (1894-1959) 于 1924 年给出的 Khinchin 重对数律。

1917 年，苏联数学家 Sergei Natanovich Bernstein (1880-1968) 构建了概率论的第一个公理体系。1919 年，奥地利数学家和空气动力学家 Richard von Mises (1883-1953) 完成了概率的频率定义和统计定义的公理化。之后，还相继出现了一些主观概率的公理体系。然而，所有这些工作都只是前奏，它们或欠缺合理性，或缺乏权威性。直到二十世纪三十年代，随着对大数律的深入研究，人们逐渐意识到概率论与测度论之间存在着深刻的联系，概率论公理化的曙光才真正出现。

1919 至 1925 年，法国数学家 Paul Pierre Lévy (1886-1971) 发展了 Lyapunov 在特征函数上的工作，得到 Lévy 反演公式和连续性定理等重要结果。利用 Lévy 连续性定理不难证得 Khinchin 弱大数律 (1929)，以及芬兰数学家 Jarl Waldemar Lindeberg (1876-1932) 于 1922 年发现的 Lindeberg-Lévy 中心极限定理等。中心极限定理的巅峰是 Lindeberg-Feller 中心极限定理，该定理给出了中心极限定理的充要条件，其中必要性是由美籍克罗地亚裔数学家 William Feller (1906-1970) 于 1935 年证得的。



1926 年，苏联数学大师 Andrey Nikolaevich Kolmogorov (1903-1987) 给出弱大数律的充要条件。在接下来的几年中，他还发现了 Kolmogorov 概率不等式，几个强大数律和 Kolmogorov 重对数律等重要结果。

二十世纪二十年代，经典统计学在频率派把持之下快速发展。概率的频率解释逐渐占了上风，这也影响了概率论的公理化。

1933 年，Kolmogorov 总结了前人的工作，在他的成名之作《概率论基础》中首次利用测度论 [94] 构建了概率论的公理化体系。该体系为大部分数学家所接受，从此概率论成为近代数学最重要的分支之一，并得到迅速的发展。忽如一夜春风来，千树万树梨花开。例如，Kolmogorov、Khinchin、Lévy、Norbert Wiener (1894-1964) 建立并发展了随机过程理论（即概率

论的动力学部分); 美国数学家 Joseph Leo Doob (1910-2004) 创立了鞅论; 日本数学家伊藤清 (Kiyoshi Itô, 1915-2008) 创立了随机积分和随机微分方程理论。

目前, 现代概率论的研究内容大致包括极限理论、独立增量过程、Markov 过程、平稳过程和时间序列、鞅论和随机微分方程、点过程等, 每个方面都积累了大量出色的成果。概率论已发展成为一个具有广泛应用背景的数学研究领域。

在信息时代, 贝叶斯主义 (Bayesianism) 重新被理解和认识。一方面, 高性能计算 (high performance computing, HPC) 让贝叶斯推断变得可行。另一方面, 人工智能 (artificial intelligence, AI) 的研究给主观概率提供了一个广阔的发展空间。可以预测, 主观概率理论在未来将大有作为。

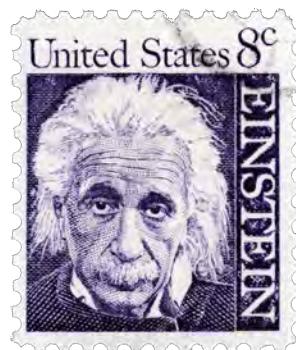
概率论的意义

概率论除了作为数理统计学的理论基础, 也是数论、图论、组合论等纯粹数学分支和金融数学、决策论、信息论、控制论、博弈论、密码学、算法、运筹学等应用数学分支常用的工具。概率论在自然科学和社会人文科学领域, 如物理学、化学、生物学、医学、心理学、语言学、经济学、社会学、教育学、政治学、法律学等, 以及所有的工程技术科学领域都能找到广泛的应用, 它的实用价值是毋庸置疑的。

“纯数学使我们能够发现概念和联系这些概念的规律, 这些概念和规律给了我们理解自然现象的钥匙。” Einstein 的这些对纯数学的盛赞之词专用于概率论也毫不为过。

同时, 概率论也深受其他学科的影响。例如, 物理学与概率论交汇产生了量子力学、随机场、随机矩阵理论、交互作用粒子系统、渗流理论、测度值随机过程等研究分支。另外, 半个多世纪以来, 计算机科学也加快了随机模拟技术的发展, 促进了随机算法的研究。

法国大数学家、直觉主义的先驱 Jules Henri Poincaré (1854-1912) 在其哲学著作《科学与假设》(1902) 的第十一章《概率计算》中说, “如果我们不是无知, 那就没有概率, 而只有让位于确定性了; 但是我们的无知不能是绝对的, 否则也将没有概率了, 因为就是要达到这种不确定的科学, 还得要借点光明才行。”对随机性的认识有助于培养科学精神, 当前国内外对古典概率的普及已经渗入到小学数学教育, 以高等数学为基础的概率论对于大学数学教育而言更是不可或缺的, 因为“生活中最重要的问题大多数都是概率问题”(Laplace, 《概率的哲学探讨》)。



推荐的课外读物

为了更好地学习并掌握概率论这一工具, 本书推荐以下课外读物, 难度由浅及深, 它们都是公认的名著。

- 美国科学院院士 W. Feller 的《概率论及其应用》上卷 [45] 对预备知识要求较少，深入浅出地介绍了概率论。Feller 对近代概率论的发展做出了卓越的贡献，他的两卷本的《概率论及其应用》都是经典著作。
- 苏联科学院院士、Kolmogorov 的学生 B. V. Gnedenko (1912-1995) 的《概率论教程》[58] 内容浓缩，书末提供了《统计学要领》和《概率论简史》。
- 王梓坤院士 (1929-) 的《概率论基础及其应用》[166] 有丰富的应用实例，如随机过程的模拟、概率论在计算方法中的一些应用、可靠性问题的概率分析等，附录还有对随机性的哲学思考。王梓坤也是 Kolmogorov 的学生。
- 苏联科学院院士、Kolmogorov 的另一学生 A. N. Shirayev (1934-) 的《概率论》[145] 是概率论的经典教材，被列为美国研究生数学丛书 GTM 第 95 号。
- 日本数学家、概率论大师伊藤清 (Kiyosi Itô, 1915-2008) 的名著《概率论基础》[161] 短小精悍却涵盖了大量的内容。美籍华裔数学家、概率论专家钟开莱 (1917-2009) 的《概率论教程》[26] 也是一本享誉世界的经典概率论教材。
- 美籍法裔概率论专家 M. Loève (1907-1979) 的《概率论》第一、二卷 [107] 被列为美国研究生数学丛书 GTM 第 45、46 号，是二十世纪七十年代以前的概率论的集大成之作。作者 Loève 是概率论大师 Lévy 的学生。作为补充，近半个世纪的概率论的发展可参阅 [40, 84]。

前三部著作无需测度论基础，非常适合初学者；后几部面向数学专业，需要一定的数学基础，建议在掌握了相当于本书概率论的主要内容之后再入手较为稳妥。这些推荐书籍的共同特点是强调了概率论的思想方法，语言优美例子丰富，读者可根据自己的需求选读这些名著。

第一章

随机事件与概率论的公理化

形而上者谓之道，形而下者谓之器，化而裁之谓之变，推而行之谓之通。

《易经·系辞》

自然界和人类社会中的现象本质上可分为确定性的和非确定性的（又称随机性的）两类。确定性的现象可以在某些条件下预言它是否发生：若发生则称之为必然事件 (certain event)，否则称之为不可能事件 (impossible event)。

必然事件有如，(i) 一个标准大气压下，纯水在 100°C 沸腾；(ii) 物体在无外力作用下速度保持不变；(iii) 光线通过引力场将发生偏移。不可能事件有如，(iv) 太阳从西方升起；(v) 欧氏几何中的三角形两边之和小于第三边。显然，必然事件的否定就是不可能事件，反之亦然。读者以往学过的数学、宏观物理学、初等化学等基本上都是研究这类确定性现象的，但宇宙万物间具有绝对确定性的现象少之又少，人类更多面对的是随机现象。

对于非确定性的现象，又称随机现象，它们在一定的条件下可能发生也可能不发生，在得知其发生与否之前，我们称之为随机事件 (random event)。如 (i) 抛一枚均匀的硬币出现正面；(ii) 一个均匀的骰子掷出奇数点；(iii) 明年三月份交货的黄金期货价格为每盎司 765.50 美元；(iv) 未来十年全球温度将持续上升；(v) 某特效药能治愈某人的胃癌等。

如何研究这些随机事件呢？传统的做法是通过多次的随机试验 (random trial) 来揭示隐藏在大量观察结果背后的规律。虽然每次试验的结果都不确定，而且少量试验也看不出什么规律，但随着试验次数的增加，那些隐藏着的“必然性”就会逐渐浮现出来。例如对“抛一枚均匀的硬币出现正面”这一随机事件，我们采用的随机试验是“在相同条件下抛该枚硬币”，只要抛足够多次，出现正面的次数与抛次之比就必然稳定在 $1/2$ 附近。

有人可能质疑：既然在相同条件下抛硬币，出现的结果应该是一样的，哪里有随机性可言？事实上，“在相同条件下”这一要求并不能完全达到，由于技术上或能

力上的局限，总有一些人为不可控制的因素影响着试验的结果，譬如地球引力的微小变化、气流的轻微扰动、抛硬币者的心理波动等，况且抛硬币动作本身也不可能达到绝对精确的重复。

为了验证抛一枚均匀硬币足够多次以后出现正面的频率会呈现一定的规律性，历史上有多位充满好奇心的学者，如英国数学家 Augustus de Morgan (1806-1871)，法国博物学家 Comte de Buffon (1707-1788)，美国数学家 William Feller (1906-1970)，英国统计学家 Karl Pearson (1857-1936) 等，都亲自做过抛硬币的随机试验，试验结果如下。

表 1.1: 历史上一些知名学者做过的抛硬币试验的结果。

试验者	抛次	正面次数	正面频率
de Morgan	2,048	1,061	0.5181
C. de Buffon	4,040	2,048	0.5069
W. Feller	10,000	4,979	0.4979
K. Pearson	12,000	6,019	0.5016
K. Pearson	24,000	12,012	0.5005

练习 1.1. 举更多有关随机现象的例子，阅读 [166] 的附录《论随机性》。

为了研究随机现象中的数量规律，需要概率论这一数学分支。为了使问题能够得到形式化的描述，概率论的研究要求

1. 随机试验 \mathcal{E} 所有可能的结果组成的集合 Ω 是已知的，我们称之为基本事件集合。 Ω 中的任一元素 ω 称为一个基本事件 (elementary event) 或样本点 (sample point)，记作 $\{\omega\}$ 或 ω (在不引起歧义的情况下)。例如，抛硬币试验的基本事件集合 $\Omega = \{\text{正面}, \text{反面}\}$ ，其中 $\{\text{正面}\}$ 和 $\{\text{反面}\}$ 都是基本事件。
2. 在相同条件下，随机试验 \mathcal{E} 可以不断重复。对于那些无法重复的随机试验，如“某特效药能治愈某人的胃癌”，可以适当地修改条件，把与此人病情、生理、生活规律、工作环境等相似的服用此药的其他胃癌患者也作为观察对象，只要试验结果对适当修改的条件不太敏感，研究者依然可以从这些“重复性试验”中寻找规律来预测该特效药能否治愈此人的胃癌。

如何用数学的方法形式地表示随机事件呢？以“掷骰子出现奇数点”为例，掷骰子的基本事件集合是所有可能出现的点数，即 $\Omega = \{1, 2, 3, 4, 5, 6\}$ ，其中出现奇数点的所有可能结果是 $A = \{1, 3, 5\}$ 。如果“掷骰子出现奇数点”这一随机事件发生了，骰子的点数必定是集合 A 中的某一个。很自然地，人们用集合 $A = \{1, 3, 5\}$ 来表示“掷骰子出现奇数点”这一随机事件，用骰子实际掷出的点数是否属于 A 来判定该随机事件是否发生（譬如，骰子被掷出的点数是 2，则随机事件“掷骰子出现奇数

点”没有发生)。像 $A = \{1, 3, 5\}$ 这样由不少于两个基本事件构成的随机事件，我们称之为复合事件 (composite event)。

任一随机事件都可用基本事件集合 Ω 的某个子集来表示，于是，集合论理所当然地成为概率论的数学基础。后文中，凡提到随机事件，都用集合来表示。

在 Ω 的所有子集中， Ω 自身和空集 \emptyset 是两个极端的例子。全集 Ω 包含了随机试验所有可能的结果，不管试验结果如何事件 Ω 总是发生的，显然 Ω 表示一个必然事件。而空集 \emptyset 不包含任何元素，所以它用来表示不可能事件。

例 1.1. 连续抛一枚均匀的硬币两次，若用 H 表示正面 (head)， T 表示反面 (tail)，则基本事件集合是 $\Omega = \{(T, T), (T, H), (H, T), (H, H)\}$ ，基本事件的个数是 $|\Omega| = 4$ 。

Ω 的幂集合 2^Ω 共有 $2^{|\Omega|} = 2^4 = 16$ 个元素。其中，元素 $\{(T, T), (H, H)\}$ 表示复合事件“两次抛出的结果相同”， $\{(T, H), (H, T)\}$ 表示复合事件“至少抛出一个正面”。请读者说一说 2^Ω 的其他元素都表示什么事件。

练习 1.2. 连续抛一枚均匀的硬币三次，请写出该试验的基本事件集合 Ω 。说明事件 A = “至少抛出一个反面” 和 B = “没有连续出现正面，并且没有连续出现反面”的关系。提示： $B \subset A$ 。

练习 1.3. 将 6 个球随机地放入标号为 $1, 2, \dots, 10$ 的盒子中，每个盒子至多放两个球，问基本事件集合的势是多少？提示： $C_{10}^6 + C_{10}^1 C_9^4 + C_{10}^2 C_8^2 + C_{10}^3 = 2850$ 。

在本质上，一个随机事件是否发生是无法预测的。在试验之前，我们不可能精确预知一个均匀骰子掷出的点数，也就不可能预知“掷出奇数点”是否发生。需要澄清的是，本质上无法预测和复杂得难以预测是截然不同的。下面两个例子有助于更好地理解随机性，它们所谈论的都不是真正意义上的随机性，虽然看上去很像。

※例 1.2. 确定的动力学系统有时也可以表现出“随机性”，即混沌 (chaos)，它是复杂系统的研究对象。混沌的本质是系统在经过长期演化后对初始条件变得敏感，以至于“差之毫厘，失之千里” [44, 125]。

譬如，在气象学方面，理论上已经证明利用动力学模型精确预报两三周后的天气情况是不可能的，一个形象的比喻是所谓的“蝴蝶效应”——北京一只蝴蝶偶尔扇动几下翅膀将引发数月后美国德克萨斯州的一场飓风。

以函数 $f(x) = 2x^2 - 1$ 的迭代为例，考虑 $[-1, 1]$ 区间上的动力系统：初值 x_0 设为 0.4 (实线) 和 0.4001 (虚线)，图 1.1 是 100 次迭代的结果的折线图。

$$x_1 = f(x_0), x_2 = f(x_1), \dots, x_{100} = f(x_{99})$$

初始值的小扰动 0.0001 在刚开始迭代的一小段时间内并没有引起函数值太大的变化。然而，在长期迭代后却引起了迭代结果貌似随机的变化，它们看上去是非周期的、不规则的和无法预测的。但这种“随机性”与掷骰子、抛硬币等有着根本的区别。

别：混沌系统的短期表现（即最初的几次迭代）是可知的，然而掷骰子在任何时候都是不能精确预测结果的。所以，混沌系统的“随机性”不是概率论意义上的随机性。

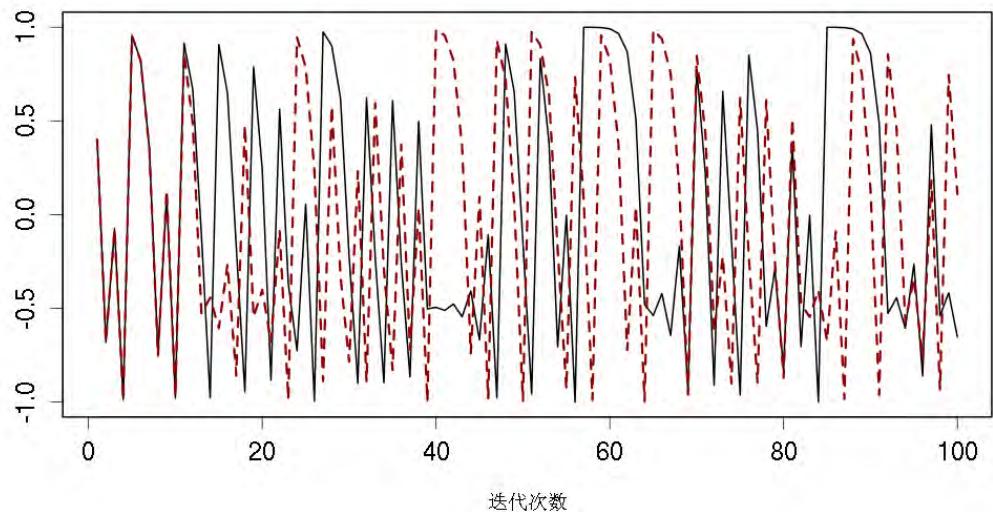


图 1.1：初值的小扰动在经过一定时间的演化后引发了混沌系统的“随机”变化。

※例 1.3 (Arnold 变换). 下述平面上的可逆周期映射被称为 Arnold 变换。

$$\Gamma : (x, y)^T \mapsto (2x + y, x + y)^T \mod 1$$

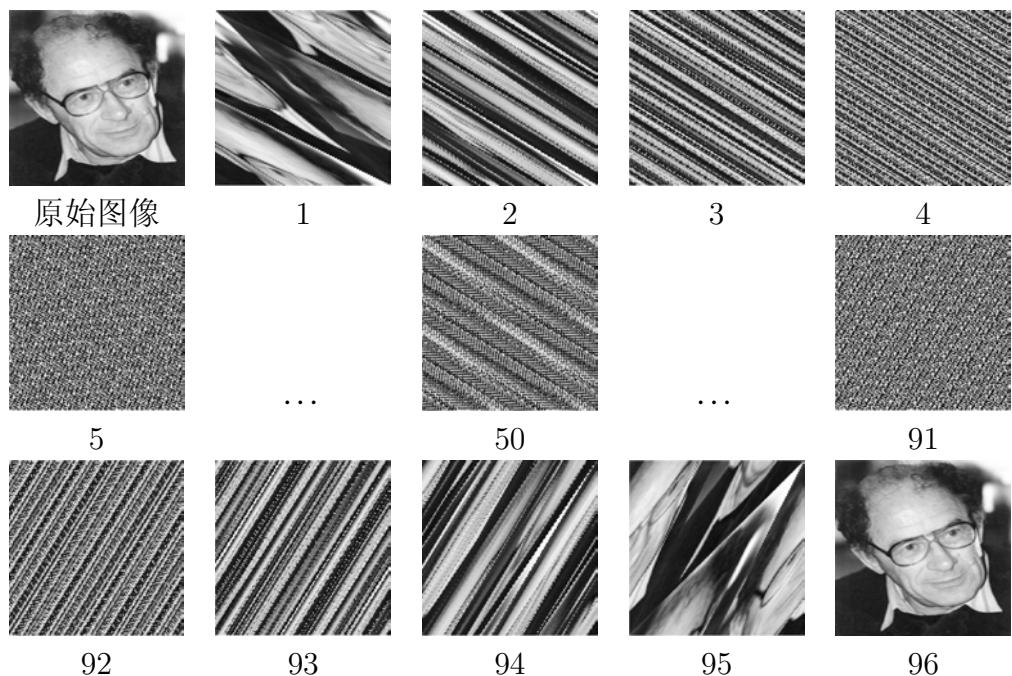
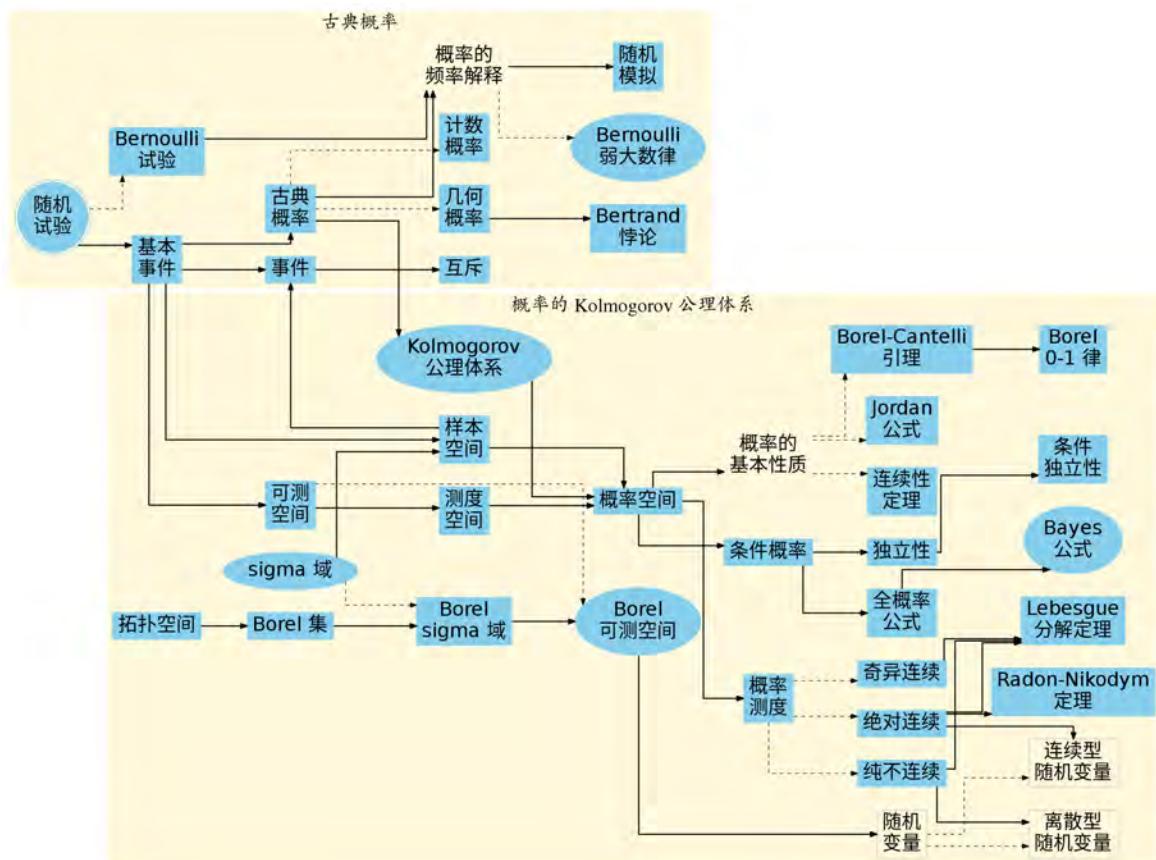


图 1.2：俄国数学家 Vladimir Igorevich Arnold (1937-2010) 的头像经过 Arnold 变换数次迭代后的结果。

不难看出，Arnold 变换可以把一幅单位正方形的图像置乱，经过若干次迭代后图像看上去就像是随机生成的，但是经过一定周期后原始图像又得以恢复，这一特点让 Arnold 变换常用于信息隐藏和图像加密。

在图 1.2 中，原始图像经过 96 次迭代后又重新得到恢复，在这一过程中，该动力系统似乎表现出“随机性”，其实不然——第 n 次迭代的结果总是第 $n - 1$ 次迭代的结果按照固定的模式分割拼装而成。换句话说，如果我们知道第 $n - 1$ 次迭代的结果，就可以准确无误地构造出第 n 次迭代的结果。

第一章的主要内容及其关系



我们用有向图来描述内容对象之间的关系，这里对象可以是概念、方法或结果，其中，圆形节点是根节点，椭圆节点是最为关键或重要的，无形节点表示一个笼统的概念。节点之间的关系 $a \rightarrow b$ 表示“ b 是 a 的一个子类（或一部分）”； $a \rightarrow b$ 表示“对象 b 基于对象 a ”，或者“对象 b 由对象 a 诱导（或定义）出”，其中 a, b 之间没有类的包含关系。这两类关系将对象联系起来，形成了一个知识图谱（knowledge graph），描绘出整个学科的全貌。

例如，上图中“Borel 可测空间”是一类“可测空间”；“独立性”由“条件概率”定义，但它不是一类“条件概率”。

1.1 古典概率模型

赌博的历史几乎与人类文明史一样久远^{*}，可为什么对它的数学研究迟滞到十七世纪才开始呢？数学史一般将之归咎于十七世纪前落后的数学符号系统无法应付复杂的组合计算，人们今天所熟悉的代数符号和计数系统直到十六世纪才逐渐成熟。例如，运算符号“+”和“-”大约出现于 1480 年，等号“=”出现于 1557 年，不等号“>”和“<”直至 1631 年才出现。



历史上第一位在其著作中以赌博为应用背景系统地考虑概率计算的学者是 Gerolamo Cardano (1501-1576)，他是文艺复兴时期意大利著名的数学家、物理学家和医生，是个百科全书式的学者，也是个狂热的赌徒和占星术士。他预言了自己的死期并在那天自杀身亡以证明预言不假。

Cardano 在他的著作《论赌博游戏》(1663) 中，着重考虑了掷骰子的概率问题。以掷两个均匀骰子为例，Cardano 明确意识到所有可能的结果是 36 个有序对 (i, j) ，其中 $i, j = 1, 2, \dots, 6$ ，而不是 21 个无序对。并且，Cardano 认为每个有序对出现的机会都是等同的，即都是 $1/36$ 。

	■	■	■	■	■	■
■	■ ■	■ ■	■ ■	■ ■	■ ■	■ ■
■	■ ■	■ ■	■ ■	■ ■	■ ■	■ ■
■	■ ■	■ ■	■ ■	■ ■	■ ■	■ ■
■	■ ■	■ ■	■ ■	■ ■	■ ■	■ ■
■	■ ■	■ ■	■ ■	■ ■	■ ■	■ ■

科学之父、意大利数学家、物理学家、天文学家、哲学家 Galileo Galilei (1564-1642) 曾被赌徒求教连续掷三次骰子，点数之和为 9 和为 10 的哪个概率大。赌徒们认为它们相等，因为点数之和为 9 的组合情况与点数之和为 10 的组合情况一样多，分别是

$$(1, 2, 6), (1, 3, 5), (1, 4, 4), (2, 2, 5), (2, 3, 4), (3, 3, 3)$$

$$(1, 3, 6), (1, 4, 5), (2, 2, 6), (2, 3, 5), (2, 4, 4), (3, 3, 4)$$

考虑三个骰子的顺序，共有 $6^3 = 216$ 个基本事件。Galileo 发现点数组合 $(3, 3, 3)$ 的情形只有 ■■■ 一种排



^{*}五千年前，骰子在印度就出现了。那时的骰子一般由石头或动物骨头制成，用于赌博和算命。

列，而 $(2, 2, 5)$ 则有 $\square\blacksquare\square$, $\square\square\blacksquare$, $\blacksquare\square\square$ 三种排列。Galileo 得到点数之和为 9 和 10 的分别有 25 和 27 种排列情形，于是其概率分别是 $25/6^3 \approx 11.6\%$ 和 $27/6^3 \approx 12.5\%$ 。具体情况如下，

点数之和为 9		点数之和为 10	
组合	数目	组合	数目
1 2 6	6	1 3 6	6
1 3 5	6	1 4 5	6
1 4 4	3	2 2 6	3
2 2 5	3	2 3 5	6
2 3 4	6	2 4 4	3
3 3 3	1	3 3 4	3
总计	25	总计	27

Cardano 和 Galileo 选对了基本事件集合，这是一个了不起的认识，要知道两百年后法国知名的数学家 Jean le Rond d'Alembert (1717-1783) 也曾在下面的简单例子上犯错。

例 1.4. 一枚均匀的硬币连续抛两次，出现正面的次数可能是 0, 1 或 2, d'Alembert 认为出现这三个结果的机会等同。果真如此吗？

解. 书中约定用 H 表示正面，用 T 表示反面，该例的基本事件集合是

$$\Omega = \{(T, T), (T, H), (H, T), (H, H)\}$$

出现 0, 1 和 2 次正面的随机事件用 Ω 的子集分别表示为 $A_0 = \{(T, T)\}$, $A_1 = \{(T, H), (H, T)\}$ 和 $A_2 = \{(H, H)\}$ ，其出现的机会分别是 $1/4$, $1/2$ 和 $1/4$ 。

 在这个例子中假定硬币是均匀的，所以我们有理由认为每个基本事件出现的机会都是等同的。为刻画随机事件 A 出现的机会，即 A 的概率 $P(A)$ ，我们用构成随机事件 A 的基本事件占整个基本事件集合 Ω 的几成来描述。显然，

$$0 \leq P(A) \leq 1, \text{ 其中 } A \subseteq \Omega$$

$$P(\emptyset) = 0, \text{ 并且 } P(\Omega) = 1$$

基本事件并不要求机会等同，如抛一枚不均匀的硬币，出现正面和反面的机会不等。再如，下图所示的转盘的例子。

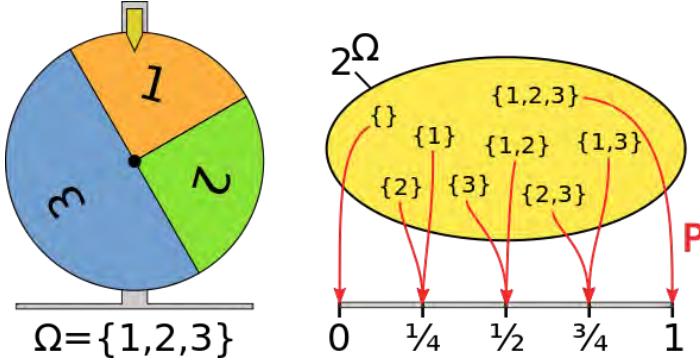


图 1.3: 转盘的基本事件集合 $\Omega = \{1, 2, 3\}$, 所有事件集合 2^Ω , 以及每个事件的概率。

除了硬币、骰子、转盘，在描述概率问题时，我们还常用到这样的道具——若干个球和一些带标号的盒子。球的状态有以下两种情形：

1. 无标号或不可分辨 (indistinguishable): 即假定这些球的大小、颜色、质地等物理特性都一样，无法对它们进行区分。
2. 可分辨的: 即可以通过球的颜色或标号对它们进行区分。有时颜色和标号都要用到，譬如约定“ n 个可分辨的黑球”意味着 n 个黑球的标号分别为 $1, 2, \dots, n$ 。

例 1.5. 随机地把两个球放入两个盒子，基本事件集合因盒子和球是否可辨而异。下面分别讨论球和盒子可分辨和不可辨时的基本事件集合 Ω 。

球 盒子	可辨	不可辨
可辨	$\Omega = \{(ab -), (a b), (b a), (- ab)\}$	$\Omega = \{[ab -], [a b]\}$
不可辨	$\Omega = \{(* * -), (* *), (- * *)\}$	$\Omega = \{[* * -], [* *]\}$

□ 随机选取 n 个人，其生日的情况相当于把 n 个可辨的球随机地放入 $N = 365$ 个可辨的盒子（一年有 365 天）。

□ 反复掷骰子 n 次，（不计次序）所掷的点数情况相当于把 n 个不可辨的球随机地放入 $N = 6$ 个可辨的盒子（骰子有六个点数）。

练习 1.4. 把 n 个球随机地放入标号为 $1, \dots, N$ 的 N 个盒子，球可分辨和不可分辨时各有多少个可能的结果？请给出 $N = n = 3$ 时各自的结果。

答案：若球可分辨，每个球有 N 个选择，共有 N^n 个可能的结果。若球不可分辨，原问题相当于从 $N + n - 1$ 个位置中选出 $N - 1$ 个位置放隔板，其余位置放球。所有可能结果的总数是 C_{N+n-1}^n ，也是不定方程 $x_1 + \dots + x_N = n$ 的非负整数解的个数（另一求解方法见例 1.15）。特别地，当 $N = n = 3$ 时，结果分别是 27 和 10 个。

因为有关随机试验 \mathcal{E} 的每个随机事件都应表示为基本事件集合 Ω 的某个子集的形式，按照严格的写法，基本事件应该记为 $\{\omega\} \subseteq \Omega$ 。在不引起歧义的情况下，基本事件 $\{\omega\}$ 有时也记作 $\omega \in \Omega$ 。

例 1.6. 若随机试验 \mathcal{E} 由两个步骤组成——抛一次硬币后再掷一次骰子，则基本事件集合为 $\Omega = \Omega_1 \times \Omega_2$ ，其中 $\Omega_1 = \{H, T\}$ 和 $\Omega_2 = \{1, 2, \dots, 6\}$ 分别是抛硬币和掷骰子的基本事件集合， Ω 是 Ω_1 和 Ω_2 的笛卡尔积 (Cartesian product)。

推而广之，如果随机试验 \mathcal{E} 由 n 个步骤组成（例如连续抛一枚硬币 n 次），步骤 k 即随机试验 \mathcal{E}_k ，所对应的基本事件集合为 $\Omega_k, k = 1, 2, \dots, n$ ，则 \mathcal{E} 的基本事件集合 Ω 是 $\Omega_1, \Omega_2, \dots, \Omega_n$ 的笛卡尔积。

$$\begin{aligned}\Omega &= \Omega_1 \times \Omega_2 \times \dots \times \Omega_n \\ &= \{(\omega^{(1)}, \omega^{(2)}, \dots, \omega^{(n)}): \omega^{(k)} \in \Omega_k, k = 1, 2, \dots, n\}\end{aligned}\quad (1.1)$$

如果基本事件集合 Ω_k 的势 (cardinality) 皆有限，不妨设 $|\Omega_k| = m_k < \infty$ ，则由式 (1.1) 得知基本事件集合 Ω 的势为

$$|\Omega| = \prod_{k=1}^n |\Omega_k| = \prod_{k=1}^n m_k$$

例 1.7. 抛一次硬币的基本事件集合是 $\tilde{\Omega} = \{H, T\}$ ，则连续抛该枚硬币 n 次的基本事件集合 Ω 是 n 个 $\tilde{\Omega}$ 的笛卡尔积。

$$\Omega = \underbrace{\tilde{\Omega} \times \dots \times \tilde{\Omega}}_{n \uparrow}$$

上式有时也简记为 $\Omega = \tilde{\Omega}^n$ 。任一基本事件 $\omega \in \Omega$ 都是 H, T 组成的长度为 n 的序列。 $n = 3$ 的情形见第 10 页的练习 1.2。

定义 1.1. 如果随机试验 \mathcal{E} 的每个基本事件 $\omega \in \Omega$ 发生的机会都是等同的，我们称之为古典概率问题，并把解决此类问题的概率模型称为古典概率模型。

Laplace 在《概率的哲学探讨》中总共给出了七条概率论的一般原理，其中原理一是概率的定义，“它是有利情况的个数与所有可能情况个数之比”。整理成集合论的语言，就是下面的性质。

性质 1.1. 对于一个古典概率问题，如果基本事件集合 Ω 是有限的，不妨设其元素个数为 n ，记作 $|\Omega| = n$ 。因为每一基本事件 $\{\omega\}$ 发生的机会等同，所以 $\{\omega\}$ 的概率，记作 $P(\{\omega\})$ ，满足以下性质。

$$\begin{aligned}P(\{\omega\}) &= \frac{1}{n} \\ \sum_{\omega \in \Omega} P(\{\omega\}) &= 1\end{aligned}$$

随机事件 $A \subseteq \Omega$ 的概率是“一个分数，分子是所有有利场合的数目，分母是所有可能场合的数目”(Laplace, 《概率的分析理论》)。

$$P(A) = \frac{|A|}{|\Omega|} = \frac{n_A}{n} = \frac{A \text{ 中基本事件的个数}}{\text{所有基本事件的个数}} \quad (1.2)$$

其中, $|A|$ 和 n_A 表示集合 A 中元素的个数, 有时候也记作 $\sharp(A)$ 。

例 1.8. 接着第 10 页的练习 1.2, 每个基本事件 $\omega \in \Omega$ 发生的机会都是 $1/8$ 。

- 事件 A = “至少抛出一个反面”由 7 个基本事件组成, 即除了 (H, H, H) 的所有基本事件, 因此 A 发生的机会是 $7/8$ 。
- 事件 B = “没有连续出现正面, 并且没有连续出现反面”即事件 $\{(H, T, H), (T, H, T)\}$, 于是 B 发生的机会是 $2/8$ 。

例 1.9. 在例 1.5 中, 当球和盒子都可辨而且被选机会都是 $1/2$ 时, 每个基本事件发生的机会都是 $1/4$ 。当球不可辨而盒子可辨且被选机会都是 $1/2$ 时, 基本事件 $(*)$ 出现的机会是 $1/2$, 其他两个基本事件出现的机会都是 $1/4$ 。

练习 1.5. 在例 1.5 中, 当球和盒子都不可辨时, 基本事件 $(*)$ 出现的机会是多大?
答案: $1/2$ 。

 古典概率要求基本事件发生的机会等同, 限制了概率论的研究对象。即便如此, 在古典概率的研究范畴里, 依然有很多问题难以解决或无法解决 (具体实例见第 33 页的例 1.25 和例 1.26), 请读者体会古典概率的局限性。后文将不断突破古典概率的狭隘, 从公理化的角度抽象地看待随机事件和概率。

对于古典概率, Laplace 解释道, “机会的理论就是把同类的所有事件化归为一定数量的等可能情况, 这里所说的等可能性, 就是对于属同一类的事件, 人们以同等程度不能判定哪个事件会发生, 并且确定在哪些情况下被考虑的事件可能发生。可能发生的次数与总的次数之比就是对该事件出现的概率的一个描述。”

同时, Laplace 还清楚地认识到, “当某一事件在所有试验中都发生, 可能事件就变成必然事件, 它的概率等于 1。在这种情况下, 可能性与确定性变得可比了, 尽管它们之间有着本质的不同……。”[165]

解决古典概率问题有时需要很高的技巧, 正如 Laplace 阐述原理二时所说, “假定所有可能情形都是等可能的, 否则要先确定它们各自的概率 (它们的精确估计正是机会理论中的难点所在), 然后将每种有利情况的概率相加就是所求的概率。”

※例 1.10 (连续三次正面问题). 抛一枚均匀的硬币 $n \geq 3$ 次, 计算事件 H_3 = “至少连续出现 3 次正面”的概率 $P_n(H_3)$ 。

解. 基本事件集合 Ω_n 是 H 和 T 构成的长度为 n 的序列的全体, 所以 $|\Omega_n| = 2^n$, 进而每个基本事件的概率为 2^{-n} 。事件“头五次抛硬币的结果是 $THTTH$ ”是以 $THTTH$ 开头的所有长度为 n 的序列的集合, 简记作 \underline{THTTH} , 它的概率是 2^{-5} 。

当 n 很大时, 通过列举 H_3 的元素来计算概率 $P_n(H_3)$ 是相当繁琐的。例如, $n = 20$ 时 $|\Omega| = 2^{20} = 1048576$, 我们有 $|H_3| = 825259$ 。

事件 H_3 可用图 1.4 所示的剪枝二叉树递归地构造出来: 树中的每个节点代表一个事件, 这些事件两两交集为空, 其并集合就是 H_3 。第 k 层节点的全体表示事件“从第 k 抛开始至少连续出现三次正面”。

用 T_k 表示第 k 层节点的个数, 则 $T_0 = 0, T_1 = T_2 = 1$ 且 $T_k = T_{k-1} + T_{k-2} + T_{k-3}$ 。序列 $T_k, k = 1, 2, \dots$ 被称为 Tribonacci 序列。该序列的前几项是 $0, 1, 1, 4, 7, 13, 24, 44, 81, 149, 274, 504, 927, 1705, 3136, 5768, 10609, 19513, 35890, \dots$, 它们恰是图 1.4 中树的各层节点数。显然, 所求的概率是

$$P_n(H_3) = \sum_{k=0}^{n-2} 2^{-2-k} T_k = \frac{1}{4} \sum_{k=0}^{n-2} \left(\frac{1}{2}\right)^k T_k$$

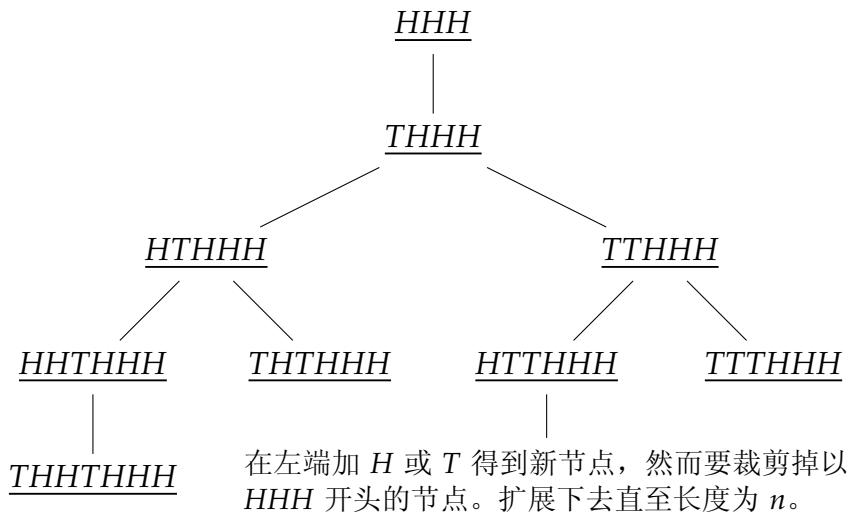


图 1.4: 树的高度为 $n - 2$, 根节点(即第 1 层节点)表示一开始就连抛三个正面, 第 k 层节点表示事件“从第 k 抛开始至少连续出现三次正面”。

不难验证 Tribonacci 序列具有如下性质:

$$\sum_{k=0}^{\infty} x^k T_k = \frac{x}{1 - x - x^2 - x^3}$$

利用该性质轻而易举地就能论证当 $n \rightarrow \infty$ 时, $P_n(H_3) \rightarrow 1$ 。即抛的次数越多, 事件 H_3 发生的机会就越大, 这几乎是显而易见的!

虽然某些古典概率问题需要解题技巧, 但本书并不强调这些星星点点的技巧,

因为它们多是初等数学的内容。使用低级工具炫耀技巧，远不如引入高级工具来得重要。所以，作者更愿意通过古典概率模型来强调概率思想，一方面为了展示概率论历史发展的原貌，把古典概率描述为通向概率公理化的必经之路，另一方面也为了使现代概率论的基本概念的引入显得更自然些。

概率论公理化之前所谓的“古典概率时期”的历史见数学史专著 [66, 67]，也可参阅数学科普名著《数学——它的内容，方法和意义》[4] 的第十一章《概率论》，或者《十九世纪的数学：数理逻辑、代数、数论和概率论》[95] 的第四章，分别由数学大师 Kolmogorov 和他的学生 Gendenko 撰写。

值得注意的是，古典概率模型无法处理由无穷次重复操作构成的随机试验，如抛一枚硬币无穷次。即便能形式地刻画基本事件，但囿于对概率的认识，古典概率模型也无法计算“至少连续出现 $t < \infty$ 次正面”的概率。数学上必须以测度论 [68] 为基础，才能真正建立起概率论的大厦。

回顾数学史，新工具、新方法往往带来更高程度的抽象，甚至改变数学分支的命运。举个例子，因为有了解析几何，古希腊数学家对圆锥曲线的研究终被历史尘封。然而，具体成果可以被遗忘，思想的延续性却不应被割裂。没有那些对圆锥曲线的系统研究，也不会诞生解析几何。类似地，如果不经历古典概率并深刻体会到它的局限性，新工具的引入也就成了无稽之谈。

本节内容

第一、二小节分别就基本事件集合 Ω 为离散的和连续的两种情形讨论了古典概率模型，前者归于排列、组合的方法，后者侧重几何概率问题及其引发的对古典概率局限性的思考。我们总结了古典概率的三条性质，为概率论的公理化积累经验。第三小节介绍了古典概率的一个重要应用——Monte Carlo 方法，借助随机模拟帮助读者粗略认识概率的频率解释。第四小节是对随机性的一些哲学思考，特别讨论了量子力学的概率解释，通过隐 Markov 模型的直观描述让读者对随机现象背后的随机性略窥一斑。

关键知识

- (1) 随机试验的基本事件集合；(2) 计数概率的“球-盒子”模型；(3) 几何概率；(4) Bertrand 悖论；(5) 随机模拟。

1.1.1 计数概率

基本事件集合 Ω 有限的这类古典概率问题，实质上就是用排列组合的方法确定 Ω 和所关心的事件 A 的势，所需的技巧也仅限于计数。很多不同应用背景的概率问题都可以化归到“球-盒子”模型加以讨论 [82]，这些简单的道具使得很多貌似不同、本质上同类的问题“原形毕露”。譬如，掷 m 个骰子的试验相当于把 m 个球随机放入 6 个盒子^{*}，再如下面的例子。

例 1.11. 把 k 个球放入 n 个盒子 ($n \geq k$)，每个盒子至多装一个球。

- 若球是可分辨的，则共有 $A_n^k = n(n-1)\cdots(n-k+1)$ 种放法（排列数）。
- 若球是不可分辨的，则共有 $C_n^k = A_n^k/k! = n!/[k!(n-k)!]$ 种放法（组合数），它是 $(1+x)^n$ 二项式展开中 x^k 的二项式系数，有时记作 $\binom{n}{k}$ 。

例 1.11 的事实也可以这样讲：从标号为 $1, 2, \dots, n$ 的球中随机无放回地抽取 k 个球，所有可能序列（或集合）的个数是 A_n^k （或 C_n^k ）。如果是有放回地抽取，所有可能序列（或集合）的个数是 n^k （或 $C_n^1 + C_n^2 + \dots + C_n^k$ ）。

性质 1.2. 在古典概率论中，最常用的组合公式有

$$C_n^k = C_n^{n-k} \quad (1.3)$$

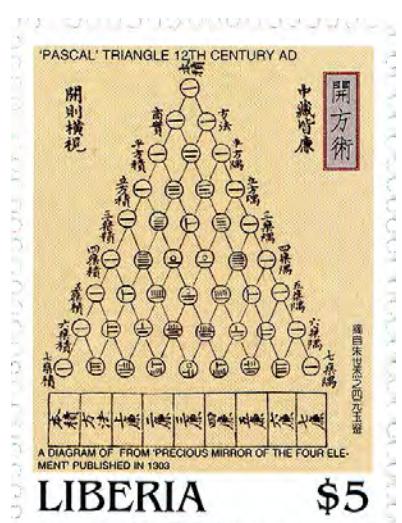
$$C_n^k = C_{n-1}^k + C_{n-1}^{k-1} \quad (1.4)$$

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \quad (1.5)$$

右图是元代数学家朱世杰 (1249-1314) 在《四元玉鉴》卷首绘制的《古法七乘方图》(1303 年)。早在十一世纪，北宋数学家贾宪就已发现结果 (1.4)。式 (1.5) 就是著名的 Stirling 公式[†]，其重要性是把阶乘 $n!$ 作了同阶的简化，常用于近似计算。Stirling 公式基于以下事实。

$$\lim_{n \rightarrow \infty} \frac{n!}{\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}} = 1$$

例 1.12 (生日问题). 假设一年有 365 天，随机选取 n 个人 ($n \leq 365$)，问至少两人生日相同的概率是多少？



*当提到“ n 个盒子”，除非明文规定它们是不可分辨的，读者一般可以缺省地认为盒子是可分辨的，即盒子的标号分别是 $1, 2, \dots, n$ 。

[†]Stirling 公式的真正发现者是 A. de Moivre，用于正态分布密度函数的推导，详见附录 A。

解. 令 A 表示 “ n 个人当中至少两人生日相同”, 先考虑随机事件 A^c , 即 “ n 个人当中没有人生日相同”, 翻译成 “球-盒子” 模型等价于: 把标号为 $1, 2, \dots, n$ 的 n 个球放入 $N = 365$ 个盒子, 每个盒子至多装一个球, 共有 A_N^n 种放法 (参见例 1.5 和例 1.11)。

而基本事件就是把 n 个球放入 N 个盒子, 共有 $|\Omega| = N^n$ 种放法 (见练习 1.4)。所以 $P(A^c) = A_N^n / N^n$, 进而原问题的解是

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - \frac{A_N^n}{N^n} \end{aligned}$$

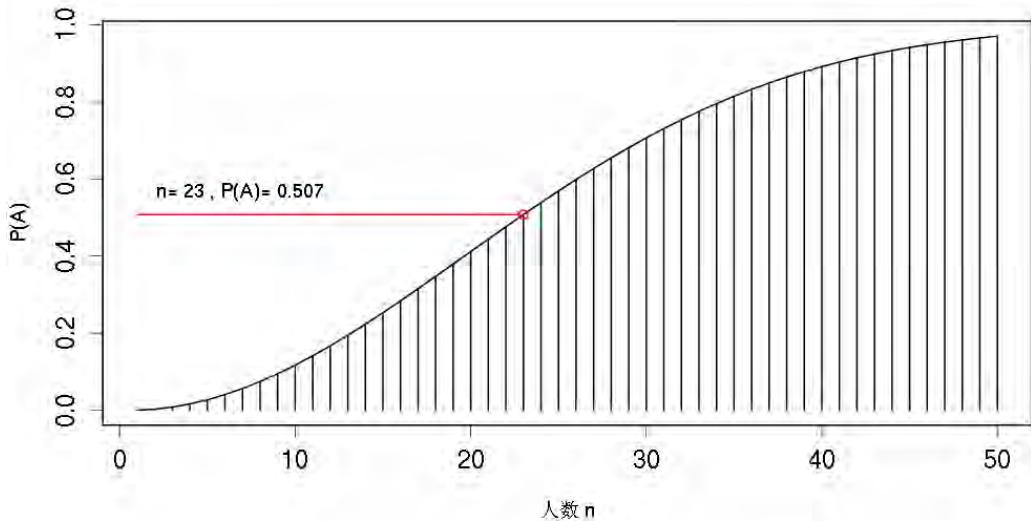


图 1.5: 在随机选取的 $n = 1, 2, \dots, 50$ 个人中至少两人生日相同的概率 $P(A)$ 。其中, $n = 23$ 是使得 $P(A) > 50\%$ 所需最少的人数。对于 $n = 50$, 概率 $P(A)$ 飙升至 97%! 联想到组合数学里的抽屉原则, 这个结果多多少少有点出乎意料。

例 1.13. 盒子里有 m 个球, 由 mp 个可分辨的黑球和 mq 个可分辨的白球组成, 其中 $p, q \in (0, 1)$ 且 $p+q = 1$ 。一次随机抽取一个球, 有放回地抽取 n 次, 试证明: 事件 $A_k =$ “恰有 k 次抽到黑球” 的概率恰是 $(p+q)^n$ 的二项式展开的第 $k+1$ 项, 即

$$P(A_k) = C_n^k p^k q^{n-k}, \text{ 其中 } k = 0, 1, \dots, n$$

证明. 随机抽取一次的基本事件集合是

$$\Omega = \{\text{黑}_1, \dots, \text{黑}_{mp}, \text{白}_1, \dots, \text{白}_{mq}\}$$



所以 $|\Omega| = m$, 有放回地抽取 n 次的基本事件集合是 Ω^n 且 $|\Omega^n| = m^n$ 。

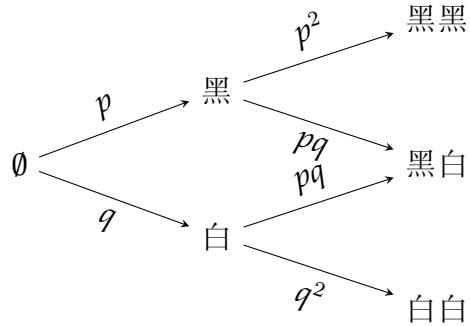


图 1.6: 到达结果 “ k 个黑球 ($n - k$ 个白球)” 的不同路径的总数是 C_n^k , 每条路径的概率都是 $p^k q^{n-k}$ 。

记集合 $\Omega_{\text{黑}} = \{\text{黑}_1, \dots, \text{黑}_{mp}\}$ 且 $\Omega_{\text{白}} = \{\text{白}_1, \dots, \text{白}_{mq}\}$ 。有序 n 元组 (t_1, \dots, t_n) 中有 k 个元素取“黑”, 有 $n - k$ 个元素取“白”, 共有 C_n^k 个无重复的 (t_1, \dots, t_n) 。事件 A_k = “恰有 k 次抽到黑球” 可分解为

$$A_k = \bigcup_{(t_1, \dots, t_n)} \Omega_{t_1} \times \dots \times \Omega_{t_n}, \text{ 由式 (1.2) 可得}$$

$$P(A_k) = \frac{|A_k|}{|\Omega^n|} = \frac{C_n^k (mp)^k (mq)^{n-k}}{m^n} = C_n^k p^k q^{n-k} \quad \square$$

例 1.13 中若给定抽取次数 n 和黑球比例 p , 则 $P(A_k)$ 是有关 $k = 0, 1, \dots, n$ 的函数, 与盒子中球的个数无关, 我们将之简记为 $P(k) \sim B(n, p)$ 。

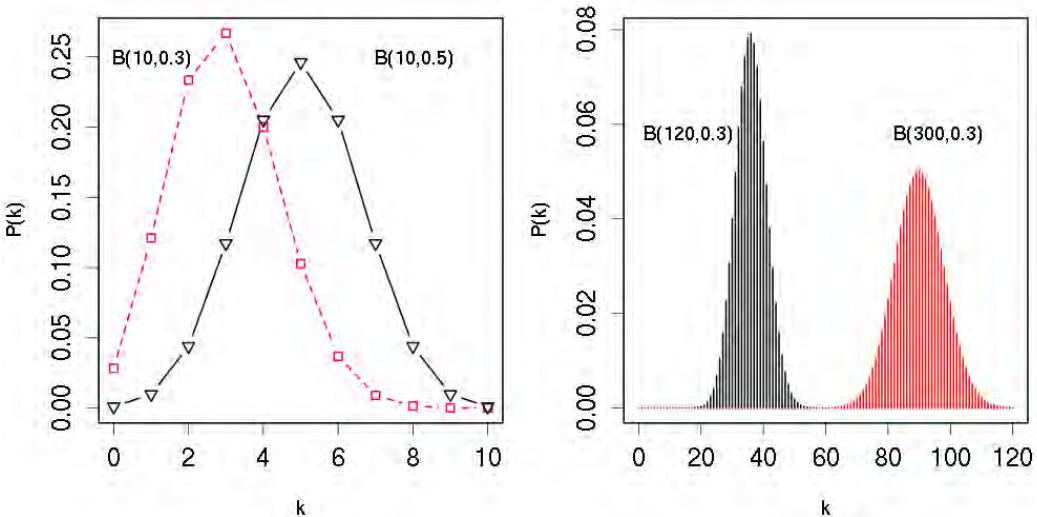


图 1.7: 左图是 $n = 10$, p 分别取 0.3 和 0.5 所对应的 $P(k) = C_n^k p^k (1 - p)^{n-k}$ 的折线图。右图是 $p = 0.3$ 而 n 很大时 $P(k)$ 的竖线图。

若 $p = 0.5$, 概率 $P(k) = C_n^k p^k (1-p)^{n-k}$ 关于实数 np 是对称的; 否则, $P(k)$ 是非对称的, 如图 1.7 中的左图所示。

由 $(p+q)^n = 1$ 不难推出 $\sum_{k=0}^n P(k) = 1$, 即所有函数值之和等于 1。或者说, 在 $P(k)$ 的竖线图中, 所有竖线长度之和归一。

然而, 当 n 很大时, 即便 $p \neq 0.5$, 概率 $P(k)$ 关于 np 也呈现出很好的对称性, 见图 1.7 中右边的竖线图。这是为什么呢? 等读者学习完第 5 章的 de Moivre-Laplace 中心极限定理就明白其中的道理了。

例 1.14. 已知 N 件产品中有 n 件次品, 其余为正品。令事件 A_k = “在随机抽取的 $m \leq n$ 件产品中恰有 k 件次品”, 试求概率 $P(A_k)$ 。

解. 翻译成球-盒子模型: 盒子里有 n 个黑球和 $N-n$ 个白球, 随机抽取 $m \leq n$ 个球, 求事件 A_k = “恰有 k 个黑球”的概率。

从 N 个球中取出 m 个球共有 C_N^m 种取法, 它是基本事件集合的势。从 n 个黑球中选出 k 个黑球有 C_n^k 种选法, 在剩下的 $N-n$ 个白球中选出 $m-k$ 个白球有 C_{N-n}^{m-k} 种选法, 于是事件 A_k 的势为 $C_n^k C_{N-n}^{m-k}$ 。因此, 事件 A_k 的概率为

$$P(A_k) = \frac{C_n^k C_{N-n}^{m-k}}{C_N^m}$$

例 1.15 (统计物理模型). 有 n 个粒子处于 N 个不同的能级上 ($n \leq N$), 求: (1) 在指定的 n 个能级上各有一个粒子的概率 p_1 , (2) 恰有 n 个能级各有一个粒子的概率 p_2 。

解. 这个问题相当于将 n 个球放入 N 个盒子中, 参考练习 1.4, 下面分别给出基于三个不同假设的粒子物理模型。

□ Maxwell-Boltzmann 模型: 如果粒子可分辨, 参考例 1.12, 有

$$p_1 = \frac{n!}{N^n} \quad p_2 = \frac{A_N^n}{N^n}$$

该模型用于描述处于热力学平衡状态下大量原子按能量的分布。但在微观粒子世界里, 这 N^n 个基本事件并非等概率的。Maxwell-Boltzmann 模型对于任何微观粒子都是不适用的, 原因是“粒子可分辨”的假设不成立。

□ Fermi-Dirac 模型: 如果粒子不可分辨, 且每个能级只能有一个粒子, 则

$$p_1 = \frac{1}{C_N^n} \quad p_2 = 1$$

该模型适合用来描述费米子, 即遵循 Pauli 不相容原理的粒子, 如中子、质子、电子等。

□ Bose-Einstein 模型：如果粒子不可分辨，根据练习 1.4 的结果，共有 C_{N+n-1}^n 种放法。下面是另外一种解法：从左至右以第一个盒子为首，把剩下的 $N-1$ 个盒子和 n 个球随机地排放在第一个盒子之后，然后将两个盒子之间的球都放入左边的盒子。这样的随机试验共有 $C_{N+n-1}^{N-1} = C_{N+n-1}^n$ 个不同的结果。



Bose-Einstein 模型认为这 C_{N+n-1}^n 个基本事件是等概率的（该假设与数学推导相悖，见例 1.9），于是

$$p_1 = \frac{1}{C_{N+n-1}^n} \quad p_2 = \frac{C_N^n}{C_{N+n-1}^n}$$

该模型适用于玻色子，即除费米子外所有的粒子，如光子、介子、胶子等。

例 1.15 说明概率模型和物理世界是有差别的，数学上讲得通的在现实中不见得行得通 [45]，Bose-Einstein 模型就是个例子。究其原因，也许是某个物理属性在概率建模中没有考虑到，大自然以某个人类尚不了解的方式操纵这些微观粒子。

例 1.16. 一个盒子里有 10 个球，标号分别为 $0, 1, \dots, 9$ 。一次随机抽取一个球，有放回地抽取 6 次，试求事件 $A_m = \text{“所抽标号之和为 } m\text{”}$ 的概率，其中 $m \in \mathbb{N} \cup \{0\}$ 。

解. 每个基本事件都形如 (x_1, \dots, x_6) ，共有 10^6 个基本事件，其中 x_j 表示第 j 次所抽取标号，满足约束条件 $0 \leq x_j \leq 9$ 。事件 A_m 所含基本事件的个数等于下述不定方程的非负整数解的个数。

$$\sum_{j=1}^6 x_j = m, \text{ 其中 } 0 \leq x_j \leq 9$$

不难看出，解的个数就是多项式 $(1 + x + x^2 + \dots + x^9)^6$ 中项 x^m 的系数。实践中，该系数可利用符号计算工具算得。例如， A_{25} 所含基本事件的个数为 53262，进而 $P(A_{25}) = 0.053262$ 。显然，当 $m > 54$ 时 $P(A_m) = 0$ ，其他取值情况见图 1.8。请读者说明为何它们关于 $m = 27$ 呈现出对称性。

1801 年，数学英雄 Leonhard Euler (1707-1783) 首先研究了多项式 $(1 + x + x^2 + \dots + x^k)^n$ 中 x^m 的系数，不妨将之记作 $\binom{n}{m}_{k+1}$ ，称作多项式系数，即

$$(1 + x + x^2 + \dots + x^k)^n = \sum_{m=0}^{kn} \binom{n}{m}_{k+1} x^m$$



令 $x = 1$, 由上式立即可得

$$\sum_{m=0}^{kn} \binom{n}{m}_{k+1} = (k+1)^n$$

多项式系数是二项式系数的推广, 可以按照如下方法递归计算 (留作习题), 或者第 102 页的式 (1.39)。

$$\binom{n}{m}_{k+1} = \sum_{i=0}^{\lfloor \frac{k-1}{k}m \rfloor} \binom{n}{m-i} \binom{m-i}{i}_k, \text{ 其中 } k = 0, 1, 2, \dots \quad (1.6)$$

多项式系数 $\binom{n}{m}_{k+1}$ 的直观含义是下述不定方程的非负整数解的个数。

$$\sum_{j=1}^n x_j = m, \text{ 其中 } 0 \leq x_j \leq k$$

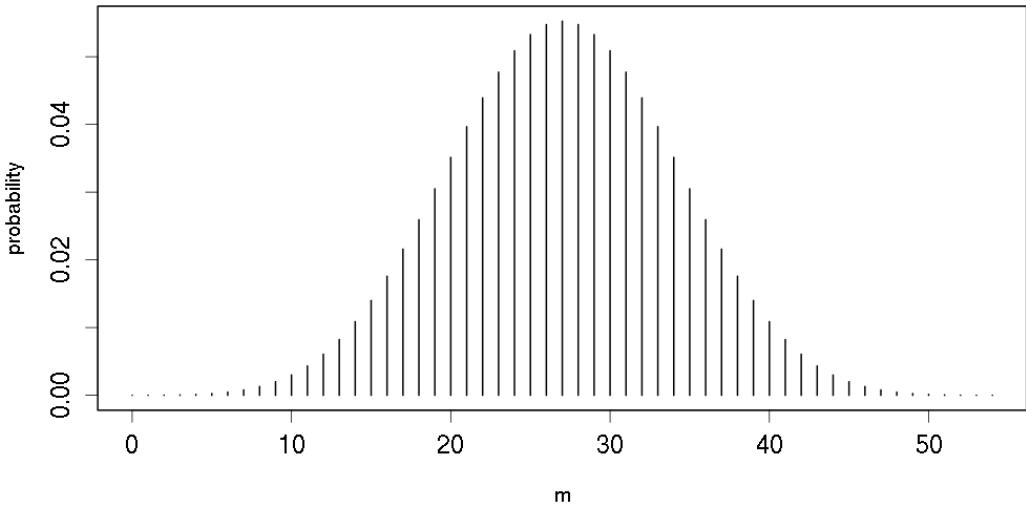


图 1.8: 有放回地抽取标号为 $0, 1, \dots, 9$ 的球 $n = 6$ 次, 标号之和为 $m = 0, 1, \dots, 54$ 的概率 “中间高两头低”, 关于 $m = 27$ 呈现出对称性 (见例 1.16)。

定义 1.2 (互斥). 如果随机事件 $A, B \subseteq \Omega$ 满足 $A \cap B = \emptyset$, 则 A 与 B 不能同时发生, 称它们是互斥的或不相容的。显然, 事件 A, A^c 是互斥的, 而且必有一个发生。

性质 1.3. 若事件 A, B 互斥, 则 $P(A \cup B) = P(A) + P(B)$ 。

证明. 因为基本事件集合 Ω 是有限的, 所以

$$P(A \cup B) = \frac{|A \cup B|}{|\Omega|} = \frac{|A| + |B|}{|\Omega|} = P(A) + P(B) \quad \square$$

例 1.17. 在**例 1.14** 中, 事件 A_0, A_1, \dots, A_m 两两互斥, 并且其并集就是基本事件集合。由 $\sum_{k=0}^m P(A_k) = 1$ 这一事实可“顺手牵羊”得到如下结果。

$$\sum_{k=0}^m C_n^k C_{N-n}^{m-k} = C_N^m \quad (1.7)$$

练习 1.6. 试证明: 对于任一自然数 $n \in \mathbb{N}$, 皆有

$$\sum_{k=0}^n (C_n^k)^2 = C_{2n}^n$$

提示: 利用式 (1.7), 令 $N = 2n, m = n$ 即得。或者用“球-盒子”模型来证明。

※例 1.18 (连续正面问题). 把**例 1.10** 一般化: 抛一枚均匀的硬币 n 次, 计算事件 $H_t =$ “至少连续出现 $t \leq n$ 次正面”的概率 $P_n(H_t)$ 。

解. 若 $t = n$, 显然 $P_n(H_n) = 2^{-n}$ 。若 $t < n$, 事件 H_t 是以下两个事件的非交并: (1) 前 $n - 1$ 抛中 H_t 发生, 概率为 $p_{n-1} = P_{n-1}(H_t)$; (2) 在第 n 抛后 H_t 才发生, 即前 $n - t - 1$ 抛中 H_t 不发生, 第 $n - t$ 抛是反面, 最后 t 抛都是正面。由此得到如下线性递归关系:

$$p_n = p_{n-1} + 2^{-t-1}(1 - p_{n-t-1}), \text{ 满足初始条件 } p_0 = p_1 = \dots = p_{t-1} = 0, p_t = 2^{-t}$$

这个递归式有解, 不过表达式过于复杂, 读者可用符号计算工具来了解 p_n 的解析表达式。当 $t = 3, 4, 5$ 时, 我们列举几个结果如下。

表 1.2: 抛一枚均匀的硬币 n 次, 至少连续出现 t 次正面的概率。

n	20	30	40	50	60	70	80	90	100
$t = 3$	0.7870	0.9078	0.9601	0.9827	0.9925	0.9968	0.9986	0.9994	0.9997
$t = 4$	0.4780	0.6391	0.7504	0.8274	0.8807	0.9175	0.9429	0.9605	0.9727
$t = 5$	0.2499	0.3682	0.4679	0.5519	0.6226	0.6821	0.7323	0.7745	0.8101

1.1.2 几何概率

古典概率并不限于有限的基本事件集合，还有连续的情形，譬如下面的投钉问题。

例 1.19 (投钉问题). 向一个固定的正方形区域 Ω 上均匀地投钉，所谓“均匀”意味着钉落在 Ω 上任意一点的机会都等同。试问钉落于一个有面积的子区域 $A \subseteq \Omega$ (图 1.9 中阴影部分) 的概率？

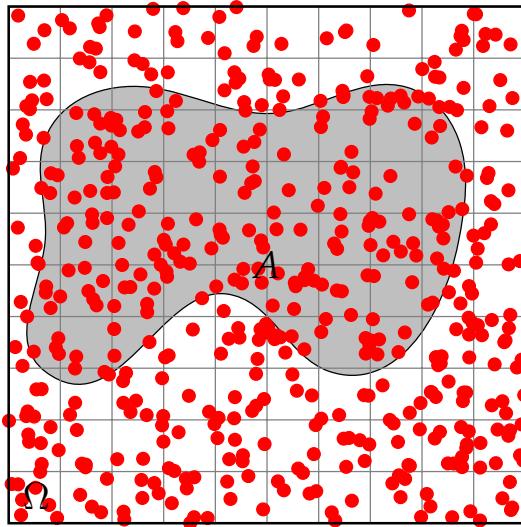


图 1.9: 均匀投钉于正方形 Ω 上，钉落在子区域 A 内的概率是 A 与 Ω 的面积之比。

解. 为求得钉落于子区域 A 的概率 $P(A)$ ，先将问题离散化：把 Ω 等分成 $N(k) = k \times k$ 个小正方形，钉落于每个小正方形的机会都等同。把所有内嵌于 A 的小正方形组成的区域记为 $A_{\text{内}}(k) \subseteq A$ ，其中小正方形的个数记为 $N_{\text{内}}(k)$ ；把那些并集恰好覆盖住 A 的小正方形组成的区域记为 $A_{\text{外}}(k) \supseteq A$ ，其中小正方形的个数记为 $N_{\text{外}}(k)$ ，显然有

$$\frac{N_{\text{内}}(k)}{N(k)} = P[A_{\text{内}}(k)] \leq P(A) \leq P[A_{\text{外}}(k)] = \frac{N_{\text{外}}(k)}{N(k)}$$

已知 A 和 Ω 的面积分别为 $\mu(A)$ 和 $\mu(\Omega)$ ，因此当 $k \rightarrow \infty$ 时，

$$\lim_{k \rightarrow \infty} \frac{N_{\text{内}}(k)}{N(k)} = \lim_{k \rightarrow \infty} \frac{N_{\text{外}}(k)}{N(k)} = \frac{\mu(A)}{\mu(\Omega)}$$

所以钉落于子区域 A 的概率^{*}为

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} \tag{1.8}$$

^{*}§5.1 将介绍 Borel 强大数律，该结果保证通过大量的投钉试验，落于子区域 A 的钉数 m 与落于 Ω 上的总钉数 n 之比 m/n 可作为 $P(A)$ 的估计值。

像这类基本事件集合 Ω 为空间中某一连续区域的古典概率问题被称为几何概率问题。在公式 (1.8) 中, 依据具体情况, $\mu(\cdot)$ 有时也可以表示长度、体积等度量。通常, 几何概率问题按下述步骤解决: (i) 确定基本事件集合 $\Omega \subset \mathbb{R}^n$; (ii) 刻画出随机事件 $A \subseteq \Omega$; (iii) 通过式 (1.8) 计算 A 的概率。

例 1.20 (见面问题). 甲乙二人约定 11 点至 12 点在某地见面, 等候 20 分钟后若对方不来就离开。如果两人可在 11 点至 12 点的任何时刻到达该地 (互不通知或影响对方), 试问两人见面的概率?

解. 以分钟为单位, 基本事件集合 $\Omega = [0, 60] \times [0, 60]$ 。设甲乙分别于 x 和 y 时刻到达, 两人能见面当且仅当 $|x - y| \leq 20$, 所以用集合 $A = \{(x, y) \in \Omega : |x - y| \leq 20\}$ 来表示随机事件“两人见面”。事件 A 即图 1.10 中的阴影部分, 令 $\mu(A)$ 和 $\mu(\Omega)$ 分别为区域 A 和 Ω 的面积。由式 (1.8) 计算得到

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = 1 - \left(\frac{2}{3}\right)^2 = \frac{5}{9}$$

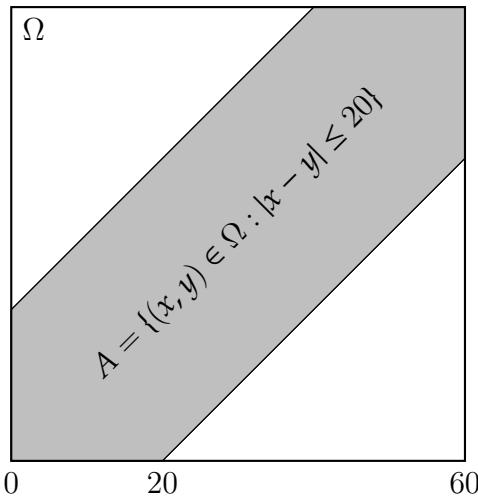


图 1.10: 阴影部分表示例 1.20 中两人能见面的情况, 占所有可能情况的 $5/9$ 。

例 1.21. 在例 1.20 中, 可以把区域 $A = \{(x, y) \in \Omega : |x - y| \leq 20\}$ 分割成可数个子块 A_1, A_2, \dots 如下。

$$A_k = \left\{ (x, y) \in A : \frac{60}{k+1} < x \leq \frac{60}{k}, \frac{60}{k+1} < y \leq \frac{60}{k} \right\}, k = 1, 2, \dots$$

显然, A_1, A_2, \dots 两两互斥且 $\bigcup_{k=1}^{\infty} A_k = A$ 。另外,

$$\sum_{k=1}^{\infty} P(A_k) = \sum_{k=1}^{\infty} \frac{\mu(A_k)}{\mu(\Omega)} = \frac{\sum_{k=1}^{\infty} \mu(A_k)}{\mu(\Omega)} = \frac{\mu(\bigcup_{k=1}^{\infty} A_k)}{\mu(\Omega)} = P\left(\bigcup_{k=1}^{\infty} A_k\right) = P(A)$$

性质 1.4. 对于古典概率问题, 若 Ω 是它的基本事件集合, 则概率满足下面的性质。

1. 非负性: 对于事件 $A \subseteq \Omega$, 总有 $P(A) \geq 0$ 。
2. 归一性: 必然事件 Ω 的概率等于 1, 即 $P(\Omega) = 1$ 。
3. 可列可加性: 如果可数 (或称可列) 个随机事件 $A_1, A_2, \dots, A_k, \dots$ 两两互斥, 则概率满足下面的“可列可加性”。

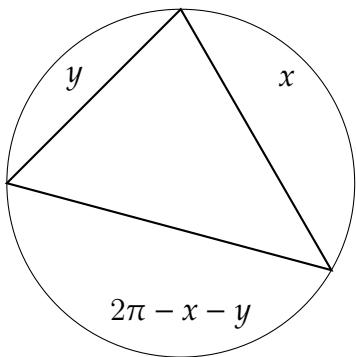
$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$$

证明. 利用式 (1.2) 或式 (1.8) 验证非负性和归一性, 可列可加性可仿照例 1.21。□

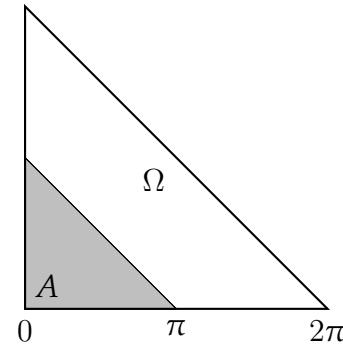
练习 1.7. $0 < k < 1$ 为常数, 在单位长度的线段上任选两点, 试问其距离小于 k 的概率? 答案: $k(2 - k)$ 。

例 1.22. 在单位圆 (半径等于 1 的圆) 的圆周上随机取三个点, 试问: 三个点的连线构成锐角三角形的概率?

解. 三个点将圆周分为三段弧, 其弧长分别为 $x, y, 2\pi - x - y$ 。基本事件集合 $\Omega = \{(x, y) : x > 0, y > 0, x + y < 2\pi\}$, “三个点的连线构成锐角三角形” 对应着区域 $A = \{(x, y) : 0 < x < \pi, 0 < y < \pi, x + y > \pi\}$, 由式 (1.8) 得 $P(A) = 1/4$ 。



(a) 单位圆的内接锐角三角形



(b) 几何概率 $P(A)$

图 1.11: 单位圆上随机三个点构成锐角三角形的概率即几何概率 $P(A) = \mu(A)/\mu(\Omega)$ 。

 在几何概率中, 如果 $\mu(A) = 0$, 则事件 A 的概率为零, 但这并不意味着 A 不会发生。换句话说, 由 $P(A) = 0$ 推导不出 $A = \emptyset$ 。譬如, 在上例中, 三个点的连线“构成直角三角形”的概率为 0。[例 1.20](#) (会面问题) 中, 事件 $A = \{(x, x) : 0 \leq x \leq 60\} \in \mathcal{S}$ 表示“两人同时到达”, 显然 $P(A) = 0$ 。

练习 1.8. 接着上例, 请读者论证, 三个点的连线“构成钝角三角形”的概率为 $3/4$ 。

例 1.23. 已知二次多项式 $f(x) = x^2 + 2ax + b$, 其中系数满足 $|a| \leq A, |b| \leq B$ 。试问: $f(x) = 0$ 有实根的概率 p ?

解. 基本事件集合 $\Omega = [-A, A] \times [-B, B]$ 且 $\mu(\Omega) = 4AB$ 。二次方程 $f(x) = 0$ 有实根当且仅当 $b \leq a^2$, 分下面两种情况讨论 (见图 1.12)。

$$p = \frac{1}{\mu(\Omega)} \left(2AB + 2 \int_0^A a^2 da \right) = \frac{1}{2} + \frac{A^2}{6B} \leq \frac{2}{3}, \text{ 如果 } B \geq A^2$$

$$p = \frac{1}{\mu(\Omega)} \left(4AB - 2 \int_0^B \sqrt{b} db \right) = 1 - \frac{\sqrt{B}}{3A} \geq \frac{2}{3}, \text{ 如果 } B \leq A^2$$

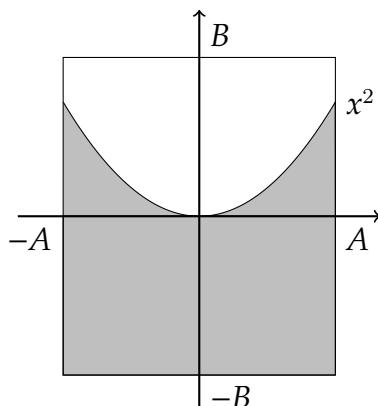
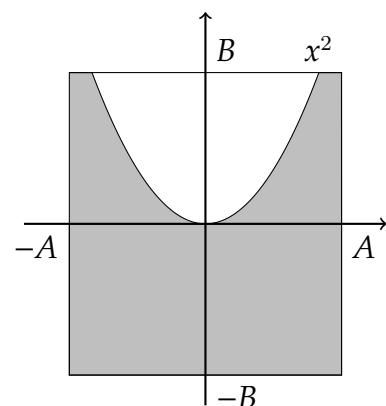
(a) 如果 $B \geq A^2$ (b) 如果 $B \leq A^2$

图 1.12: 满足条件 $|a| \leq A, |b| \leq B$ 的二次多项式 $f(x) = x^2 + 2ax + b$ 具有实根的概率。

在几何概率问题中, 基本事件集合 Ω 是一个无限集合, 而数学里遇到无限的情况通常要倍加小心。历史上对几何概率的批评当数 Bertrand 悖论最为引人瞩目, 它是法国数学家 Joseph Louis François Bertrand (1822-1900) 在其著作《概率计算》(1889) 中构造的, 具体内容见下面的例 1.24。数学史上, 大凡出现悖论, 概念或认识的严谨化就紧随而至*, Bertrand 悖论也终将唤出概率论的公理化 (§1.2 将介绍概率论的 Kolmogorov 公理体系), 促使概率论明确定义“随机事件”及其概率, 从此走上严格化的道路。



*例如, Pythagoras 悖论导致人们对数的认识从有理数域扩展到实数域 (Hippasus 发现 $\sqrt{2}$ 是无理数), Berkeley 悖论促使数学家把分析基础变得更加严密, Russell 悖论所引发的危机让 Cantor 的朴素集合论发展成各种各样公理化的集合论。

*例 1.24 (Bertrand 悖论). 在单位圆内随机取一条弦，试求：该弦长度超过 $\sqrt{3}$ (即单位圆内接正三角形边长) 的概率 p ?

解. 至少有下面三种不同的解法导出的三个不同的答案。读者认为哪种选弦方式最合理？我们将在 §1.2 介绍完概率论的 Kolmogorov 公理体系之后回顾 Bertrand 悖论，进一步讨论它的成因。

1. 若预设了弦的方向，则有唯一的直径垂直于该方向。设该直径的 $1/4$ 和 $3/4$ 分点分别为 E, F ，只有交于线段 EF 的弦才能使弦长超过内接正三角形的边长，见图 1.13 之 (a)。因为 EF 的长度是 1，所以 $p = 1/2$ 。
2. 不妨将弦的一端固定在圆周的某一点 A 上，连同另外两点 B, C ，圆周被三等分。弦的另一端只有离开 A 点 $1/3$ 圆周，即落于图 1.13 之 (b) 中的 \widehat{BC} 弧上，才能使得其长度大于内接正三角形的边长，所以 $p = 1/3$ 。
3. 除了落在圆心上，弦的中点唯一决定了弦的位置。若想让弦长超过 $\sqrt{3}$ ，弦的中点 M 必须落于半径为 $1/2$ 的同心圆内，见图 1.13 之 (c)，此时 $p = 1/4$ 。

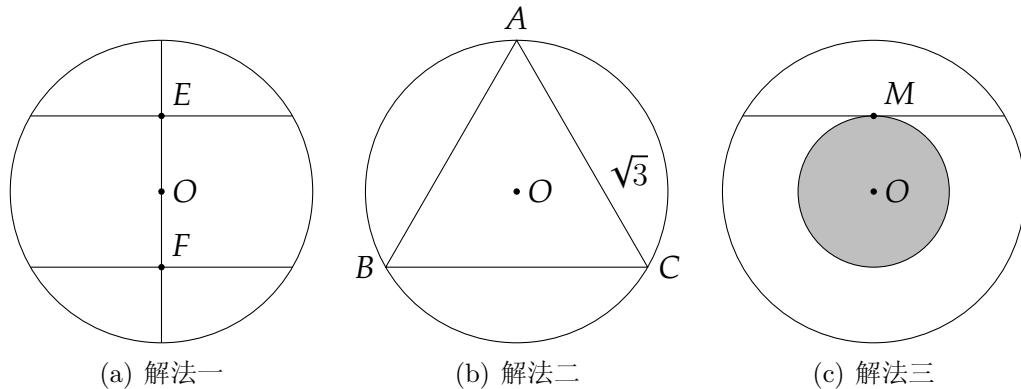


图 1.13: Bertrand 问题三种不同解法的直观图示。所有解法听起来都蛮有道理，但答案只能有一个，到底哪个解法是正确的呢？

Bertrand 悖论是几何概率的隐患，它的“症结”在哪里呢？Laplace 在《概率的哲学探讨》中曾经说过，“一般来说，不同的意见往往产生于对已有的信息的理解和使用上。在概率论中往往考虑一些如此微妙的问题，以至于由同样的信息得到不同的结果也是很平常的，特别对于那些复杂的问题尤为如此。”

通过比较，读者不难发现症结出在对“随机”一词有不同的理解上，也就是说，“在单位圆内随机取一条弦”有歧义，不同的“随机”取弦方式决定了不同的解。上述三种解法的取弦方式分别是：

1. 按照方向区间 $I = [0, 2\pi)$ 上的均匀分布（即 I 上的任意一点被选择的机会都是等同的，均匀分布的严格定义详见 §2.1）随机选定一个角度 θ ，再按照 $(0, 1)$ 上的均匀分布随机选定一个极径长度 ρ ，构造过点 $(\rho \cos \theta, \rho \sin \theta)$ 的弦。
2. 按照弧度区间 $[0, 2\pi)$ 上的均匀分布在圆周上选定两个不同的弧度（即选定两个不同的点），构造以此二点为端点的弦。
3. 按照开圆盘 $D = \{(x, y) : x^2 + y^2 < 1\}$ 上的均匀分布，在 D 内选定一点，构造以此点为中点的弦。

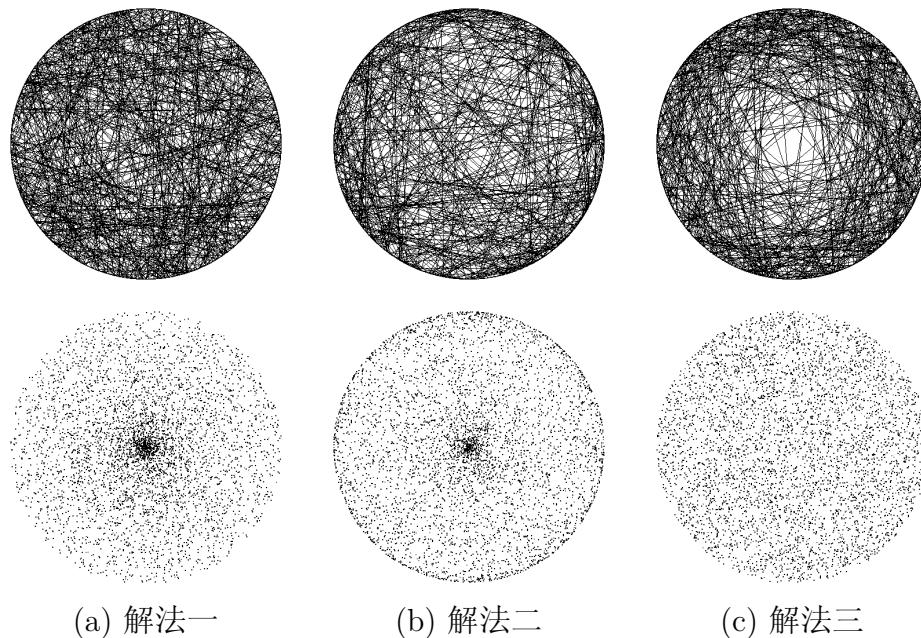


图 1.14: 在 Bertrand 问题中，分别按照三种不同的取弦方式随机生成 $n = 5000$ 条弦（为了能清楚显示，图中只画出了前 $m = 500$ 条弦）。然后，画出每条弦的中点。不难发现，这些中点在圆心周围的疏密程度不同。

除了有悖论的麻烦，古典概率在很多问题上显得力所不及，譬如下面的概率问题。数学家逐渐认识到，要说清楚概率是什么以及如何求得，必须有更先进的工具才行。

例 1.25. 考虑连续抛一枚均匀的硬币直至第一次出现正面的随机试验，其基本事件集合 $\Omega = \{H*, TH*, TTH*, TTTH*, \dots\}$ ，其中 * 是通配符，在这里表示 H, T 构成的长度无穷的任意字符串。显然，每个基本事件的概率都是零。

令 A_n 表示事件“抛了 n 次才第一次出现正面”，则 $A_1 = \{H*\}, A_2 = \{TH*\}, \dots$ 。不难看出，事件 $A_1^c = \{T*\}$ 的概率与事件 A_1 的相同，都是 $1/2$ 。事件 A_2 的概率和事件 $\{HH*\}, \{HT*\}, \{TT*\}$ 的概率相同，都是 $1/4$ 。以此类推，

事件	A_1	A_2	\cdots	A_n	\cdots
概率	$\frac{1}{2}$	$\left(\frac{1}{2}\right)^2$	\cdots	$\left(\frac{1}{2}\right)^n$	\cdots

事件 $A_1, A_2, \dots, A_n, \dots$ 之间两两不交且 $\Omega = A_1 \cup A_2 \cup \dots \cup A_n \cup \dots$, 不难验证 $P(\Omega) = 1$ 。令 A 表示事件“第一个正面出现之前已经抛过偶数次”，其概率为

$$P(A) = \sum_{n=1}^{\infty} P(A_{2n-1}) = \sum_{n=1}^{\infty} \frac{1}{2^{2n-1}} = \frac{2}{3}$$

※例 1.26 (连续正面问题的极限版). 这是例 1.18 的升级版：连续抛一枚均匀的硬币无穷次，计算事件 $H_t = \text{“至少连续出现 } t < \infty \text{ 次正面”}$ 的概率 $P(H_t)$ 。

解. 任一基本事件都是 H 和 T 构成的无穷长度的字符串，把它当作输入，事件 H_t 的概率即是下面的确定有限状态自动机 A_t 停机的概率。

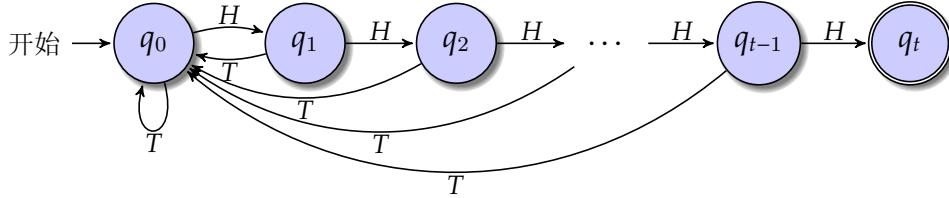


图 1.15: 有限状态自动机 A_t : 输入字母表是 $\Sigma = \{H, T\}$, q_0 是初始状态, q_t 是接受状态(亦称终止状态)。不难发现该自动机具有以下特点: (1)一旦出现 T , 立刻回到初始状态 q_0 。(2)如果当前状态是 $q_k, k = 0, \dots, t$, 则之前的输入恰是 k 个连续的 H 。

基本事件集合 Ω 是 H 和 T 构成的无穷序列的全体。我们把 H, T 无穷序列中的 H 改为 1, 把 T 改为 0, 再在开头添上 0., 则原序列就改造成单位区间 $I = [0, 1]$ 上的一个二进制小数。可惜, 这个改造不是 Ω 和 I 之间的一一对应*, 譬如, 序列 $HHTTT\dots$ 和 $HTHHH\dots$ 都对应着有理数 $3/4$ 。

如果我们能够说明上述这个“翻译”不影响概率计算, 那么 Ω 不妨视为 $I = [0, 1]$ 。令事件 H_t 对应着点集 $A \subset [0, 1]$, 原问题化归成一个几何概率问题, 所求概率 $P(H_t)$ 就是 A 的“长度” $\mu(A)$ 。我们需要测度论这一现代数学的工具才能算出 $\mu(A)$, 例 1.26 的问题尚未解决, 这儿先留下一个伏笔, 待工具备好后我们再来。

“如果未成功解决一个数学问题, 原因往往是我们没能认识到更一般的观点, 在该观点之下, 当前问题知识一串相关问题中的某个环节。在找到该观点之后, 不仅问题更容易解决, 同时我们还获得能应用于相关问题的普遍方法。”

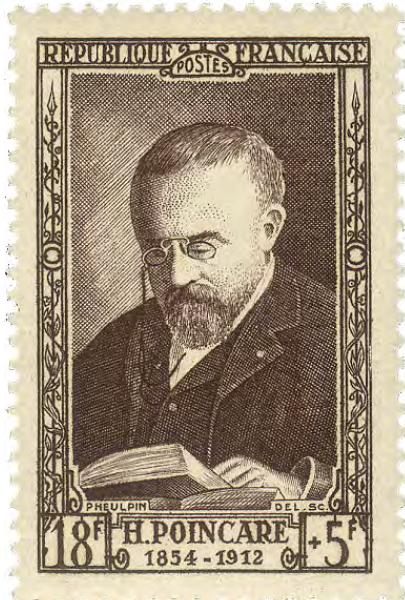
— 1900 年, Hilbert 在国际数学家大会上的报告《数学问题》

*无理数的二进制表示是惟一的。然而对于任意非零有理数, 都有两个二进制表示。例如 0.5 可以表示为二进制小数 0.1 或 0.011…。通常为了表示的唯一性, 约定不能用有限小数。

练习 1.9. 把例 1.25 中的 H, T 字符串“翻译”成 $[0, 1]$ 上的二进制小数，解释为何 $P(A_1) = P(A_1^c) = 1/2, P(A_2) = 1/4, \dots$ 。

法国数学大师、理论物理学家、哲学家 Jules Henri Poincaré (1854-1912) 是数学史上最后一位通才，研究工作涉猎几乎所有的数学分支，包括概率论。Poincaré 写过一本《概率计算》的书，不过没有流行起来。在哲学著作《科学与假设》一书的第十一章 [122]，Poincaré 讨论了 (1) 概率问题的分类，(2) 数学中的概率，(3) 物理学中的概率，(4) 随机游戏，(5) 原因的概率，(6) 误差理论等话题，他谈到了 Bertrand 悖论并对概率缺乏严格的定义表示担忧，也明显地意识到了主观概率和客观概率的区别。Poincaré 也谈到了公理化，“为了进行概率计算，甚至要使这计算有个意义，我们必须承认某个略带主观性的假设或者公约作为出发点。”

Poincaré 认识到要想把概率论变成一个数学分支，必须对它进行公理化。遗憾的是，像 Poincaré 这样伟大的数学家，对概率论的贡献也只是在哲学和教学方法上。虽然十九世纪末至二十世纪初，测度论之花已经在另外两位法国数学家 Émile Borel (1871-1956) 和 Henri Lebesgue (1875-1941) 的栽培下绽放，但它仍在等待那个将它采摘献给概率论公理化的有缘人。



1.1.3 Monte Carlo 方法

考虑图 1.9 所示投钉问题的反问题：已知区域 Ω 的面积为 $\mu(\Omega)$ ，而子区域 $A \subset \Omega$ （阴影部分）的面积 $\mu(A)$ 是未知的，如何近似地求解 $\mu(A)$? 根据式 (1.8)，可以通过大量的投钉试验来估计 $\mu(A)$ 。显然，

$$\mu(A) \approx \frac{m}{n} \cdot \mu(\Omega)$$

其中 m/n 是落于子区域 A 和区域 Ω 的钉数之比。这种通过大量随机试验给出实际问题的数值近似解的计算方法统称为 Monte Carlo 方法，也称为随机模拟方法。

首位使用随机模拟方法的是法国博物学家、数学家 Comte de Buffon (1707-1788)。他是一位百科全书式的学者，代表作是 44 卷的巨著《自然史》。1733 年，Buffon 曾考虑过下面例 1.27 所描述的投针试验，它是一个很有趣的概率问题。Buffon 于 1777 年在《自然史》的增刊中纂文《或然算术试验》给出了该问题的解，这是最早对随机模拟方法和几何概率的研究。通过 Buffon 投针试验，可以得到圆周率 π 的估计值。

例 1.27 (Buffon 投针试验). 随机地向一个宽度为 D 的带状区域上投针，已知针长为 $L < D$ 。只考虑针与该区域有接触的情形，试问针与区域 D 的上下边界（即两条平行线）之一相交的概率是多少？

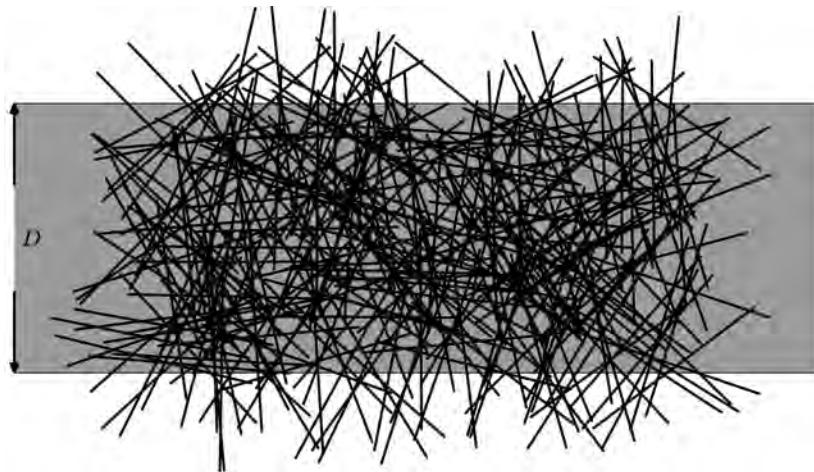


图 1.16: Buffon 投针试验：设定 $D = 1, L = 0.75$ ，投针 $n = 300$ 次。

解. 令针的中点 M 到两平行线的最短距离为 y ，针与边界的夹角为 θ ，即以 M 为中心顺时针旋转至与两平行线平行所扫过的角度，见图 1.17(a)。针与平行线相交当且仅当 $y \leq \frac{L}{2} \sin \theta$ 。

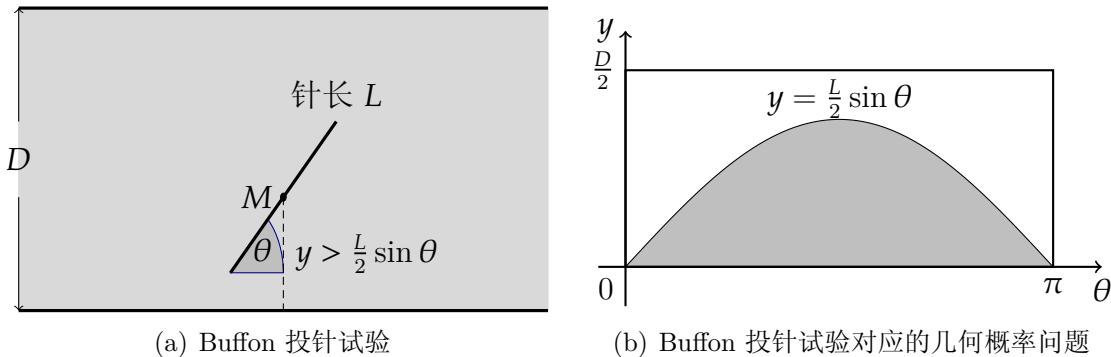


图 1.17: (a) Buffon 投针试验中针的状态。针与两平行线之一相交当且仅当 $y \leq \frac{L}{2} \sin \theta$, 其中 $\theta \in [0, \pi]$, y 是针的中点 M 到两平行线的最短距离。(b) Buffon 投针试验所求概率即为往长方形 $\Omega = [0, \pi] \times [0, D/2]$ 上随机投钉落于阴影区域的概率。

显然, 针的位置 (θ, y) 的变化范围是长方形 $\Omega = [0, \pi] \times [0, D/2]$ 。Buffon 投针试验相当于向区域 Ω 内投钉, 考察落于子区域 $0 \leq y \leq \frac{L}{2} \sin \theta$ 的概率的随机试验, 见图 1.17(b)。于是, 针与平行线相交的概率是

$$p = \frac{1}{D\pi/2} \int_0^\pi \frac{L}{2} \sin \theta d\theta = \frac{2L}{D\pi} \quad (1.9)$$

算法 1.1 (估算圆周率). 设在 n 次 Buffon 投针试验中有 m 次交于平行线, 将针与平行线相交的频率 m/n 视作式 (1.9) 中 p 的近似值, 于是

$$\pi \approx \hat{\pi} = \frac{2Ln}{Dm}$$

练习 1.10. 在 Buffon 投针试验次数给定的情况下, 针长 L 的选择是否影响对圆周率 π 的近似精度? 为什么?

如果仅想估算圆周率 π , 我们可以设计更简单的投钉试验, 譬如投钉区域 Ω 为单位正方形, 子区域 A 为该正方形的内接圆。模拟试验产生 n 个“钉”均匀地分布于正方形区域 $\Omega = [-1/2, 1/2] \times [-1/2, 1/2]$ 上 (见图 1.18), 判断它们中哪些落到内接圆盘 A 上, 设其个数为 m 个, 则圆周率 π 的估计值即为

$$\hat{\pi} = \frac{4m}{n}$$

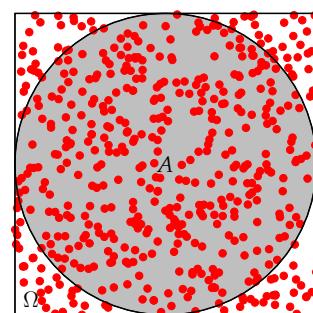


图 1.18: 通过随机投钉来近似计算 π 。

在学过大数律之后, 我们将理解投钉次数越多, 越有可能得到 π 的精确估计值。然而, 圆周率的投钉估算法贵在启发性, 而不是实用性,

因为该方法收敛速度奇慢、稳健性欠佳，远不如级数法效率高。

$$\pi = \sqrt{12} \sum_{k=0}^{\infty} \frac{(-3)^{-k}}{2k+1} = \sqrt{12} \left(1 - \frac{1}{3 \cdot 3} + \frac{1}{5 \cdot 3^2} - \frac{1}{7 \cdot 3^3} + \dots \right)$$

1910 年，印度天才数学家 S. Ramanujan (1887-1920) 发现收敛速度更快的无穷级数。

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{k=0}^{\infty} \frac{(1103 + 26390k) \cdot (4k)!}{396^{4k} \cdot (k!)^4}$$

例 1.28. 利用 Buffon 投针试验所得的圆周率估计值 $\hat{\pi}_1, \dots, \hat{\pi}_n$ 围绕在真实值 π 的周围，试问它们的绝对误差 $|\hat{\pi}_1 - \pi|, \dots, |\hat{\pi}_n - \pi|$ 有无规律可循？

解. 取 $\epsilon = 0.01, 0.02, 0.03$ ，计算估计值 $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_n$ 落于开区间 $(\pi - \epsilon, \pi + \epsilon)$ 内的频率。模拟试验结果如图 1.19 所示。不难发现，投针次数越多，估计值越可能以高比例集中在真实值的“不远处”。

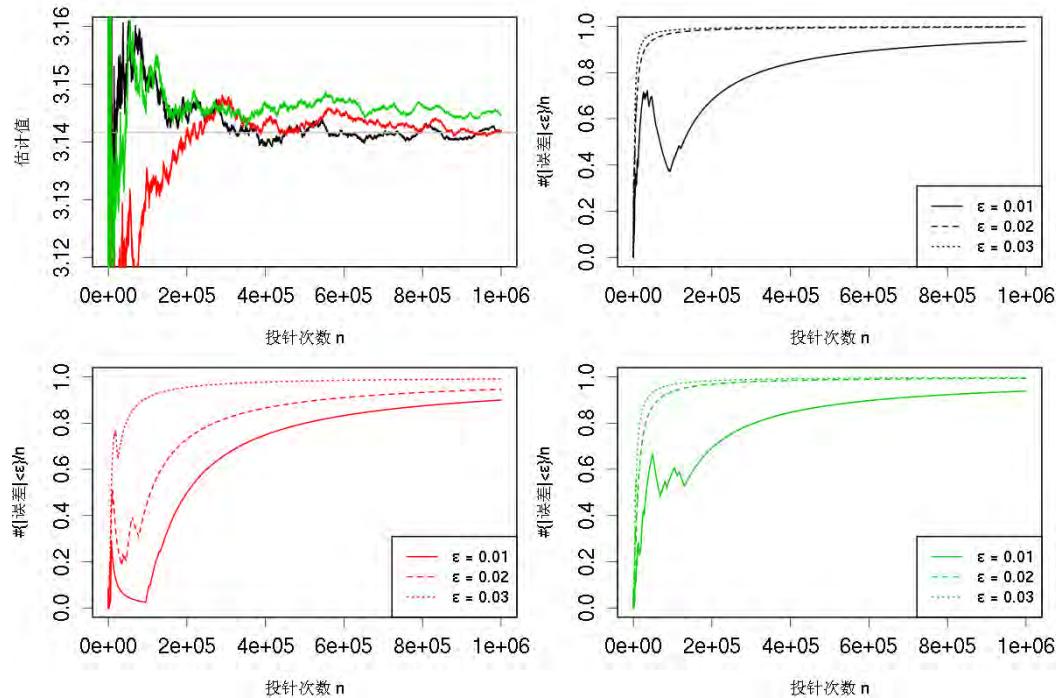


图 1.19：利用 Buffon 投针试验估算圆周率，所得估计值 $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_n$ 落于开区间 $(\pi - \epsilon, \pi + \epsilon)$ 内的频率随着 $n \rightarrow \infty$ 有逼近 1 的趋势，其中 $\epsilon = 0.01, 0.02, 0.03$ 。

利用随机模拟的方法来估算圆周率，试验次数并非总是多多益善。在不知道真实值的前提下，该如何判定何时停止呢？数学里有最优停止理论 (Optimal

Stopping Theory) 来指导模拟试验的规模 [25]，该话题不在本书范围之内。

在计算机诞生之前，随机模拟方法只能停留在 Buffon 投针试验这样的趣味游戏的阶段。计算机为数值计算而生，给数值计算带来了一场革命。有些实际问题中的计算无法或难以用确定性的算法实现，但却可以利用随机模拟来近似，于是人们对随机模拟方法进行了系统的研究。有关随机模拟技术及其算法的详细介绍见第 15 章。

最初的一些随机模拟算法是由 Stanislaw Marcin Ulam (1909-1984) 等几位美国物理学家在“Manhattan 计划”中提出来的，因为此方法与概率论中伪随机数 (pseudo random number) 的产生有关，方法的提出者们^{*}考虑到原子弹研制中重要技术的保密要求就采用了摩纳哥最有名的赌场 Monte Carlo 来命名它。

随机模拟中要用到大量的随机数，其中，如何产生 $[0, 1]$ 上的均匀分布的随机数最为重要。曾几何时，人们通过制定随机数表和人工查表的笨拙方法产生随机数。现如今，几乎所有的编程语言都自带伪随机数产生器，造随机数不再是繁琐之事。



另外，随机数常作为数据源用于检验计算机算法的有效性，它们对随机化算法也是至关重要的。为此，算法大师 D. Knuth 在《计算机程序设计艺术》的第二卷《半数值算法》的第三章《随机数》花了大量的篇幅讨论线性同余法来产生 $[0, 1]$ 上的均匀分布的伪随机数，以及对伪随机数的统计检验 [91]。

总之，随机模拟是计算机擅长的，随着计算技术的进步，随机模拟的应用与日俱增，它在数值分析里的地位也今非昔比。例如，Monte Carlo 方法在工程与运筹（如任务调度、路径优化等）、物理过程与结构（如模拟核试验、分子构型等）、经济与金融（如风险管理、股票期货预测等）、人工智能（如 Monte Carlo 树搜索等）、计算统计学（如贝叶斯分析等）中的应用另辟蹊径，突破了传统的研究套路。因此，我们将在第 4 章介绍常见分布的伪随机数产生算法，它们原本是统计计算的基础内容之一。



^{*}另外几个主要成员也都来自美国 Los Alamos 国家实验室，他们是 Enrico Fermi (1901-1954), John von Neumann (1903-1957) 和 Nicholas Metropolis (1915-1999)。

1.1.4 对随机性的思考

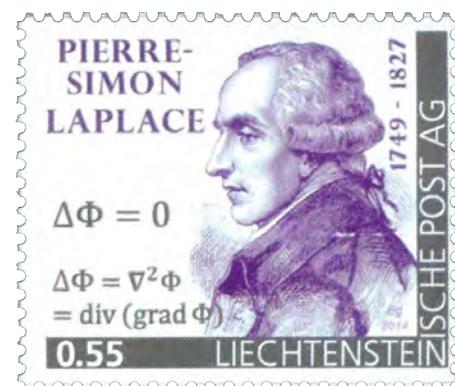
人们很早就开始思考自然界、人类社会中随处可见的随机性。《易经》最初用于占卜和预测天气，后来发展成一套理论来描述状态的简易、变易和不易，它既承认状态时时处于变化之中，同时又认为有某种不变的规律在其中^{*}。《易经》所含的哲学思想是深刻的，它影响了中国文化。十七世纪末，德国数学家、哲学家 G. W. Leibniz (1646-1716) 甚至用二进制来解释六十四卦。《易经》试图用一种符号系统来预测未来或解释自然，它是一次伟大的尝试，但它毕竟不是严谨的数学，也没有启发 Leibniz 思考随机数学。事实上，Leibniz 在 1672-1676 年旅居巴黎期间受到 Pascal、Fermat、Huygens 等人概率论研究成果的影响，开始研究这门“新逻辑学”。



图 1.20: Leibniz 和六十四卦中的“履”。

十八至十九世纪，决定论 (determinism) 统治着欧洲科学界。那时的数学和自然科学的迅速崛起让数学家和科学家充满了信心，要接近上帝洞悉自然的规律。Laplace 把概率部分地归结于人类的无知。我们知道，即使人类的知识可以不断地增长，也不会突破人类认识能力的范围。所以，Laplace 所说的那种“无知”也许是永远不可改变的。有趣的是，Laplace 认为概率“部分地依赖于我们的知识”，这句话表白了他的主观贝叶斯主义。比如，“明天下雨”的概率，不同的预测模型或许给出不同的答案，个人的主观预测也因气象知识的差异而不尽相同。

在《概率的哲学探讨》中，Laplace 清楚地表述了这样的科学观 [165]，“……假设存在这样一种超常的智慧，它能了解使自然界运转的全部动力，以及构成自然界中每个物体的各自的位置，并且还能对它所知道的这些情况进行加工和分析，以至于用一个公式就可以把宇宙中最大的物体连同最小的原子的运动都给出完整的描述，那么在它看来，未来要发生的事情跟过去已经发生的事情一样都是一清二楚的。当然若是这样的话，也就没



^{*}东汉郑玄在《易论》中指出，“易名而含三义：简易一也；变易二也；不易三也。”

有什么是不可确定的了。……所有这些探索真理的努力使人类的智慧逐渐地接近于上面所说的大智慧。可是最终达到它是不可能的。……其实，单个空气或水蒸气分子的运动轨迹和行星运行的轨道一样是有规律可循的，只不过人类缺乏对前者足够的认识。概率部分地与这种无知有关，部分地依赖于我们的知识。”



苏格兰物理学家、数学家 James Clerk Maxwell (1831-1879) 认为概率论是“真正的逻辑”，他说，“现实的逻辑科学目前只精通两种东西：必然的、不可能的或完全值得怀疑的，(幸运的是)它们毋须争辩。因此，这个世界真正的逻辑是概率的计算，它考虑的是，或者应该说是，一个理性人类思维中可能性的大小。”

二十世纪初，量子力学诞生，发现波粒二象性 (wave-particle duality) 是微观粒子的基本特性之一。1926 年，奥地利物理学家 Erwin Schrödinger (1887-1961) 提出 Schrödinger 波动方程。德国物理学家 Max Born (1882-1970) 给出波函数的概率诠释，后来发展成为量子力学的 Copenhagen 诠释。量子力学奠基者之一、大物理学家 Albert Einstein (1879-1955) 对这个诠释很不满意，与量子力学另一位奠基者 Niels Bohr (1885-1962) 展开了多年的论战。

1926 年，Einstein 在给 Born 的信中写道，“无论如何，我确信上帝不掷骰子”，这表明了他对量子力学概率解释的反对态度。Einstein 坚信随机性反映了人类对现实世界基本性质的无知，他认为 Werner Heisenberg (1901-1976) 的不确定性原理（也称测不准原理）只是权宜之计。

这似乎是命中注定，人类对自身主观认识能力的迷惑过去有，现在有，将来还会有^{*}。Einstein-Bohr 之争在哲学上是决定论的是非之争，孰是孰非历史已经给出了答案：Einstein 拒不接受的不确定性原理已被多数物理学家接受并成为量子力学的基石之一。

退一步说，即便“上帝不掷骰子”，当系统复杂到无法精确描述所有粒子的运动，莫不如假设存在大量的随机现象，用概率统计来研究系统的宏观性质。Occam 刃 (Occam's razor)[†]留下的是随机数学模型。譬如，十九世纪发展起来的统计力学 (statistical mechanics)，就是联系微观物理状态和宏观物理量统计规律的学问。



^{*}二十世纪中叶，早期的人工智能 (Artificial Intelligence, AI) 研究者多以强人工智能 (strong AI) 为目标，即智能机器将具备甚至超越人类的智慧，表现出人类的所有智能行为。随着人工智能子领域，如自然语言理解、知识表示、演绎推理、机器学习、情感计算、机器人学、机器知觉等在研究上受阻而停滞不前，有些研究者开始倾向于弱人工智能，即不可能造出智能机器，所谓的“智能”只是局部地看上去像，而非真正意义上的智能，机器并不具有自主意识。

[†]科学上，Occam 刃常用于模型或假说的选择：若两个模型或假说解释同样的现象，简单为上。

Schrödinger 不喜欢波粒二象性这样的二元解释，试图只用波来解释量子力学，同时他对 Copenhagen 诠释也不满意。1935 年，Schrödinger 在与 Einstein 的通信讨论中受 Einstein 提到的不稳定的火药桶会处于爆炸和不爆炸的叠加状态的启发，提出一个理想实验：一只可怜的猫被放在密室里，内有致命的氰化物。在打开密室之前，氰化物有 50% 的机会被释放，猫的状态对观察者来说是未知的，死、活两种本征态的概率都是 50%。Schrödinger 想通过这个实验从宏观阐述量子叠加原理，并求证观测介入时量子的存在形式。其实，对一次试验而言，不管观测与否，猫的状态都是存在在那里的，只是观测者无法知道而已。至今，量子力学也没有一个令所有人信服的解释。



图 1.21: Schrödinger 的猫：对“上帝”而言，猫的状态是唯一的，要么死要么活。对观测者而言，猫的状态是未知的，主观上可以认为它的状态是生死各半。

- 从主观概率的角度理解，由认知的局限性所导致的未知状态“可视为”具有随机性，Schrödinger 的猫生或死都各有一半可能。
- 从客观概率的角度理解，就是在大量独立进行的重复试验中，Schrödinger 的猫有频率 50% 处于活的状态，有频率 50% 处于死的状态。

英国物理学家 Stephen Hawking (1942-2018) 甚至说，“Einstein 说‘上帝不掷骰子’时，他是双重地错了。……上帝不但掷骰子，有时还把骰子掷到无法被看到的地方。”Hawking 的这番话原是谈论黑洞发射粒子，人们所能预言的只是某些粒子的发射概率。当然，也可以泛泛地将这句话理解为大自然并没有把所有的随机现象都展示在人类面前。即便如此，很多情况下人们依然有办法透过可观察到的随机现象或多或少地了解到一些自然的本质和现象背后的事。要做到这一点，非得借助概率模型不可。读者可通过例 1.29 加深理解 Hawking 的观点。



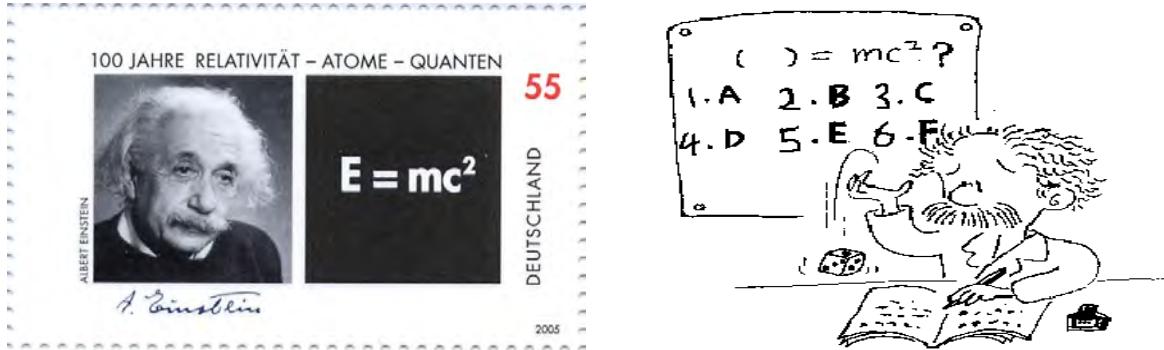


图 1.22: 上帝不掷骰子, 但我们掷……。

※例 1.29 (隐 Markov 模型的直观描述). 上帝和 Einstein 玩一个游戏, 道具是 n 个盒子 (标号分别为 $1, 2, \dots, n$), 每个盒子里都有 m 种不同颜色的球。Einstein 知道一些信息, 如, 第 k 种颜色的球在盒子 i 中所占的比例为 b_{ik} 。显然, 它们满足

$$\sum_{k=1}^m b_{ik} = 1, \text{ 其中 } i = 1, 2, \dots, n$$

另外, 还有 $n+1$ 个 n 面骰子。其中, 标号为 0 的 “ n 面骰子”, 其点数 $i = 1, 2, \dots, n$ 出现的概率为 π_i , 显然这些概率满足归一性。即,

$$\sum_{i=1}^n \pi_i = 1$$



标号为 i 的 “ n 面骰子”的各点数出现的概率依次为 a_{ij} , 其中 $i, j = 1, 2, \dots, n$ 。我们把矩阵 $A = (a_{ij})$ 称为转移矩阵, 显然该矩阵逐行满足归一性。即,

$$\sum_{j=1}^n a_{ij} = 1, \text{ 其中 } i = 1, 2, \dots, n$$

游戏规则: 上帝先掷 0 号骰子, 设所掷点数是 i , 按点数上帝选取盒子 i , 然后在该盒子里随机抽取一个球, 汇报该球的颜色后将球放回盒内。接着, 上帝再掷 i 号骰子, 按掷出的点数再选取下一个盒子, 重复刚才的过程……。游戏要求上帝掷骰子的过程对 Einstein 来说是不可见的 (如 Hawking 所言, 骰子被掷到无法看到的地方), 即所选盒子的序列对 Einstein 来说是不可观察的, Einstein 能得到的观测数据就是上帝汇报的颜色序列。

试问: 聪明的 Einstein 该如何猜出上帝掷骰子的点数序列, 即颜色序列 $x_{1:T} = x_1 x_2 \cdots x_T$ 所对应的盒子序列 $z_{1:T} = z_1 z_2 \cdots z_T$?

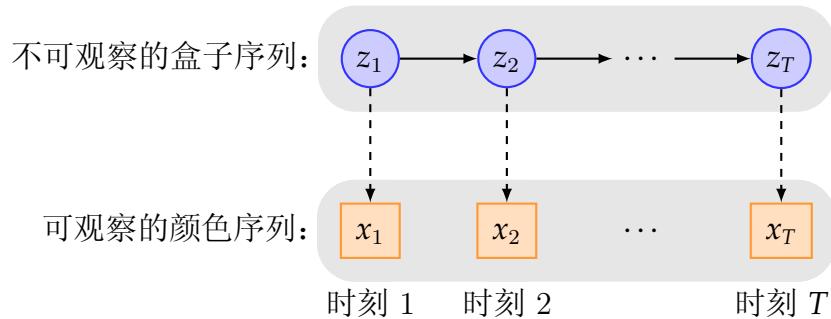
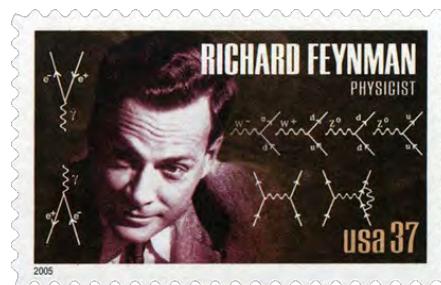


图 1.23: 在观察到的颜色序列 $x_1 x_2 \cdots x_T$ 的基础上, 人们关心的是不可观察的盒子序列 $z_1 z_2 \cdots z_T$ 中概率最大者, 因为它揭示了现象背后隐藏着的事实。

解. 可以利用隐 Markov 模型的 Viterbi 算法 (细节见 §13.1) 找到最有可能的盒子序列, 这是一个动态规划算法, 广泛地应用于语音识别 [128]、生物序列分析 (如蛋白质结构预测)、词性标注 (part-of-speech tagging, POS tagging)^{*} 等实际问题。

不确定性在人类的认知中比比皆是, 概率作为人类探索自然的工具不会被确定性颠覆。美国物理学家 Richard Feynman (1918-1988) 说过 [47], “现代物理学在尽可能多地了解自然的努力中发现, 某些事物是无法确定地 ‘搞清楚’ 的。我们的许多知识必须始终保持不确定。我们可以知道的最多的就是概率。”



^{*}中文常见的词性有名词、代词、动词、数词、量词、形容词、副词、介词、连词、助词、语气词等。在自然语言处理中, 词性在句法重写规则中充当着非终结符, 因此如何“教会”机器自动识别词性是句法分析的一个基本问题, 直接影响机器翻译、语义分析等后续计算。

1.2 概率论的公理化

起源于古希腊，公理化方法 (axiomatic method) 是指从尽可能少的基本概念和一组不加证明的公理出发，通过逻辑推理构建一个演绎系统的方法。公理化方法有助于清晰地描述理论体系，减少歧义。该方法并不局限于数学领域，例如，法国哲学家、数学家 René Descartes (1596-1650) 在《哲学原理》中，英国物理学家、数学家 Isaac Newton (1642-1727) 在《自然哲学的数学原理》中，都主张并自觉地使用了公理化方法。



公理化数学最早的范例是古希腊数学家、几何学之父 Euclid (约公元前 325-公元前 265) 的著作《几何原本》，对其中的第五公设是否独立的研究历经了两千年终于在十九世纪结成正果——非欧几何学诞生了。“建立几何的公理和探究它们之间的联系，是一个历史悠久的问题；关于这问题的讨论，从 Euclid 以来的数学文献中，有过难以计数的专著，这问题实际就是要把我们的空间直观加以逻辑的分析。”

——引自 Hilbert 《几何基础》序言

1830 年左右，俄罗斯数学家 N. I. Lobachevsky (1792-1856) 和匈牙利数学家 János Bolyai (1802-1860) 分别独自创立了非欧几何中的双曲几何，但在当时并未得到数学界的认可，属于他们的荣誉在他们死后才姗姗来迟。





1899 年，伟大的德国数学家 David Hilbert (1862-1943) 发表了公理化思想的传世之作《几何基础》，第一次给出了完备的欧氏几何公理体系。“本书中的研究，是重新尝试着来替几何建立一个完备的，而又尽可能简单的公理系统；要根据这个系统推证最重要的几何定理，同时还要使我们的推证能明显地表出各类公理的含义和个别公理的推论的含义。”Hilbert 坚信，“我们必定可以用桌子、椅子、啤酒杯来代替点、线、面”，于是他舍弃了点、线、面的直观意义而把它们看作不加定义的纯粹抽象物，并明确指出几何学关心的是点、线、面之间的关系，这样建立的几何公理系统具有最大的一般性。

《几何基础》是划时代的，对后世产生了深远的影响，其后公理化方法渗透到几乎所有的纯数学领域，Hilbert 因此被公认为现代公理化方法的奠基人。读者可参阅《古今数学思想》的第四十二章去大致了解 Hilbert 《几何基础》的构架和基本思想。

1900 年，Hilbert 在巴黎第二届国际数学家大会上提出了 23 个关键的数学问题，其中第六问题是物理学的公理化。“对几何基础的研究促成此问题：借助公理用同样的方法处理那些深受数学影响的物理科学，首推概率论和力学。”虽然 Hilbert 第六问题至今仍未完全解决，但概率论的公理化却由一位苏联数学家实现。

1933 年，伟大的苏联天才数学家 Andrey Nikolaevich Kolmogorov (1903-1987) 出版了专著《概率论基础》[93]，在总结前人工作的基础上以测度论为工具完成了概率论的公理化。基于概率的频率解释，Kolmogorov 公理体系得到了大部分数学家的认可，形成了频率派（或称经典学派），由此蓬勃发展起来的概率论已成为传统数理统计学的基础。本书的绝大多数内容都是基于 Kolmogorov 公理体系的。

与此同时，也有一些学者致力于非传统概率论的研究，如贝叶斯学派的主观概率*（见本书第 12 章）。



* 贝叶斯学派认为，随机事件 A 的概率仅是个体主观认为 A 会发生的信念度 (belief degree)。例如，我认为“Einstein 在 1945 年 8 月 6 日早上掷过骰子”的概率是 90%，显然没有可重复的随机试验能考察此事，它仅仅表达了我对这个陈述的相信程度。

总所周知，集合论是现代数学的基础，也是通用的数学语言。本书假定读者已经掌握了 Cantor 朴素集合论，我们将以它为工具定义一个关键的概念——样本空间，并在此基础上描述 Kolmogorov 公理体系。在引出样本空间这一概念之前，我们先列出需要用到的集合论中的一些定义和结果，并约定

- 在不引起歧义的情况下， $A \cap B, A \setminus B$ 和 $\bigcap_{j=1}^{\infty} A_j$ 等集合运算也常记作 $AB, A - B$ 和 $\prod_{j=1}^{\infty} A_j$ 等算术运算。
- 集合组成的类用英文或德文手写体字母表示，如 $\mathcal{S}, \mathfrak{A}, \mathfrak{B}$ 等。

~性质 1.5 (de Morgan 律). 令 \mathfrak{A} 是某些集合组成的类，则

$$\begin{aligned}\left(\bigcup_{A \in \mathfrak{A}} A\right)^c &= \bigcap_{A \in \mathfrak{A}} A^c \\ \left(\bigcap_{A \in \mathfrak{A}} A\right)^c &= \bigcup_{A \in \mathfrak{A}} A^c\end{aligned}$$

定义 1.3 (划分). 回顾第 29 页的例 1.21。非空集合 A 的一个划分就是 A 的某些非空子集构成的类 \mathfrak{P} ，使得对任意 $a \in A$ ，仅落在一个子集中。换句话说，

$$\emptyset \notin \mathfrak{P}, \mathfrak{P} \text{ 中的元素两两不交，且 } \bigcup_{S \in \mathfrak{P}} S = A$$

定义 1.4 (等价关系). 设 R 是集合 A 上的一个二元关系，如果它满足自反性、对称性和传递性，则称该二元关系是 A 上的一个等价关系。即 $\forall x, y, z \in A$ ，皆有

$$xRx, \text{ 且 } xRy \Rightarrow yRx, \text{ 及 } xRy, yRz \Rightarrow xRz$$

~性质 1.6. 集合 A 上的划分与 A 上的等价关系是一一对应的。

例 1.30. 集合 $\Omega = \{1, 2, \dots, 7\}$ 上的等价关系 $R : x \equiv y \pmod{3}$ 决定了商集 $\Omega/R = \{\{1, 4, 7\}, \{2, 5\}, \{3, 6\}\}$ ，它就是 Ω 的一个划分。

例 1.31. 语言学从词语的用法上定义了各种词性，然而词性分类并不是划分。虽然语言学家会尽量把它做得像个划分，但依然有很多的兼类词。

※例 1.32. 如果集合 A 的势为 $n < \infty$ ，它有多少个不同的划分？

解. 令势为 n 的集合的所有不同划分的个数为 B_n ，显然 $B_0 = 1, B_1 = 1, B_2 = 2$ 。读者不难验证 B_n 满足如下的递归关系。

$$B_{n+1} = \sum_{k=0}^n C_n^k B_k \quad (1.10)$$

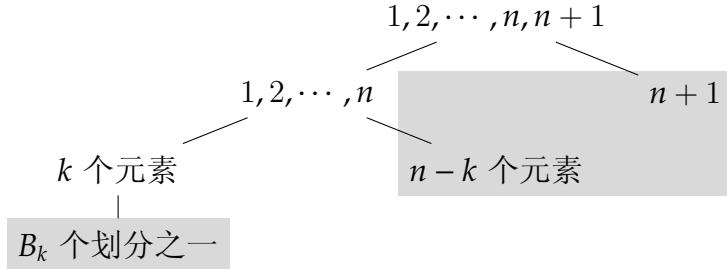


图 1.24: 集合 $\{1, 2, \dots, n, n+1\}$ 的一个划分可以这样给出: 从元素 $1, 2, \dots, n$ 中选取 $n-k$ 个元素, 将它们与元素 $n+1$ 组成一个子集, 再给出剩下的 k 个元素的一个划分。因为 k 个元素有 B_k 个不同的划分, 因此才有递归式 (1.10)。

我们称 B_n 为第 n 个 Bell 数, 以纪念美国数学家 Eric Temple Bell (1883-1960)。Bell 数增长得很快, 例如 $B_3 = 5, B_4 = 15, B_5 = 52, B_6 = 203, \dots, B_{18} = 682076806159, \dots$ 。

性质 1.7. 已知 $A_1, A_2, \dots, A_n, \dots$ 是一个集合的序列, 则

$$\bigcup_{n=1}^{\infty} A_n \supseteq \dots \supseteq \bigcup_{n=k}^{\infty} A_n \supseteq \dots \quad \text{且} \quad \bigcap_{n=1}^{\infty} A_n \subseteq \dots \subseteq \bigcap_{n=k}^{\infty} A_n \subseteq \dots \quad (1.11)$$

$$\text{元素 } a \text{ 属于无穷多个 } A_n \Leftrightarrow a \in \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n \quad (1.12)$$

$$\text{元素 } a \text{ 仅不属于有限多个 } A_n \Leftrightarrow a \in \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n \quad (1.13)$$

证明. 往证 (1.13): “元素 a 仅不属于有限多个 A_n ” 当且仅当 $\exists k \in \mathbb{N}$ 使得 $\forall n \geq k$, 皆有 $a \in A_n$, 进而 $a \in \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n$ 。反之亦然。结果 (1.12) 留作练习。 \square

定义 1.5. 对于集合序列 $\{A_n\}$, 有时分别将 $\bigcup_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$ 和 $\bigcap_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n$ 简记作 $\overline{\lim}_{n \rightarrow \infty} A_n$ 和 $\underline{\lim}_{n \rightarrow \infty} A_n$, 或者 $\limsup_{n \rightarrow \infty} A_n$ 和 $\liminf_{n \rightarrow \infty} A_n$, 并将之分别称作该集合序列的上极限 (limit superior) 和下极限 (limit inferior)。它们分别类似于实数序列 $\{s_n\}$ 的上极限 $\limsup_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} (\sup_{m \geq n} s_m)$ 和下极限 $\liminf_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} (\inf_{m \geq n} s_m)$ 。实数序列 $\{s_n\}$ 的上极限和下极限有时也记作 $\overline{\lim}_{n \rightarrow \infty} s_n$ 和 $\underline{\lim}_{n \rightarrow \infty} s_n$ 。

例 1.33. 若 $s_n = (-1)^n(1 - 1/n)$, 则 $\limsup_{n \rightarrow \infty} s_n = 1$ 且 $\liminf_{n \rightarrow \infty} s_n = -1$ 。已知 $A_n = (-1 - 1/n, s_n)$, 则

$$\overline{\lim}_{n \rightarrow \infty} A_n = [-1, 1]$$

$$\underline{\lim}_{n \rightarrow \infty} A_n = \{-1\}$$

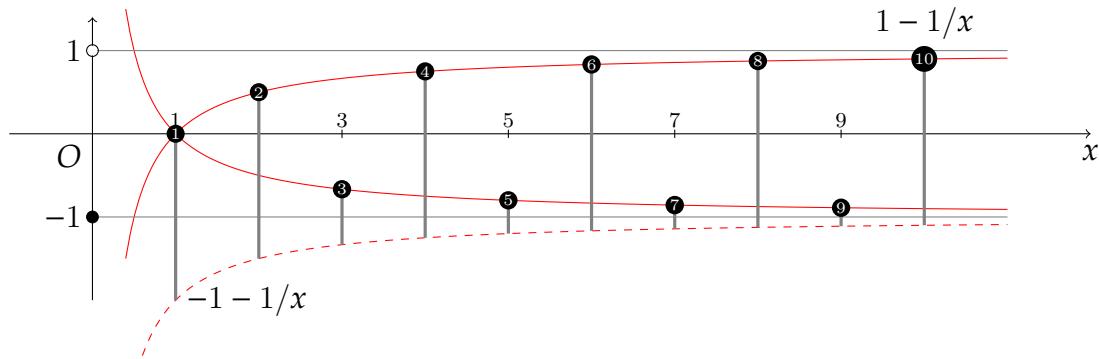


图 1.25: 竖直的粗线描绘的是例 1.33 中的集合序列 $\{A_n\}$, 带标号的黑点表示的是 A_n 的右端点 s_n 。请读者通过求 $\liminf_{n \rightarrow \infty} A_n$ 和 $\limsup_{n \rightarrow \infty} A_n$ 来理解性质 1.7。

练习 1.11. 显然, $\liminf_{n \rightarrow \infty} s_n \leq \limsup_{n \rightarrow \infty} s_n$, 其中等号成立当且仅当序列 $\{s_n\}$ 收敛。请根据性质 1.7 简要说明 $\liminf_{n \rightarrow \infty} A_n \subseteq \overline{\limsup_{n \rightarrow \infty} A_n}$ 。

本节内容

出于对集合可数个并、交运算的封闭性要求, 第一小节引入 σ 域的概念, 并用它定义了样本空间和随机事件这两个最重要的概念。另外, 我们还介绍了生成的 σ 域、Borel 集和 Borel σ 域等概念, 从样本空间角度回顾了 Bertrand 悖论。第二小节的主题是 Kolmogorov 公理体系, 简而言之, 概率就是一个定义在样本空间上的满足非负性、归一性和可列可加性的实值集函数 (set function)。从测度论的角度看, “概率空间 = 样本空间 + 概率测度”。我们将通过连续正面问题及其极限版这一暗线了解测度论对概率论的重要性, 透过 Bernoulli 弱大数律及大量 Bernoulli 试验进一步认识概率的频率解释。第三小节介绍了概率的一些重要的性质, 如概率的连续性定理、Borel-Cantelli 引理等。

关键知识

(1) σ 域、样本空间、随机事件、概率测度; (2) 概率的 Kolmogorov 公理体系; (3) 概率的连续性定理、Borel-Cantelli 引理等; (4) Bernoulli 弱大数律。

1.2.1 σ 域与样本空间

给定非空集合 Ω , 它的幂集合记为 2^Ω 或 $\mathcal{P}(\Omega)$ 。已知 $\mathcal{F} \subseteq 2^\Omega$ 是 Ω 的某些子集组成的非空类, 显然它的元素之间可以有交、并、补等集合运算。如果 \mathcal{F} 及其元素间的集合运算是人们关注的对象, 集合运算的封闭性就显得非常必要, 否则谈论 \mathcal{F} 上的集合运算就没有多大意义。

定义 1.6. 设 \mathcal{F} 是非空集合 Ω 的某些子集构成的非空类, 如果 \mathcal{F} 中的元素经过有限次交、并、补等集合运算结果仍属于 \mathcal{F} , 则称 \mathcal{F} 是 Ω 上的一个域 (field) 或代数 (algebra)。

性质 1.8. 已知 \mathcal{F} 是非空集合 Ω 上的一个域, 则 \mathcal{F} 含有空集 \emptyset 和全集 Ω , 即 $\emptyset, \Omega \in \mathcal{F}$ 。

证明. 因为 \mathcal{F} 是非空类, 设 Ω 的某个子集 $A \in \mathcal{F}$, 则 $A^c \in \mathcal{F}$, 进而 $A \cap A^c = \emptyset$ 和 $A \cup A^c = \Omega$ 都属于 \mathcal{F} 。 \square

例 1.34. 显然, $\mathcal{F} = 2^\Omega$ 是非空集合 Ω 上的一个域。再如, $\Omega = \{1, 2, 3, 4\}$, 则 $\mathcal{F}_1 = \{\emptyset, \Omega\}$ 和 $\mathcal{F}_2 = \{\emptyset, \{1, 2\}, \{3, 4\}, \Omega\}$ 都是 Ω 上的域。

定义 1.7 (σ 域). 如果非空集合 Ω 上的域 \mathcal{S} 对可数个并运算封闭, 即

$$\forall A_1, A_2, \dots \in \mathcal{S}, \text{ 皆有 } \bigcup_{j=1}^{\infty} A_j \in \mathcal{S} \quad (1.14)$$

则称 \mathcal{S} 是 Ω 上的一个 σ 域 (σ -field) 或 σ 代数 (σ -algebra)。二元组 (Ω, \mathcal{S}) 称作可测空间 (measurable space)^{*}, \mathcal{S} 中的元素称作可测集 (measurable set)。

例 1.35. 连续抛一枚均匀的硬币 n 次, 基本事件集合 $\Omega_n = \{(\omega_1, \dots, \omega_j, \dots, \omega_n) : \omega_j = 0 \text{ 或 } 1\}$ 。令 Ω 是抛该枚硬币无穷次的基本事件集合, 下面我们定义 Ω 上的一个有限 σ 域 \mathcal{S}_n , 其元素皆为如下集合

$$A = \{(\omega_1, \dots, \omega_n, \dots) : (\omega_1, \dots, \omega_n) \in E \text{ 且 } E \in 2^{\Omega_n}\}$$

\mathcal{S}_n 的元素个数为 2^{2^n} , 且满足

$$\mathcal{S}_1 \subset \mathcal{S}_2 \subset \mathcal{S}_3 \subset \dots$$

下面我们说明 $\mathcal{F} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \dots$ 是一个域, 但不是 σ 域。

*在数学里, 所谓“空间”常指一个非空集合上赋予了某种结构。例如, 非空集合 T 的某些子集构成的类 \mathcal{T} 如果满足以下条件, 则称 (T, \mathcal{T}) 是一个拓扑空间, 称 \mathcal{T} 为 T 上的一个拓扑 (topology), 称 \mathcal{T} 中的元素为 (T, \mathcal{T}) 的开集。(1) $T, \emptyset \in \mathcal{T}$; (2) 若 $A, B \in \mathcal{T}$, 则 $A \cap B \in \mathcal{T}$; (3) \mathcal{T} 中任意多 (可数或者不可数) 元素之并集合仍在 \mathcal{T} 中。显然, 此处的条件 (3) 比条件 (1.14) 要强许多。

- 首先, $\Omega \in \mathcal{S}_1$, 于是 $\Omega \in \mathcal{F}$ 。
- 若 $E \in \mathcal{F}$, 则必存在 n 使得 $E \in \mathcal{S}_n$, 于是 $E^c \in \mathcal{S}_n$, 进而 $E^c \in \mathcal{F}$ 。
- 若 $A, B \in \mathcal{F}$, 则必存在 i, j 使得 $A \in \mathcal{S}_i, B \in \mathcal{S}_j$, 取 $n = \max(i, j)$, 我们有 $A \cup B \in \mathcal{S}_n$, 因此 $A \cup B \in \mathcal{F}$ 。

按照定义, \mathcal{F} 是一个域。下面说明它不是 σ 域: 考虑单点集合 $E = \{(1, 1, 1, \dots)\}$, 显然 $E \notin \mathcal{F}$, 然而

$$E = \bigcap_{j=1}^{\infty} \{(\omega_1, \dots, \omega_j, \dots) : \omega_1 = \dots = \omega_j = 1\}$$

性质 1.9. 若 \mathcal{S} 是集合 Ω 上的一个 σ 域, 则 \mathcal{S} 中的元素对可数个交运算封闭, 即

$$\forall A_1, A_2, \dots \in \mathcal{S}, \text{ 皆有 } \bigcap_{j=1}^{\infty} A_j \in \mathcal{S}$$

证明. $\forall A_1, A_2, \dots \in \mathcal{S}$, 因为 \mathcal{S} 是一个 σ 域, 所以 $A_1^c, A_2^c, \dots \in \mathcal{S}$, 进而

$$\bigcup_{j=1}^{\infty} A_j^c \in \mathcal{S}$$

由 de Morgan 律 (性质 1.5), 不难得到

$$\bigcap_{j=1}^{\infty} A_j = \left(\bigcup_{j=1}^{\infty} A_j^c \right)^c \in \mathcal{S}$$

□

※例 1.36. 在图 1.9 所示的投钉问题中, 不妨设正方形区域 $\Omega = [0, 1] \times [0, 1]$, 令 \mathcal{S} 是由 Ω 的所有具有面积的子集所构成的类。利用实分析的知识可以证明, \mathcal{S} 是 Ω 上的一个 σ 域。但是, $\Omega \cap (\mathbb{Q} \times \mathbb{Q}) \notin \mathcal{S}$, 因为它没有面积 [54], 其中 \mathbb{Q} 是有理数集合。

样本空间的概念由奥地利数学家 Richard von Mises (照片见右) 于 1931 年引入, 首次明确定义了什么是随机事件。

※定义 1.8 (样本空间). 已知 \mathcal{S} 是随机试验的基本事件集合 Ω 上的一个 σ 域, 特称可测空间 (Ω, \mathcal{S}) 为一个样本空间 (sample space), 称 \mathcal{S} 中的任一元素为一个随机事件 (random event) 或事件。

如果 Ω 有限, 则称 (Ω, \mathcal{S}) 为有限样本空间。如果 Ω 至多可数, 则称 (Ω, \mathcal{S}) 为离散样本空间。如果 Ω 不可数, 则称 (Ω, \mathcal{S}) 为不可数样本空间, 特别地, 当 $\Omega = \mathbb{R}^k$ 时称 (Ω, \mathcal{S}) 为连续样本空间。



定义 1.9 (并事件). 由样本空间的定义知, 若 $A, B \in \mathcal{S}$, 则 $A \cup B \in \mathcal{S}$, 即 $A \cup B$ 也是一个随机事件, 我们称之为 A, B 的并事件或者和事件。

如果事件 A_1, A_2, \dots, A_n 两两互斥, 我们习惯用 $A_1 + A_2$ 表示 $A_1 \cup A_2$, 用 $\sum_{j=1}^n A_j$ 表示 $\bigcup_{j=1}^n A_j$, 同时在上下文中也会不厌其烦地提醒读者它们是“非交并”。

例 1.37. 令 Ω 是抛两次硬币随机试验的基本事件集合, $\mathcal{S} = 2^\Omega$ 使得 (Ω, \mathcal{S}) 构成一个样本空间。集合 $A = \{(T, H), (H, T)\} \in \mathcal{S}$ 表示事件“抛得一正一反”; $B = \{(T, H), (H, T), (H, H)\} \in \mathcal{S}$ 表示事件“抛得至少一个正面”。

□ 显然 $A \subset B$, 意味着“若 A 发生, 则 B 也发生”。

□ $\{(T, T)\} \cup \{(H, H)\} = \{(T, T), (H, H)\}$ 表示事件“两次抛出的结果相同”等同于“两次都抛出反面或者两次都抛出正面”。

定义 1.10. 类似地, 可以定义 A, B 的交事件 $A \cap B$ (也记作 AB , 称为积事件)、差事件 $A \setminus B$ (也记作 $A - B$)、对称差事件 $A \Delta B = (A \setminus B) \cup (B \setminus A)$, 下面一一画出它们的 Venn 图。请读者画出 A 的补事件 (或称余事件、对立事件) A^c 的 Venn 图。

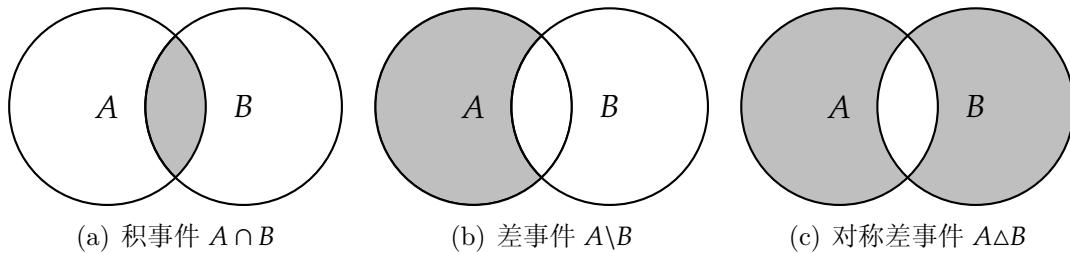


图 1.26: (a) 若 $A \cap B$ 发生, 则 A, B 同时发生。(b) 若 $A \setminus B$ 发生, 则 A 发生且 B 不发生。(c) 若 $A \Delta B$ 发生, 则 A 发生或 B 发生, 但 A, B 不同时发生。

例 1.38. 已知随机事件 $A, B, C \in \mathcal{S}$, 请用集合运算来描述下面的随机事件: (1) A, B, C 都不发生; (2) A, B, C 至少有一个发生; (3) A, B, C 至少有两个发生; (4) A, B, C 恰有一个发生; (5) A, B, C 至多有一个发生; (6) A, B, C 至多有两个发生。

解. (1) $A^c B^c C^c$; (2) $A \cup B \cup C$; (3) $AB \cup BC \cup CA$; (4) $AB^c C^c \cup BA^c C^c \cup CA^c B^c$; (5) $(AB \cup BC \cup CA)^c$; (6) $(ABC)^c$, 即 $A^c \cup B^c \cup C^c$ 。

性质 1.10. 已知事件 $A_1, A_2, \dots, A_n, \dots \in \mathcal{S}$, 则对于任意 $n \in \mathbb{N}$, 事件 $\bigcup_{j=1}^n A_j$ 和 $\bigcup_{j=1}^{\infty} A_j$ 分别具有如下的非交分解。

$$\begin{aligned} \bigcup_{j=1}^n A_j &= A_1 + A_1^c A_2 + A_1^c A_2^c A_3 + \cdots + A_1^c \cdots A_{n-1}^c A_n \\ \bigcup_{j=1}^{\infty} A_j &= \sum_{n=1}^{\infty} A_1^c \cdots A_{n-1}^c A_n \end{aligned}$$

等式右边意味着“ A_1 发生”或“ A_1 不发生 A_2 发生”或“ A_1, A_2 都不发生 A_3 发生”或……。

现在，我们回过头来谈论一下如何构造基本事件集合 Ω 上的 σ 域。若 Ω 有限，其 σ 域的构造是简单的。若 Ω 是连续的，如在 Buffon 投针试验（例 1.27）中， $\Omega = [0, \pi] \times [0, D/2]$ ，我们该如何构造其 σ 域呢？

首先，我们可以考虑 Ω 的某些子集组成的非空类 \mathcal{A} ，这些子集是我们所关心的研究对象。例如，对于 Buffon 投针试验， $\mathcal{A} = \{(a, b] \times (c, d] : 0 \leq a < b \leq \pi, 0 \leq c < d \leq D/2\}$ 。然而， \mathcal{A} 对集合运算不封闭怎么办呢？我们可以对 \mathcal{A} 稍加扩充使之变成一个 σ 域，见下面的定义。

定义 1.11 (生成的 σ 域). 已知类 $\mathcal{A} \subseteq 2^\Omega$ 非空，显然，所有包含 \mathcal{A} 的 σ 域之交 \mathcal{S}_0 仍是 σ 域，它是包含 \mathcal{A} 的唯一最小的 σ 域，称之为由 \mathcal{A} 生成的 σ 域，记作 $\mathcal{S}_0 = \sigma(\mathcal{A})$ 。

例 1.39. 已知 $A \subset \Omega$ 非空，则 $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$ 。

例 1.40. 欧氏空间 \mathbb{R}^n 上所有形如 $(a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_n, b_n]$ 的“长方体”组成的类所生成的 σ 域简记作 \mathfrak{B}_n 。特别地， \mathfrak{B}_1 是实数轴 \mathbb{R} 上所有形如 $(a, b]$ 的左开右闭区间组成的类所生成的 σ 域，它包含所有单点集合和所有的区间，这是因为

$$\begin{aligned}\{x\} &= \bigcap_{n=1}^{\infty} (x - 1/n, x] & (x, y) &= (x, y] - \{y\} & [x, y] &= (x, y] + \{x\} \\ [x, y) &= \{x\} + (x, y] - \{y\} & (x, \infty) &= (-\infty, x]^c\end{aligned}$$

练习 1.12. 试证明 \mathfrak{B}_1 是 \mathbb{R} 上所有形如 $(-\infty, b]$ 的区间组成的类所生成的 σ 域。提示： $(a, b] = (-\infty, b] - (-\infty, a]$ 。

1898 年，测度论的奠基者之一、法国数学家 Émile Borel (1871-1956) 在拓扑空间的基础上引入 Borel 集这一极为重要的概念，此概念关乎现代概率论的基础，将被用于 Borel 可测函数和随机变量的定义。

定义 1.12 (Borel 集和 Borel σ 域). 拓扑空间 (X, \mathcal{T}) 的开集经过可数个交、并、补运算得到的集合被称为 Borel 集或 Borel 可测集。显然， X 的所有 Borel 集形成 X 上的一个 σ 域，被称为 Borel σ 域，它是包含 (X, \mathcal{T}) 所有开集的最小的 σ 域——这正是 Borel σ 域的特殊之处。



练习 1.13. 试证明 \mathfrak{B}_n 是 \mathbb{R}^n 上的 Borel σ 域，其势为 \aleph_0 。我们把 $(\mathbb{R}^n, \mathfrak{B}_n)$ 称作 n 维 Borel 可测空间。

练习 1.14. 已知 $\Omega \in \mathbb{R}^n$, 试证明 $\mathcal{S} = \{A \cap \Omega : A \in \mathfrak{B}_n\}$ 是 Ω 上的一个 σ 域。该 σ 域被称为 Ω 的 Borel σ 域, 简记作 $\Omega \cap \mathfrak{B}_n$ 。

到此为止, 样本空间的构造已不成问题。最后要提醒注意的是, 实数集合 \mathbb{R} 的许多子集不在 \mathfrak{B}_1 中。例 1.41 给出了一个非 Borel 集的例子, 其构造过程并不平凡。

※例 1.41. 任何实数 x 皆可唯一地表示为连分数 $[a_0; a_1, a_2, \dots]$, 即如下的形式。

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}, \text{ 其中, } a_0 \text{ 为整数, } a_1, a_2, \dots \text{ 为正整数}$$

有理数的连分数表示一定是有限的, 而无理数的连分数表示一定是无限的。例如, $\sqrt{2} = [1; 2, 2, 2, \dots]$, $\pi = [3; 7, 15, 1, \dots]$, 黄金分割点 $\frac{\sqrt{5}-1}{2} = [0; 1, 1, 1, \dots]$ 。



通过连分数, 人们可以轻易求得实数 x 的第 n 个渐近分数 $f_n = [a_0; a_1, \dots, a_n]$ 。在所有分母不超过 f_n 的分母的分数当中, f_n 是实数 x 的最佳逼近。譬如, 圆周率 $\pi = 3.1415926 \dots$ 的约率和密率*。

$$\text{约率: } \frac{22}{7} = [3; 7]$$

$$\text{密率: } \frac{355}{113} = [3; 7, 15, 1]$$



* 我国古代数学家祖冲之 (429-500) 早于欧洲一千年算得密率, 由于其数学著作《缀术》不幸失传, 他的密率算法至今仍是个谜。我国著名数学家华罗庚 (1910-1985) 曾撰文《从祖冲之的圆周率谈起》, 讨论约率和密率的内在意义。华罗庚认为约率和密率提出了“用有理数最佳逼近实数”的问题 [162], 祖冲之很有可能掌握了连分数的技巧, 这是个了不起的成就。



1927 年, 苏联数学家 Nikolai Luzin (1883-1950) 首次构造了一个非 Borel 集的例子, 即所有满足下述条件的无理数 $x = [a_0; a_1, a_2, \dots]$ 的全体: a_0, a_1, a_2, \dots 有无穷子序列 $a_{k_0}, a_{k_1}, a_{k_2}, \dots$ 使得 a_{k_j} 整除 $a_{k_{j+1}}, j = 0, 1, 2, \dots$ 。例如,

$$e = [2; 1, 2, 1, 1, 4, 1, \dots, 1, 2n, 1, \dots]$$

※例 1.42 (回顾 Bertrand 悖论). 在例 1.24 中, 令 E 表示事件“弦长大于单位圆内接正三角形边长”。按照对“随机取一条弦”的不同理解, 得到三个不同的基本事件集合, 进而三个不同的样本空间 (Ω, \mathcal{S}) 。其中, \mathcal{S} 是 Ω 的 Borel σ 域, 即 $\mathcal{S} = \{B \cap \Omega : B \in \mathfrak{B}_2\}$ 。

1. $\Omega = [0, 2\pi) \times (0, 1)$, 则 $E = [0, 2\pi) \times (0, 1/2)$ 且 $P(E) = 1/2$ 。
2. $\Omega = [0, 2\pi) \times [0, 2\pi) - \{(x, x) : x \in [0, 2\pi)\}$, 则 $E = \{(x, y) : 2\pi/3 < |x - y| < 4\pi/3\}$ 且 $P(E) = 1/3$ (请仿照例 1.20 验证之)。
3. $\Omega = \{(x, y) : x^2 + y^2 < 1\}$, 则 $E = \{(x, y) : x^2 + y^2 < 1/4\}$ 且 $P(E) = 1/4$ 。

三种不同的随机取弦方法对应着三个不同的样本空间, 它们对随机事件的定义是不同的。换句话说, 随机事件 E 在不同的样本空间里有着不同的语义, 其概率不同就不令人诧异了。Bertrand 悖论应该采用哪种随机取弦的方法至今尚无定论, 每种方法似乎都能找到合适的理由。

1973 年, 美国物理学家、贝叶斯学派学者 E. T. Jaynes (1922-1998) 撰文 [80] 指出弦的分布应该独立于圆的位置与半径, 如果承认这一点, 第一种方法就是唯一的答案。对 Bertrand 悖论更深入的讨论不在此书的范围之内, 感兴趣的读者可以参阅 [80]。

1.2.2 Kolmogorov 公理体系

在可测空间 (Ω, \mathcal{S}) 上定义了测度便得到了测度空间，测度论是研究测度空间一般性质的数学理论，也是概率论的严格基础。然而，测度论对于计算机实践并无增益，所以本书不打算以测度论为工具。尽管如此，我们仍坚持以两种等价的方式来定义概率空间：一是把它视为某类特殊的测度空间，二是直接公理化。前者的好处是可以调用测度论的很多结果，后者的好处是把概率论当作一门独立的数学分支。

定义 1.13. 给定可测空间 (Ω, \mathcal{S}) ，定义在 \mathcal{S} 上的集函数 μ 若满足以下条件，则称 $(\Omega, \mathcal{S}, \mu)$ 为测度空间 (measure space)，称 μ 为测度。

① 非负性： $\mu(A) \geq 0$ ，其中任意的 $A \in \mathcal{S}$ 。

② 可列可加性：若 $A_j \in \mathcal{S}, j = 1, 2, \dots$ 两两不交，则

$$\mu\left(\sum_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j)$$

③ 空集测度为零： $\mu(\emptyset) = 0$ 。

测度空间 $(\Omega, \mathcal{S}, \mu)$ 上测度为零的集合称为 μ -零测集，简称零测集。把零测集的所有子集加入到 \mathcal{S} 所生成的 σ 域记作 \mathcal{S}_μ ，我们称 $(\Omega, \mathcal{S}_\mu, \mu)$ 为一个完备的测度空间 (complete measure space)。

例 1.43 (计数测度). 已知 Ω 是有限集合，可测空间 $(\Omega, 2^\Omega, \mu)$ 中测度 $\mu(A)$ 定义为集合 $A \subseteq \Omega$ 的势，称为计数测度。



希腊数学家 Constantin Carathéodory (1873-1950) 在函数论、变分法、测度论上有许多贡献，一个著名的结果是 Carathéodory 扩张定理。该定理利用如下定义的外侧度 (outer measure) 和内测度 (inner measure) 对测度空间 $(\Omega, \mathcal{S}, \mu)$ 进行了扩张，重新定义了可测集，进而得到了一个完备的测度空间，它的测度限制在 $(\Omega, \mathcal{S}, \mu)$ 上就是 μ 。

外测度： $\mu^*(A) = \inf\{\mu(E) : A \subseteq E, \forall E \in \mathcal{S}\}$

内测度： $\mu_*(A) = \sup\{\mu(E) : E \subseteq A, \forall E \in \mathcal{S}\}$

构造 $\mathcal{S}' = \{A \subseteq \Omega : \mu^*(A) = \mu_*(A)\}$ 。一般地， $\mu_*(A) \leq \mu^*(A)$ ，使等号成立的 A 称为 μ -可测的，其测度为 $\mu(A) = \mu^*(A) = \mu_*(A)$ 。 \mathcal{S}' 也可等价地定义为

$$\mathcal{S}_{\mu^*} = \{E \subseteq \Omega : \mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c), \forall A \subseteq \Omega\}$$

Carathéodory 扩张定理保证 \mathcal{S}_{μ^*} 也是 Ω 上的一个 σ -域，使得 $\mathcal{S} \subseteq \mathcal{S}_{\mu^*}$ ，进而 $(\Omega, \mathcal{S}_{\mu^*}, \mu^*)$ 也是一个（完备的）测度空间。我们把测度空间 $(\Omega, \mathcal{S}_{\mu^*}, \mu^*)$ 称为 $(\Omega, \mathcal{S}, \mu)$ 的 Carathéodory 扩张。

奥地利数学家 Hans Hahn (1879-1934, 照片见右) 证得若 μ 是 σ -有限的^{*}（这个条件在很多场合下是容易满足的），则 Carathéodory 扩张是唯一的。我们有时把这两个结果合称为 Carathéodory-Hahn 定理。当我们谈论一个测度空间时，都缺省地指经过 Carathéodory 扩张后得到的那个测度空间。如此一来，零测集都可以忽略不计，方便了讨论与计算，见例 1.44 和例 1.46。



例 1.44. Lebesgue 测度是欧氏空间 \mathbb{R}^n 里区间长度、长方形面积、长方体体积的一般化。例如，Borel 可测空间 $(\mathbb{R}, \mathfrak{B}_1)$ 上的 Lebesgue 测度定义如下。

1. 若开集 $(a_j, b_j), j = 1, 2, \dots$ 两两不交，且 $\bigcup_{j=1}^{\infty} (a_j, b_j) \subset [a, b]$ ，则

(a) 开集 $A = \bigcup_{j=1}^{\infty} (a_j, b_j)$ 的 Lebesgue 测度为

$$m(A) = \sum_{j=1}^{\infty} (b_j - a_j)$$

(b) 闭集 $B = [a, b] - \bigcup_{j=1}^{\infty} (a_j, b_j)$ 的 Lebesgue 测度为

$$m(B) = (b - a) - \sum_{j=1}^{\infty} (b_j - a_j)$$

2. 对于一般的有界集合 $L \subseteq [a, b]$ ，它的任意开覆盖（即包含 L 的开集）都具有 Lebesgue 测度，其下确界定义为 L 的外测度，记作 $m^*(L)$ 。显然， L 的“大小”不会超过 $m^*(L)$ 。令 $L^c = [a, b] - L$ ，如果 $m^*(L) + m^*(L^c) = b - a$ ，则称 L 是 Lebesgue 可测集，其 Lebesgue 测度定义为 $m(L) = m^*(L)$ 。

性质 1.11. 下面不加证明地介绍 Lebesgue 测度的几个性质，详情见实变函数理论。

□ 可以把 \mathbb{R} 上 Lebesgue 测度很自然地推广到 \mathbb{R}^n 上 [68]：可测空间 $(\mathbb{R}^n, \mathfrak{B}_n)$ 上存在唯一的测度 μ （称为积测度）使得

$$\mu(S_1 \times \cdots \times S_n) = \prod_{j=1}^n m(S_j), \text{ 其中 } S_1, \dots, S_n \in \mathfrak{B}_1$$

*对于测度空间 $(\Omega, \mathcal{S}, \mu)$ ，如果 Ω 能表达为可数个测度有限的 \mathcal{S} 的元素之并集，即 $\Omega = \bigcup_{j=1}^{\infty} A_j$ ，其中 $A_j \in \mathcal{S}$ 且 $\mu(A_j) < \infty$ ，则称 μ 是 σ -有限的，或称该测度空间具有 σ -有限测度。

□ Borel 集都是 Lebesgue 可测的，但反之不成立，见第 54 页的例 1.41。

□ 外测度为零的集合及其子集都是可测的，其测度为零。

例 1.45. \mathbb{R} 上零测集的典型例子如：(1) 在 \mathbb{R} 上处处稠密的有理数集 \mathbb{Q} ；(2) 无处稠密却具有连续统 \aleph_0 的 Cantor 集 C 。



$$C = \left\{ \sum_{j=1}^{\infty} \frac{x_j}{3^j} : x_j = 0 \text{ 或 } 2 \right\}$$

定义 1.14 (概率). 当测度空间 (Ω, \mathcal{S}, P) 中 (Ω, \mathcal{S}) 是一个样本空间并且 $P(\Omega) = 1$ 时，特称 P 为概率测度，简称概率，称 (Ω, \mathcal{S}, P) 为一个概率测度空间或概率空间。

定义 1.15. 给定概率空间 (Ω, \mathcal{S}, P) ，对于任意 $A \subset \Omega$ ，若 $A \notin \mathcal{S}$ ，定义

$$P^*(A) = \inf\{P(E) : A \subseteq E, \forall E \in \mathcal{S}\}$$

根据 Carathéodory-Hahn 定理，我们可唯一得到 (Ω, \mathcal{S}, P) 的 Carathéodory 扩张 $(\Omega, \mathcal{S}_{P^*}, P^*)$ ，其中 \mathcal{S}_{P^*} 的每个元素都是 P -可测的。当我们谈论概率空间时，缺省都是指那个扩张后的完备的概率空间。

※例 1.46. 接着例 1.26，测度空间 $(I, I \cap \mathfrak{B}_1, m)$ 是一个概率空间，其中 m 是 Lebesgue 测度。经过扩张得到新的概率空间 (I, \mathcal{S}, m) ，其中 \mathcal{S} 是单位闭区间 $I = [0, 1]$ 上所有 Lebesgue 可测集合的全体。

因为有理数集合的测度为零，所以 $I = [0, 1]$ 上被“多对一”的那些点的集合的测度为零，不影响概率计算。下面，我们仿照例 1.10 的做法计算 $m(A)$ ：首先，把 H_t 拆成一些非交并，然后逐一进行“翻译”，其结果也是两两不交的。

$$\begin{aligned} & \underbrace{H \cdots H}_t * \rightarrow 0.1 \underbrace{\cdots 1}_t * \\ & T \underbrace{H \cdots H}_t * \rightarrow 0.0 \underbrace{1 \cdots 1}_t * \\ & HT \underbrace{H \cdots H}_t * \rightarrow 0.10 \underbrace{1 \cdots 1}_t * \end{aligned}$$

形如 $0.\underbrace{1\cdots 1}_t*$ 的二进制实数的全体是一个长度为 2^{-t} 的区间，以此类推，于是

$$m(A) = \sum_{n=0}^{\infty} 2^{1-t-n} F_n^{(t)}$$

其中， $F_n^{(t)} = \sum_{i=1}^t F_{n-i}^{(t)}$ ，满足 $F_0^{(t)} = 0, F_1^{(t)} = F_2^{(t)} = 1$

上式中，序列 $F_n^{(t)}$ 被称为 n -步 Fibonacci 序列。当 $n = 2$ 时，就是 Fibonacci 序列；当 $n = 3$ 时，就是 Tribonacci 序列。不难验证 n -步 Fibonacci 序列具有如下性质。

$$\sum_{n=0}^{\infty} x^n F_n^{(t)} = \frac{x}{1-x-x^2-\cdots-x^t}$$

利用该性质，下面往证 $m(A) = 1$ ，进而 $P(H_t) = 1$ 。

$$\begin{aligned} m(A) &= 2^{1-t} \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n F_n^{(t)} \\ &= 2^{1-t} \frac{2^{-1}}{1-2^{-1}-2^{-2}-\cdots-2^{-t}} = 1 \end{aligned}$$

抛一枚均匀的硬币无穷次，几乎必然抛出任意给定的有限长度的 H, T 序列。用 H, T 构成的长度为 16 的字符串对汉字进行编码，小说《红楼梦》就是一个有限 H, T 序列。上述结果说明，只要坚持下去，迟早会“抛”出一部《红楼梦》！如果继续抛下去，还能再“抛”出 n 部《红楼梦》。最早发现这一结果的是法国数学家 Émile Borel，他在 1913 年用勤奋打字的猴子作比喻来形象地阐述该结果：只要一直努力随机敲下去，早晚敲出一部 Hamlet。

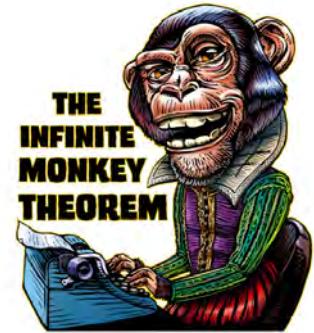


表 1.3: 概率论中的某些基本概念在测度论中对应的解释。

概率论中的概念	测度论中的解释
概率空间 (Ω, \mathcal{S}, P)	测度空间 (Ω, \mathcal{S}, P) 满足 $P(\Omega) = 1$
基本事件 $\omega \in \Omega$	Ω 中的点 $\omega \in \Omega$
所有随机事件的集合 \mathcal{S}	Ω 的 σ 域 \mathcal{S}
随机事件 $A \in \mathcal{S}$	Ω 的子集 $A \in \mathcal{S}$
随机事件 A 的概率 $P(A)$	集合 A 的测度 $P(A)$

从形式上，全部概率的数学理论就是加了“ Ω 的测度为 1”这一限制的测度论。“尽管如此，依照所解决的问题的实质看来，概率论仍是一门独立的数学分支；某些结果（如大数律和极限定理）对于概率论来说是基本的，但从纯粹测度论的观点看却似乎是人为制造出来的，似乎是用不着的。这样看待问题不仅使得概率数学理论的形式结构显得非常清晰，而且还使得概率论本身及形式结构与之相近的其他数学理论都获得了非常实际的进展。……随便什么地方，只要那里概率论公理能够成立，那里就可以引用这些公理的推论，即便是那里和现实的随机性没有任何共同点也可以不管。”

— A. N. Kolmogorov

1933 年，现代概率论奠基人 A. N. Kolmogorov 给出概率的 Kolmogorov 公理体系。参考古典概率的性质 1.4，现代概率论的公理这样给出是水到渠成的。概率论于是成为一门大学问，而不是测度论的一个旁门左道。在《概率论基础》一书中，Kolmogorov 是这样定义概率空间的 [93]。

④定义 1.16. 已知随机试验的样本空间 (Ω, \mathcal{S}) ，如果 \mathcal{S} 上的实值集函数 $P(\cdot)$ 满足以下三个条件，则称 P 为概率或概率测度，称 (Ω, \mathcal{S}, P) 为一个概率空间。

- ① 非负公理：对于任一随机事件 $A \in \mathcal{S}$ ，总有 $P(A) \geq 0$ 。非负实数 $P(A)$ 称为事件 A 的概率。
- ② 归一公理： $P(\Omega) = 1$ ，即必然事件 Ω 的概率等于 1。
- ③ 有限可加公理：若事件 A, B 互斥（即 $A \cap B = \emptyset$ ），则

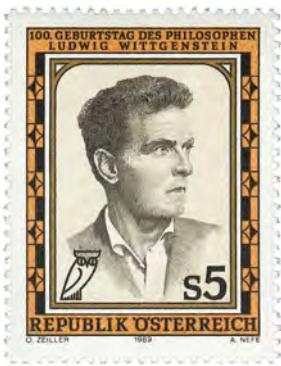
$$P(A + B) = P(A) + P(B)$$

Kolmogorov 论证了 ①②③ 给出的概率公理体系是和谐的，此类的概率测度被称为有限可加概率，譬如常见的古典概率。应研究和应用的需求，有限可加公理 ③ 被补充加强为如下的可列可加公理 ③。

- ③ 可列可加公理：若事件 $A_j \in \mathcal{S}, j = 1, 2, \dots$ 两两互斥，则

$$P\left(\sum_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j) \quad (1.15)$$

根据测度论中的 Carathéodory-Hahn 定理，若有可列可加公理这个加强条件，概率空间 (Ω, \mathcal{S}, P) 可唯一地被扩张到完备的概率空间 $(\Omega, \mathcal{S}_P, P^*)$ （见定义 1.15）。满足可列可加公理的概率测度被称为可列可加概率或完全可加概率。



“概率”是什么？如果有人问“围棋”是什么，最好的回答是“围棋”的游戏规则，“围棋”之所以有别于“五子棋”也是因为它们的游戏规则不同。英籍奥地利裔哲学家、语言哲学的奠基人、二十世纪伟大的哲学家 Ludwig Wittgenstein (1889-1951) 认为“意义即用法” (Meaning is use)。作为类比，这里“概率”的含义就是上述三条公理 ①②③ 及其蕴含的性质，整个经典概率论的大厦就是建基于此的^{*}，所以数学家的工作就是去发现蕴含在这些基本假设之后的那些推论。

日本概率大师伊藤清 (1915-2008) 在《概率论基础》[161] 一书中说，“仅依靠形式的推理是不能导出可列可加性的。将概率的可列可加性作为基础来假设，是数学上的理想化模式。你渐渐地便能理解这种理想化不是与实际相悖的，反而是与其一致的。”表面上如何理解 Kolmogorov 的这三条公理呢？

- 归一性约定必然事件 Ω 的概率是 1，由可列可加性可得 $P(\emptyset) = 0$ ，即不可能事件 \emptyset 的概率是 0。再由概率 $P(\cdot)$ 的非负性易证 P 的取值范围是闭区间 $[0, 1]$ (证明见下一小节)。
- 集合求并 $\bigcup_{j=1}^{\infty} A_j$ 与次序无关，概率 $P(\cdot)$ 的非负性保证了级数 $\sum_{j=1}^{\infty} P(A_j)$ 绝对收敛，求和与次序无关 (若级数不是绝对收敛，Riemann 证明可调整求和次序使级数收敛至任意给定的值)。而式 (1.15) 则具体给出了求非交并事件 $\sum_{j=1}^{\infty} A_j \in \mathcal{S}$ 的概率的方法。

Kolmogorov 公理体系要求对随机事件及其概率的讨论是在给定的概率空间 (Ω, \mathcal{S}, P) 上进行的，任意随机事件 $E \in \mathcal{S}$ 都是明确定义好了的。概率的公理化并没有指明如何构造 (Ω, \mathcal{S}, P) 最合理，它只是约定好一个起点，只要从这个起点出发就不会遇到悖论。

事实上，公理化和直觉洞察对于概率论都是需要的。伊藤清说，“因为空间中的点可表示为三个实数，空间图形的所有几何性质可用实数的方式表述。于是仅靠分析就能够理论上理解几何学。然而，对几何真正的鉴赏不仅需要分析技术，还需要几何对象的直观。同样对概率论也是如此。现代概率论通过测度和积分被形式化，因此从逻辑观点看它是现代分析的一部分。但若要真正玩透概率论，须用对随机现象的直觉洞察力抓住这个理论发展的主方向才行。”[79]

有了计算机这个好的辅助工具，对概率论的一部分直觉洞察可以通过大量的随机试验来获得。这是现代数学研究的福音，以前的数学家只有笔和纸，现在计算机能帮我们做很多事情，未来人工智能走入数学也不是不可能的。

^{*}匈牙利数学家 Alfréd Rényi (1921-1970) 在 1955 年发表的论文《论概率的一个新公理体系》和 1970 年出版的遗作《概率论》中所描述的贝叶斯概率公理体系 [132] 使 Kolmogorov 公理体系成为其特殊情形，不仅保留了频率派的所有经典结果，也为贝叶斯理论奠定了基础。第 12 章第二节将简介 Rényi 公理体系（见第 644 页的定义 12.7）。

随机事件的概率有一个直观但不严格的描述：大量重复实验中该事件出现的相对频率。频率派认为，概率是大量同类随机现象中固有的属性，与认识主体无关，是一种物理属性。为区别于贝叶斯学派的主观概率，频率派所认同的概率称为客观概率。大多数物理学家所接受的概率都是客观概率^{*}。所谓“随机事件概率的频率解释”，形象地描述就是该随机事件在大量重复的随机试验中出现的频率随着试验次数的增加越来越有“资格”充当概率的近似。对它更准确的刻画需要用到 Jacob Bernoulli (1654-1705) 于十七世纪末发现的弱大数律（其证明见第 5 章），它揭示了随机现象中蕴藏的客观规律。



定义 1.17 (Bernoulli 试验). 像抛硬币这种只出现两个非此即彼结果（常比作“成功”与“失败”，或者“正面”与“反面”）的随机试验被称为 Bernoulli 试验，而像连续抛某硬币 n 次这样独立重复的 Bernoulli 试验则被称为 n 重 Bernoulli 试验。

例 1.47 (n 重 Bernoulli 试验). 假设抛一枚硬币出现正面的概率是 p ，连续抛该硬币 n 次，试问：“恰好出现 k 次正面”的概率 $P(k)$ ？

解. 该问题翻译成球-盒子模型即第 22 页的例 1.13 所述，请读者确认之。所以，

$$P(k) = C_n^k p^k (1-p)^{n-k}, \text{ 其中 } k = 0, 1, \dots, n$$

定理 1.1 (Bernoulli 弱大数律). 已知随机事件 A 的概率 $P(A) = p$ ，在 n 重 Bernoulli 试验中 A 出现了 m 次，则对于任一给定的正数 ϵ ，恒有

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{m}{n} - p\right| \leq \epsilon\right\} = 1 \quad (1.16)$$

Bernoulli 在《猜度术》中是这样论述的，“所要探讨的是，是否随着观测次数的增加，记录下来的赞成与不赞成例数的比值接近真实比值的概率也随之不断增加，使得这个概率最终将超过任意确信度。……假如从现在直至永远，所有事件都被连续地观测到，将会发现世界上每个事件的发生都有着明确的原因并遵循着明确的法则，甚至是看起来相当偶然的事件……。”通俗地讲，就是当试验次数足够多时，频率和概率以接近 1 的大概率充分地接近。

试验次数 $n \rightarrow \infty$ 时，根本无法谈论 m/n 的极限，因为 m 是不确定的，可以取 $0, 1, \dots, n$ 中的任何值。那种称“频率的极限为概率”的说法是完全错误的。正确的理解应该是，给定 ϵ ，大的 n 让事件 $|m/n - p| < \epsilon$ 以接近 1 的大概率发生。

^{*}美国物理学家 E. T. Jaynes (1922-1998) 是个例外，他在遗著《概率论：科学的逻辑》中致力发展贝叶斯概率和贝叶斯统计推断。

为了更好地理解弱 Bernoulli 弱大数律，在下面的例 1.48 中，我们将通过大量的随机试验探究频率与概率的关系。

例 1.48. 抛一枚不均匀的硬币，假设正面出现的概率为 $P(H) = 0.6$ 。连续抛该硬币 n 次，在这 n 重 Bernoulli 试验中出现了 m 次正面。下面，我们设计随机试验来揭示频率 $f = m/n$ 与概率 $P(H)$ 之间的关系：把 n 重 Bernoulli 试验看作一次随机试验 \mathcal{E} 考察正面频率 f ，我们独立重复地做 k 次 \mathcal{E} 来看 f 的散落情况。

1. 分别独立地抛硬币 $n_1 = 10^3$ 次、 $n_2 = 10^4$ 次和 $n_3 = 10^5$ 次，把出现正面的频率 f_1, f_2, f_3 记录下来。
2. 重复此过程 $k = 1000$ 次，分别得到三组频率数据 $\{f_{i1}, f_{i2}, \dots, f_{ik}\}$ ，其中 $i = 1, 2, 3$ 。
3. 基于此三组数据得到下面三个直方图^{*}。

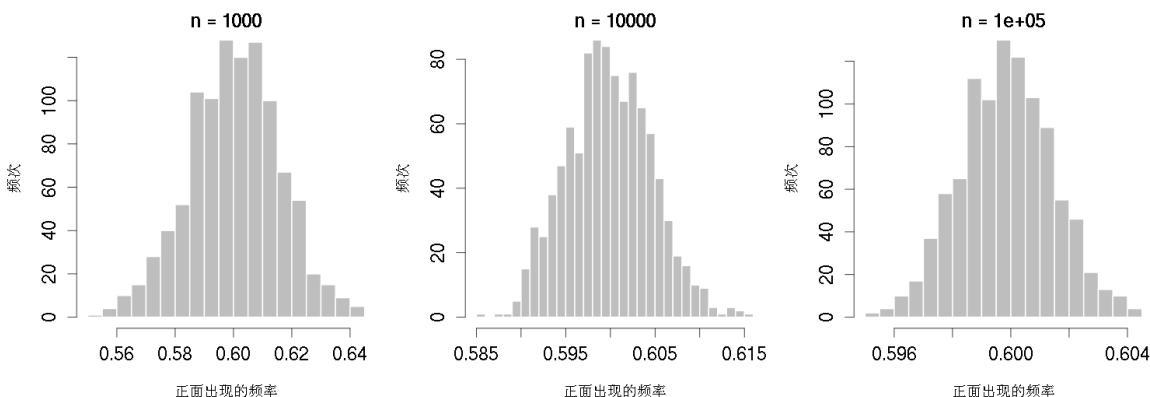


图 1.27: 分别抛硬币 $n_1 = 10^3$ 、 $n_2 = 10^4$ 和 $n_3 = 10^5$ 次，记录下出现正面的频率，算完成一次随机试验。重复 $k = 1000$ 次这样的随机试验，得到正面频率的直方图。

通过考察和对比这些直方图，读者不难发现下面的事实。

- 出现正面的频率 $f = m/n$ 基本围绕在出现正面的概率 $P(H)$ 的两侧且偏离 $P(H)$ 越远越少有发生。
- 随着试验次数 n 的增加，出现正面的频率 $f = m/n$ 越来越紧密地凝聚在 $P(H)$ 的周围，误差 $|m/n - P(H)|$ 小于给定正数 ϵ 的机会也越来越大。

^{*}直方图 (histogram) 是一种常见的数据图示方法。把实数轴划分为几个区间，对有限区间不要求区间长度一定相等。以每个区间为底边画一个矩形，用该矩形的面积表示落于此区间里的观察值的比例便得到了直方图。显然，这些小矩形的面积之和等于 1。直方图显示观察值在何处聚集以及疏散程度等，常用来探索分析数据的分布情况。

 细心的读者也许发现, Bernoulli 弱大数律用接近 1 的概率确保随机事件的概率可以由大量独立重复试验中该事件出现的频率来近似, 这是在用概率的语言来解释频率和概率之间的关系。客观概率都能用例 1.48 的手法用大量独立重复试验来解释, 它揭示了随机现象背后的客观规律。

■**定义 1.18 (高斯函数).** 在概率统计中, 如下定义的 $x \in \mathbb{R}$ 的实值函数 $\phi(x|\mu, \sigma^2)$ 非常之重要, 有关它的由来可在学习完第 2 章后参阅附录 A。

$$\phi(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \text{ 其中参数 } \mu \in \mathbb{R}, \sigma > 0 \quad (1.17)$$

为了记述的方便, 我们简记 $\phi(x|0, 1)$ 为 $\phi(x)$, 即

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \text{ 其中 } x \in \mathbb{R} \quad (1.18)$$

约定用符号“ \propto ”表示函数之间的正比关系: $f(x) \propto g(x)$ 意味着这两个函数之间相差一个非零的常数因子 c , 即 $f(x) = cg(x)$ 。所有满足 $f(x) \propto \phi(x|\mu, \sigma^2)$ 的函数 $f(x)$ 都被称为高斯函数 (Gaussian function)。例如,

$$f(x) = \exp(-\alpha x^2), \text{ 其中 } \alpha > 0$$

$\phi(x|\mu, \sigma^2)$ 是一个在概率论里无处不在的函数, 被印在钱币上、邮票上, 再也找不到第二个函数有此“殊荣”。它被称为参数是 μ, σ^2 的正态分布 (或高斯分布) 的密度函数。



图 1.28: 德国马克上的 Gauss 头像和充满魔力的正态密度函数 $\phi(x|\mu, \sigma^2)$ 的曲线。其实, 法国数学家 Laplace 更早地发现了这个概率密度函数。

练习 1.15. 仿照图 1.7 用计算机绘图来体验：当 n 很大时，

$$\begin{aligned} C_n^k p^k (1-p)^{n-k} &\approx \phi(k|np, np(1-p)), \text{ 其中 } k = 0, 1, \dots, n \\ &= \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left\{-\frac{(x-np)^2}{2np(1-p)}\right\} \end{aligned}$$

性质 1.12. 1774 年, Laplace 证明了

$$\int_{-\infty}^{+\infty} \phi(x) dx = 1$$

证明. 令 $\phi(x)$ 在 \mathbb{R} 上的积分等于 m , 则

$$\begin{aligned} m^2 &= \int_{-\infty}^{+\infty} \phi(x) dx \int_{-\infty}^{+\infty} \phi(y) dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left\{-\frac{x^2+y^2}{2}\right\} dx dy \\ &\stackrel{y=\rho \sin \theta}{=} \frac{1}{2\pi} \int_0^{2\pi} \int_0^{+\infty} \rho \exp\left\{-\frac{\rho^2}{2}\right\} d\rho d\theta = 1 \quad \square \end{aligned}$$

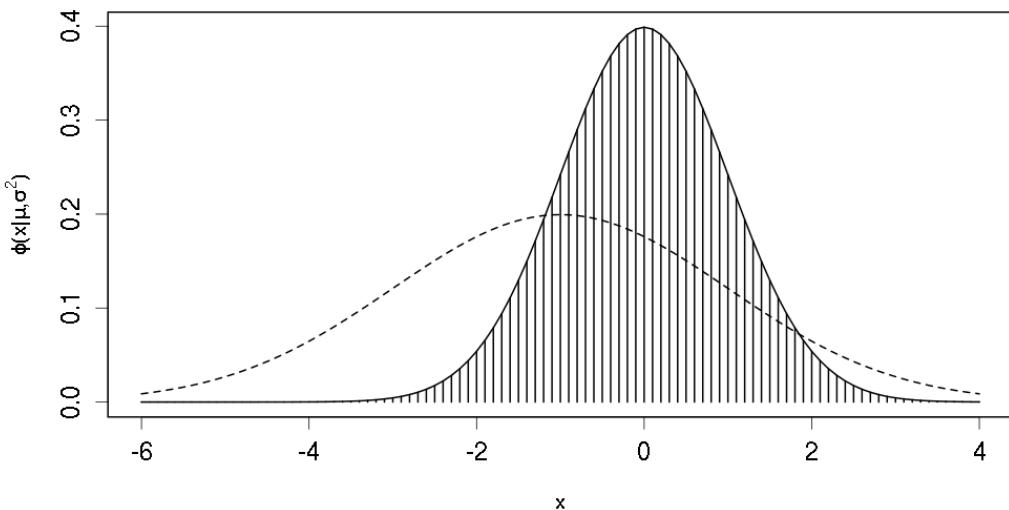


图 1.29: 实线是 $\phi(x)$, 虚线是 $\phi(x-1, 4)$ 。函数 $\phi(x|\mu, \sigma^2)$ 的曲线呈钟型, 关于 $x = \mu$ 对称, 我们称 μ 为位置参数。参数 σ^2 刻画的是 $\phi(x|\mu, \sigma^2)$ 的“体形”: σ^2 越小曲线越“高瘦”, 被称为尺度参数。**性质 1.12** 保证阴影部分的面积等于 1。

例 1.49. 试证明如下定义的实值集函数 P 是样本空间 $(\mathbb{R}, \mathfrak{B}_1)$ 上的概率测度。

$$P(A) = \int_A \phi(x) dx, \text{ 其中 } A \in \mathfrak{B}_1$$

证明. 下面逐一验证实值集函数 $P(A)$ 满足概率测度的三条公理。由积分的性质，显然 $P(A) \geq 0$ 且 P 满足可列可加性。而**性质 1.12** 正说明 $P(\Omega) = 1$ 。 \square

练习 1.16. 接着**例 1.49**，试证明：实值集函数 $P(A) = \int_A \phi(x|\mu, \sigma^2)dx$ 也是 $(\mathbb{R}, \mathfrak{B}_1)$ 上的一个概率测度。提示：利用下面的关系式，

$$\phi(x|\mu, \sigma^2) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \quad (1.19)$$

历史上首次明确提出正态密度函数的是法国数学家、天文学家 Pierre-Simon Laplace (1749-1827, 肖像见右)。统计之父 K. Pearson 在《对相关性的历史注记》[119] 开篇提到，早在 1783 年，Laplace 就在文章里使用 $\phi(x)$ 并建议为下面的概率积分制表，他才是真正的“正态分布之父”。

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}x^2} dx$$

1799 至 1825 年，Laplace 出版了五卷本的经典著作《天体力学》，被誉为法国的 Newton。1806 年，Laplace 成为法兰西第一帝国的伯爵。

K. Pearson 为避免优先权之争而称之为正态分布，如他自己所说，这个叫法有个缺憾就是误导人们以为其他分布都是异常的 (abnormal)。天才的 Gauss 把正态曲线用于最小二乘法误差的分布，他对正态分布的贡献也是巨大的，所以正态分布也称作 Laplace-Gauss 分布。



Laplace

***例 1.50.** 已知参数 τ^2, μ, σ^2 ，利用 $\phi(\theta|\cdot, \cdot)$ 的性质证明下面的结果。

$$\int_{-\infty}^{+\infty} \phi(x|\theta, \tau^2) \phi(\theta|\mu, \sigma^2) d\theta = \phi(x|\mu, \sigma^2 + \tau^2) \quad (1.20)$$

证明. 式 (1.20) 左边是一个关于 x 的概率密度函数，这是因为

$$\int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} \phi(x|\theta, \tau^2) \phi(\theta|\mu, \sigma^2) d\theta \right\} dx = 1$$

令 $\rho = \sigma^{-2} + \tau^{-2}$, 利用正比关系简化式 (1.20) 中的被积函数, 我们有

$$\begin{aligned}\phi(x|\theta, \tau^2)\phi(\theta|\mu, \sigma^2) &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\sigma^2} + \frac{(x-\theta)^2}{\tau^2}\right]\right\} \\ &= \exp\left\{-\frac{\rho}{2}\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\sigma^2} + \frac{x}{\tau^2}\right)\right]^2 - \frac{(x-\mu)^2}{2(\sigma^2 + \tau^2)}\right\} \\ &\propto \phi\left(\theta \left| \frac{1}{\rho}\left(\frac{\mu}{\sigma^2} + \frac{x}{\tau^2}\right), \frac{1}{\rho}\right. \right) \phi(x|\mu, \sigma^2 + \tau^2)\end{aligned}$$

因此, $\int_{-\infty}^{+\infty} \phi(x|\theta, \tau^2)\phi(\theta|\mu, \sigma^2)d\theta \propto \phi(x|\mu, \sigma^2 + \tau^2)$, 进而式 (1.20) 得证。 \square

以上配方法仅凭借 $\phi(x|\cdot, \cdot)$ 在 \mathbb{R} 上积分为 1 这一事实来定性地求解积分, 这一技巧在贝叶斯分析中经常使用。例 1.50 的结果也可以通过符号计算取得。

下面我们将讨论三种不同类型的概率测度, 该分类直接导致离散型和连续型随机变量的概念。初学可暂时略过, 学到随机变量的分类时再回头看这部分。

定义 1.19. 对概率空间 (Ω, \mathcal{S}, P) 而言, 如果单点事件 $\{\omega\}$ 的概率 $P(\{\omega\}) > 0$, 则称 ω 是 P 的不连续点。

□ 没有不连续点的概率测度 P 称为连续概率测度。特别地, 当 $\Omega = \mathbb{R}^n, \mathcal{S} = \mathfrak{B}_n$, 令 m 是 (Ω, \mathcal{S}) 上的 Lebesgue 测度,

- ① 若 $\forall A \in \mathcal{S}, m(A) = 0$ 皆蕴含 $P(A) = 0$, 则称 P 关于 m 绝对连续, 简称绝对连续 (absolutely continuous), 记作 $P \leq m$ 。
- ② 若 $\exists A \in \mathcal{S}, m(A) = 0$ 使得 $P(A) = 1$, 则称 P 关于 m 奇异连续, 简称奇异连续 (singular continuous) 或奇异概率测度。

□ 如果对于不连续点的全体 D , 若有 $P(D) = 1$, 则称 P 为纯不连续的概率测度。

例 1.51. 不难看出例 1.25 和例 1.47 中的概率测度都是纯不连续的。高斯函数所定义的概率测度 (见练习 1.16) 是连续的, 因为其下没有概率非零的单点事件。另外, 它也是绝对连续的。

练习 1.17. 试证明: 概率测度不連續点的全体至多可数。

提示: 利用结果 “不可数个正数之和为 ∞ ” (若对任意 $n \in \mathbb{N}$, 大于 $\frac{1}{n}$ 的加数为有限个, 则指标集可数。所以, 必然存在 N 使得大于 $\frac{1}{N}$ 的加数无限多, 得证)。

※例 1.52 (奇异概率测度). 将单位闭区间 $I = [0, 1]$ 三等分, 中间那段记作 $I_1 = [1/3, 2/3]$ (见第 58 页中 Cantor 集的图示), 此为第一轮挑选区间。定义 $h(x) = 1/2, \forall x \in I_1$ 。

再将 $I - I_1$ 的两个区间分别三等分，其中间两段分别记作 I_2, I_3 ，即 $I_2 = [1/3^2, 2/3^2], I_3 = [7/3^2, 8/3^2]$ ，此为第二轮挑选区间。定义

$$h(x) = \begin{cases} 1/2^2 & \text{若 } x \in I_2 \\ 3/2^2 & \text{若 } x \in I_3 \end{cases}$$

在第 n 轮中，我们得到 2^{n-1} 个区间 $I_{2^{n-1}}, I_{2^{n-1}+1}, \dots, I_{2^n-1}$ 。定义

$$h(x) = \frac{1}{2^n} + \frac{k}{2^{n-1}} - 1, \quad \text{其中 } x \in I_k, k = 2^{n-1}, 2^{n-1} + 1, \dots, 2^n - 1$$

不难看出，函数 $h(x)$ 在 $\bigcup_{k=1}^{\infty} I_k$ 上是非减的，并且一致连续。这是因为

$$|x_1 - x_2| < \frac{1}{3^n} \Rightarrow |h(x_1) - h(x_2)| < \frac{1}{2^n}$$

根据 $\bigcup_{k=1}^{\infty} I_k$ 在 I 上稠密这一事实，以及完备一致空间的重要性质，函数 $h(x)$ 可唯一扩张成 I 上一个非减的连续函数。利用 h 定义 I 上的一个概率测度如下，

$$P(A) = m(h(A)), \quad \text{其中 } A \in \mathfrak{B}_1 \cap I$$

显然， $h(\bigcup_{k=1}^{\infty} I_k)$ 是可数的，所以其 Lebesgue 测度为零。按照 P 的定义，有

$$P\left\{\bigcup_{k=1}^{\infty} I_k\right\} = 0$$

此例的概率测度 P 是一个奇异概率测度，这是因为 Cantor 集 $C = I - \bigcup_{k=1}^{\infty} I_k$ 的 Lebesgue 测度为零，然而它的概率却是

$$P(C) = P(I) - P\left\{\bigcup_{k=1}^{\infty} I_k\right\} = 1$$

下面不加证明地介绍两个定理，它们有助于理解概率测度的分类和概率密度函数（见第 123 页的定义 2.12）的由来。

1. 概率测度的 Lebesgue 分解定理：它揭示了一般概率测度可分解为纯不连续、绝对连续和奇异连续三种概率测度的加权平均。
2. Radon-Nikodym 定理^{*}：绝对连续的概率测度可表示为某个几乎处处 (almost everywhere, a.e.) 非负的 Lebesgue 可测函数的积分。一个性质在测度空间上

^{*}奥地利数学家 Johann Radon (1887-1956) 于 1913 年证明了 \mathbb{R}^n 的特例，波兰数学家 Otto M. Nikodym (1887-1974) 于 1930 年证明了一般情况。

“几乎处处”成立，意味着使得此性质不成立的仅是一个零测集。例如，“几乎处处收敛”即是，除了一个零测集之外，在其他点上都收敛。某函数“几乎处处为零”意味着该函数不为零的点集为零测集。

定理 1.2 (Lebesgue 分解). 对于概率空间 (Ω, \mathcal{S}, P) ，若存在连续测度 m 使得 Ω 是可数个 m -测度有限的集合的并，则存在非负实数 $\alpha_1, \alpha_2, \alpha_3$ 和绝对连续、纯不连续和奇异连续概率测度 P_1, P_2, P_3 使得 P 可分解为

$$P = \alpha_1 P_1 + \alpha_2 P_2 + \alpha_3 P_3, \text{ 其中 } \alpha_1 + \alpha_2 + \alpha_3 = 1$$

定理 1.3 (Radon-Nikodym 定理). 如果概率空间 (Ω, \mathcal{S}, P) 满足下面两个条件，

- ① Ω 可以表示成可数个 Lebesgue 测度（记作 μ ）有限的集合的并集，
- ② P 关于 μ 绝对连续，

则 $\forall E \in \mathcal{S}$ ，总存在一个 Ω 上的一个几乎处处非负的、可积的可测函数 $f : \Omega \rightarrow [0, +\infty)$ 使得 P 可表示为 f 关于 μ 的 Lebesgue 积分（见附录 D），即

$$P(E) = \int_E f d\mu \tag{1.21}$$

定理 1.3 给函数 $P : \mathcal{S} \rightarrow [0, 1]$ 一个统一的积分表示，其中性质如此好的函数 f 被称为（关于 μ 绝对连续的）概率测度 P 的概率密度函数，或称为 P 关于 μ 的 Radon-Nikodym 导数，记作 $f = \frac{dP}{d\mu}$ 。显然，概率密度函数总满足

$$\int_{\Omega} f d\mu = 1$$

1.2.3 概率的一些基本性质

已知概率空间 (Ω, \mathcal{S}, P) , 根据 Kolmogorov 的三条公理, 不难推导出概率 $P(\cdot)$ 的一些基本性质如下。

性质 1.13. 对于任意事件 $A \in \mathcal{S}$, 皆有 $P(A^c) = 1 - P(A)$ 。

证明. 由非交分解 $\Omega = A + A^c$ 和概率的可列可加性可得 $P(\Omega) = P(A) + P(A^c)$, 再由归一性推得 $P(A^c) = 1 - P(A)$ 。 \square

推论 1.1. 不可能事件的概率为零, 即 $P(\emptyset) = 0$ 。

 由 $P(A) = 0$ 推导不出 $A = \emptyset$ 。类似地, 也不能由 $P(B) = 1$ 断言 B 必然发生。所以, Laplace 在《概率的哲学探讨》说到, “当人们说某件事情是事实时, 它就被认为是实实在在的, 而当人们说某事件发生的概率为 1, 那就意味着对它发生的断言还有修正的可能。”即便如此, 说某事件“以概率 1”发生, 在概率论里也可算作很强烈的语气了, 常用几乎必然 (almost surely, a.s.) 来作它的同义语 (见下面的例子)。相反, 一个事件的概率若为零, 我们常用“几乎必然不发生”来描述它。

例 1.53. 将 $(0, 1)$ 内的实数都以十进制表示成无限小数, 为了表示的唯一性, 不允许以 9 的循环结尾。假设每个小数 $x \in (0, 1)$ 被选中的机会等同, 令 $S_n(x)$ 表示十进制小数 $x \in (0, 1)$ 的小数点后面前 n 位数字中 0 的个数, $S_n(x)/n$ 就是 x 的前 n 位小数中 0 的频率。

 1909 年, Borel 发现事件 $\lim_{n \rightarrow \infty} S_n(x)/n = 1/10$ 几乎必然发生, 简记作 $S_n(x)/n \xrightarrow{\text{a.s.}} 1/10$ 或 $\lim_{n \rightarrow \infty} S_n(x)/n \stackrel{\text{a.s.}}{=} 1/10$, 即

$$P \left\{ \lim_{n \rightarrow \infty} \frac{S_n(x)}{n} = \frac{1}{10} \right\} = 1$$

如果把十进制改为 k 进制, 上述性质只需把 $1/10$ 改为 $1/k$ 即可, 对其他数字 $1, 2, \dots, k-1$ 也有相同的结果。

 1922 年, 苏联数学家 A. Ya. Khinchin 证得了更精细的结果 (参见定理 5.13)。

$$P \left\{ \limsup_{n \rightarrow \infty} \frac{|S_n(x) - \frac{n}{10}|}{\sqrt{n \ln \ln n}} = \frac{3\sqrt{2}}{10} \right\} = 1$$

性质 1.14. 如果 $A, B \in \mathcal{S}$ 且 $A \subseteq B$, 则 $P(A) \leq P(B)$ 。

证明. 由非交分解 $B = A + (B - A)$ 和概率的可列可加性知 $P(B) = P(A) + P(B - A)$, 再由概率的非负性知 $P(B - A) \geq 0$, 得证。 \square

推论 1.2. 对于任意事件 $A \in \mathcal{S}$, 皆有 $0 \leq P(A) \leq 1$ 。

性质 1.15. 已知概率空间 (Ω, \mathcal{S}, P) , 若 $\{A_j \in \mathcal{S} : j = 1, 2, \dots\}$ 是 Ω 的一个划分, 则 $\sum_{j=1}^{\infty} P(A_j) = 1$ 且对于任意事件 $B \in \mathcal{S}$ 皆有

$$P(B) = \sum_{j=1}^{\infty} P(BA_j) \quad (1.22)$$

证明. 利用非交分解 $B = \sum_{j=1}^{\infty} BA_j$ 即可证得。 \square

~定理 1.4 (和事件的概率). 如果 $A, B \in \mathcal{S}$, 则有如下的加法法则。

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

证明. 事件 $A \cup B$ 有非交分解 $A \cup B = (A - B) + (B - A) + AB$, 所以

$$P(A \cup B) = P(A - B) + P(B - A) + P(AB)$$

同理, $P(A) = P(A - B) + P(AB)$ 且 $P(B) = P(B - A) + P(AB)$ 。将这几个等式联立即可证得结果。 \square

推论 1.3. 对任意事件 $A, B \in \mathcal{S}$, 总有

$$P(A \cup B) \leq P(A) + P(B)$$

$$\text{并且 } P(A - B) = P(A) - P(AB)$$

练习 1.18. 若 $P(B) = 0$, 则 $P(A \cup B) = P(A)$ 。

提示: $P(A) \leq P(A \cup B) \leq P(A) + P(B) = P(A)$ 。

~练习 1.19 (Boole 不等式). 对于随机事件 $A_1, A_2, \dots \in \mathcal{S}$, 总有

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} P(A_k) \quad (1.23)$$

$$\text{等价地, } P\left(\bigcap_{k=1}^{\infty} A_k\right) \geq 1 - \sum_{k=1}^{\infty} P(A_k^c)$$

练习 1.20. 令 A 是概率等于 1 的事件 $A_j, j = 1, 2, \dots$ 的交集, 试证明 $P(A) = 1$ 。

将定理 1.4 进行一般化, 用数学归纳法不难验证如下结果。

定理 1.5 (Jordan 公式). 对于任意事件 $A_1, \dots, A_n \in \mathcal{S}$, 总有下面的等式成立。

$$\begin{aligned} P\left(\bigcup_{k=1}^n A_k\right) &= \sum_{k=1}^n P(A_k) - \sum_{k_1 < k_2} P(A_{k_1}A_{k_2}) + \\ &\quad \sum_{k_1 < k_2 < k_3} P(A_{k_1}A_{k_2}A_{k_3}) + \dots + (-1)^{n+1}P\left(\prod_{k=1}^n A_k\right) \end{aligned} \quad (1.24)$$

例 1.54. 有限集合 $A = \{a_1, a_2, \dots, a_n\}$ 上的一个置换即 A 到自身的一个一一映射。随机选取 A 的一个置换, 试问该置换没有不动点的概率?

解. 随机选取 A 的一个置换相当于把标号为 $1, 2, \dots, n$ 的 n 个球随机放入 n 个盒子, 每个盒子只能放一个球, 总共有 $n!$ 种不同的放法。令 A_k 表示事件“第 k 个盒子里装着第 k 号球”, 则

$$P(A_k) = \frac{(n-1)!}{n!}, \text{ 并且 } \sum_{k=1}^n P(A_k) = \frac{(n-1)!}{n!} \cdot C_n^1 = 1$$

第 k_1 个盒子装着第 k_1 号球, 且第 k_2 个盒子装着第 k_2 号球的概率 $P(A_{k_1}A_{k_2}) = (n-2)!/n!$, 于是

$$\sum_{k_1 < k_2} P(A_{k_1}A_{k_2}) = \frac{C_n^2(n-2)!}{n!} = \frac{1}{2!}$$

依次类推, $P(A_{k_1}A_{k_2}A_{k_3}) = (n-3)!/n!$ 且 $\sum_{k_1 < k_2 < k_3} P(A_{k_1}A_{k_2}A_{k_3}) = C_n^3(n-3)!/n! = 1/3!, \dots$ 。Jordan 公式 (1.24) 右侧的每一个求和项都能算出, 于是

$$\begin{aligned} P(\text{至少有一个不动点}) &= P\left(\bigcup_{k=1}^n A_k\right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n-1} \frac{1}{n!} \rightarrow 1 - \frac{1}{e} \\ P(\text{没有不动点}) &= 1 - P\left(\bigcup_{k=1}^n A_k\right) = \frac{1}{2!} - \frac{1}{3!} + \dots + (-1)^n \frac{1}{n!} \rightarrow \frac{1}{e} \end{aligned}$$

于是, 当 n 很大时随机选取的置换没有不动点的概率约为 e^{-1} 。

定义 1.20. 随机事件的序列 $A_1, A_2, \dots, A_n, \dots$ 若满足关系

$$A_1 \supseteq A_2 \supseteq \dots \supseteq A_n \supseteq \dots \supseteq A = \bigcap_{n=1}^{\infty} A_n \in \mathcal{S}$$

则称之为降序或单调减序列, 简记作 $A_n \downarrow A$ 。事件 A 称为降序极限。

随机事件的序列 $A_1, A_2, \dots, A_n, \dots$ 若满足关系

$$A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq \dots \subseteq A = \bigcup_{n=1}^{\infty} A_n \in \mathcal{S}$$

则称之为升序或单调增序列，简记作 $A_n \uparrow A$ 。事件 A 称为升序极限。

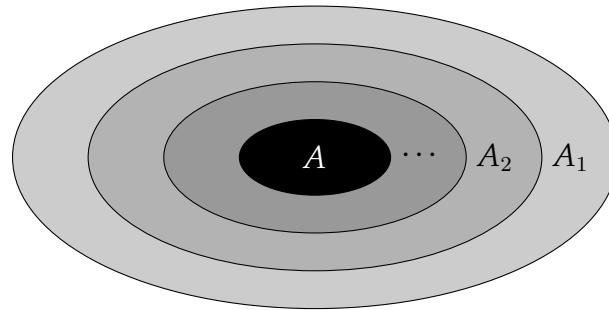


图 1.30: 若随机事件序列 $A_n \downarrow A$, 则 $A = \bigcap_{n=1}^{\infty} A_n$ 是 A_1, A_2, \dots 共有的“核”。

在数学分析里, 实函数 f 在 $x \in \mathbb{R}$ 处连续当且仅当对于任意收敛于 x 的单调序列 $x_1, x_2, \dots, x_n, \dots$, 皆有 $\lim_{n \rightarrow \infty} f(x_n) = f(x)$ 。概率 P 也有类似连续函数的如下性质。

定理 1.6 (概率的连续性定理). 对于随机事件的降序和升序, 其对应的概率序列收敛于降序极限的概率和升序极限的概率, 即

$$\text{若 } A_n \downarrow A, \text{ 则 } \lim_{n \rightarrow \infty} P(A_n) = P(A) \quad (1.25)$$

$$\text{若 } B_n \uparrow B, \text{ 则 } \lim_{n \rightarrow \infty} P(B_n) = P(B) \quad (1.26)$$

证明. 见图 1.30, 事件 $A_k A_{k+1}^c$ 就像一个套一个的环, 两两不交。由非交分解

$$A_n = \sum_{k=n}^{\infty} A_k A_{k+1}^c + A$$

我们有

$$\begin{aligned} P(A_n) &= P\left(\sum_{k=n}^{\infty} A_k A_{k+1}^c\right) + P(A) \\ &= \sum_{k=n}^{\infty} P(A_k A_{k+1}^c) + P(A) \end{aligned}$$

因为正项级数 $\sum_{k=1}^{\infty} P(A_k A_{k+1}^c)$ 收敛，所以余项极限为零，即

$$\lim_{n \rightarrow \infty} \sum_{k=n}^{\infty} P(A_k A_{k+1}^c) = 0$$

进而得证 $P(A) = \lim_{n \rightarrow \infty} P(A_n)$ 。结果 (1.26) 留作练习。 \square

\nwarrow 引理 1.1 (Borel-Cantelli^{*}, 1909, 1917). 若随机事件的序列 $A_1, A_2, \dots, A_n, \dots$ 满足 $\sum_{n=1}^{\infty} P(A_n) < \infty$ ，则 $P(\text{无穷多个 } A_n \text{ 发生}) = 0$ 。有时，我们把“无穷多个 A_n 发生”简记作“ $A_n \text{ i.o.}$ ”，其中 *i.o.* 表示 infinitely often。

证明. 已知级数 $\sum_{n=1}^{\infty} P(A_n)$ 收敛，所以 $\lim_{k \rightarrow \infty} \sum_{n=k}^{\infty} P(A_n) = 0$ 。如果无穷多个 A_n 发生，则对于任意的 $k < \infty$ 都有 $B_k = \bigcup_{n=k}^{\infty} A_n$ 发生，于是下述事件 A 发生。

$$A = \bigcap_{k=1}^{\infty} B_k = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n = \overline{\lim}_{n \rightarrow \infty} A_n$$

反之，若事件 A 发生，由结果 (1.12) 可知，有无穷多个 A_n 发生。综上所述，无穷多个 A_n 发生当且仅当 A 发生。对于降序 $B_k \downarrow A$ ，利用定理 1.6 和式 (1.23) 可以得出

$$P(A) = \lim_{k \rightarrow \infty} P(B_k) \leq \lim_{k \rightarrow \infty} \sum_{n=k}^{\infty} P(A_n) = 0 \quad \square$$

^{*}此结果由法国数学家 Émile Borel (1871-1956) 和意大利数学家 Francesco Paolo Cantelli (1875-1966) 分别于 1909 年和 1917 年得到。

1.3 条件概率与随机事件的独立性

实践中，人们常遇到这样的问题：吸烟者患肺癌的机会有多大？一般的作法是随机调查 n 个人，发现在 m 个吸烟者中有 k 个人患肺癌。问题的答案近似为

$$\frac{k}{m} = \frac{k/n}{m/n} = \frac{\text{吸烟的肺癌患者占总人群的比例}}{\text{吸烟者在总人群中的比例}}$$

此处，“吸烟者”是一个条件，对“患肺癌”这个随机事件的考察要受到该条件的制约。毫不夸张地讲，科学实践中几乎所有与概率有关的问题都是带条件的，条件可以是假设，可以是观察到的数据，也可以是验前已知的信息等等。

譬如天气预报，研究者总是根据当前对大气状况或特定气象指标（如气压、温度、湿度等）的观察来预测未来某一时间各种天气情况的可能性，这些当前的观察结果就是条件。

例 1.55. Laplace 在《概率的哲学探讨》中曾举过这样一个例子：仅仅知道盒子 A, B, C 中有两个盒子装着白球，一个盒子装着黑球，则从 C 盒子摸出黑球的概率为 $1/3$ 。如果已知 A 盒子装着白球，则从 C 盒子摸出黑球的概率就是 $1/2$ 。Laplace 借此说明已知条件对概率的影响。

条件概率的计算依靠全概率公式和 Bayes 公式，后者归功于英国数学家、长老会牧师 Thomas Bayes (1701?-1761)。Bayes 留给后世的资料很少，甚至右边这幅他唯一的画像也可能是假的。Bayes 生前发表过的两篇文章都与概率论无关，但他的遗作《论有关机遇问题的求解》(1763) 却给他带来了无尽的荣耀，在这篇论文中他推导出了逆概率公式，即著名的 Bayes 公式。

很难评述 Bayes 本人对概率的哲学认识，他的学说被后继者们赋予了更广泛、更深刻的理解，以至发展成为贝叶斯学派，甚至贝叶斯主义 (Bayesianism)。如今人们只能从 Bayes 的这篇重要论文中探究他的思想，多数研究者将他归为主观贝叶斯主义者 [8, 148]。

根据 Laplace 在《概率的哲学探讨》中构建的基于概率的归纳推理模型来看，他也是一个贝叶斯主义者。Laplace 说，“当某一事件发生的概率尚属未知时，可以假定它是从 0 到 1 的任何一个值……。如果发现一个事件已经持续发生了若干次，那么它下次再发生的概率等于一个跟发生次数有关的数。考虑从五千年前或 1826213 天前开始算起，在此期间，太阳每隔 24 小时就要重新升起，那么赌明天太阳照常升起的胜算肯定是 1826214 比 1。”为何胜算是这样一个比例，要说清楚它需要一些统计学的知识，详情见第 661 页的式 12.2.3，这是后话。



在概率统计发展史中，频率派一直占据主导地位，贝叶斯学派的学说算不上主流，但近些年来情况有所改变。

- 经过多年的概率哲学基础之争 [105]，频率派从贝叶斯学派那里不断汲取营养，二十世纪五十年代频率派中兴起的经验 Bayes 方法 (empirical Bayes method) 就是一个很好的例证。
- 人们不再满足于哲学上的思辨，更多看重的是算法和实践的效果 [55]，随机模拟技术的进步和对小样本分析的需求让越来越多的学者关注贝叶斯学派。第 12 章将介绍贝叶斯统计学和贝叶斯数据分析的常用方法。

频率派和贝叶斯学派对 Bayes 公式的理解是不同的，于是用它来作推断的手法也不相同。Bayes 公式是贝叶斯推断的核心，Laplace 称之为“最基本原理”。贝叶斯推断常利用事件的验后概率（即试验之后的概率，也称后验概率，posterior probability），或比较它与验前概率（也称先验概率，prior probability）的差异来揭示数据是否支持该事件的发生。而频率派无验前概率一说，常直接利用条件概率的比较来作推断。

有些频率派的学者对 Bayes 公式的使用持非常谨慎的态度，如英国著名统计学家 R. A. Fisher (1890-1962) 等。

条件概率引发了对随机事件独立性和条件独立性的思考。读者谨记，世间万物都是有联系的，独立性是概率测度的性质，而不是事件本身的性质！

概率模型有时通过简化研究对象间的关系来降低算法复杂度，会对某些事件做出（条件）独立性假设。这样的假设非常之强烈，需慎重使用。

揭示研究对象之间的不独立也是很重要的。例如，通过随机调查发现吸烟人群中患肺癌的比例远高于所有人群中肺癌的患病率，可以断言“吸烟”和“患肺癌”之间不是相互独立的。至于二者之间是否有直接的因果关系，还需要进行因果推断 [118]，本书不予涉猎。

本节内容

第一小节在 Kolmogorov 公理体系下诱导出条件概率的定义并讨论它的性质。第二小节的重点是条件概率的两个经典结果——全概率公式和 Bayes 公式。我们介绍频率派和贝叶斯学派对 Bayes 公式的不同理解，以及如何利用它进行推断。第三、四小节分别讨论随机事件的独立性、条件独立性及其应用，如 Borel 0-1 律、即时雇佣问题、概率数论中的 Chebyshev 问题、随机上下文无关文法、多专家决策系统、贝叶斯垃圾邮件过滤等。

关键知识

(1) 条件概率；(2) 全概率公式和 Bayes 公式；(3) 贝叶斯推断；(4) 独立性、条件独立性；(5) Borel 0-1 律。

1.3.1 条件概率及其性质

先从下面简单的例子出发，引出条件概率的定义。

例 1.56. 掷两个均匀的骰子，基本事件集合是 $\Omega = \{(i, j) : i, j = 1, 2, \dots, 6\}$ ，样本空间为 $(\Omega, 2^\Omega)$ 。令 A 表示随机事件“点数相同”，令 B 表示随机事件“点数之和小于 6”，则 $P(A) = 1/6, P(B) = 5/18, P(AB) = 1/18$ 。现在已知事件 B 发生了，问事件 A 发生的概率？

解. 已知事件 B 发生了，所以掷双骰子的试验结果只可能是下述集合的某个元素。

$$\tilde{\Omega} = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (4, 1)\}$$

进而，得出 A 发生的概率是 $1/5$ ，数值上等于 $P(AB)/P(B)$ 。

Laplace 在《概率的哲学探讨》中这样描述概率论的一般原理之五，“如果我们预先已经计算出某观察到的事件的概率以及由该事件和另一待考察事件复合而成的事件的概率，则第二个概率值除以第一个概率值就得出了待考察事件的概率。”

换句话说，已知事件 B 发生的前提下，比值 $P(AB)/P(B)$ ，即 $P(B)$ 中 $P(AB)$ 占几成，刻画了事件 A 发生的机会大小。用这样的比值可以定义一个概率测度吗？

定理 1.7 (条件概率). 已知概率空间 (Ω, \mathcal{S}, P) ，并且事件 $B \in \mathcal{S}$ 满足 $P(B) > 0$ 。在事件 B 已经发生的情况下，对任意事件 $A \in \mathcal{S}$ ，定义

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

则 $P_B(\cdot)$ 也是 \mathcal{S} 上的一个概率测度，即 $(\Omega, \mathcal{S}, P_B)$ 也是一个概率空间，我们称之为条件概率空间。概率 $P_B(A)$ 称为 B 发生后事件 A 的条件概率 (conditional probability)，通常记作 $P(A|B)$ 。

证明. 下面依次验证 P_B 满足 Kolmogorov 的三条公理。

1. $\forall A \in \mathcal{S}$ ，显然有 $P_B(A) = P(AB)/P(B) \geq 0$ ，非负性成立。
2. $P_B(\Omega) = P(\Omega B)/P(B) = 1$ ，归一性成立。
3. 如果 $A_1, A_2, \dots \in \mathcal{S}$ 两两不交，则

$$P_B\left(\sum_{j=1}^{\infty} A_j\right) = \frac{1}{P(B)} P\left(B \sum_{j=1}^{\infty} A_j\right) = \frac{1}{P(B)} \sum_{j=1}^{\infty} P(BA_j) = \sum_{j=1}^{\infty} P_B(A_j)$$

综上所述， $(\Omega, \mathcal{S}, P_B)$ 构成一个概率空间。 □

 读者可以把概率空间 $(\Omega, \mathcal{S}, P_B)$ 看作是由 (Ω, \mathcal{S}, P) 和事件 B 诱导出来的概率空间，用来考察 B 已发生的情况下事件 $A \in \mathcal{S}$ 的概率。记号 $P_B(A)$ 和 $P(A|B)$ 的使用依语境而定，一般教科书里直接按下面的方式定义条件概率。

$$P(A|B) = \frac{P(AB)}{P(B)} \quad (1.27)$$

例 1.57. 袋子里装着三枚硬币，其中一枚是正常的硬币，两面的图案分别是 H 和 T 。另外两枚硬币是错币，一枚两面都是 H ，一枚两面都是 T 。现在，从袋子中随机地摸出一枚硬币，观察到图案 H ，分别求另一面图案是 T 和 H 的概率。

解. 三枚硬币被摸到的机会等同，则观察到 HH, HT, TH, TT 的概率分别为

观察结果	HH	HT	TH	TT
概率	$1/3$	$1/6$	$1/6$	$1/3$

根据式 1.27，另一面图案是 H 和 T 的概率分别是

$$\begin{aligned} P(H|H) &= \frac{P(HH)}{P(H)} = \frac{1/3}{1/2} = \frac{2}{3} \\ P(T|H) &= \frac{P(HT)}{P(H)} = \frac{1/6}{1/2} = \frac{1}{3} \end{aligned}$$

练习 1.21. 若恰有两个孩子的家庭（ B 代表男孩， G 代表女孩，长幼有序）服从分布 $\frac{1}{4}\langle BB \rangle + \frac{1}{4}\langle GB \rangle + \frac{1}{4}\langle BG \rangle + \frac{1}{4}\langle GG \rangle$ ，已知家中一个孩子是男孩，问另一个是女孩的概率是多少？（答案： $2/3$ 。在已知条件下，只需考虑家庭类型 GB, BG, BB ）

练习 1.22. 在恰有两个孩子的家庭中随机找一个男孩（不妨叫他 A ），问他有兄弟的概率是多少？（答案： $1/2$ 。所有可能的情况是 AB, BA, AG, GA ）

推论 1.4. 条件如定理 1.7 所述，若 $A_1, A_2, \dots, A_j, \dots$ 是 Ω 的一个划分，则总有

$$\sum_{j=1}^{\infty} P(A_j|B) = 1$$

练习 1.23. 已知概率空间 (Ω, \mathcal{S}, P) ，且事件 $B \in \mathcal{S}$ 满足 $P(B) > 0$ 。请验证：

1. $P(A|B) = 1$ ，其中 $B \subseteq A \in \mathcal{S}$ ，并请读者给出直观的解释。
2. 对于定理 1.7 定义的条件概率空间 $(\Omega, \mathcal{S}, P_B)$ ，如果事件 $C \in \mathcal{S}$ 满足 $P_B(C) > 0$ ，则 $P_B(A|C) = P(A|BC)$ 。
3. $\mathcal{S}_B = \mathcal{S} \cap B = \{E \cap B : E \in \mathcal{S}\}$ 也是 B 上的 σ 域。

4. $\forall C \in \mathcal{S}_B$, 定义 $P_B(C) = P(C)/P(B)$, 则 (B, \mathcal{S}_B, P_B) 构成一个概率空间。

例 1.58. 已知 $P(B) > 0$, 则

$$P(A|B) \geq \frac{P(A) + P(B) - 1}{P(B)}$$

证明. 由 $P(A \cup B) \leq 1$ 和定理 1.4 得, $P(B) - P(AB) \leq P(A^c)$ 或者 $P(A|B) \geq 1 - P(A^c)/P(B) = [P(A) + P(B) - 1]/P(B)$, 得证。 \square

~定理 1.8. 如果 $P(AB) > 0$, 则有下面的结果成立。

① 积事件 AB 的概率*有如下的分解:

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \quad (1.28)$$

Laplace 在《概率的哲学探讨》中把式 (1.28) 列为概率论的一般原理之四——“如果两个事件彼此相关, 那么它们同时发生的概率等于第一个事件发生的概率乘以在第一个事件已发生的前提下, 第二个事件发生的概率。”

② 如果 $P(A|B) > P(A)$, 则 $P(B|A) > P(B)$ 。也就是说, 如果 B 的发生利于 A 的发生, 则 A 的发生也利于 B 的发生。

证明. 因为 $AB \subseteq A$ 且 $P(AB) > 0$, 所以 $P(A) > 0$, 同理 $P(B) > 0$ 。由条件概率的定义式 (1.27) 可以得到式 (1.28)。 \square

推论 1.5 (积事件的概率). 把式 (1.28) 推广到 $A_1A_2 \cdots A_n$ 的情形: 已知概率空间 (Ω, \mathcal{S}, P) , 且事件 $A_1, A_2, \dots, A_n \in \mathcal{S}$ 满足 $P(A_1A_2 \cdots A_{n-1}) > 0$, 则有乘法法则如下。

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots P(A_n|A_1A_2 \cdots A_{n-1}) \quad (1.29)$$

证明. 由已知条件和 $A_1A_2 \cdots A_{n-1} \subseteq A_1A_2 \cdots A_{n-2} \subseteq \cdots \subseteq A_1A_2 \subseteq A_1$ 可得 $0 < P(A_1A_2 \cdots A_{n-1}) \leq P(A_1A_2 \cdots A_{n-2}) \leq \cdots \leq P(A_1A_2) \leq P(A_1)$ 。因此,

$$\begin{aligned} P(A_1A_2 \cdots A_n) &= P(A_1A_2 \cdots A_{n-1})P(A_n|A_1A_2 \cdots A_{n-1}) \\ &= P(A_1A_2 \cdots A_{n-2})P(A_{n-1}|A_1A_2 \cdots A_{n-2})P(A_n|A_1A_2 \cdots A_{n-1}) \end{aligned}$$

将 $P(A_1A_2 \cdots A_{n-2})$ 继续分解下去, 经过有限步便得到式 (1.29)。 \square

例 1.59. 一批零件共 100 个, 次品率为 5%。不放回地随机抽取零件, 问第三次才取得合格品的概率?

*积事件 AB 的概率也称为事件 A, B 的联合概率, 有时记作 $P(A, B)$ 。类似地, 条件概率 $P(A|BC)$ 有时也记作 $P(A|B, C)$, 都表示“ B, C 同时发生的条件下 A 的概率”。

解. 令 A_k 表示事件“第 k 次抽取的零件是次品”， $k = 1, 2, 3$ 。问题所求概率是 $P(A_1 A_2 A_3^c)$ ，利用乘法法则，即式 (1.29) 可得

$$P(A_1 A_2 A_3^c) = P(A_1)P(A_2|A_1)P(A_3^c|A_1 A_2) = \frac{5}{100} \times \frac{4}{99} \times \frac{95}{98} \approx 0.002$$

例 1.60. 盒子里有一黑一白两个球，一次抽取一个球，直至抽到黑球。若抽到白球，除了放回还要再补充两个白球回盒子，问前 n 次抽取中黑球不出现的概率 $P(n)$ ？

解. 前 n 次抽球中每次都是白球，所以

$$P(n) = \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdots \frac{2n-1}{2n} = \frac{(2n)!}{2^{2n}(n!)^2}$$

利用 Stirling 公式 (1.5)，当 n 充分大时， $P(n) \approx 1/\sqrt{\pi n}$ 。事实上，当 $n \geq 10$ 时， $P(n)$ 与其近似 $1/\sqrt{\pi n}$ 就已经非常接近了。

例 1.61. 盒子里有 m 个黑球和 n 个白球 ($m \geq n$)，不放回地连续抽取 n 次，每次抽取两个球，试问：每次抽取都是一黑一白的概率 p ？

解. 令 A_k 表示事件“第 k 次抽取了一黑一白”， $k = 1, 2, \dots, n$ 。

$$\begin{aligned} p &= \frac{C_m^1 C_n^1}{C_{m+n}^2} \times \frac{C_{m-1}^1 C_{n-1}^1}{C_{m+n-2}^2} \times \cdots \times \frac{C_{m-(n-1)}^1 C_{n-(n-1)}^1}{C_{m+n-2(n-1)}^2} \\ &= \frac{2^n n! m!}{(n+m)!} \end{aligned}$$

例 1.62. 受存储条件的限制，我们只能依次读取集合 $S = \{x_1, \dots, x_n\}$ 中的元素，最多记录下其中的 k 个元素，其中 n 未知， k 已知。问如何从 S 不放回地随机抽取 k 个元素，保证它们被抽中的概率是 $1/C_n^k$ ？

解. 初始化结果为 $r_1 = x_1, \dots, r_k = x_k$ 。下面考虑 S 的第 $j \in \{k+1, \dots, n\}$ 个元素的去留问题：

- 从 $1, 2, \dots, j$ 中等概率地抽取一个数 u^* 。
- 若 $u^* \leq k$ ，则 $r_{u^*} \leftarrow x_j$ 。否则，令 $j \leftarrow j+1$ ，重复上述步骤直至最后一个元素。结果 r_1, \dots, r_n 便是问题所求。

上述算法称为水库抽样 (reservoir sampling)^{*}：以概率 k/j 放走水库中的某个元素，同时放入新元素 x_j 。第 $j \in \{k+1, \dots, n\}$ 个元素被抽中的概率是

$$\frac{k}{j} \times \left(\frac{j}{j+1} \times \cdots \times \frac{n-1}{n} \right) = \frac{k}{n}$$

^{*}水库抽样是统计学家赵民德于 1982 年提出的在线无放回抽样通用方法（赵方法）的一个特例。

类似地，第 $j \in \{1, \dots, k\}$ 个元素被抽中的概率是

$$\frac{k}{k+1} \times \left(\frac{k+1}{k+2} \times \cdots \times \frac{n-1}{n} \right) = \frac{k}{n}$$

于是，任意 k 个元素被抽中的概率都是

$$\frac{k}{n} \times \frac{k-1}{n-1} \times \cdots \times \frac{1}{n-k+1} = \frac{1}{C_n^k}$$

练习 1.24. 把式 (1.29) 做一个推广：已知概率空间 (Ω, \mathcal{S}, P) ，事件 $B, A_1, A_2, \dots, A_n \in \mathcal{S}$ 满足 $P(BA_1A_2 \cdots A_{n-1}) > 0$ ，则

$$P(A_1A_2 \cdots A_n | B) = P(A_1 | B) \prod_{k=2}^n P(A_k | BA_1A_2 \cdots A_{k-1}) \quad (1.30)$$

特别地，我们得到了式 (1.27) 和式 (1.28) 的“条件概率版本”。

$$\begin{aligned} P(A_2 | BA_1) &= \frac{P(A_1A_2 | B)}{P(A_1 | B)} \\ P(A_1A_2 | B) &= P(A_2 | BA_1)P(A_1 | B) \end{aligned} \quad (1.31)$$

1.3.2 全概率公式与 Bayes 公式

本节的主要内容是概率计算的两个重要公式——全概率公式和 Bayes 公式及其应用，先通过两个练习题来了解这两个公式的产生背景。

练习 1.25. 在盒子 A_1 中有 3 个白球和 2 个黑球，在盒子 A_2 中有 1 个白球和 4 个黑球。试验者被蒙上双眼，先选盒子，再从盒子里摸球。如果选中 A_1 和 A_2 的机会等同，即 $P(A_1) = P(A_2) = 1/2$ ，问摸到白球的概率？提示：利用下面的全概率公式。

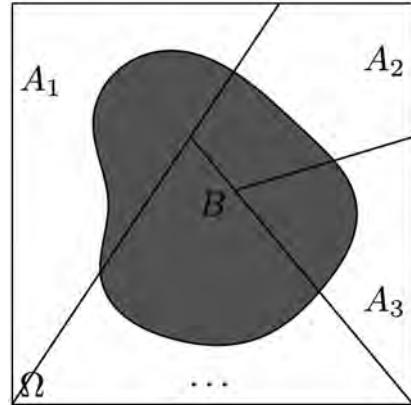
→**定理 1.9 (全概率公式).** 已知概率空间 (Ω, \mathcal{S}, P) 且 $\{A_j \in \mathcal{S} : P(A_j) \neq 0, j = 1, 2, \dots\}$ 是 Ω 的一个划分，则对任一事件 B 皆有

$$P(B) = \sum_{j=1}^{\infty} P(A_j)P(B|A_j) \quad (1.32)$$

证明. 由式 (1.22) 和式 (1.28) 可证。 □

Laplace 在《概率的哲学探讨》中把**定理 1.9** 列为概率论的一般原理之七，“未来事件的概率等于所有导致该事件发生的原因的概率乘以在该原因下该事件发生的概率的总和。”

例 1.63. 接着第 47 页的**例 1.30**，假设 Ω 是基本事件集合，每个基本事件发生的概率都是 $1/7$ ，则事件 B = “素数”的概率是 $4/7$ 。令 $A_1 = \{1, 4, 7\}$, $A_2 = \{2, 5\}$, $A_3 = \{3, 6\}$ 。下面计算式 (1.32) 的右边，显然



$$\begin{aligned} & P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) \\ &= \frac{3}{7} \cdot \frac{1}{3} + \frac{2}{7} \cdot 1 + \frac{2}{7} \cdot \frac{1}{2} = \frac{4}{7} \end{aligned}$$

图 1.31: **定理 1.9** 的直观解释：事件 B 被划分成多个碎片，其概率就是这些碎片概率之和。

将**定理 1.9** 用于上面的**练习 1.25**，不难得到摸到白球的概率是

$$P(\text{白球}) = P(A_1)P(\text{白球}|A_1) + P(A_2)P(\text{白球}|A_2) = \frac{1}{2} \left(\frac{3}{5} + \frac{1}{5} \right) = \frac{2}{5}$$

由**性质 1.6**，全概率公式 (1.32) 中， $P(B|A_j)$ 就是从等价类 A_j 的角度看到事件 B 发生的概率。利用全概率公式，我们可以把一个复杂问题划分成几个简单的小问题，各个击破后再拼出原问题的结果。

练习 1.26. 接着**练习 1.25**，如果摸到了白球，问该白球从 A_1 盒子和 A_2 盒子中摸出的概率各是多少？提示：利用下面的 Bayes 公式。

条件概率 $P(A_j|B)$ 的直观含义是 B 的碎片 BA_j 的概率 $P(BA_j)$ 占 $P(B)$ 的几成 (见图 1.31)。在定理 1.9 的条件之下, 它可以用来考察在事件 B 发生的前提下, 哪个类别是最有可能的。

\rightsquigarrow 定理 1.10 (Bayes 公式或逆概率公式). 在定理 1.9 的条件之下, 对于任一事件 $B \in \mathcal{S}$, 如果 $P(B) > 0$, 我们有

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{j=1}^{\infty} P(A_j)P(B|A_j)} \quad (1.33)$$

证明. 由 $P(A_j|B) = P(A_j)P(B|A_j)/P(B)$ 和全概率公式 (1.32) 可证。 \square

早在 1786 年, Laplace 就指出逆概率的理论是基于关系 $P(A|B) = P(AB)/P(B)$ 。在《概率的哲学探讨》中 Laplace 提出概率论的第六条原理, “如果导致某一被考虑的事件的发生有各种原因, 每种原因的可能性设想为导致该事件发生的概率, 某种原因成立的概率是个分数, 分子是该原因下该事件发生的概率, 分母是所有相应原因下该事件发生的概率总和。”

Laplace 所说的“某一被考虑的事件”就是式 (1.33) 中的 B , 而“各种原因”就是式 (1.33) 中的 $A_j, j = 1, 2, \dots$, 如果各种原因是等可能的, 则有 Bayes 公式 (1.33) 的一个简化版本。

$$P(A_j|B) = \frac{P(B|A_j)}{\sum_{j=1}^{\infty} P(B|A_j)}, \text{ 其中 } A_1, A_2, \dots \text{ 是等可能的}$$



Laplace 进一步论述, “如果先验地考虑一些非等可能的原因时, 那么在计算分母时, 不用该事件在每个原因下发生的概率, 而必须用这一概率乘以该原因自身的可能性。”这便是 Bayes 公式 (1.33), 历史上 Laplace 首次清楚地阐述了 Bayes 的工作, Laplace 也是贝叶斯学派的先驱人物。

例 1.64. 在练习 1.25 中, 把有关 $P(A_1), P(A_2)$ 的条件放宽为 $P(A_1) + P(A_2) = 1$, 如果摸到白球, 问该白球从哪个盒子摸出的可能性大?

解. 不妨设 $P(A_1) = p, P(A_2) = 1 - p$, 其中 $0 < p < 1$ 。

$$\begin{aligned} P(A_1|\text{白球}) &= \frac{P(\text{白球}|A_1)P(A_1)}{P(\text{白球}|A_1)P(A_1) + P(\text{白球}|A_2)P(A_2)} = \frac{3p}{2p+1} \\ P(A_2|\text{白球}) &= 1 - P(A_1|\text{白球}) = \frac{1-p}{2p+1} \end{aligned}$$

直觉上，该白球从 A_1 盒子摸出的可能性更大些，否则一个小概率事件便发生了 (A_2 盒子中摸到白球的概率是 $1/5$)。而在人们的理念中，更倾向于相信发生了的事件是一个大概率事件。

下面将介绍两个截然不同的推断方法，它们有时能得出相同的结论，有时不能。无所谓对与错，推断模式本来就不是唯一的，其基础是哲学而不是数学。请读者自己评判在你的观念中哪个更合理些，或者更容易接受些。

- 频率派：只有当 Bayes 公式中的每个概率项都有频率解释时， $P(A_1|\text{白})$ 和 $P(A_2|\text{白})$ 才具有客观意义，进而利用 Bayes 公式作推断才是合理的。按照不充分理由原则 (principle of insufficient reason)*，我们假定 $P(A_1) = P(A_2) = 1/2$ (可以想像 A_1, A_2 两个盒子装在一个大箱子内，每个盒子被随机选中的概率都是 50%)。由 Bayes 公式得 $P(A_1|\text{白球}) = 3/4 > P(A_2|\text{白球}) = 1/4$ ，所以该白球从 A_1 盒子摸出的可能性大。
- 贝叶斯学派：如果把概率理解为“相信某不确定事件会发生”的信念度，则观察到白球之后“相信从 A_1 盒子摸球”的信念度就是 $P(A_1|\text{白球})$ ，它的大小依赖于试验之前的信念度 $P(A_1)$ 和观察结果。贝叶斯学派允许对 $P(A_1)$ 表达个人意志，可以偏执地认为 $P(A_1) = 1/10$ ，并由 $P(A_1|\text{白球}) = 1/4 < P(A_2|\text{白球}) = 3/4$ 坚持认为该白球从 A_2 盒子摸出的可能性大。

 贝叶斯学派允许个体的先验知识介入推断，符合人类思维的常规。先入之见的确会影响判断，对待同一事物才会有不同的“经验之谈”。另外，试验前后信念度的差别 $P(A_1|\text{白球}) - P(A_1)$ ，不正刻画了观测数据“白球”给信念度带来的变化吗？

$$\begin{aligned} P(A_1|\text{白球}) - P(A_1) &= \frac{3p}{2p+1} - p > 0 \\ P(A_2|\text{白球}) - P(A_2) &= \frac{1-p}{2p+1} - (1-p) < 0 \end{aligned}$$

不管验前、验后对 A_1 和 A_2 的信念度孰大孰小，所有的贝叶斯主义者都不得不接受这样的事实——观测数据支持答案“ A_1 ”。根据后验概率的大小，推断者依然可以坚持选 A_2 ，但必须承认当前的观察结果降低了该信念的程度。通过考察验前、验后信念度的改变来评判观测数据支持哪个论断，这个推断模式不依赖于推断者的主观先验，反映出贝叶斯推断 (Bayesian inference) 中客观的一面。

例 1.65. 我们换一个角度解释上例中的“球-盒子”模型：把“白球”理解为“医院诊断有 Z 病”，把“黑球”理解为“医院诊断没 Z 病”，把“ A_1 盒子”理解为

*这一原则最早由 Jacob Bernoulli 提出，1921 年经济学家 John Maynard Keynes (1883-1946) 将之称为无差别原则 (principle of indifference)：当我们对基本事件的概率一无所知的时候，每个基本事件都不比其他基本事件优先发生，可以假定所有基本事件都是等概率的。该约定虽然不能用逻辑来证明，但与经验相吻合，所以被广泛接受。

“患有 Z 病”，把 “ A_2 盒子” 理解为 “未患 Z 病”。已知 Z 病的患病率不高，譬如 $P(A_1) = 1/10, P(A_2) = 9/10$ 。通常情况下，对医学诊断的评价有两个常见指标：

1. 敏感度 (sensitivity)，又称真阳性率 (true positive rate, TPR)，即患有 Z 病者被诊断为 “有 Z 病” 或 “阳性”的概率 $P(\text{白球}|A_1)$ ，此值越大诊断越灵敏。此例中， $P(\text{白球}|A_1) = 3/5$ 表明医院诊断有 Z 病的正确率为 $3/5$ 。
2. 特异度 (specificity)，又称真阴性率 (true negative rate, TNR)，即未患 Z 病者被诊断为 “没 Z 病” 或 “阴性”的概率 $P(\text{黑球}|A_2)$ ，此值越大诊断越精确。此例中， $P(\text{黑球}|A_2) = 4/5$ 说明医院诊断没 Z 病的正确率为 $4/5$ 。

因为 $P(A_1|\text{白球}) = 1/4 < P(A_2|\text{白球}) = 3/4$ ，这意味着，医院诊断有 Z 病，实际未患 Z 病的可能性更大些。所以，两个学派都选 “ A_2 ”，即 “未患 Z 病”。但它们对 $P(A_1) = 1/10, P(A_2) = 9/10$ 的理解是完全不同的。

- 频率派的患者会乐观地想，首先患 Z 病的可能性小，其次即使患有 Z 病，医院诊断的敏感度也不高，我怎么就不幸中招了呢？一定是医院诊断有误！
- 贝叶斯学派的患者也可以乐观，但不得不承认该诊断结果支持 “患有 Z 病”。这似乎更符合常人的心态，相信医院的诊断具有一定的说服力。

练习 1.27. 那些非此即彼的分类问题被称为二分类 (binary classification) 问题，如 [例 1.64](#)。二分类的性能评估一般基于下面的混淆矩阵 (confusion matrix)。

表 1.4: 二分类问题的混淆矩阵：表中括号里的字符是满足所述性质的样本个数。譬如，假阴性表示 “错误地判定为阴性”，其个数是 FN 。

判定结果	真实情况	实际阳性 (P)	实际阴性 (N)
理论阳性 (P')	真阳性 (TP)	假阳性 (FP)	
理论阴性 (N')	假阴性 (FN)	真阴性 (TN)	

请读者解释下述二分类性能指标的含义。

$$\begin{aligned} \text{准确率: } ACC &= \frac{TP + TN}{P + N} \\ \text{真阳性率: } TPR &= \frac{TP}{P} = \frac{TP}{TP + FN} \\ \text{假阳性率: } FPR &= \frac{FP}{N} = \frac{FP}{FP + TN} \\ \text{真阴性率: } TNR &= \frac{TN}{N} = \frac{TN}{FP + TN} = 1 - FPR \end{aligned}$$

例 1.66. 有枚硬币，抛出正面的概率 $p = 0.8$ 或者 $p = 0.4$ 。抛 10 次该硬币出现了 8 次正面，问 $p = 0.8$ 和 $p = 0.4$ 哪个更有可能？

解. 令 D 表示事件“抛 10 次该硬币出现了 8 次正面”。令 A_1 表示事件“ $p = 0.8$ ”， A_2 表示事件“ $p = 0.4$ ”，无观察数据时，二者是等可能的，即 $P(A_1) = P(A_2) = 0.5$ 。我们只需比较 $P(A_1|D)$ 和 $P(A_2|D)$ 孰大孰小。

$$\begin{aligned} P(A_1|D) &= \frac{P(D|A_1)P(A_1)}{P(D|A_1)P(A_1) + P(D|A_2)P(A_2)} \\ &= \frac{C_{10}^8 \cdot 0.8^8 \cdot 0.2^2 \cdot 0.5}{C_{10}^8 \cdot 0.8^8 \cdot 0.2^2 \cdot 0.5 + C_{10}^8 \cdot 0.4^8 \cdot 0.6^2 \cdot 0.5} \\ &\approx 0.9660377 \\ P(A_2|D) &= 1 - P(A_1|D) \\ &\approx 0.0339623 \end{aligned}$$

显然抛出正面的概率更有可能是 $p = 0.8$ 。事实上，我们无需计算 $P(A_1|D)$ 和 $P(A_2|D)$ ，只需比较 $P(D|A_1)$ 和 $P(D|A_2)$ 即可。这是因为 $P(A_1|D) = P(D|A_1)P(A_1)/P(D)$ 与 $P(A_2|D) = P(D|A_2)P(A_2)/P(D)$ 中 $P(A_1) = P(A_2)$ 。

定义 1.21. 在条件 A 下观察到数据 D 的概率 $P(D|A)$ 被称为似然 (likelihood)。如上例所示，似然常被频率派用来比较哪个条件更有可能。

练习 1.28. 试用似然来解决例 1.64 的问题。

1.3.3 随机事件的独立性

如果 $P(B) > 0$, $P(A|B) = P(A)$ 意味着事件 B 发生与否丝毫不影响事件 A 发生的概率, 也意味着 $P(AB) = P(A)P(B)$, 由此引出了随机事件之间独立性的定义, 它是概率测度的特殊性质, 与通常表达“两不相干”意思的词语“独立”是不同的。

定义 1.22 (独立性). 已知概率空间 (Ω, \mathcal{S}, P) , 事件 $A, B \in \mathcal{S}$ 相互独立 (independent) 当且仅当 $P(AB) = P(A)P(B)$, 记作 $A \perp\!\!\!\perp B$ 或者 $\perp\!\!\!\perp \{A, B\}$ 。

更一般地, 事件 $A_1, A_2, \dots, A_n \in \mathcal{S}$ 独立, 记作 $\perp\!\!\!\perp \{A_1, A_2, \dots, A_n\}$, 当且仅当对于 $1, 2, \dots, n$ 的任意子序列 $k_1 < k_2 < \dots < k_s$,

$$P(A_{k_1} A_{k_2} \cdots A_{k_s}) = P(A_{k_1})P(A_{k_2}) \cdots P(A_{k_s}) \quad (1.34)$$

Laplace 在《概率的哲学探讨》中把式 (1.34) 列为概率论的一般原理之三, “如果几个事件是相互独立的, 那么它们同时发生的概率就是它们单个发生时的概率的乘积。”

Laplace 在论述该原理时还提到, “一个可能性很大的事件在多次重复的过程中连续发生就变得十分不可能。设想某件事情由二十个人依次相传, 设每次传递的可信度为 $9/10$, 经过二十次后可信度将不及 $1/8$ 。……历史学家对这种经过多年之后事件可靠性的降低却很少给予足够的重视。基于这种事实, 很多一直用最肯定的语气来描述的历史事件至少应该成为疑问。”

例 1.67. 在练习 1.21 中, 很多人给出答案 $1/2$, 理由是两个孩子的性别是独立的。事实上, 我们考虑的是家庭类型, 它与给定的性别信息不是独立的, 例如

$$P(BB|B) = \frac{1}{3}, \text{ 然而 } P(BB) = \frac{1}{4}$$

性质 1.16 (独立与互斥). 如果事件 A 与 B 独立, 且 $P(A) > 0, P(B) > 0$, 则 A, B 不互斥, 即 $AB \neq \emptyset$ 。我们也经常用它的逆否命题: 如果 A, B 互斥, 且 $P(A) > 0, P(B) > 0$, 则 A 与 B 不独立。

性质 1.17. 事件组对 $\{A, B\}, \{A, B^c\}, \{A^c, B\}, \{A^c, B^c\}$ 中任何一个组对独立, 都能推导出其他组对也是独立的。

证明. 假设 $\perp\!\!\!\perp \{A, B\}$, 则 $P(AB^c) = P(A - AB) = P(A) - P(AB) = P(A)[1 - P(B)] = P(A)P(B^c)$, 于是 $\perp\!\!\!\perp \{A, B^c\}$ 得证。其他的证明类似。 \square

练习 1.29. 如果事件 A_1, A_2, \dots, A_n 独立, 则 $A_1^c, A_2^c, \dots, A_n^c$ 独立。

练习 1.30. 已知概率空间 (Ω, \mathcal{S}, P) , 并且事件 $A, B, C \in \mathcal{S}$ 满足 $P(A) > 0, P(B) > 0$ 。试证明:

① 如果 A, B 独立, 则 $P(C|A) = P(B)P(C|AB) + P(B^c)P(C|AB^c)$ 。

提示: 利用全概率公式 $P_A(C) = P_A(B)P_A(C|B) + P_A(B^c)P_A(C|B^c)$ 。

② 若上式成立并满足 $P(C|AB) \neq P(C|A)$ 且 $P(C) > 0$, 则 A, B 独立。

提示: 由条件 $P(C|AB) \neq P(C|A)$ 可证得 $P_A(C|B) \neq P_A(C|B^c)$, 从 $[P(B) - P_A(B)][P_A(C|B) - P_A(C|B^c)] = 0$ 得出 $P(B) = P_A(B)$ 。

例 1.68. (1) 事件 A_1, A_2, A_3 两两独立, 问 A_1, A_2, A_3 是否一定独立? (2) 如果 $P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3)$, 问 A_1, A_2, A_3 是否一定独立?

解. 这两个问题的答案都是“否”, 分别构造反例如下。

1. Bernstein 反例: 盒子里装有四个球, 标号分别为 110, 101, 011 和 000。从盒子中随机抽取一球, 令 A_k 表示“标号的第 k 个位置是 1”, 其中 $k = 1, 2, 3$, 则

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{2} \text{ 且 } P(A_1A_2A_3) = 0$$

从 $P(A_1|A_2) = P(A_1) = 1/2$ 判定 $A_1 \perp\!\!\!\perp A_2$ 。类似地, $A_1 \perp\!\!\!\perp A_3$ 且 $A_2 \perp\!\!\!\perp A_3$ 。然而 $P(A_1)P(A_2)P(A_3) = 1/8$, 故 A_1, A_2, A_3 不独立。

2. Kac 反例: 设基本事件集合 $\Omega = \{1, 2, 3, 4\}$ 满足 $P(\{1\}) = \sqrt{2}/2 - 1/4$, $P(\{2\}) = P(\{4\}) = 1/4$, $P(\{3\}) = 3/4 - \sqrt{2}/2$ 。令 $A_1 = \{1, 3\}, A_2 = \{2, 3\}, A_3 = \{3, 4\}$, 则

$$P(A_1) = P(\{1\}) + P(\{3\}) = \frac{1}{2} \text{ 且 } P(A_2) = P(A_3) = 1 - \frac{\sqrt{2}}{2}$$

此例满足 $P(A_1A_2A_3) = P(\{3\}) = 3/4 - \sqrt{2}/2 = P(A_1)P(A_2)P(A_3)$, 然而 $P(A_1A_2) \neq P(A_1)P(A_2)$ 说明 A_1, A_2, A_3 不独立。



图 1.32: 瑞典绘图大师、不可能图形之父 Oscar Reutersvärd (1915-2002) 的作品: 眼见不一定为实, 直觉有时带有欺骗性, 从局部到整体需要理性的判断。

\nwarrow 定理 1.11. 已知随机事件 $A_1, A_2, \dots, A_n, \dots$ 相互独立且 $\sum_{n=1}^{\infty} P(A_n) = \infty$, 则几乎必然有无穷多个 A_n 发生, 即 $P(A_n \text{ i.o.}) = 1$ 。

证明. 对于任意有限的 k , 皆有 $\sum_{n=k}^{\infty} P(A_n) = \infty$ 。

$$\begin{aligned} P\left(\bigcap_{n=k}^{\infty} A_n^c\right) &= \prod_{n=k}^{\infty} P(A_n^c) = \prod_{n=k}^{\infty} [1 - P(A_n)], \text{ 由不等式 } 1 - x \leq e^{-x} \text{ 得出} \\ &\leq \prod_{n=k}^{\infty} \exp\{-P(A_n)\} = \exp\left\{-\sum_{n=k}^{\infty} P(A_n)\right\} = 0 \end{aligned}$$

由第 74 页的引理 1.1 的证明可知, 无穷多个 A_n 发生当且仅当事件 $\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$ 发生, 从而

$$\begin{aligned} P\{A_n \text{ i.o.}\} &= \lim_{k \rightarrow \infty} P\left(\bigcup_{n=k}^{\infty} A_n\right) \\ &= \lim_{k \rightarrow \infty} \left[1 - P\left(\bigcap_{n=k}^{\infty} A_n^c\right)\right] = 1 - \lim_{k \rightarrow \infty} P\left(\bigcap_{n=k}^{\infty} A_n^c\right) = 1 \quad \square \end{aligned}$$

推论 1.6 (Borel 0-1 律). 对于一列相互独立的随机事件 $A_1, A_2, \dots, A_n, \dots$, 总有

$$P\left(\overline{\lim}_{n \rightarrow \infty} A_n\right) = \begin{cases} 0 & \text{当且仅当级数 } \sum_{n=1}^{\infty} P(A_n) \text{ 收敛} \\ 1 & \text{当且仅当级数 } \sum_{n=1}^{\infty} P(A_n) \text{ 发散} \end{cases} \quad (1.35)$$

证明. 由 Borel-Cantelli 引理 1.1 及定理 1.11 可证。 \square

定义 1.23. 如果一个随机事件 A 要么几乎必然发生, 要么几乎必然不发生, 则称之为尾事件 (tail event)。即, 尾事件 A 的概率只有 $P(A) = 1$ 和 $P(A) = 0$ 两种情况。

 判定准则式 (1.35) 也被称为 Borel 0-1 准则, 按照级数 $\sum_{n=1}^{\infty} P(A_n)$ 是否发散来判定尾事件 $\overline{\lim}_{n \rightarrow \infty} A_n$ 的概率为 1 还是为 0。其中, 由 Borel-Cantelli 引理 1.1 可知, 判定该尾事件的概率为 0 时不需要 $\{A_n\}$ 的独立性假设。Borel 0-1 准则将在 §5.1.2 用于证明强大数律, 也常用于证明“以概率 1”成立的性质, 见下面给出的尾事件的例子。

例 1.69. 已知硬币正面的概率 $P(H) = p$, 抛这枚硬币无穷次, 令 A_n 表示事件“至少 n 个连续的正面发生在第 2^n 抛和第 $2^{n+1} - 1$ 抛之间”。试证明:

$$P(A_n \text{ i.o.}) = \begin{cases} 1 & \text{如果 } p \geq \frac{1}{2} \\ 0 & \text{如果 } p < \frac{1}{2} \end{cases}$$

证明. 显然, $A_1, A_2, \dots, A_n, \dots$ 两两独立。估计 $P(A_n)$ 的上下界,

$$1 - (1 - p^n)^{\lfloor 2^n/n \rfloor} \leq P(A_n) < (2p)^n$$

□ 若 $p < 1/2$, 则

$$\sum_{n=1}^{\infty} P(A_n) < \sum_{n=1}^{\infty} (2p)^n = \frac{2p}{1-2p} < \infty$$

□ 若 $p \geq 1/2$, 则

$$\ln(1 - P(A_n)) \leq \left\lfloor \frac{2^n}{n} \right\rfloor \ln(1 - p^n) \leq -\frac{(2p)^n}{n} \leq -\frac{1}{n}$$

于是, $P(A_n) \geq 1 - e^{1/n} \sim \frac{1}{n}$, 进而, $\sum_{n=1}^{\infty} P(A_n) = \infty$

根据 Borel 0-1 律, 若 $p < 1/2$, 有 $P(A_n \text{ i.o.}) = 0$; 否则, 有 $P(A_n \text{ i.o.}) = 1$ 。 □

练习 1.31. 抛一枚均匀的硬币无穷次, 则事件“连续出现 100 次正面的事件出现无限多次”是一个尾事件。

在很多实际问题中, 人们经常假设某些事件之间存在独立性来简化概率计算。独立性假设有时是合理的, 有时不那么合理也被采纳。尤其在权衡了模型的有效性与复杂性之后, 简单而高效的模型往往更能赢得青睐, 哪怕有一点不合理的假设也无所谓。下面, 我们通过解决几个有趣的概率问题来加深对独立性的理解。

例 1.70. 已知电话在长度为 t 的时间段内被呼叫 k 次的概率是

$$P_t(k) = \frac{(\alpha t)^k}{k!} e^{-\alpha t}, \text{ 其中 } \alpha \text{ 是单位时间内平均被呼叫的次数, } k = 0, 1, 2, \dots$$

如果不相交的两段时间内被呼叫的次数是独立的, 试求: 在时间段 $2t$ 内被呼叫 s 次的概率 $P_{2t}(s)$? 其中 $s = 0, 1, 2, \dots$ 。

解. 令 A_t^k 表示事件“在时间段 t 内被呼叫 k 次”, 则“在时间段 $2t$ 内被呼叫 s 次”

有非交分解 $A_{2t}^s = A_t^0 A_t^s + A_t^1 A_t^{s-1} + \cdots + A_t^s A_t^0$ 。

$$\begin{aligned} P_{2t}(s) &= P(A_{2t}^s) = \sum_{j=0}^s P(A_t^j A_t^{s-j}), \text{ 根据独立性假设, 进而有} \\ &= \sum_{j=0}^s P_t(j)P_t(s-j) \\ &= \sum_{j=0}^s \frac{(\alpha t)^s e^{-2\alpha t}}{j!(s-j)!} \\ &= \frac{(2\alpha t)^s}{s!} e^{-2\alpha t} \end{aligned}$$

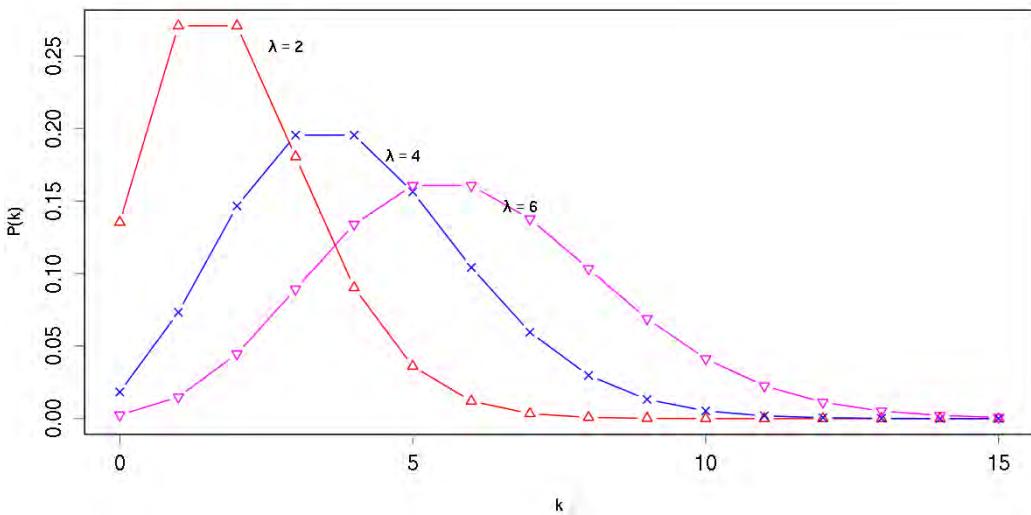


图 1.33: 在例 1.70 中, $\lambda = \alpha t$ 分别取 2, 4, 6 时 $P(k) = \lambda^k e^{-\lambda} / k!$ 的折线图: λ 越大折线图越呈现中间高两头低的对称性。

※例 1.71 (即时雇佣问题 [27]). 某公司为招募一名员工而依次面试 n 个应聘者, 假设面试成绩没有相同的, 每次面试都要即时地决定是否雇佣, 公司采用下述雇佣算法: (1) 记录前 k 名应聘者的最优成绩 bestscore 并拒绝他们, (2) 雇佣剩下的 $n - k$ 人中成绩首个超过 bestscore 的, 若没有超过 bestscore 的则雇佣最后一位。

令 S 表示事件“雇佣到成绩最优者”, 对 $k = 1, 2, \dots, n - 1$, 分别求该公司雇佣到成绩最优者的概率 $P(S)$, 并由此决定最优的 k 。

解. 令 S_i 表示事件“第 i 个应聘者是成绩最优者且被公司雇佣到”, 则 S_1, S_2, \dots, S_n 是两两互斥的且 $P(S) = \sum_{i=1}^n P(S_i)$ 。

按照雇佣算法, 若最优者是前 k 个应聘者之一, 则最优者不会被雇佣到, 即 $P(S_i) = 0$, 其中 $i = 1, 2, \dots, k$ 。

令 B_i 表示事件“第 i 个应聘者是成绩最优者”, 则 $P(B_i) = 1/n$ 。令 D_i 表示事件

“第 $k+1$ 至第 $i-1$ 个应聘者都未被雇佣”，即前 $i-1$ 个应聘者中成绩最好的必是前 k 个应聘者之一，于是 $P(D_i) = k/(i-1)$ 。显然，事件 B_i 与 D_i 相互独立且 $S_i = B_i D_i$ 。

$$\begin{aligned} P(S) &= \sum_{i=k+1}^n P(S_i) = \sum_{i=k+1}^n P(B_i)P(D_i) \\ &= \sum_{i=k+1}^n \frac{1}{n} \cdot \frac{k}{i-1} = \frac{k}{n} \sum_{i=k}^{n-1} \frac{1}{i} \end{aligned}$$

根据下面的不等式可以得到 $P(S)$ 的上下界如下，

$$\begin{aligned} \int_k^n \frac{1}{x} dx &\leq \sum_{j=k}^{n-1} \frac{1}{j} \leq \int_{k-1}^{n-1} \frac{1}{x} dx \\ \Downarrow \\ \frac{k}{n} \ln \frac{n}{k} &\leq P(S) \leq \frac{k}{n} \ln \frac{n-1}{k-1} \end{aligned}$$

函数 $\frac{k}{n} \ln \frac{n}{k}$ 是关于 k 的单峰函数，在 $k = ne^{-1}$ 处取得最大值 e^{-1} ，即 $P(S) \geq e^{-1}$ 。

※例 1.72 (Chebyshev 问题). 任选一个分数，它不可约的概率是多少？

解. 一个分数不可约当且仅当该分数的分子、分母互素。该问题等价于：任选两个自然数构成二元组 (a, b) ，求它们互素的概率？

□ 概率空间 $(\Omega_n, \mathcal{S}_n, P_n)$ 定义为 $\Omega_n = \{1, 2, \dots, n\} \times \{1, 2, \dots, n\}$, $\mathcal{S}_n = 2^{\Omega_n}$ 且 $\forall \omega \in \Omega_n, P_n(\omega) = 1/n^2$ 。定义概率 $P(n)$ 如下：

$$P(n) = P_n\{\text{自然数对 } (a, b) \in \Omega_n \text{ 互素}\}, \text{ 其中 } n \geq 2 \quad (1.36)$$

如果原问题有解，则意味着 $n \rightarrow \infty$ 时 $P(n)$ 的极限存在。为此，我们考察 $P(10^k), k = 1, 2, \dots, 6$ ，看是否有规律。

k	$P(10^k)$
1	0.63
2	0.6087
3	0.608383
4	0.60794971
5	0.6079301507
6	0.607927104783

□ 设 $\{2, \dots, n\}$ 里所有的素数为 $p_1 < \dots < p_{r_n}$, 若分数 a/b 可约, 则存在素数 $p \in \{p_1, \dots, p_{r_n}\}$ 是 a, b 公因子, 即 a, b 选自 $p, 2p, \dots, kp$, 其中 $k = \lfloor n/p \rfloor$.

所以, $\frac{(k-1)^2}{n^2} < P_n\{\text{素数 } p \text{ 是 } a, b \text{ 的公因子}\} = \frac{k^2}{n^2}$, 进而

$$1 - \frac{1}{p^2} \leq P_n\{\text{素数 } p \text{ 不是 } a, b \text{ 的公因子}\} = 1 - \frac{k^2}{n^2} < 1 - \frac{1}{p^2} + \frac{4}{np}$$

显然, 当 $n \rightarrow \infty$ 时, $P_n\{\text{素数 } p \text{ 不是 } a, b \text{ 的公因子}\} \rightarrow 1 - 1/p^2$.

□ 令事件 A_j 表示“第 j 个素数不是 a, b 的公因子”, 则积事件 $A_1A_2 \cdots A_{r_n}$ 表示“ a, b 互素”。下面考察 $P_n(A_iA_j)$, 其中 $i \neq j$.

$$\begin{aligned} P_n(A_iA_j) &= 1 - P_n(A_i^c) - P_n(A_j^c) + P_n(A_i^cA_j^c) \\ &= 1 - P_n(A_i^c) - P_n(A_j^c) + \frac{k_{ij}^2}{n^2}, \text{ 其中 } k_{ij} = \left\lfloor \frac{n}{p_i p_j} \right\rfloor \\ &\rightarrow 1 - \frac{1}{p_i^2} - \frac{1}{p_j^2} + \frac{1}{p_i^2 p_j^2}, \text{ 当 } n \rightarrow \infty \text{ 时} \\ &= P_n(A_i)P_n(A_j) \end{aligned}$$

□ 类似地, 利用 Jordan 公式 (1.24) 可证: 当 n 充分大时, 近似地有 A_1, \dots, A_{r_n} 独立。于是, 任选一个分数不可约的概率是

$$\begin{aligned} P\{\text{分数不可约}\} &= \lim_{n \rightarrow \infty} P(n) \\ &= \lim_{n \rightarrow \infty} \prod_{j=1}^{r_n} P_n(A_j) \\ &= \prod_{p \text{ 是素数}} \left(1 - \frac{1}{p^2}\right) \end{aligned}$$

要完成上式的计算, 需要用到下面的数论结果, 它由瑞士数学家 Leonhard Euler (1707-1783) 首次发现: 当实数 $x > 1$ 时,

$$\sum_{n=1}^{\infty} \frac{1}{n^x} = \prod_{p \text{ 是素数}} \left(1 - \frac{1}{p^x}\right)^{-1} \quad (1.37)$$

上式中, 有关 x 的函数 $\sum_{n=1}^{\infty} n^{-x}$ 被称为 Riemann ζ



函数, 记作 $\zeta(x)$ 。于是,

$$\begin{aligned} P\{\text{分数不可约}\} &= \frac{1}{\zeta(2)} \\ &= \left[\sum_{n=1}^{\infty} \frac{1}{n^2} \right]^{-1}, \text{ 引用那个印在钱币上的数学公式} \\ &= \frac{6}{\pi^2} \\ &\approx 0.60792710185403 \end{aligned}$$

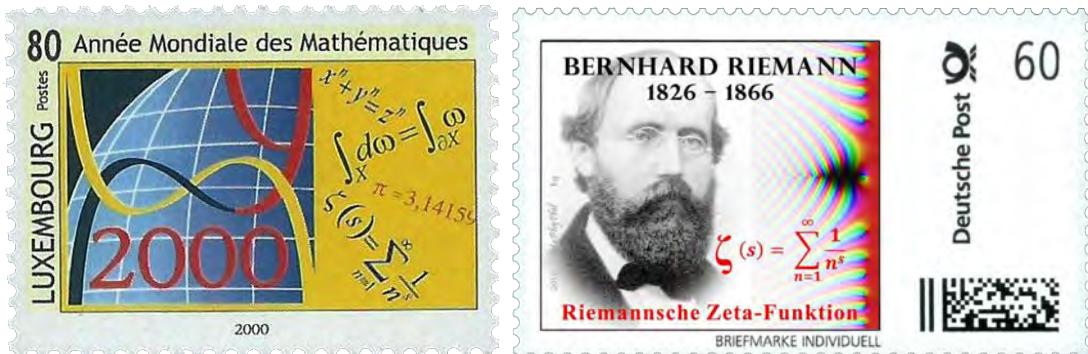
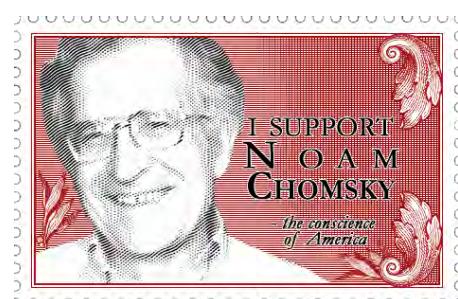


图 1.34: 1859 年, Riemann 在论文《论不大于一个给定值的素数个数》论证了 ζ 函数可解析延拓到整个复平面, 并证明了当 $\Re(s) < 0$ 时 $s = -2, -4, -6, \dots$ 等负偶数是 ζ 函数的平凡零点。Riemann 猜想: ζ 函数非平凡零点的实部都是 $1/2$ 。Riemann 猜想是数学史上最伟大的猜想之一, 至今未解。

1957 年, 美国著名语言学家 Noam Chomsky (1928-) 出版了《句法结构》一书, 提出了语言的生成模型, 句法被形式化为一组重写规则 (rewriting rules) 或产生式 (productions)。Chomsky 的形式文法谱系包含四种文法, 分别是

- 无限制文法 (0-型文法)
- 上下文相关文法 (1-型文法)
- 上下文无关文法 (2-型文法)
- 正则文法 (3-型文法)



上下文无关文法 (context-free grammar, CFG) 的产生式左边只有一个非终结符, 这是“上下文无关”的含义, 它所产生的语言称为上下文无关语言, 该语言能够被非确定下推自动机识别。

表 1.5: Chomsky 形式文法谱系。该表产生式中，大写英文字母表示非终结符，小写英文字母表示终结符，希腊字母 α, β, γ 表示包含终结符和非终结符的字串，其中 α, β 可以是空串， γ 不是空串。

形式文法	重写规则	自动机
无限制文法	$\alpha \rightarrow \beta$	Turing 机
上下文无关文法	$\alpha A \beta \rightarrow \alpha \gamma \beta$	线性有界非确定 Turing 机
上下文相关文法	$A \rightarrow \gamma$	非确定下推自动机
正则文法	$A \rightarrow aB$ 或者 $A \rightarrow a$	有限状态自动机

每个文法的产生式可以赋予概率形成随机文法。譬如，随机正则文法（等价于隐 Markov 模型，见第 43 页的例 1.29，详细内容见第 13 章）和随机上下文无关文法 (stochastic CFG, SCFG)，见下面的例子。

※例 1.73. 下面的 SCFG 是一个玩具模型，其中重写规则的概率，以及终结符的概率都是虚构的。在自然语言处理中，利用大规模语料来估计这些概率并非易事。

```
## 句子 S 由 名词短语 NP 和动词短语 VP 构成
S -> NP + VP [1]
```

```
## 名词短语 NP 的构成
NP -> Pronoun [0.1]
| Name [0.1]
| Noun [0.15]
| Det + Noun [0.6]
| NP + PP [0.05]
```

```
## 动词短语 VP 的构成
VP -> Verb [0.6]
| VP + NP [0.2]
| VP + PP [0.2]
```

```
## 介词短语 PP 的构成
PP -> Prep + NP [1]
```

```
## 终结符的分布
Name -> Mike [0.002] | Jane [0.003] | ...
Noun -> telescope [0.001] | microscope [0.001] | boy [0.01] | girl [0.01] | ...
Verb -> saw [0.02] | study [0.01] | ...
Pronoun -> I [0.2] | you [0.1] | it [0.3] | ...
Det -> the [0.3] | a [0.35] | every [0.02] | ...
Prep -> in [0.2] | to [0.3] | on [0.04] | with [0.1] | ...
```

- 随机上下文无关文法要求每个非终结符都必须满足从它导出的重写规则的概率之和为 1 (归一性), 譬如, 名词短语 NP 的所有产生式概率之和等于 1, 等等。另外, 为了简化计算, 假设所有的重写规则是独立的。譬如, 重写规则 $S \rightarrow NP + VP [1]$, $NP \rightarrow Det + Noun [0.6]$ 和 $VP \rightarrow VP + PP [0.2]$ 的联合概率是 $1 \times (0.6 \times 0.2) = 0.12$ 。
- 也许, 语言学家并不认可规则 $NP \rightarrow NP + PP$ 在条件 $S \rightarrow NP + VP$ 和条件 $VP \rightarrow VP + NP$ 之下的概率都是一样的。独立性假设带来的计算上的方便, 与模型在正确性上的牺牲, 往往是实践需要权衡的。

按照给定的重写规则*, 句子 *the boy saw a girl with a microscope* 有两个不同的编译结果, 其中之一见图 1.35, 请读者给出另外一个结果。当编译结果不唯一时, 机器通过句法树 (syntactic tree) 的产生概率来挑出最有可能的那个。

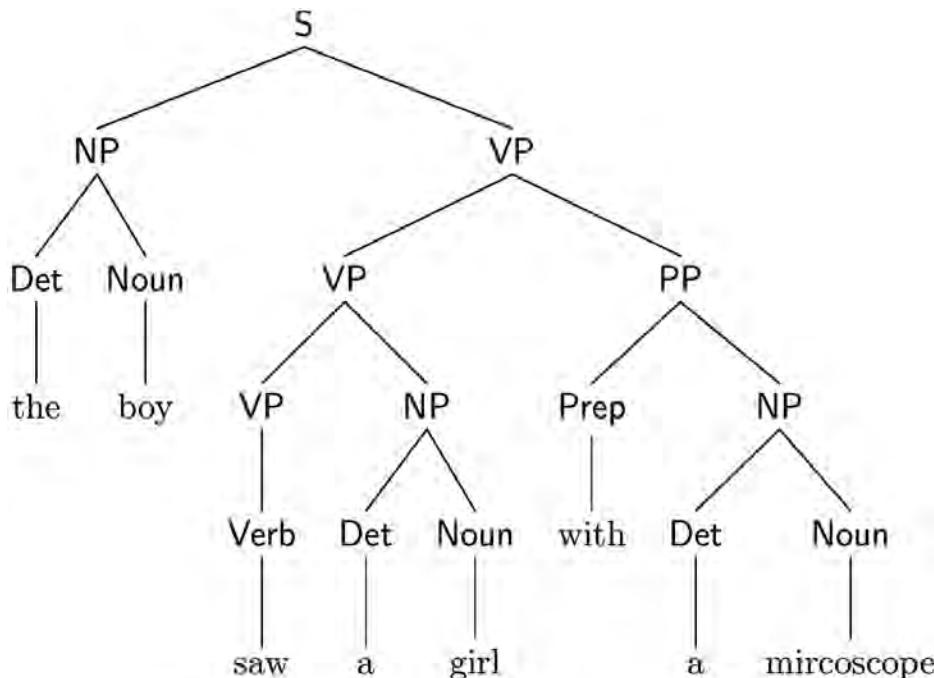


图 1.35: 利用例 1.73 给定的重写规则分析句子 *the boy saw a girl with a microscope*, 可以得到两个截然不同的句法树。当前结果中, 介词短语 *with a microscope* 作状语。

图 1.35 所示句法树的产生概率是 $p = 1 \times (0.6 \times 0.2) \times (0.3 \times 0.01 \times 0.2 \times 1) \times (0.6 \times 0.6 \times 0.1 \times 0.6) \times (0.02 \times 0.35 \times 0.01 \times 0.35 \times 0.001) = 3.81024 \times 10^{-14}$, 而另一

*该例的重写规则中有递归定义, 如左递归结构 $NP \rightarrow NP + PP$, 会给某些编译算法带来麻烦, 引起冲突或降低效率。因为本书不谈论具体的编译算法, 所以未对该例的重写规则做优化处理。

棵句法树的产生概率是 9.5256×10^{-15} (请读者给出计算细节)。经过比较, 机器选择图 1.35 所示的句法树为编译结果, 然而, 该句法树导出的语义^{*}是错误的!

练习 1.32. 某随机上下文无关文法能产生很多句子, 显然一个句子上所有可能的句法树的概率之和小于 1 (见例 1.73)。对于所有该文法能产生的句子, 其句法树概率之和是否等于 1? 为什么?

形式文法被句法模式识别 (syntactic pattern recognition) 用来刻画模式, 付京孙 (1930-1985) 是这方面的知名学者 [52]。后来, 形式文法和句法模式识别又启发瑞典统计学家 Ulf Grenander (1923-2016) 和美国数学家 D. B. Mumford (1937-) 创立和发展了模式理论 [65, 112], 成为现代人工智能坚实的数学基础。上下文无关文法具有足够强的表达能力从而成为大多数程序设计语言的语法, 同时也被用作描述工具广泛应用于自然语言处理、生物信息学等研究领域。

^{*}自然语言的语义分析通向自然语言理解的必经之路, 是计算语言学里一个尚未解决的难题。虽然根据产生概率的大小来选择句法分析结果有时会犯错误, 但在某些对自然语言理解要求不高的实践中, 若能绕开语义分析, 找到一个统计意义上效果不错的普适的句法分析方法也是难能可贵的。

1.3.4 条件独立性及其性质

定义 1.24 (条件独立性). 已知概率空间 (Ω, \mathcal{S}, P) , 且事件 $B \in \mathcal{S}$ 满足 $P(B) > 0$ 。事件 A_1, A_2 关于 B 条件独立 [31, 118], 记作 $A_1 \perp\!\!\!\perp_B A_2$, 有时也记作 $\perp\!\!\!\perp_B \{A_1, A_2\}$ 或 $A_1 \perp\!\!\!\perp A_2|B$ 或 $\perp\!\!\!\perp \{A_1, A_2\}|B$ 等, 当且仅当

$$P(A_1 A_2|B) = P(A_1|B)P(A_2|B)$$

$A_1 \perp\!\!\!\perp_B A_2$ 的另一等价定义是: 在由定理 1.7 定义的条件概率空间 $(\Omega, \mathcal{S}, P_B)$ 上, 事件 $A_1, A_2 \in \mathcal{S}$ 相互独立, 即

$$P_B(A_1 A_2) = P_B(A_1)P_B(A_2)$$

更一般地, 事件 $A_1, A_2, \dots, A_n \in \mathcal{S}$ 关于 B 条件独立, 记作 $\perp\!\!\!\perp_B \{A_1, A_2, \dots, A_n\}$, 当且仅当对于 $1, 2, \dots, n$ 的任意子序列 $k_1 < k_2 < \dots < k_s$, 皆有

$$P(A_{k_1} A_{k_2} \cdots A_{k_s}|B) = P(A_{k_1}|B)P(A_{k_2}|B) \cdots P(A_{k_s}|B)$$

性质 1.18. 与性质 1.16 类似, 如果 A, B 互斥, 即 $AB = \emptyset$, 且 $P(A_1|B) > 0, P(A_2|B) > 0$, 则 A_1, A_2 关于 B 不是条件独立的。

例 1.74. 由 $A_1 \perp\!\!\!\perp_B A_2$ 能推导出 $A_1 \perp\!\!\!\perp A_2$ 或者 $A_1 \perp\!\!\!\perp_{B^c} A_2$ 吗?

解. 不能! 下面构造一个反例。下图是一个 6×6 的格子棋盘, 每个小格子代表一个基本事件, 选中它的概率是 $1/36$ 。事件 A_1 (灰色部分), 事件 A_2 (斜线部分) 和事件 B (粗框部分) 如图所示。

事件 A_1, A_2 关于 B 条件独立, 因为

$$\begin{aligned} P(A_1 A_2|B) &= \frac{2}{12} = \frac{1}{6} \\ P(A_1|B)P(A_2|B) &= \frac{6}{12} \cdot \frac{4}{12} = \frac{1}{6} \end{aligned}$$

显然, $P(A_1 A_2) = 1/6 < P(A_1)P(A_2)$, 即 A_1, A_2 并不独立。

事件 A_1, A_2 关于 B^c 也不是条件独立的, 这是因为 $P(A_1 A_2|B^c) = 1/6$, 但 $P(A_1|B^c) = 13/24$, $P(A_2|B^c) = 9/24$ 。

练习 1.33. 若事件 B 满足 $P(B) > 0$, 由 $A_1 \perp\!\!\!\perp_B A_2$ 能推导出 $A_1 \perp\!\!\!\perp_B A_2$ 吗? 提示: 不能。参考上例, 适当地修改图 1.36。

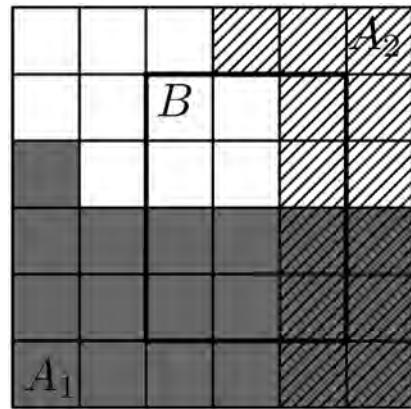


图 1.36: 条件独立 $\not\Rightarrow$ 独立。

性质 1.19. 给定条件 B , 事件组对 $\{A_1, A_2\}, \{A_1, A_2^c\}, \{A_1^c, A_2\}, \{A_1^c, A_2^c\}$ 中任何一个组对条件独立, 都能推导出其他组对也是条件独立的。

例 1.75 (兼听则明). 令 G 表示市场利于投资, 已知 $P(G) = 0.7$ 。某投资人总是听取三个理财顾问 I_1, I_2, I_3 中的多数意见。

- $P(I_1|G) = 0.95$ 表示市场利于投资的条件下顾问 I_1 建议投资的概率, $P(I_1|G^c) = 0.2$ 则表示市场不利于投资的条件下顾问 I_1 建议投资的概率, 显然它们刻画了顾问 I_1 的理财水平。其他两个理财顾问的水平情况是: $P(I_2|G) = 0.75, P(I_2|G^c) = 0.1, P(I_3|G) = 0.8, P(I_3|G^c) = 0.25$ 。
- 假设三个理财顾问独立工作、互不影响, 即 $\perp\!\!\!\perp_G \{I_1, I_2, I_3\}$ 且 $\perp\!\!\!\perp_{G^c} \{I_1, I_2, I_3\}$ 。

试问: 投资人决策正确的概率 $P(D)$?

解. 由全概率公式得 $P(D) = P(D|G)P(G) + P(D|G^c)P(G^c)$, 其中,

$$\begin{aligned} P(D|G) &= P(I_1 I_2 I_3|G) + P(I_1 I_2 I_3^c|G) + P(I_1^c I_2 I_3|G) + P(I_1^c I_2 I_3^c|G) \\ P(D|G^c) &= P(I_1^c I_2^c I_3^c|G^c) + P(I_1 I_2^c I_3^c|G^c) + P(I_1^c I_2^c I_3|G^c) + P(I_1^c I_2 I_3^c|G^c) \end{aligned}$$

由条件独立性假设 $\perp\!\!\!\perp_G \{I_1, I_2, I_3\}$ 和 $\perp\!\!\!\perp_{G^c} \{I_1, I_2, I_3\}$ 可得

$$\begin{aligned} P(I_1 I_2 I_3|G) &= P(I_1|G)P(I_2|G)P(I_3|G) = 0.95 \cdot 0.75 \cdot 0.8 = 0.57 \\ P(I_1^c I_2^c I_3^c|G^c) &= P(I_1^c|G^c)P(I_2^c|G^c)P(I_3^c|G^c) = 0.8 \cdot 0.9 \cdot 0.75 = 0.54 \end{aligned}$$

其他项可类似计算, 最终算得 $P(D|G) = 0.9325, P(D|G^c) = 0.915$, 进而得到投资人决策正确的概率为 $P(D) = 0.92725$ 。

事实上, 无论市场如何风云变化, 即无论 $P(G)$ 如何取值, $P(D)$ 总是高于任何一个理财顾问的水平。这个例子应了“兼听则明”的古语。

 在统计机器学习和模式识别中, 决策问题的结果往往就是在多个学习机*或分类器(类似多个专家)的投票中占多数者, 一般情况下其效果比依靠单个专家的要好些。但是, 这并不意味着“三个臭皮匠凑个诸葛亮”——如果臭皮匠的水平实在太臭, 再多也没用。如何为决策问题选择“专家团”是集成学习(ensemble learning)关注的话题, 一般来说“专家”之间要相互独立、各有所长等等。

练习 1.34. 一群学生中男女各半, 令事件 M = “所选学生为男生”, 则 $P(M) = P(M^c) = 1/2$ 。令事件 B = “所选学生喜欢打篮球”, C = “所选学生喜欢电玩”。

*学习机(learner)就是实现某一特定任务的算法或程序, 它能够通过经验修改自身以求达到某给定标准之下更好的效果。如, 人工神经网络、支持向量机(support vector machine, SVM)等[15, 16, 135]。

□ 已知 $3/4$ 的男生和 $1/4$ 的女生喜欢打篮球，一半的男生和一半的女生喜欢电玩。即， $P(B|M) = 3/4, P(B|M^c) = 1/4, P(C|M) = P(C|M^c) = 1/2$ 。

□ 假设男女生喜欢打篮球和喜欢电玩是独立的，即 $\perp\!\!\!\perp_M \{B, C\}, \perp\!\!\!\perp_{M^c} \{B, C\}$ 。

试证明： B, C 独立，但 B, C, M 不独立。提示：由全概率公式可求得 $P(B) = P(C) = 1/2, P(BC) = 1/4$ 。

练习 1.35. 一般情况下能从 $\perp\!\!\!\perp_M \{B, C\}$ 且 $\perp\!\!\!\perp_{M^c} \{B, C\}$ 推导出 $\perp\!\!\!\perp \{B, C\}$ 吗？

提示：不能。把**练习 1.34** 中的条件“一半的男生和一半的女生喜欢电玩”改为“ $3/4$ 的男生和 $1/4$ 的女生喜欢电玩”，其他条件不变，则 B, C 并不独立。

※例 1.76 (贝叶斯垃圾邮件过滤). 垃圾邮件或垃圾短信的识别在通讯日益发达的今天显得尤为重要。贝叶斯垃圾邮件过滤 (Bayesian spam filtering) 实质就是一个二分类器 (binary classifier)，通过样本的训练可用来推断给定的新邮件是垃圾 (S) 或不是垃圾 (S^c)。一般步骤是：

- 给定一封新邮件，对它不做任何句法或语义的分析，邮件内容被简化为一个实词的词表（也可以仅考虑名词，或者用户感兴趣的词集） $L = (w_1, w_2, \dots, w_n)$ 。
- 随机收集一定规模的邮件样本，其中垃圾邮件被贴上 S 的“标签”。在包含词 w_j 的所有邮件中，先统计出垃圾邮件的频率是 $p_j \in [0, 1]$ 。只要邮件样本的规模足够大，我们有理由假设 $P(S|w_j) = p_j$ ，进而 $P(S^c|w_j) = 1 - p_j$ 。
- 为了降低计算复杂度，假设词语关于 S 和关于 S^c 都是条件独立的，即

$$\begin{aligned} P(L|S) &= P(w_1|S) \cdots P(w_n|S) = \frac{1}{[P(S)]^n} \prod_{j=1}^n p_j \prod_{j=1}^n P(w_j) \\ P(L|S^c) &= P(w_1|S^c) \cdots P(w_n|S^c) = \frac{1}{[P(S^c)]^n} \prod_{j=1}^n (1 - p_j) \prod_{j=1}^n P(w_j) \end{aligned}$$

这个条件独立假设显然不符合语言学的事实，因为不管是否垃圾邮件，词语之间也不可能都是独立的。但是为了使算法可行，模型粗糙一点儿也是迫不得已。

□ 由无差别原则，假定垃圾邮件和非垃圾邮件的验前概率为 $P(S) = P(S^c) = 1/2$ ，并利用 Bayes 公式计算 S 的验后概率

$$P(S|L) = \frac{P(L|S)}{P(L|S) + P(L|S^c)} = \frac{\prod_{j=1}^n p_j}{\prod_{j=1}^n p_j + \prod_{j=1}^n (1 - p_j)}$$

再利用**例 1.64** 介绍的推断方法来判定是 S 的可能性大，还是 S^c 的可能性大。经过用户确认，垃圾邮件被贴上 S 的标签加入到样本中去，如此循环，以便提高识别的精度并改进个性化。

1.4 习题

- 1.1. 将 n 个球随机地放入 n 个盒子，求每个盒子都恰有一个球的概率 p_n ? 当 $n = 10$ 的时候，这个概率多大?
- 1.2. 将 n 个球放入 n 个盒子，问恰有一个盒子空着的概率?
- 1.3. 某射手在 3 次射击中至少命中 1 次目标的概率为 0.875，求该射手在 1 次射击中命中目标的概率 p ?
- 1.4. 电灯泡使用寿命在 1000 小时以上的概率为 0.2，则 3 个灯泡在使用 1000 小时后，最多只有 1 个坏了的概率为多少?
- 1.5. k 个盒子各装 n 个球，标号为 $1, 2, \dots, n$ ，从每个盒子中各取一个球，计算 $A_m =$ “取到的 k 个球中最大标号是 m ” 的概率，其中 $1 \leq m \leq n$ 。
- 1.6. 从 10 双不同的鞋中任选 4 只，问至少配成一双的概率?
- 1.7. 已知 10 人中有一对夫妇，他们随机地围坐在一张圆桌周围聊天，求事件 $A =$ “这对夫妇坐在一起”的概率?
- 1.8. 盒子里有 n 球，标号为 $1, 2, \dots, n$ ，现有标号为 $1, 2, \dots, n$ 的 n 个人分别随机地从盒子中取走一个球。当 n 很大时，求至少有一个人拿到相同标号的概率?
- 1.9. 例 1.14 中的抽取是无放回的，当 N 充分大而 m 不大时， $P(A_k) \approx C_m^k p^k (1-p)^{m-k}$ ，其中 $p = n/N$ ，即有放回的抽取和无放回的抽取相差无几。
- 1.10. 盒子里装有 w 个白球， b 个黑球。不放回地一次一个地抽取，试求：(1) 同色球可分辨；(2) 同色球不可分辨这两种情况下，第 k 次取出白球的概率，其中 $1 \leq k \leq w + b$ 。
- ☆ 1.11. 用球-盒子模型证明：若非负整数 m, n, k, r 满足 $k \leq \min(m, n)$ 且 $r + k \leq n$ ，则

$$\sum_{j=0}^k C_m^{k-j} C_n^j C_{n-j}^r = C_{m+n-r}^k C_n^r$$

- 1.12. 利用符号计算工具“证明”李善兰恒等式*：

$$\sum_{j=0}^n (C_n^j)^2 C_{m+2n-j}^{2n} = (C_{m+n}^n)^2, \text{ 其中非负整数 } m, n \text{ 满足 } n \leq m \quad (1.38)$$

*李善兰 (1810-1882)，字竞芳，号秋纫，清末著名数学家、天文学家、翻译家和教育家，在其著作《垛积比类》(写于 1859-1867 年间) 中给出这一著名的恒等式。

1.13. 试证明多项式系数的结果 (1.6) 以及下面的结果，并类比二项式系数的性质 1.2。

$$\begin{aligned} \binom{n}{m}_{k+1} &= \sum_{j=0}^k \binom{n-1}{m-j}_{k+1} = \binom{n}{kn-m}_{k+1} \\ &= \sum_{j=0}^{\lfloor \frac{m}{k+1} \rfloor} (-1)^j \binom{n}{j} \binom{n+m-(k+1)j-1}{n-1} \end{aligned} \quad (1.39)$$

- 1.14. 在长为 1 的线段 AD 上任取两点 B, C 并在 B, C 处折断而得三个线段，求这三个线段能构成三角形的概率。
- 1.15. 在区间 $(0, 1)$ 中随机抽取两个数，求事件“两个数之和小于 $6/5$ ”的概率？
- 1.16. 假定情报员能否破译密码是相互独立的，每位情报员破译出密码的概率都为 0.6。试问：至少要用几位情报员才能使得破译出一份密码的概率大于 95%？
- 1.17. 盒子里有 n 个球，标号依次是 $1, 2, \dots, n$ 。独立有放回地均匀抽取 n 次，若 n 很大，求某个球从未被抽中的概率？
- 1.18. 已知随机事件 A, B, C 的概率为 $P(A) = P(B) = P(C) = 1/4$ 且 $P(AB) = P(AC) = 0, P(BC) = 1/8$ ，试用集合运算表示下述事件并求出它们相应的概率：
 (1) A, B, C 同时发生；(2) A, B, C 中至少有一个事件发生；(3) A, B 不发生， C 发生；(4) A, B, C 中至少有两个事件发生。
- 1.19. Banach 问题：喜好抽烟的波兰数学家 Stefan Banach (1892-1945) 左右口袋各装着一盒火柴，每盒皆有 n 根火柴。Banach 以概率 p 选左口袋的火柴，以概率 $q = 1 - p$ 选右口袋的火柴，并且他不记忆所剩火柴的数量。问他发现一盒空而此时另一盒剩 k 根火柴 ($k = 0, 1, \dots, n$) 的概率 P_k ？
- ☆ 1.20. 甲、乙举行射击比赛，每比一场胜者得一分。在每次射击中，甲取胜的概率为 α ，乙取胜的概率为 β 。设 $\alpha > \beta$ 且 $\alpha + \beta = 1$ 。各场比赛独立，直到有一人超过对方两分获得奖牌为止，分别求出甲、乙获得奖牌的概率。
- 1.21. 盒子里有 $n - 1$ 只黑球和 1 只白球，每次从盒中随机抽取一球，然后换入一只黑球，这样继续下去，求第 k 次取到黑球的概率。
- 1.22. 不论随机事件 A 的概率 $P(A) > 0$ 如何地小，随着试验次数的增加，试证明： A 迟早发生的概率是 1。
- 1.23. 掷一个均匀的骰子直至第一次出现 6 点，请读者写出概率空间 (Ω, \mathcal{S}, P) 。用 A_k 表示“第 k 次掷出首个 6 点”， $k = 1, 2, \dots$ ；用 A 表示“将掷出 6 点”。试求： $P(A_k)$ 和 $P(A)$ 。

- 1.24. Kolmogorov 公理体系中的归一性和可列可加性等价于：若 $\{B_k \in \mathcal{S} : k = 1, 2, \dots\}$ 是 Ω 的一个划分，则 $\sum_{k=1}^{\infty} P(B_k) = 1$ 。
- 1.25. 反复地抛一枚均匀的硬币，若头 100 次都是正面，问第 101 次抛该硬币出现反面的概率？有人认为正面的机会大些，因为前 100 次都是正面。而有人认为反面的机会大些，因为 Bernoulli 弱大数律说，当抛次趋向无穷时，正面频率以大概率接近 $P(H) = 1/2$ 。你如何认为？
- ☆ 1.26. 已知 A, B 是两个随机事件，试证明： $|P(AB) - P(A)P(B)| \leq 1/4$ 。
- ☆ 1.27. 餐馆衣帽台彻底弄乱了客人们的 n 顶帽子和它们的号牌，求恰有 k 人取到自己帽子的概率 $P_{n,k}$ ？
- 1.28. 已知随机事件序列 A_1, A_2, \dots 满足 $\sum_{n=1}^{\infty} P(A_n) < \infty$ ，证明并解释：
- $$P\left\{\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c\right\} = 1$$
- 1.29. 设 Bernoulli 试验中事件 A 的概率 $P(A) = 1/2$ 。在 3 重 Bernoulli 试验中，若已知 A 至少出现 1 次，求 A^c 至少出现一次的概率。
- 1.30. 设 A, B 是任意两个事件，其中 A 的概率不等于 0 和 1，试证明： $P(B|A) = P(B|A^c)$ 是事件 A 与 B 独立的充分必要条件。
- 1.31. 假设两枚硬币都是均匀的，定义事件 A = “抛 1 号硬币出现正面”，事件 B = “抛 2 号硬币出现正面”，事件 C = “抛 1、2 号硬币只出现一个正面”。试说明：事件 A, B, C 两两独立，而事件 A, B, C 并不独立。
- 1.32. 有朋自远方来，所乘交通工具为火车、轮船、汽车、飞机的概率分别为 $0.3, 0.2, 0.1, 0.4$ ，而这些交通工具迟到的概率分别为 $1/4, 1/3, 1/12, 0$ 。求：(1) 朋友迟到的概率；(2) 若朋友迟到了，朋友乘火车的概率。
- 1.33. 设某学生期中考试及格的概率为 p 。若期中考试及格，则期末考试及格的概率也为 p ；若期中考试不及格，则期末考试及格的概率为 $p/2$ 。(1) 求至少有一次考试及格的概率；(2) 若期末考试及格，求期中考试及格的概率。
- ☆ 1.34. 给定非空集合 Ω ，记 \mathbb{R}_+ 为非负实数集合，满足以下三个条件的集函数

$\mu^* : 2^\Omega \rightarrow \mathbb{R}_+ \cup \{\infty\}$ 称为 Carathéodory 外测度。

$$\begin{aligned}\mu^*(\emptyset) &= 0 \\ A \subset B \Rightarrow \mu^*(A) &\leq \mu^*(B) \\ \mu^*\left(\bigcup_{j=1}^{\infty} A_j\right) &\leq \sum_{j=1}^{\infty} \mu^*(A_j)\end{aligned}$$

(1) 试证明 $\mu^*(B) \leq \mu^*(B \cap A) + \mu^*(B \cap A^c)$; (2) 若 $\forall B \subseteq \Omega$, 皆有 $\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c)$, 则称 A 是 Carathéodory μ^* -可测的。令 \mathcal{S} 是所有 Carathéodory μ^* -可测的子集的全体, 求证 $(\Omega, \mathcal{S}, \mu^*)$ 是一个完备的测度空间。

☆ 1.35. 已知概率空间 (Ω, \mathcal{S}, P) , 则对任意事件 $A_1, \dots, A_n \in \mathcal{S}$ 有

$$P\left(\bigcup_{k=1}^n A_k\right) \geq \sum_{k=1}^n P(A_k) - \sum_{1 \leq i < j \leq n} P(A_i A_j)$$

☆ 1.36. 以 A_t 表示“某分子在时间段 $(0, t]$ 内不与其它分子碰撞”, 在 A_t 发生的条件下, 该分子在时间段 $(t, t + \Delta t]$ 内与其它分子发生碰撞”的概率为 $\lambda \Delta t + o(\Delta t)$, 其中 $\lambda > 0$ 为常数, 求 $P(A_t)$ 。

1.37. 商标“MAXIMA”中有 2 个字母脱落, 有人捡起它们随意放回, 问放回后仍为“MAXIMA”的概率?

1.38. 已知概率空间 (Ω, \mathcal{S}, P) , 事件 $A_j \in \mathcal{S}, j = 1, 2, \dots$ 是 Ω 的一个划分, 且事件 C 满足 $P(C A_j) > 0$, 试证明: 对任一事件 B 皆有

$$P(B|C) = \sum_{j=1}^{\infty} P(B|C, A_j) P(A_j|C)$$

1.39. 盒子里装有一球, 可能是白球也可能是黑球。放一个白球到盒子中, 然后再从中随机取出一球。若取出的球是白球, 问盒子里剩下的球也是白球的概率。

☆ 1.40. 一条生产线连续生产 n 件产品不出故障的概率为 $\lambda^n e^{-\lambda} / n!$, 其中 $n = 0, 1, 2, \dots$ 。假设产品的正品率为 $0 < p < 1$, 并且各产品是否为正品相互独立。(1) 求两次故障之间共生产 k 件正品的概率, 其中 $k = 0, 1, 2, \dots$; (2) 若已知在某两次故障之间生产了 k 件正品, 求生产线共生产 m 件产品的概率。

★ 1.41. 有 $N+1$ 个盒子 A_0, A_1, \dots, A_N , 假设 N 非常之大。盒子 A_k 有 k 个黑球, $N-k$ 个白球, $k = 0, 1, \dots, N$ 。从这 $N+1$ 个盒子中随便取一个盒子, 从该盒中有放回地抽取 n 次球, 结果全为黑球, 求下一次抽取还是黑球的概率。

☆ 1.42. 令 $A_0 = \emptyset$, 若对于 $j = 1, 2, \dots$ 皆有 $A_j A_{j-1}^c \cdots A_0^c$ 与 B_j 独立, 则

$$P\left\{\bigcup_{j=1}^{\infty} A_j B_j\right\} \geq \alpha P\left\{\bigcup_{j=1}^{\infty} A_j\right\}, \text{ 其中 } \alpha = \inf_j P(B_j)$$

1.43. 如果 $A_1 \perp\!\!\!\perp_B A_2$, 试证明 $P(A_1|B) = P(A_1|BA_2)$ 。

☆ 1.44. 设事件 $A_j, j = 1, 2, \dots$ 满足 $P(A_j) = 1$, 试证明: $P(\bigcap_{j=1}^{\infty} A_j) = 1$ 并给出解释。

第二章

随机变量及其数字特征

欲穷千里目，更上一层楼。

王之涣《登鹳雀楼》

在数学发展史上，对变量的认识曾带来观念和方法的革命^{*}，概率论亦是如此。在此之前，我们谈论随机事件，只有“发生”和“不发生”两个状态。而随机变量，粗略地讲，则可以有多个可能的取值。当随机变量的概念被引入到概率论中，具体问题被抽象和提炼成具有广泛代表性的一般问题，现代概率论才得以蓬勃发展。

历史上，法国数学家、物理学家 S. D. Poisson (1781-1840) 可能是首位意识到随机变量的一般概念并将之从具体问题中剥离出来的概率专家。在此之前的数学家，像 Laplace、Gauss 等人也使用随机变量，但都与具体的观察结果或者误差问题有关。Poisson 眼中的离散型随机变量是以相应概率取值的“任何东西”(une chose quelconque)，他试图以同样的方式考虑连续型随机变量及其分布函数。然而，Poisson 对随机变量的认识并未突破他的时代，真正严格的定义是在 Kolmogorov 概率公理化之后才姗姗而至。

日本概率论大师伊藤清 (Kiyoshi Itô, 1915-2008) 在《我的六十年概率之路》回忆了这段历史，“从我当学生起，我就被看上去完全随机的现象中存在统计规律这一事实吸引。我知道概率论是描述这类现象的手段，在大三我开始阅读概率论方面的文章和书籍。逐渐地，我明白统计规律带有数学本质，不过我对当时概率论的文章并不满意，因为它们连概率论里最基本的东西——随机变量——也未清晰地定义。

尽管现在人们理所当然地认为数学系统应该建立在每个概念严格定义的基础之上，整个领域也只是最近才达成这个共识。譬如，微分和积分运算在十九世纪末实



^{*}变量数学始于十七世纪上半叶，其标志是法国著名哲学家、数学家 René Descartes (1596-1650) 发表了解析几何学的奠基之作《几何学》(1637)，使微积分的创立成为可能。具体内容请参阅美国数学史专家、科普作家 Morris Kline (1908-1992) 的名著《古今数学思想》[89]。

数概念的清晰定义被形式化之后才发展成为严谨的数学系统。……然而，那时候的概率论文章和著作还像十九世纪的微积分一样用着直观的描述。”

例如，按照 Poisson 的理解，随机变量的取值是不确定的，总是以某一概率来取值，这是它与普通变量不一样的地方。

例 2.1. 接着**例 1.47**：在 n 重 Bernoulli 试验中，随机事件“出现正面”的次数 X 是一个随机变量，它的取值范围是 $\{0, 1, 2, \dots, n\}$ ，其中 X 的取值为 k 的概率是

$$P(X = k) = C_n^k p^k (1 - p)^{n-k}, \text{ 其中 } k = 0, 1, 2, \dots, n$$

我们可以方便地探讨 X 的性质，而不必计较它的具体取值，这就是变量数学的好处。譬如，实函数 $F_X(x) = P(X \leq x)$ 关于 x 是非减的。

例 2.1 对随机变量的定义虽然朴素直观，但没法推广到一般情况。一方面，脱离了概率空间谈论随机变量无法交代清楚随机变量和随机事件的关系。另一方面，在实践中，随机试验的结果往往不是数域，如抛硬币、摸扑克牌、词典中任取一词、钢琴上弹出一个音符等。譬如，下面的五线谱就是乱弹琴的结果。



图 2.1: 按照已定的节拍随机产生的 MIDI 序列。电脑音乐的制作经常用到随机方法。

伊藤清回忆道，“我找到俄国数学家 Kolmogorov 写的一本书。我意识到这正是我想要的，于是一口气读完了。在这本 1933 年用德文写的《概率论基础》里，Kolmogorov 试图把随机变量定义为概率空间上的函数，并且用测度论将概率论系统化。我觉得这本书仿佛清除了阻挡我视线的雾霾，引领我最终坚信概率论可被建成现代数学的一个领域。在我心中，概率论的基础从此建立起来了。”要定义随机变量，必须想办法将随机事件与实数域 \mathbb{R} 建立起联系。

例 2.2. 接着**例 2.1**，更合理地定义随机变量：连续抛 n 次硬币，基本事件集合 Ω 是所有由 H, T 构成的长度为 n 的序列。定义单值函数 $X : \Omega \rightarrow \mathbb{R}$ 如下，

$$X(\omega) = \omega \text{ 中 } H \text{ 的个数, 其中 } \omega \in \Omega$$

显然，作为函数，随机变量 X 的值域是 $\{0, 1, 2, \dots, n\}$ ，它取值 $k \in \{0, 1, 2, \dots, n\}$ 的概率，记作 $P(X = k)$ ，定义为随机事件 $X^{-1}(k)$ 的概率，即

$$P(X = k) = P\{X^{-1}(k)\} = C_n^k p^k (1 - p)^{n-k}$$

这样定义的随机变量 X , 完全被 $P(X = k), k = 0, 1, 2, \dots, n$ 精确描述, 我们可以抛开它的原始定义 $X : \Omega \rightarrow \mathbb{R}$ 来谈论它的性质。

回顾第 58 页的例 1.46, 我们曾利用单值映射把原问题的样本空间“翻译”成 Borel 可测空间。为了“故技重施”, 我们需要做一些一般化的工作, 其中要用到下述有关单值映射^{*}逆像的重要性质。

性质 2.1. 已知单值映射 $g : \Omega \rightarrow \Delta$, 定义 $A \subseteq \Delta$ 的逆像为

$$g^{-1}(A) = \{\omega \in \Omega : g(\omega) \in A\}$$

$g^{-1}(A)$ 可能为空集, 所以 g^{-1} 不一定是映射。对于 $A, A_k \subseteq \Delta$, 其中 k 属于某个指标集 K (可以是不可数的), 求逆像 g^{-1} 可以和 Δ 的子集的交、并、补运算交换次序, 即

$$g^{-1}(A^c) = [g^{-1}(A)]^c \quad (2.1)$$

$$g^{-1}\left(\bigcap_{k \in K} A_k\right) = \bigcap_{k \in K} g^{-1}(A_k) \quad (2.2)$$

$$g^{-1}\left(\bigcup_{k \in K} A_k\right) = \bigcup_{k \in K} g^{-1}(A_k) \quad (2.3)$$

证明. 下面证结果 (2.2), 其他结果留给读者证明。

$$\forall x \in g^{-1}\left(\bigcap_{k \in K} A_k\right) \Leftrightarrow g(x) \in \bigcap_{k \in K} A_k \Leftrightarrow g(x) \in A_k, \forall k \in K \Leftrightarrow x \in g^{-1}(A_k), \forall k \in K \quad \square$$

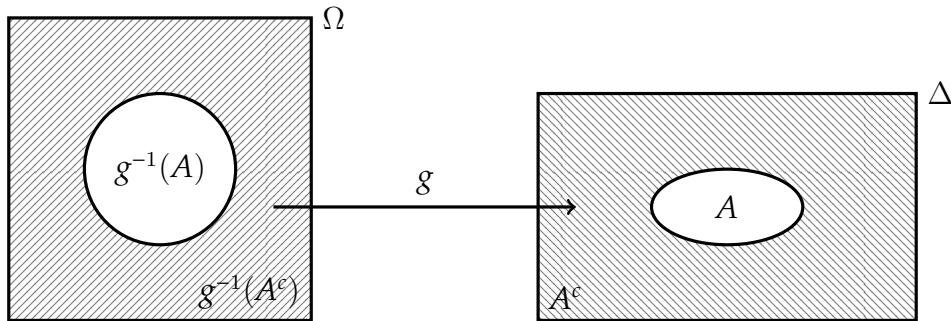


图 2.2: 结果 (2.1) 的直观解释: A 的补集的逆像即是 A 的逆像的补集。换句话说, 求逆像和求补集可以交换次序。

*单值映射 (single-valued mapping) 把定义域中任一元素对应到值域中一个元素, 多值映射 (multivalued mapping) 则可以把一个元素对应到若干个元素。本书几乎不涉及多值映射, 所以提到的映射缺省都为单值映射。注意: 单值映射和单射 (injection) 是两个不同的概念。

定义 2.1 (可测映射). 设映射 $g : \Omega_1 \rightarrow \Omega_2$ 是定义在两个可测空间 $(\Omega_1, \mathcal{S}_1)$ 和 $(\Omega_2, \mathcal{S}_2)$ 之间的单值映射, 我们称 g 为一个可测映射*, 如果它满足条件

$$\forall S \in \mathcal{S}_2, \text{ 皆有 } \{\omega \in \Omega_1 : g(\omega) \in S\} \in \mathcal{S}_1$$

 显然, 单值映射 g 是可测映射当且仅当任意可测集的逆像仍为可测集。可测映射 g 的好处是: 对 $(\Omega_2, \mathcal{S}_2)$ 的可测集进行可数个集合运算的操作, 通过求逆像回到 $(\Omega_1, \mathcal{S}_1)$, 根据**性质 2.1**, 如同遥控着在 $(\Omega_1, \mathcal{S}_1)$ 上做一样的操作, 而且从不用担心运算的封闭性。

性质 2.2. 若 $g : (\Omega_1, \mathcal{S}_1) \rightarrow (\Omega_2, \mathcal{S}_2)$ 和 $h : (\Omega_2, \mathcal{S}_2) \rightarrow (\Omega_3, \mathcal{S}_3)$ 都是可测映射, 则合成映射 $h \circ g : (\Omega_1, \mathcal{S}_1) \rightarrow (\Omega_3, \mathcal{S}_3)$ 也是可测的。即, 可测映射的合成依然可测。

证明. 对于任意 $S \in \mathcal{S}_3$, 有 $(h \circ g)^{-1}(S) = g^{-1}(h^{-1}(S)) \in \mathcal{S}_1$. □

定义 2.2 (Borel 可测函数). 如果 $g : \Omega \rightarrow \mathbb{R}^n$ 是定义在可测空间 (Ω, \mathcal{S}) 与 $(\mathbb{R}^n, \mathcal{B}_n)$ 之间的可测映射, 则称 g 为可测函数。特别地, 当 (Ω, \mathcal{S}) 是 Borel 可测空间时, 这样的可测函数 g 被称为 Borel 可测函数或 Borel 函数。我们平常用到的实函数多是 Borel 函数, 它有很多好的性质, 例如 Borel 函数把随机变量依然映为随机变量。



图 2.3: 可测映射的合成依然是可测映射。

可测函数是一类很广泛的函数, 具有很好的运算封闭性。**附录 D** 列举了可测函数的一些常用性质, 读者也可以参阅 W. Rudin 的名著《数学分析原理》[140] 的第十一章或测度论 [68]、函数论 [94, 163] 教材。下面, 我们不加证明地介绍可测函数的一个重要性质, 该性质将用于定义随机变量和随机向量[†]。

性质 2.3. 定义在可测空间 (Ω, \mathcal{S}) 上的实值函数 $g : \Omega \rightarrow \mathbb{R}^n$ 是可测函数当且仅当对任意列向量 $\mathbf{x} = (X_1, X_2, \dots, X_n)^\top \in \mathbb{R}^n$ 皆有

$$\{\omega \in \Omega : g(\omega) \in (-\infty, \mathbf{x}]\} \in \mathcal{S}$$

*类似于两个拓扑空间之间连续映射的定义, 即任意开集的逆像仍是开集。

[†]随机向量就是取值为向量的随机变量, 在本质上随机向量和随机变量可以不作区分。但为了便于初学者理解, 我们还是把随机向量的内容单列一节。

即，任意 Borel 集 $(-\infty, x]$ 的逆像仍是可测集，其中 $(-\infty, x]$ 表示笛卡尔积 $(-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n] \subseteq \mathbb{R}^n$ 。

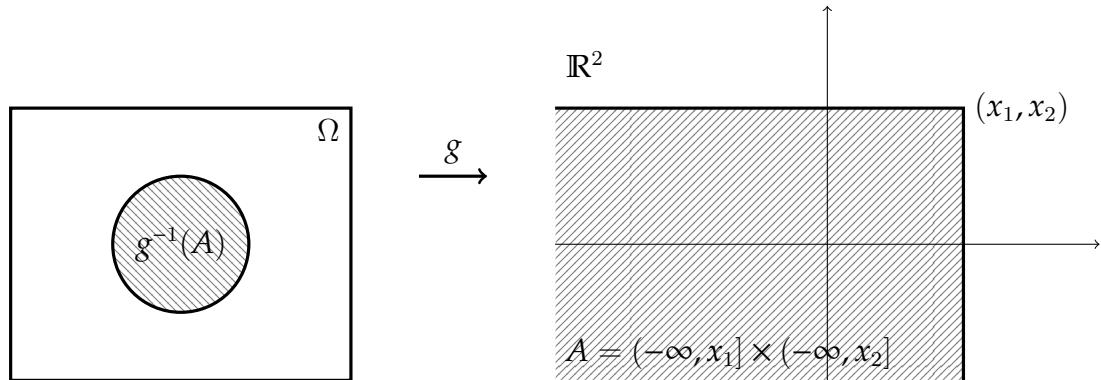
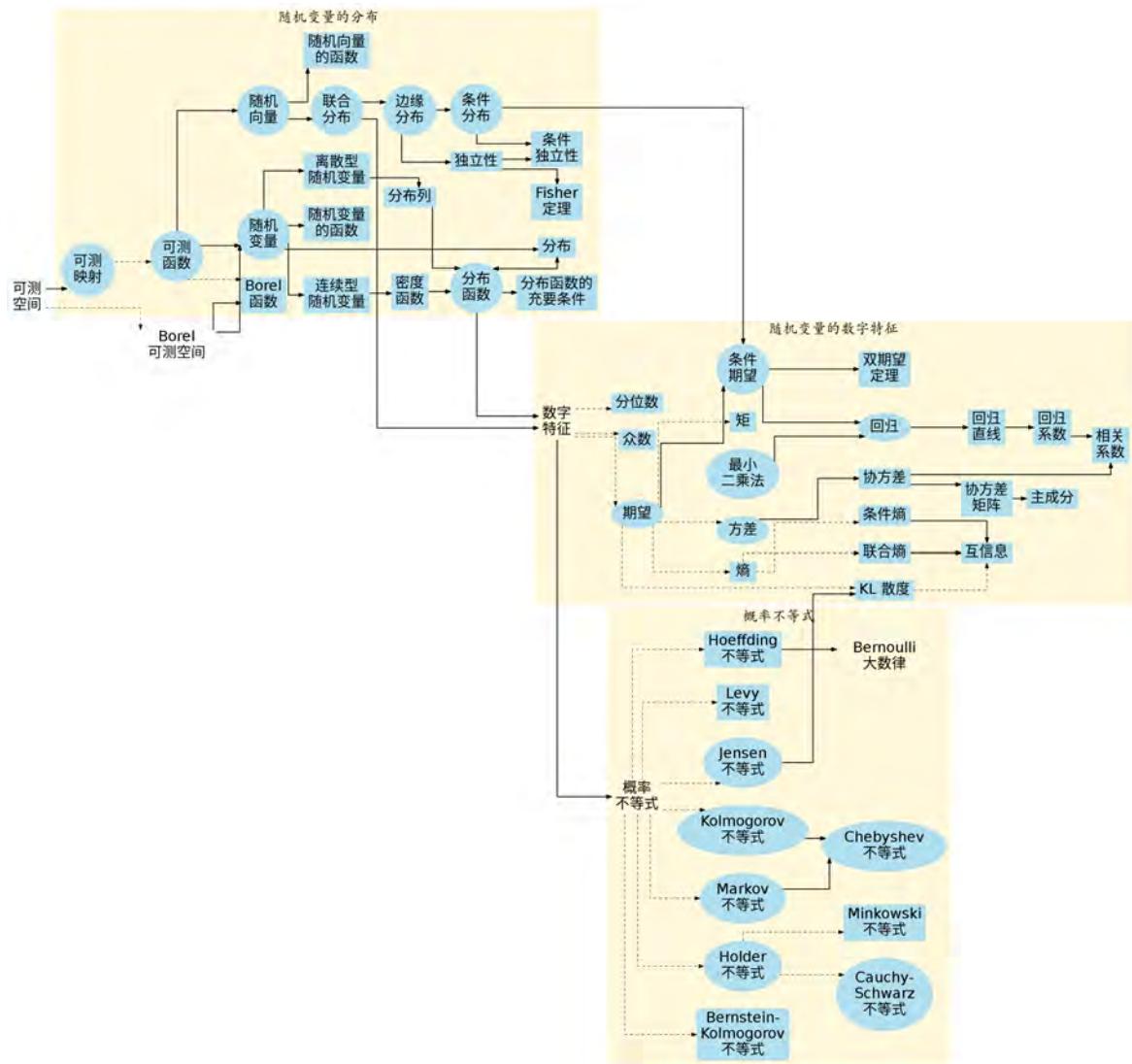


图 2.4: 实值函数 $g : \Omega \rightarrow \mathbb{R}^2$ 是可测函数当且仅当对任意 $A = (-\infty, x_1] \times (-\infty, x_2]$, 其逆像 $g^{-1}(A)$ 都是一个事件。

有了 Borel 可测函数，我们就可以定义随机变量。随机变量概念的引入，把概率论代入了变量数学的世界。

第二章的主要内容及其关系



2.1 随机变量及其基本性质

为了在样本空间 (Ω, \mathcal{S}) 和 Borel 空间 $(\mathbb{R}, \mathfrak{B}_1)$ 之间搭建桥梁，自然联想到令单值函数 $X : \Omega \rightarrow \mathbb{R}$ 为可测函数（见**定义 2.2**），即对于任意 Borel 集 $B \in \mathfrak{B}_1$ ，其逆像是 (Ω, \mathcal{S}) 的某一随机事件，即 $X^{-1}(B) \in \mathcal{S}$ ，这样我们才好谈论 $X^{-1}(B)$ 的概率。

定义 2.3 (随机变量). 定义在样本空间 (Ω, \mathcal{S}) 上的随机变量 (random variable, rv) X 就是样本空间 (Ω, \mathcal{S}) 到一维 Borel 空间 $\mathcal{X} = (\mathbb{R}, \mathfrak{B}_1)$ 的可测函数 $X : \Omega \rightarrow \mathbb{R}$ 。

自此以后，我们约定采用大写的英文字母或者小写的希腊字母表示随机变量，如 X, ξ 等。另外，小写的希腊字母还用来表示参数，如 θ, μ, σ 等。

定理 2.1 (随机变量的等价定义). 定义在样本空间 (Ω, \mathcal{S}) 上的单值函数 $X : \Omega \rightarrow \mathbb{R}$ 是一个随机变量当且仅当 $\forall x \in \mathbb{R}$ ，区间 $(-\infty, x]$ 的逆像是一个随机事件，即

$$X^{-1}(-\infty, x] = \{\omega : X(\omega) \leq x\} \in \mathcal{S}$$

证明. “ \Rightarrow ” 是显然的，因为 $(-\infty, x] \in \mathfrak{B}_1$ 。下面往证 “ \Leftarrow ”：由于 \mathbb{R} 的任一 Borel 集可通过对形如 $(-\infty, x]$ 的左开右闭集合进行可数个交、并、补运算得到，由**性质 2.1**，该 Borel 集的逆像是 (Ω, \mathcal{S}) 的一个随机事件，“ \Leftarrow ” 得证。 \square

为方便起见，我们把 $\{\omega : X(\omega) \in B\}$ 简记为 $\{X \in B\}$ ，在不引起歧义的情况下有时也记作 $X \in B$ 。由随机变量的定义知， $\{X \in \{x\}\}$ 或 $\{X = x\}$ 、 $\{X \in (a, b)\}$ 或 $\{a < X < b\}$ 、 $\{a < X \leq b\}$ 、 $\{a \leq X < b\}$ 、 $\{a \leq X \leq b\}$ 都是随机事件。随机变量更一般的定义是概率空间 (Ω, \mathcal{S}, P) 到 Borel 空间 $(\mathbb{R}, \mathfrak{B}_1)$ 的映射 X ，使得 $\forall A \in \mathfrak{B}_1, X^{-1}(A)$ 都是 P -可测的（见**定义 1.15**）。**定义 2.3** 显然是上述定义的特款。因为本书几乎不涉及测度论，我们对基于测度论的概率理论浅尝辄止，不再细究数学上的严谨。

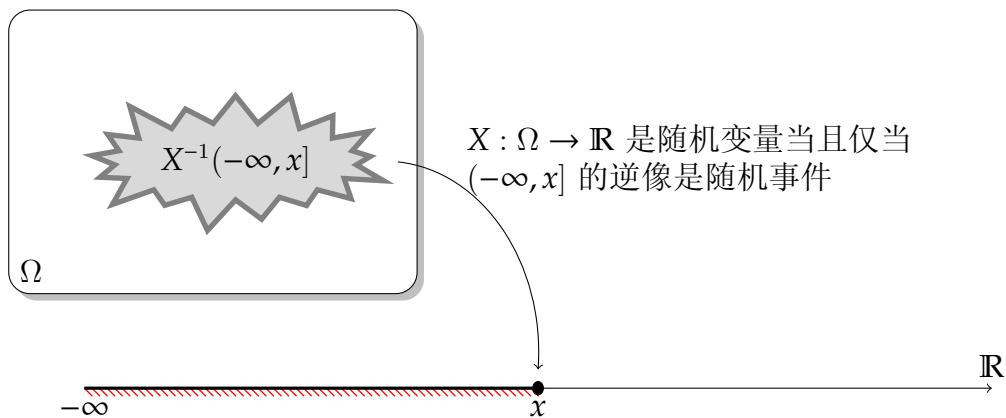


图 2.5: 单值函数 $\Omega \xrightarrow{X} \mathbb{R}$ 是一个随机变量当且仅当任意一个左开右闭区间 $(-\infty, x]$ 的逆像为某一随机事件 $\{X \leq x\} \in \mathcal{S}$ 。

例 2.3. 掷骰子的基本事件集合是 $\Omega = \{1, 2, 3, 4, 5, 6\}$, 定义 σ 域 $\mathcal{S} = 2^\Omega$ 。考虑如下定义的单值函数 $X : \Omega \rightarrow \mathbb{R}$, 满足 $X(k) = k$, 其中 $k = 1, 2, \dots, 6$ 。容易验证

$$\{\omega \in \Omega : X(\omega) \leq x\} = \{X \in (-\infty, x]\} = \begin{cases} \emptyset \in \mathcal{S} & \text{当 } x < 1 \\ \{1\} \in \mathcal{S} & \text{当 } 1 \leq x < 2 \\ \vdots \\ \Omega \in \mathcal{S} & \text{当 } x \geq 6 \end{cases}$$

按照定理 2.1, X 是定义在样本空间 (Ω, \mathcal{S}) 上的随机变量。类似地, 在同一样本空间上, 我们可以定义其他的随机变量, 譬如, 容易验证 $Y : \Omega \rightarrow c$ (其中 $c \in \mathbb{R}$ 是常数) 也是一个定义在样本空间 (Ω, \mathcal{S}) 上的随机变量, 这是因为

$$\{\omega \in \Omega : Y(\omega) \leq y\} = \{Y \in (-\infty, y]\} = \begin{cases} \emptyset \in \mathcal{S} & \text{当 } y < c \\ \Omega \in \mathcal{S} & \text{当 } y \geq c \end{cases}$$

练习 2.1. 接着例 2.3, 令 $a < b$ 为实数, 定义映射 $Z : \Omega \rightarrow \{a, b\}$ 满足

$$Z(\omega) = \begin{cases} a & \text{当 } \omega = 1, 3, 5 \\ b & \text{当 } \omega = 2, 4, 6 \end{cases}$$

请读者验证 Z 是定义在样本空间 (Ω, \mathcal{S}) 上的随机变量。

例 2.4. 抛一枚均匀的硬币两次, 令随机变量 $X : \Omega \rightarrow \{0, 1, 2\}$ 表示正面的次数, 其中 $\Omega = \{TT, HT, TH, HH\}$ 。显然,

$$\begin{aligned} \{X(\omega) = 0\} &= \{TT\} \\ \{X(\omega) = 1\} &= \{HT, TH\} \\ \{X(\omega) = 2\} &= \{HH\} \end{aligned}$$

例 2.5. 抛一枚硬币出现正面的概率是 $P(H) = p$, 令 a, b 是两个实数。定义随机变量 $X : \{H, T\} \rightarrow \mathbb{R}$ 如下,

$$X(\omega) = \begin{cases} a & \text{若 } \omega = H \\ b & \text{若 } \omega = T \end{cases}$$

显然, $P\{X(\omega) = a\} = p$, $P\{X(\omega) = b\} = 1 - p$ 。并且, $P\{X(\omega) = x\} = 0$, 其中 $x \notin \{a, b\}$ 。

定义 2.4 (指示函数). 集合 A 的指示函数 (indicator function) I_A 定义为

$$I_A(a) = \begin{cases} 1 & \text{当 } a \in A \\ 0 & \text{当 } a \notin A \end{cases} \quad (2.4)$$

性质 2.4. I_A 是定义在样本空间 (Ω, \mathcal{S}) 上的随机变量当且仅当 $A \in \mathcal{S}$, 因为

$$\{\omega : I_A(\omega) \leq x\} = \begin{cases} \emptyset \in \mathcal{S} & \text{当 } x < 0 \\ A^c \in \mathcal{S} & \text{当 } 0 \leq x < 1 \\ \Omega \in \mathcal{S} & \text{当 } x \geq 1 \end{cases}$$

显然, 随机变量 I_A 取值 1, 0 的概率分别为

$$\begin{aligned} P(I_A = 1) &= P(A) \\ P(I_A = 0) &= 1 - P(A) \end{aligned}$$

练习 2.2. 请读者证明指示函数的以下性质:

① $A \subseteq B$ 当且仅当 $I_A \leq I_B$ 。特别地, $A = B$ 当且仅当 $I_A = I_B$ 。

② $I_{A^c} = 1 - I_A$, $I_{AB} = I_A I_B$ 且 $I_{A \cup B} = I_A + I_B - I_{AB}$ 。

性质 2.5. 已知 X 是定义在样本空间 (Ω, \mathcal{S}) 上的随机变量, 则对于任意常数 $a, b \in \mathbb{R}$, $aX + b$ 也是一个定义在 (Ω, \mathcal{S}) 上的随机变量。

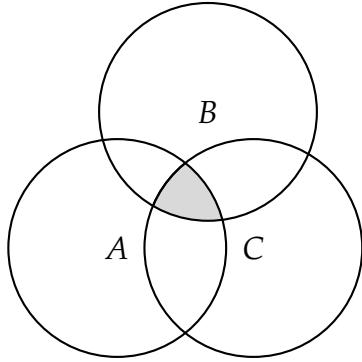
证明. 往证 $\{\omega : aX(\omega) + b \leq x\} = \{aX + b \leq x\} \in \mathcal{S}$, 事实上

$$\{aX + b \leq x\} = \begin{cases} \{X \leq (x - b)/a\} \in \mathcal{S} & \text{当 } a > 0 \\ \{X \geq (x - b)/a\} \in \mathcal{S} & \text{当 } a < 0 \\ \Omega \in \mathcal{S} & \text{当 } a = 0 \text{ 且 } x \geq b \\ \emptyset \in \mathcal{S} & \text{当 } a = 0 \text{ 且 } x < b \end{cases} \quad \square$$

性质 2.6. 已知样本空间 (Ω, \mathcal{S}) , 令事件 $A_1, A_2, \dots, A_n \in \mathcal{S}$ 是 Ω 的一个划分, 则简单函数 (见第 760 页的式 D.1) $X = \sum_{j=1}^n x_j I_{A_j}$ (其中 $x_j \in \mathbb{R}$ 有限) 为一个随机变量, 我们称之为简单随机变量, 它是 I_{A_1}, \dots, I_{A_n} 的线性组合。

证明. $\forall x \in \mathbb{R}$, 集合 $\{\omega : \sum_{j=1}^n x_j I_{A_j}(\omega) \leq x\} \in \mathcal{S}$, 因为它是 A_1, A_2, \dots, A_n 中某些事件的并集, 自然也是一个事件。 \square

例 2.6. 由事件 $A, B, C \in \mathcal{S}$ 定义的实函数 $X = I_A - 3I_B + 2I_C - 1$ 是一个随机变量, 共有 7 个可能的取值 $\{-4, -3, -2, -1, 0, 1, 2\}$ 。下面列出所有的情况, 请读者验证。



事件	I_A	$-3I_B$	$2I_C$	X
$A^cB^cC^c$	0	0	0	-1
A^cB^cC	0	0	2	1
A^cBC^c	0	-3	0	-4
A^cBC	0	-3	2	-2
AB^cC^c	1	0	0	0
AB^cC	1	0	2	2
ABC^c	1	-3	0	-3
ABC	1	-3	2	-1

图 2.6: 事件 A, B, C 之间通过集合运算可得到 8 个基本结果。譬如, 若 $\omega \in A^cB^cC^c$, 则 $X(\omega) = -1$ 。反之, $\{X = -1\} = A^cB^cC^c + ABC$ 。

随机变量就是定义在样本空间 (Ω, \mathcal{S}) 上的(可测)函数, 通常情况下与我们过去对函数的理解并无二异。从函数逼近的角度看, 任何随机变量 X 都可以由简单随机变量来近似。

为简单起见, 考虑有界的随机变量 $X \in (a, b]$ 。不妨设 $a = c_0 < c_1 < \cdots < c_{n-1} < c_n = b$ 是 $(a, b]$ 的任意一个分割, 令 $S_j = (c_{j-1}, c_j]$, 其中 $j = 1, 2, \dots, n$, 则 $E_j = X^{-1}(S_j)$, $j = 1, 2, \dots, n$ 是 Ω 的一个划分。构造简单随机变量 X_n 如下,

$$X_n = \sum_{j=1}^n x_j I_{E_j}, \text{ 其中 } x_j \in S_j \quad (2.5)$$

只要区间分割得足够细致, 可用简单随机变量 X_n 来逼近随机变量 X 。

定义 2.5. 类似函数之间大小的比较, 定义在样本空间 (Ω, \mathcal{S}) 上的随机变量 X 和 Y 满足 $X \leq Y$ 当且仅当 $\forall \omega \in \Omega$, 皆有 $X(\omega) \leq Y(\omega)$ 。

练习 2.3. 如果在式 (2.5) 中, x_j 选为区间 S_j 的右端点, 则 $X \leq X_n$ 。

性质 2.7. 若 X, Y 都是定义在样本空间 (Ω, \mathcal{S}) 上的随机变量, 则 $|X|^\alpha, X \pm Y, XY$ 也是随机变量, 其中 $\alpha > 0$, 并且满足不等式

$$|X \pm Y| \leq |X| + |Y| \quad (2.6)$$

$$|XY| \leq \frac{|X|^r}{r} + \frac{|Y|^s}{s}, \text{ 其中 } r > 1 \text{ 且 } \frac{1}{r} + \frac{1}{s} = 1 \quad (2.7)$$

证明. 不难验证下述集合都是随机事件。

$$\begin{aligned}\{|X|^\alpha \leq z\} &= \begin{cases} \emptyset & \text{若 } z < 0 \\ \{X \leq z^{1/\alpha}\} \cap \{X \geq -z^{1/\alpha}\} & \text{若 } z \geq 0 \end{cases} \\ \{X + Y \leq z\} &= \bigcup_{x \in \mathbb{R}} \{X \leq x\} \cap \{Y \leq z - x\} \\ \{|XY| \leq z\} &= \begin{cases} \emptyset & \text{若 } z < 0 \\ \{X = 0\} \cup \{Y = 0\} \cup \bigcup_{x > 0} \{|X| \leq x\} \cap \{|Y| \leq z/x\} & \text{若 } z \geq 0 \end{cases}\end{aligned}$$

显然, $\forall \omega \in \Omega$, 皆有 $|X(\omega) + Y(\omega)| \leq |X(\omega)| + |Y(\omega)|$ 。利用 Young 不等式 (F.2) 容易验证不等式 (2.7)。 \square

本节内容

第一小节讨论随机变量的分布函数及其性质, 对随机变量的研究落实在它的分布函数上。第二小节分别定义了离散型和连续型两类随机变量, 并举例给出非离散型也非连续型的随机变量。初步介绍了单点分布、两点分布、二项分布、均匀分布、正态分布等常见分布, 有关它们的更多的细节见第 4 章。第三小节探讨了如何由已知随机变量 X 和已知函数 g 构造新的随机变量 $Y = g(X)$, 并求 Y 的分布列或密度函数。

关键知识

(1) 分布函数及其充要条件; (2) 离散型随机变量: 单点分布、两点分布、二项分布; (3) 连续型随机变量: 均匀分布、正态分布等。

2.1.1 随机变量的分布与分布函数

已知定义在概率空间 (Ω, \mathcal{S}, P) 上的随机变量 $X : \Omega \rightarrow \mathbb{R}$, 如何构造 Borel 空间 $(\mathbb{R}, \mathfrak{B}_1)$ 上相应的概率测度, 使得谈论 Borel 集 $B \in \mathfrak{B}_1$ 的概率测度就如同谈论随机事件 $\{X \in B\} \in \mathcal{S}$ 的概率? 我们需要“提炼”出下面的概念。

定义 2.6 (分布). 对于任意 $B \in \mathfrak{B}_1$, 如下定义的集函数 $F_X : \mathfrak{B}_1 \rightarrow [0, 1]$ 被称为随机变量 X 在概率空间 (Ω, \mathcal{S}, P) 上的概率分布 (probability distribution), 简称分布。

$$F_X(B) = P(\{X \in B\}), \text{ 其中 } B \in \mathfrak{B}_1$$

定理 2.2. 定义 2.6 所描述的集函数 $F_X(\cdot)$ 是定义在样本空间 $(\mathbb{R}, \mathfrak{B}_1)$ 上 (由 X 诱导出) 的概率测度, 即 $(\mathbb{R}, \mathfrak{B}_1, F_X)$ 是一个概率空间。

$$(\Omega, \mathcal{S}, P) \xrightarrow{X} (\mathbb{R}, \mathfrak{B}_1, F_X)$$

证明. 往证 $F_X(\cdot)$ 满足定义 1.16: 若 $B_j \in \mathfrak{B}_1, j = 1, 2, \dots$ 两两不交, 则

$$F_X\left(\bigcup_{j=1}^{\infty} B_j\right) = P\left(\left\{X \in \bigcup_{j=1}^{\infty} B_j\right\}\right) = P\left(\bigcup_{j=1}^{\infty} \{X \in B_j\}\right) = \sum_{j=1}^{\infty} P\{X \in B_j\} = \sum_{j=1}^{\infty} F_X(B_j)$$

显然 $\forall B \in \mathfrak{B}_1$, 皆有 $0 \leq F_X(B) \leq 1$ 且 $F_X(\mathbb{R}) = 1$ 。 \square

定义 2.6 和定理 2.2 描述概率空间 (Ω, \mathcal{S}, P) 与 $(\mathbb{R}, \mathfrak{B}_1, F_X)$ 的关系如下。

$$\begin{array}{ccc} (\Omega, \mathcal{S}, P) & \xrightarrow{X} & (\mathbb{R}, \mathfrak{B}_1, F_X) \\ \Downarrow & & \Downarrow \\ \{X \in B\} & \xleftarrow{X^{-1}} & B \\ \downarrow & & \downarrow \\ P(\{X \in B\}) & = & F_X(B) \end{array}$$

利用分析工具直接讨论分布这一集函数并不是很方便, 这促使我们去寻找它的替代品。由练习 1.12, 所有形如 $(-\infty, x]$ 的区间生成了 \mathfrak{B}_1 , 于是有了下面的概念。

定义 2.7 (分布函数). 如下定义的实值函数 $F_X : \mathbb{R} \rightarrow [0, 1]$ 被称为随机变量 X 的累积分布函数 (cumulative distribution function, CDF), 以后简称之为分布函数*。

$$F_X(x) = P\{\omega : X(\omega) \leq x\} = P\{X^{-1}(-\infty, x]\} \quad (2.8)$$

*某些概率论著作对分布函数的定义是 $P(X < x)$, 这导致分布函数是左连续的 [58, 107], 而按照现在流行的定义, 分布函数 $P(X \leq x)$ 是右连续的 [46, 137]。差别由约定俗成引起, 多数结果不受影响。

 利用测度论可以证明分布函数与分布之间是一一对应的 [40, 46, 107]，所以在后文中二者不再严格区分，有时也简称分布函数为“分布”。在不引起歧义的情况下， X 的分布函数也简记作 $F(x) = P(X \leq x)$ 或 $P\{X \in (-\infty, x]\}$ 。本章的多数内容都与这个函数有关，随机变量的那些事儿都从它开始讲起。

例 2.7. 定义在同一概率空间 (Ω, \mathcal{S}, P) 上的不同的随机变量可通过分布函数加以区分。接着第 113 页的**例 2.3**，如果骰子是均匀的，则 X 和 Y 的分布函数分别是

$$F_X(x) = \begin{cases} 0 & \text{当 } x < 1 \\ \frac{1}{6} & \text{当 } 1 \leq x < 2 \\ \frac{1}{3} & \text{当 } 2 \leq x < 3 \\ \vdots & \\ 1 & \text{当 } x \geq 6 \end{cases} \quad F_Y(y) = \begin{cases} 0 & \text{当 } y < c \\ 1 & \text{当 } y \geq c \end{cases}$$

如果分布函数不同，随机变量一定是不同的。但反之不成立，即两个不同的随机变量可以拥有相同的分布函数！所以，当我们说一些随机变量具有相同的分布的时候，千万不要认为它们是相同的函数。

W. Feller 在《概率论及其应用》中这样解释为何我们要使用随机变量 [45]，“原则上，我们能够把概率论限制在由随机变量的概率分布所定义的样本空间上，这样做就避免涉及抽象的样本空间，也避免涉及‘试验’、‘实验结果’等术语。把概率论缩影于随机变量便于使用分析知识且简化了理论的诸多方面。”这是引入随机变量对概率论有利的一面。“可是，它也有遮蔽概率背景的弊端。随机变量仍然容易被含糊地当作‘某个以不同的概率取不同的值的东西’，然而随机变量就是通常的函数，而函数的概念绝不是概率论所特有的。”

 **性质 2.8.** 已知随机变量 X 的分布函数为 $F(x)$ ，设 $a < b$ ，则 X 落于区间 $(a, b]$ 上的概率是 $F(b) - F(a)$ ，即

$$P\{X \in (a, b]\} = F(b) - F(a)$$

证明. 因为 $a < b$ ，所以 $X^{-1}(-\infty, a] \subseteq X^{-1}(-\infty, b]$ 。进而，

$$\begin{aligned} P\{X \in (a, b]\} &= P\{X^{-1}(a, b]\} \\ &= P\{X^{-1}(-\infty, b] \cap (X^{-1}(-\infty, a])^c\} \\ &= P\{X^{-1}(-\infty, b]\} - P\{X^{-1}(-\infty, a]\} \\ &= F(b) - F(a) \end{aligned} \quad \square$$

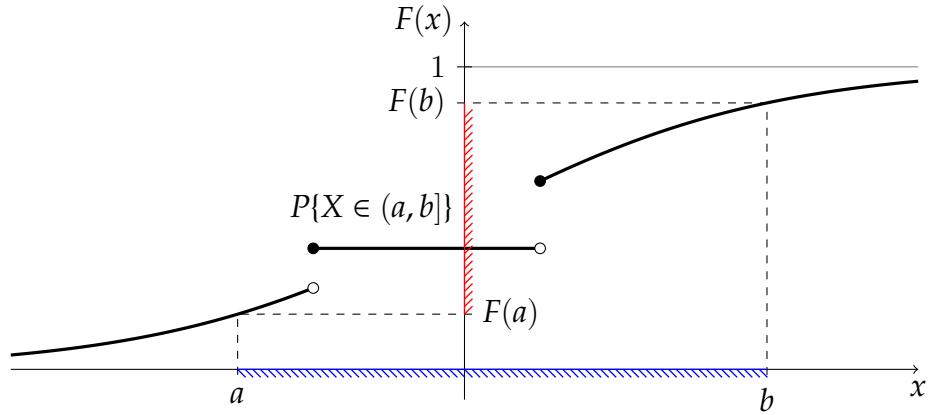


图 2.7: 性质 2.8 的直观解释: 随机变量 X 落于左开右闭区间 $(a, b]$ 的概率等于 $F(b) - F(a)$ 。空心点表示“抠掉”此点, 实心点表示“包含”此点。

练习 2.4. 接着性质 2.8 和图 2.7, 请读者验证并直观地解释

$$P\{X \in (a, b)\} = F(b-) - F(a)$$

其中 $F(b-)$ 表示 $F(x)$ 在 $x = b$ 点的左极限。另外还有,

$$\begin{aligned} P\{X \in [a, b]\} &= F(b-) - F(a-) \\ P\{X \in [a, b]\} &= F(b) - F(a-) \end{aligned}$$

→**定理 2.3** (分布函数的充要条件). 实值函数 $F(x)$ 是某概率空间上定义的随机变量 X 的分布函数, 见式 (2.8), 当且仅当 $F(x)$ 满足以下三条性质:

- ① 函数 $F(x)$ 是非减的。
- ② 函数 $F(x)$ 是右连续的, 即 F 在 x 点的右极限 $F(x+)$ 等于 $F(x)$ 。
- ③ 函数 $F(x)$ 满足 $F(-\infty) = 0$ 和 $F(+\infty) = 1$ 。

※证明. 往证 “ \Rightarrow ”: 已知 $F(x)$ 是定义在样本空间 (Ω, \mathcal{S}) 上的随机变量 X 的分布函数,

- (1) 如果 $x_1 > x_2$, 则事件 $\{\omega : X(\omega) \leq x_1\} \supseteq \{\omega : X(\omega) \leq x_2\}$, 进而

$$F(x_1) = P\{\omega : X(\omega) \leq x_1\} \geq F(x_2) = P\{\omega : X(\omega) \leq x_2\}$$

所以, 分布函数 $F(x)$ 是非减的。

- (2) 要证明 F 右连续, 只需验证对任意收敛到 x 的递减序列 $\{x_n\}$, 皆有 $F(x_n) \rightarrow F(x)$ 。令 $A_n = \{X \in (x, x_n]\}$, 则 $A_n \downarrow \emptyset$ 。由第 73 页的定理 1.6 有 $\lim_{n \rightarrow \infty} P(A_n) = 0$, 由性

质 2.8 即得

$$\lim_{n \rightarrow \infty} F(x_n) - F(x) = 0$$

(3) 令序列 $\{x_n\}$ 单调升趋于 $+\infty$, 则 $\{X \leq x_n\} \uparrow \Omega$ 。由定理 1.6, 有

$$F(+\infty) = \lim_{x_n \rightarrow +\infty} P(\{X \leq x_n\}) = 1$$

类似可证 $F(-\infty) = 0$, 留给读者。

往证 “ \Leftarrow ”: 在样本空间 $(\mathbb{R}, \mathcal{B}_1)$ 上, 定义 $P\{(-\infty, x]\} = F(x)$, 易证 P 是 $(\mathbb{R}, \mathcal{B}_1)$ 上的概率测度。于是, $F(x)$ 是定义在概率空间 $(\mathbb{R}, \mathcal{B}_1, P)$ 上的随机变量 $\mathbf{1} : \mathbb{R} \rightarrow \mathbb{R}$ (即 \mathbb{R} 到自身的恒等映射) 的分布函数。 \square

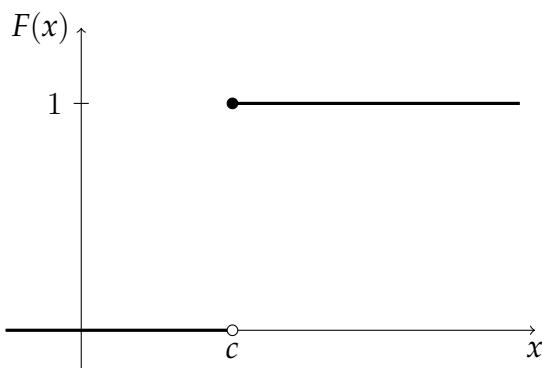
定义 2.8. 两个定义在概率空间 (Ω, \mathcal{S}, P) 上的随机变量 X 和 Y 等价当且仅当 $P(X \neq Y) = 0$, 记作 $X \stackrel{P}{\sim} Y$ 。

性质 2.9. 若两个随机变量等价, 则其分布函数相等。

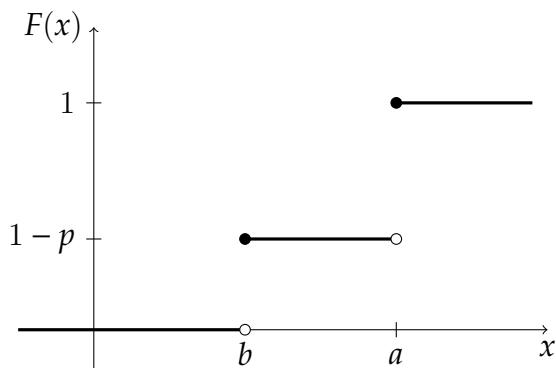
例 2.8 (单点分布). 在概率空间 (Ω, \mathcal{S}, P) 上, 如果事件 $\{X = c\}$ 的概率为 1, 即 X 几乎必然等于常数 c , 则称 X 服从单点分布 (one-point distribution), 记作 $X \sim \langle c \rangle$ 或者 $X \stackrel{a.s.}{=} c$ 或者 $P(X = c) = 1$ 。譬如, 常值函数 $X : \Omega \mapsto c$ 。

由式 (2.8), 可得随机变量 X 的分布函数 $F(x)$ 如下。显然, 单点分布的分布函数满足定理 2.3 所描述的三条性质。

$$F(x) = \begin{cases} 0 & \text{当 } x < c \\ 1 & \text{当 } x \geq c \end{cases} \quad (2.9)$$



(a) 单点分布 $X \sim \langle c \rangle$ 的分布函数



(b) 两点分布 $X \sim p\langle a \rangle + (1-p)\langle b \rangle$ 的分布函数

图 2.8: 单/两点分布的分布函数是有一/两个跳跃点的阶梯函数。

例 2.9 (两点分布). 参考例 2.5, 不妨设实数 $b < a$, 定义在概率空间 (Ω, \mathcal{S}, P) 上的随机变量 X 如果满足 $P(X = a) = p, P(X = b) = 1 - p$, 其中 $0 < p < 1$, 则称 X 服从两点分布 (two-point distribution), 记作 $X \sim p\langle a \rangle + (1-p)\langle b \rangle$ 。其分布函数 $F(x)$ 也是阶梯函数, 跳跃点是 $x = a$ 和 $x = b$, 具体如下。

$$F(x) = \begin{cases} 0 & \text{当 } x < b \\ 1-p & \text{当 } b \leq x < a \\ 1 & \text{当 } x \geq a \end{cases}$$

例 2.10 (0-1 分布). 我们称两点分布 $p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 为 0-1 分布。例如, 定义 2.4 刻画的指示函数 $I_A \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 其中 $p = P(A)$, A 是一个随机事件。

定理 2.4. 分布函数 $F(x)$ 的不连续点都是跳跃点且至多可数。

证明. 因为 \mathbb{R} 上单调增函数的不连续点都是第一类的且至多可数。 \square

练习 2.5. 利用性质 2.6 说明服从单点和两点分布的随机变量都是简单随机变量。

※例 2.11. 在第 67 页的例 1.52 中, 定义随机变量 $X : (I, \mathcal{B}_1 \cap I) \rightarrow (\mathbb{R}, \mathcal{B}_1)$ 为 $X(i) = i, \forall i \in I$ 。其分布函数为

$$F(x) = \begin{cases} 0 & \text{若 } x < 0 \\ h(x) & \text{若 } 0 \leq x < 1 \\ 1 & \text{若 } x \geq 1 \end{cases}$$

$F(x)$ 是一个连续函数, 其导函数几乎处处为零, 被称为奇异型分布 (singular distribution), 因为它是一个奇异概率测度 (见第 67 页的定义 1.19)。

2.1.2 离散型与连续型随机变量

定义 2.9 (分布列). 已知概率空间 (Ω, \mathcal{S}, P) 上定义的随机变量 X , 如果 $X(\Omega)$ 是可数的, 则称 X 为离散的随机变量。例如, 单点分布、两点分布等简单随机变量。

不妨设 $X(\Omega) = \{x_1, x_2, \dots\}$, 我们称 $P(X = x_j) = p_j$ 为概率质量函数 (probability mass function, pmf) 或概率函数, 并用下面的分布列来描述它。有时候, 也用 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \dots + p_j\langle x_j \rangle + \dots$ 来表示。

表 2.1: 离散型随机变量 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \dots + p_j\langle x_j \rangle + \dots$ 的分布列。

X	x_1	x_2	...	x_j	...
概率	p_1	p_2	...	p_j	...

性质 2.10. 已知随机变量 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \dots + p_j\langle x_j \rangle + \dots$, 总有

$$\sum_{j=1}^{\infty} p_j = 1$$

证明. 事件 $\{X = x_j\}, j = 1, 2, \dots$ 是基本事件集合 Ω 的一个划分, 所以 $\sum_{j=1}^{\infty} P(X = x_j) = 1$, 即 $\sum_{j=1}^{\infty} p_j = 1$ 。 \square

定义 2.10. 离散型随机变量 $X \sim p_0\langle 0 \rangle + p_1\langle 1 \rangle + \dots + p_k\langle k \rangle + \dots + p_n\langle n \rangle$ 被称为服从参数是 $(p_0, p_1, \dots, p_n)^T$ 的类别分布 (categorical distribution), 记作 $X \sim \text{Cat}(p_0, p_1, \dots, p_n)$ 。其中, p_k 即是 X 取第 k 个类的概率 $P(X = k), k = 0, 1, \dots, n$ 。

定义 2.11. 若概率函数 $P(X = k) = C_n^k p^k (1 - p)^{n-k}, k = 0, 1, \dots, n$, 则称 X 服从参数为 n, p 的二项分布 (binomial distribution), 记作 $X \sim B(n, p)$ 。见第 62 页的例 1.47。

显然, 单点分布、两点分布、二项分布都是类别分布的特款。“二项分布”一词源于 $C_n^k p^k (1 - p)^{n-k}$ 是二项式展开 $1 = [p + (1 - p)]^n = \sum_{k=0}^n C_n^k p^k (1 - p)^{n-k}$ 的通项。

性质 2.11. 离散型随机变量 X 也可用指示函数来表示,

$$X(\omega) = \sum_{j=1}^{\infty} x_j I_{\{X=x_j\}}(\omega)$$

其中, X 的分布函数为一个简单函数 (见附录 D) 如下,

$$F(x) = \sum_{x_j \leq x} [F(x_j) - F(x_j - 0)] = \sum_{x_j \leq x} P(X = x_j) = \sum_{j=1}^{\infty} p_j J(x - x_j)$$

此处，函数 $J(\cdot)$ 称为非负判定函数，定义为

$$J(x) = \begin{cases} 0 & \text{当 } x < 0 \\ 1 & \text{当 } x \geq 0 \end{cases} \quad (2.10)$$

定义 2.12 (密度函数). 概率空间 (Ω, \mathcal{S}, P) 上定义的随机变量 X 称为连续的，如果存在非负函数 $f(x)$ 使得 X 的分布函数 $F(x)$ 为

$$F(x) = \int_{-\infty}^x f(t)dt, \text{ 其中 } x \in \mathbb{R} \quad (2.11)$$

其中， $f(x)$ 称为随机变量 X 的概率密度函数 (probability density function, pdf) 或简称密度函数，有时为区别其他随机变量的密度函数，也记作 $f_X(x)$ 。若 $F(x)$ 关于 Lebesgue 测度绝对连续，Randon-Nikodym 定理（见第 69 页的定理 1.3）保证了密度函数的存在。

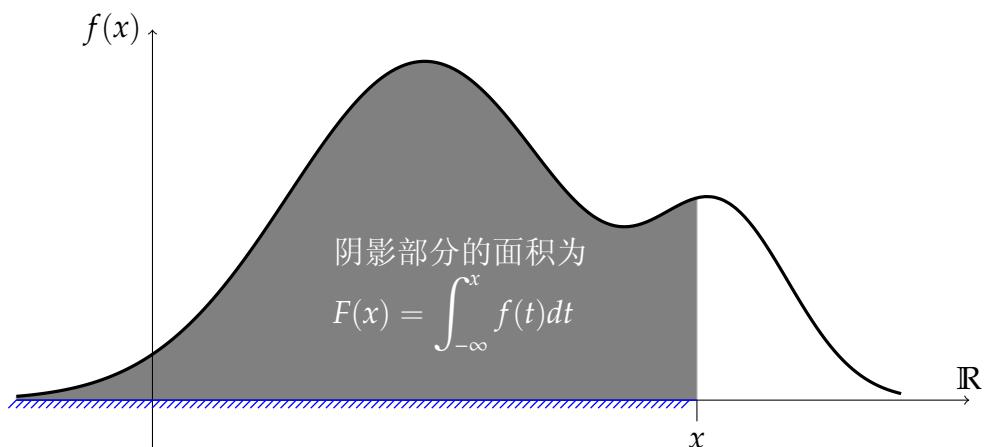


图 2.9: 连续型随机变量 X 的密度函数曲线 $f(x)$ 在上半平面。 X 落于区间 $(-\infty, x]$ 的概率为 $P\{X \in (-\infty, x]\} = F(x) = \int_{-\infty}^x f(t)dt$ ，即图中阴影部分的面积。

若 x 是密度函数 $f(x)$ 的连续点，则 $F'(x) = f(x)$ 。显然，密度函数曲线 $f(x)$ 与 x 轴围成的面积为 1，即

$$\int_{-\infty}^{+\infty} f(x)dx = F(+\infty) - F(-\infty) = 1$$

更一般地，

$$\int_a^b f(x)dx = F(b) - F(a) = P(a < X \leq b)$$

定理 2.5. 连续型随机变量 X 的分布函数 $F(x)$ 和密度函数 $f(x)$ 具有以下关系:

$$F'(x) = f(x) \quad (2.12)$$

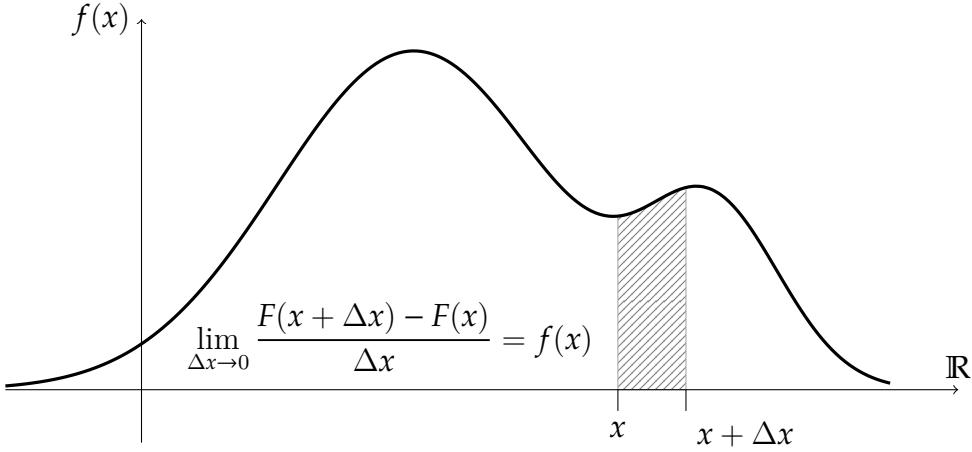


图 2.10: 结果 (2.12) 的直观解释: 阴影部分的面积是 $F(x + \Delta x) - F(x)$, 约等于 $f(x) \cdot \Delta x$, 其中 Δx 很小。

定理 2.6. 已知 $f(x)$ 是定义在 \mathbb{R} 上的非负的、可积的实函数, 则 $f(x)$ 是某个随机变量的密度函数当且仅当

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

证明. “ \Rightarrow ” 是显然的, 因为 $F(+\infty) = 1$ 。下面往证 “ \Leftarrow ”: 令 $F(x) = \int_{-\infty}^x f(t)dt$, 则 $F(x)$ 非减且 $F(-\infty) = 0, F(+\infty) = 1$ 。请读者验证 $F(x)$ 右连续。由定理 2.3, $F(x)$ 是某随机变量的分布函数。□

练习 2.6. 已知密度函数 $f_1(x), \dots, f_n(x)$ 和离散分布 $Y \sim p_1\langle 1 \rangle + \dots + p_n\langle n \rangle$, 试证明: $g(x) = p_1f_1(x) + \dots + p_nf_n(x)$ 也是一个密度函数。

练习 2.7. 若 X 是一个连续型随机变量, 则 $\forall x \in \mathbb{R}$, 皆有 $P(X = x) = 0$ 。

例 2.12 (均匀分布). 如果一个连续型随机变量 X 有如下的分布函数, 则称该随机变量服从 $[a, b]$ 上的均匀分布 (uniform distribution), 记作 $X \sim U[a, b]$ 。

$$F(x) = \begin{cases} 0 & \text{当 } x < a \\ \frac{x-a}{b-a} & \text{当 } a \leq x < b \\ 1 & \text{当 } x \geq b \end{cases}$$

练习 2.8. 验证均匀分布 $X \sim U[a, b]$ 的密度函数是

$$f(x) = \frac{1}{b-a} I_{[a,b]}(x)$$

其中 $I_{[a,b]}(x)$ 是式 (2.4) 定义的集合 $[a, b]$ 的指示函数。

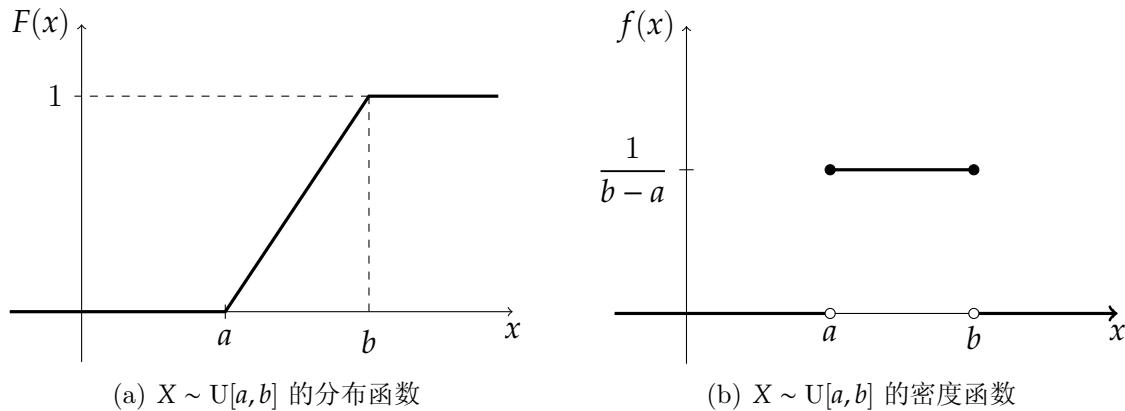


图 2.11: 均匀分布 $X \sim U[a, b]$ 的密度函数 $f(x)$ 和分布函数 $F(x)$ 的图像。

定义 2.13 (正态分布). 若连续型随机变量 X 的密度函数为式 (1.17) 所描述的 $\phi(x|\mu, \sigma^2)$, 则称 X 服从参数为 (μ, σ^2) 的正态分布 (normal distribution), 记作 $X \sim N(\mu, \sigma^2)$, 其分布函数简记作 $\Phi(x|\mu, \sigma^2)$ (见图 1.28 硬币中的增函数)。

$$\Phi(x|\mu, \sigma^2) = \int_{-\infty}^x \phi(z|\mu, \sigma^2) dz$$

特别地, $X \sim N(0, 1)$ 称为标准正态分布 (standard normal distribution), 其密度函数就是式 (1.18), 分布函数简记作 $\Phi(x)$ 。

正态分布也称高斯分布 (Gaussian distribution)^{*} 或 Gauss-Laplace 分布, 在概率论中扮演着十分重要的角色。在概率统计、偏微分方程、信号处理、机器学习中有着广泛的应用的误差函数 $\text{erf}(x)$ 也可用 $\Phi(x)$ 来定义, 即

$$\begin{aligned} \text{erf}(x) &= \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \\ &= 2\Phi(\sqrt{2}x) - 1 \end{aligned}$$



^{*}德国数学家 C. F. Gauss 曾利用函数 $\phi(x|\mu, \sigma^2)$ 分析过天文数据的观测误差, 但他不是首个发现此函数及其重要价值的人。正态分布的历史回顾参见附录 A 和 §4.2.3。

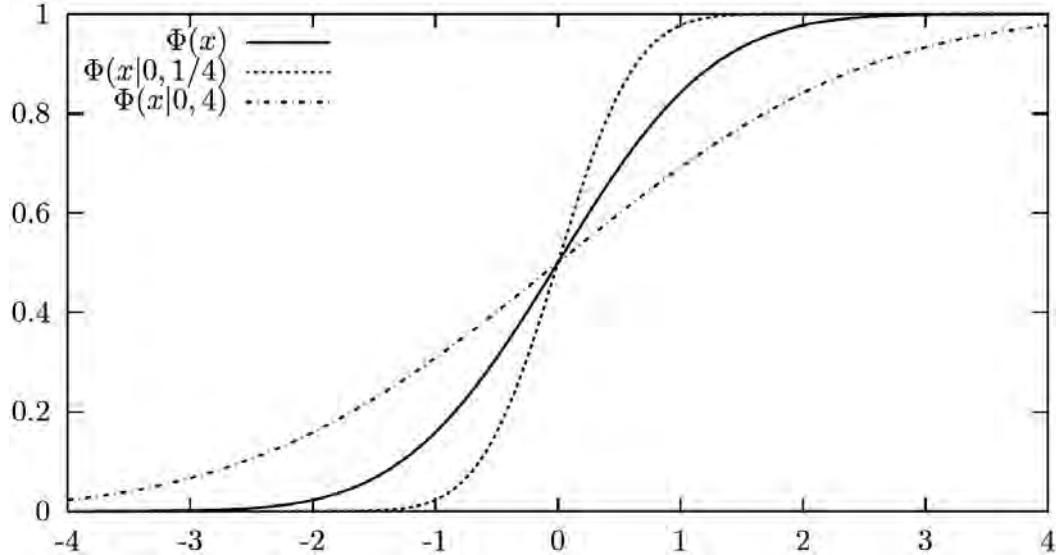


图 2.12: 正态分布 $X \sim N(0, \sigma^2)$ 的分布函数曲线 $\Phi(x|0, \sigma^2)$, 其中实线是 $\Phi(x)$ 。不难发现尺度参数 σ^2 越小, 曲线越“陡”。

性质 2.12. 正态分布 $X \sim N(\mu, \sigma^2)$ 与标准正态分布之间有下面的关系。

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \quad (2.13)$$

其中, 变换 $(X - \mu)/\sigma$ 被称为随机变量 X 的标准化。

证明. 随机变量 Z 的分布函数是

$$P(Z \leq z) = P\left(\frac{X - \mu}{\sigma} \leq z\right) = P(X \leq \sigma z + \mu) = \int_{-\infty}^{\sigma z + \mu} \phi(x|\mu, \sigma^2) dx \stackrel{y = \frac{x - \mu}{\sigma}}{=} \int_{-\infty}^z \phi(y) dy$$

上式右边即是 $Z \sim N(0, 1)$ 的分布函数, 得证。 \square

性质 2.13. 标准化把对 $\Phi(x|\mu, \sigma^2)$ 的计算“转嫁”到 $\Phi(\cdot)$ 上, 即

$$\Phi(x|\mu, \sigma^2) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad (2.14)$$

证明. 利用性质 2.12, 不难得到

$$\Phi(x|\mu, \sigma^2) = P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad \square$$

性质 2.14. 直观上, 分布函数 $\Phi(x)$ 是曲线 $\phi(x)$ 与 $(-\infty, x]$ 所围成的面积。由于 $\phi(x)$

关于 $x = 0$ 对称，所以对任意 $x \in \mathbb{R}$ 皆有

$$\Phi(-x) = 1 - \Phi(x) \quad (2.15)$$

$$P(|X| \leq |x|) = 2\Phi(|x|) - 1, \text{ 其中 } X \sim N(0, 1)$$

函数 $\Phi(x)$ 没有显式表达。过去人们制作了它的数值表，通过查表完成计算，现在这些琐事可以交给计算机来做了。例如，利用 Taylor 级数展开来近似计算 $\Phi(x)$ 。

$$\Phi(x) = \frac{1}{2} + \phi(x) \left\{ x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \cdots + \frac{x^{2n-1}}{(2n-1)!!} + \cdots \right\}$$

其中 $(2n-1)!! = \prod_{j=1}^n (2j-1)$ 称为“双阶乘”

当对 $\Phi(x)$ 的精度要求不高时，也可以使用下面的 Pólya 近似 (1945)，其最大误差不超过 0.00315。

$$\Phi(x) \approx \frac{1}{2} \left\{ 1 + \text{sign}(x) \sqrt{1 - \exp\left(-\frac{2x^2}{\pi}\right)} \right\}, \text{ 其中 sign 是符号函数} \quad (2.16)$$

例 2.13. 令 r 是非负实数，利用性质 2.12 和性质 2.14，可得到随机变量 $X \sim N(\mu, \sigma^2)$ 落于区间 $(\mu - r\sigma, \mu + r\sigma)$ 的概率是

$$P(|X - \mu| \leq r\sigma) = P\left(-r \leq \frac{X - \mu}{\sigma} \leq r\right) = 2\Phi(r) - 1 \quad (2.17)$$

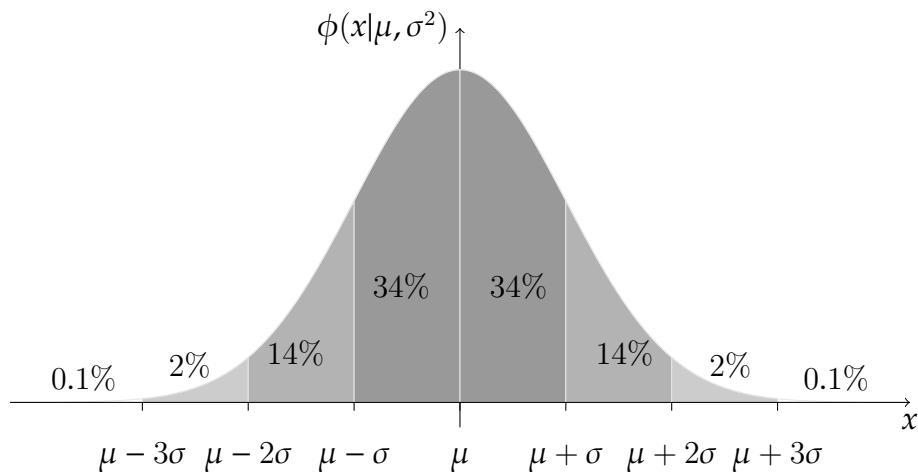


图 2.13: “ 3σ 原则” 可粗略地表述为 $X \sim N(\mu, \sigma^2)$ 以大概率落在 $\mu \pm 3\sigma$ 的范围里。

特别地, 当式 (2.17) 中 $r = 3$ 时, 我们有

$$P(|X - \mu| \leq 3\sigma) = 2\Phi(3) - 1 > 99.7\%$$

上式断言 $X \sim N(\mu, \sigma^2)$ 以接近 1 的概率落于区间 $(\mu - 3\sigma, \mu + 3\sigma)$ 内, 这一事实被称为“ 3σ 原则”, 见图 2.13。而 $X \sim N(\mu, \sigma^2)$ 落于区间 $(\mu - 2\sigma, \mu + 2\sigma)$ 内的概率也不小于 95.4%, 股票分析中常用的 Bollinger 带就是基于这些事实而定的。

※例 2.14. 试证明正态分布函数 $\Phi(x|\mu, \sigma^2)$ 具有下面的性质。

$$\int_{-\infty}^{+\infty} \{\Phi(x|\mu_1, \sigma_1^2) - \Phi(x|\mu_2, \sigma_2^2)\} dx = \mu_2 - \mu_1$$

证明. 利用式 (2.14) 和交换积分次序的方法, 我们有

$$\begin{aligned} \text{左边} &= \int_{-\infty}^{+\infty} \left\{ \int_{\frac{x-\mu_2}{\sigma_2}}^{\frac{x-\mu_1}{\sigma_1}} \phi(y) dy \right\} dx \stackrel{\substack{\text{交换} \\ \text{积分次序}}}{=} \int_{-\infty}^{+\infty} \left\{ \int_{\sigma_1 y + \mu_1}^{\sigma_2 y + \mu_2} \phi(y) dx \right\} dy \\ &= \int_{-\infty}^{+\infty} \phi(y)[(\sigma_2 - \sigma_1)y + \mu_2 - \mu_1] dy = \mu_2 - \mu_1 \end{aligned} \quad \square$$

练习 2.9. 试证明下面的结果, 然后用符号计算工具验证之, 并尝试发现更多的结果。

$$\begin{aligned} \int_{-\infty}^{+\infty} \phi(x|0, \sigma_1^2) \Phi(x|0, \sigma_2^2) dx &= \frac{1}{2} \\ \int_{-\infty}^{+\infty} x^2 \phi(x|0, \sigma_1^2) \Phi(x|0, \sigma_2^2) dx &= \frac{\sigma_1^2}{2} \\ \int_{-\infty}^{+\infty} \phi^2(x|0, \sigma_1^2) \Phi(x|0, \sigma_2^2) dx &= \frac{1}{4\sigma_1 \sqrt{\pi}} \\ \int_{-\infty}^{+\infty} x \phi(x|0, \sigma_1^2) \Phi^2(x|0, \sigma_2^2) dx &= \frac{\sigma_1^2}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \end{aligned}$$

例 2.15. 存在既不是离散型的也不是连续型的随机变量。例如, 假设某地气温 $T \sim N(\mu, \sigma^2)$, 某温度计的最大读数是 t_{\max} , 最小读数是 t_{\min} , 满足 $t_{\min} < \mu < t_{\max}$ 。用该温度计测量该地的温度, 其读数为随机变量

$$X = \begin{cases} t_{\min} & \text{当 } T \leq t_{\min} \\ T & \text{当 } t_{\min} < T < t_{\max} \\ t_{\max} & \text{当 } T \geq t_{\max} \end{cases}$$

按照定义, 随机变量 X 既不是离散型的, 也不是连续型的 (事实上, 它是二者

的混合)。 X 的分布函数为

$$F_X(x) = \begin{cases} 0 & \text{当 } x < t_{\min} \\ \Phi[(x - \mu)/\sigma] & \text{当 } t_{\min} \leq T < t_{\max} \\ 1 & \text{当 } x \geq t_{\max} \end{cases}$$

和纯不连续、绝对连续和奇异连续三种概率测度相对应，有三种类型的分布函数(或者随机变量)：离散型、连续型和奇异型，使得任何分布函数皆是这三种分布函数的加权平均(类似定理 1.2)。

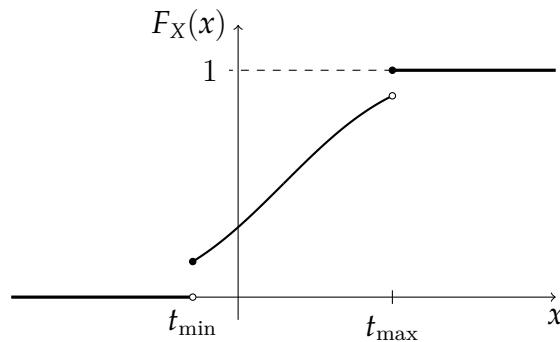


图 2.14: 例 2.15 中随机变量 X 的分布函数 $F_X(x)$ 的示意图。

定理 2.7 (分布函数的表示). 任何分布函数 $F(x)$ 可唯一地分解为三个分布函数的加权平均: $F(x) = \alpha_1 F_1(x) + \alpha_2 F_2(x) + \alpha_3 F_3(x)$, 其中非负实数 $\alpha_1, \alpha_2, \alpha_3$ 满足 $\alpha_1 + \alpha_2 + \alpha_3 = 1$, 并且 $F_1(x)$ 是某连续型随机变量的分布函数; $F_2(x)$ 是某离散型随机变量的分布函数; $F_3(x)$ 是一个奇异型分布(见例 2.11)。

或者更简单地, 任何分布函数 $F(x)$ 都可以分解为绝对连续分布函数 $G(x)$ 和导数几乎处处为零的分布函数 $H(x)$ 的加权平均, 即

$$F(x) = \alpha G(x) + (1 - \alpha)H(x), \text{ 其中 } 0 \leq \alpha \leq 1$$

 因为有定理 2.7 这样的结果, 离散型和连续型的随机变量是两类最基本的随机变量。除非有特殊的声明, 本书只考虑离散型和连续型的随机变量, 当提到随机变量时都缺省地指代这两种类型。另外, 随机变量的“分布”或“概率分布”在特定的语境里有时也指代分布函数、分布列或密度函数, 并无歧义, 读者很容易鉴别它们。

2.1.3 随机变量的函数

已知在概率空间 (Ω, \mathcal{S}, P) 上定义的随机变量 X 具有分布函数 $F_X(x)$, 由 X 可以构造出新的随机变量 $Y = g(X)$, 其中 g 是某一给定的映射。人们很自然地要问什么样的映射 g 能使得 Y 依然是 (Ω, \mathcal{S}, P) 上定义的随机变量?

\nwarrow 定理 2.8. 若 X 是概率空间 (Ω, \mathcal{S}, P) 上定义的随机变量, Borel 可测函数 (见定义 2.2) g 使得 $Y = g(X)$ 依然是 (Ω, \mathcal{S}, P) 上定义的随机变量。

证明. 因为 g 是 Borel 可测函数, 所以 $g^{-1}(-\infty, y] \in \mathcal{B}_1$, 进而 $\{Y \leq y\} = \{g(X) \leq y\} = \{X \in g^{-1}(-\infty, y]\} \in \mathcal{S}$, 由定义 2.3 得证。 \square

在下文中, 所讨论的随机变量的函数都是 (或缺省地假定) 是 Borel 可测函数 (附录 D 简介了可测函数的一些性质)。

性质 2.15. 设离散型随机变量 X 具有如表 2.1.2 所示的分布列, 则离散型随机变量 $Y = g(X)$ 的分布列中 $Y = g(x_j)$ 的概率是这样计算的: 令 x_{j_1}, x_{j_2}, \dots 是所有 $g(x_j)$ 的逆像, 则 $P\{Y = g(x_j)\} = p_{j_1} + p_{j_2} + \dots$ 。

证明. $P\{Y = g(x_j)\} = P(X \in \{x_{j_1}, x_{j_2}, \dots\}) = p_{j_1} + p_{j_2} + \dots$ 。 \square

例 2.16. 从离散型随机变量 X 的分布列得到 $Y = X^2$ 的分布列。

X	-2	-1	0	1	$\xrightarrow{Y=X^2}$	Y	0	1	4
概率	1/4	3/16	1/2	1/16		概率	1/2	1/4	1/4

连续型随机变量 X 的函数 $Y = g(X)$ 的分布函数 $F_Y(y)$ 可由原始定义 $P(Y \leq y) = P(g(X) \leq y)$ 得到, 性质 2.12 的证明已经用过这个方法。在某些限制条件之下, 也可以用下面的方法求得 $F_Y(y)$ 。

\nwarrow 定理 2.9. 已知一一映射 $y = g(x)$ 有连续导数且 $x = h(y)$ 为 $y = g(x)$ 的逆映射, 连续型随机变量 X 的密度函数为 $f_X(x)$, 则随机变量 $Y = g(X)$ 的密度函数为

$$f_Y(y) = |h'(y)| \cdot f_X[h(y)] \quad (2.18)$$

证明. 由已知条件知 g 为单调函数, 不妨设 g 单调不减 (g 单调不增时的证明也是类似的)。对任意的实数 y , 则 Y 的分布函数为

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\{g(X) \leq y\} = P\{X \leq h(y)\} \\ &= \int_{-\infty}^{h(y)} f_X(x) dx = \int_{-\infty}^y h'(y) \cdot f_X[h(y)] dy \end{aligned}$$

由密度函数的定义证得 $f_Y(y) = F'_Y(y) = |h'(y)| \cdot f_X[h(y)]$ 。 \square

例 2.17. 已知随机变量 X 的密度函数为 $f_X(x)$, 分布函数为 $F_X(x)$ 。

□ 线性映射 $Y = aX + b$ 的逆映射为 $X = (Y - b)/a$, 其中 $a \neq 0$, 利用式 (2.18), 则 Y 的密度函数为

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

例如, 令 $\sigma > 0$, 若 $X \sim N(0, 1)$, 则随机变量 $Y = \sigma X + \mu$ 的密度函数是 $\frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right)$, 即 $Y = \sigma X + \mu \sim N(\mu, \sigma^2)$ 。

□ 非线性映射 $Y = X^2$ 不是一一映射, 求 Y 的密度函数不能直接利用定理 2.9。可先求得 Y 的分布函数

$$F_Y(y) = P(Y \leq y) = \begin{cases} 0 & \text{如果 } y \leq 0 \\ P(-\sqrt{y} \leq X \leq \sqrt{y}) & \text{如果 } y > 0 \\ = F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \end{cases}$$

于是, Y 的密度函数为

$$f_Y(y) = F'_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$$

□ 类似地, 随机变量 $Y = |X|$ 的分布函数为

$$F_Y(y) = \begin{cases} 0 & \text{如果 } y \leq 0 \\ F_X(y) - F_X(-y) & \text{如果 } y > 0 \end{cases}$$

进而, 其密度函数为 $f_Y(y) = f_X(y) + f_X(-y)$ 。

例 2.18. 若 $X \sim N(0, 1)$, 试求随机变量 $Y = X^2$ 的密度函数 $f_Y(y)$ 。

解. 根据例 2.17 的结果, 随机变量 $Y = X^2$ 的密度函数是

$$f_Y(y) = \frac{\phi(\sqrt{y}) + \phi(-\sqrt{y})}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}} \exp\left\{-\frac{y}{2}\right\}$$

我们称 Y 服从自由度为 1 的 χ^2 分布, 记作 $Y \sim \chi_1^2$ 。

练习 2.10. 令 $X \sim U(0, 1)$, 请读者验证随机变量 $Y = e^X$ 的密度函数为

$$f_Y(y) = \begin{cases} 1/y & \text{当 } 1 < y < e \\ 0 & \text{其他} \end{cases}$$

例 2.19. 设随机变量 $X \sim U(-\pi/2, \pi/2)$, 求 $Y = \cos X$ 的密度函数 $f_Y(y)$ 。

解. 随机变量 $X \sim U(-\pi/2, \pi/2)$ 的密度函数为

$$p(x) = \begin{cases} 1/\pi & \text{当 } |x| < \pi/2 \\ 0 & \text{当 } |x| \geq \pi/2 \end{cases}$$

随机变量 Y 的分布函数为 $F_Y(y) = P\{\cos X \leq y\}$, 因为 $\cos x$ 是偶函数, 所以当 $0 \leq y \leq 1$ 时,

$$\begin{aligned} F_Y(y) &= P\left\{-\frac{\pi}{2} < X \leq -\arccos y\right\} + P\left\{\arccos y < X \leq \frac{\pi}{2}\right\} \\ &= \int_{-\pi/2}^{-\arccos y} p(x)dx + \int_{\arccos y}^{\pi/2} p(x)dx \\ \text{因此, } f_Y(y) &= F'_Y(y) = \frac{p(-\arccos y)}{\sqrt{1-y^2}} + \frac{p(\arccos y)}{\sqrt{1-y^2}} = \frac{2}{\pi \sqrt{1-y^2}} \end{aligned}$$

2.2 随机向量及其基本性质

描述 这些随机现象经常需要随机向量。譬如，炮弹落点（平面坐标）、身体状况（心率、血压、血糖等）、各科目的学习成绩、股票行情（成交量、开盘价、收盘价等）等。另外，随机向量这一工具也便于我们研究随机变量之间的各种关系。

定义 2.14 (随机向量). 定义在概率空间 (Ω, \mathcal{S}, P) 上的一个 n 维随机向量 \mathbf{X} 就是一个从样本空间 (Ω, \mathcal{S}) 到 n 维 Borel 空间 $(\mathbb{R}^n, \mathfrak{B}_n)$ 的可测函数 $\mathbf{X}: \Omega \rightarrow \mathbb{R}^n$ ，有时也称作 n 维随机变量。显然，随机向量就是取值为向量的随机变量。

令 $X_j = p_j \circ \mathbf{X}$, $j = 1, 2, \dots, n$, 其中 $p_j: \mathbb{R}^n \rightarrow \mathbb{R}$ 是 n 维向量向其第 j 个分量的投射，则 \mathbf{X} 是 Borel 可测函数当且仅当 X_1, X_2, \dots, X_n 都是 Borel 可测函数（见习题 2.1）。

所以，随机向量 \mathbf{X} 可表示为 n 维列向量 $(X_1, X_2, \dots, X_n)^\top$ ，其中 X_1, X_2, \dots, X_n 是定义在同一概率空间 (Ω, \mathcal{S}, P) 上的 n 个随机变量，记作 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 。

不妨把随机向量 \mathbf{X} 理解为映射 $\omega \xrightarrow{\mathbf{X}} (X_1(\omega), X_2(\omega), \dots, X_n(\omega))^\top$ ，使得 $\forall B \in \mathfrak{B}_n$ 皆有 $\mathbf{X}^{-1}(B) \in \mathcal{S}$ 。特别地，对于事件 $\mathbf{X}^{-1}(-\infty, \mathbf{x}] = \bigcap_{j=1}^n \{X_j \leq x_j\} \in \mathcal{S}$ ，我们约定用 $\{\mathbf{X} \in (-\infty, \mathbf{x}]\}$ 或 $\mathbf{X} \in (-\infty, \mathbf{x}]$ 来表示。

定义 2.15. 下面的 n 元函数称为随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的分布函数，有时也称之为随机变量 X_1, \dots, X_n 的联合分布函数。

$$F_{\mathbf{X}}(\mathbf{x}) = P\{\mathbf{X} \in (-\infty, \mathbf{x}]\} = P\left(\bigcap_{j=1}^n \{X_j \leq x_j\}\right) \quad (2.19)$$

定理 2.10. 类似定理 2.8，已知 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 是一个 n 维随机向量，若 $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ 是 Borel 可测函数，则 $g(\mathbf{X})$ 是一个 m 维的随机向量。

定义 2.16 (Kolmogorov 距离). 两个分布函数 $F(\mathbf{x})$ 和 $G(\mathbf{x})$ (其中 $\mathbf{x} \in \mathbb{R}^n$) 之间的 Kolmogorov 距离定义为 $d(F, G) = \sup_{\mathbf{x} \in \mathbb{R}^n} |F(\mathbf{x}) - G(\mathbf{x})|$ ，用来刻画二者的差异。

性质 2.16. 以二维随机向量 $(X, Y)^\top$ 为例，它的分布函数 $F(x, y) = P(X \leq x, Y \leq y)$ 具有以下的性质。

① 对 x 来说， $F(x, y)$ 是非减、右连续的。对 y 来说，亦是如此。

② $F(-\infty, y) = F(x, -\infty) = 0$ 并且 $F(+\infty, +\infty) = 1$ 。

③ 二维随机向量 $(X, Y)^\top$ 落于区域 $D = (x_1, x_2] \times (y_1, y_2]$ 里的概率是

$$P\{(X, Y)^\top \in D\} = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \quad (2.20)$$

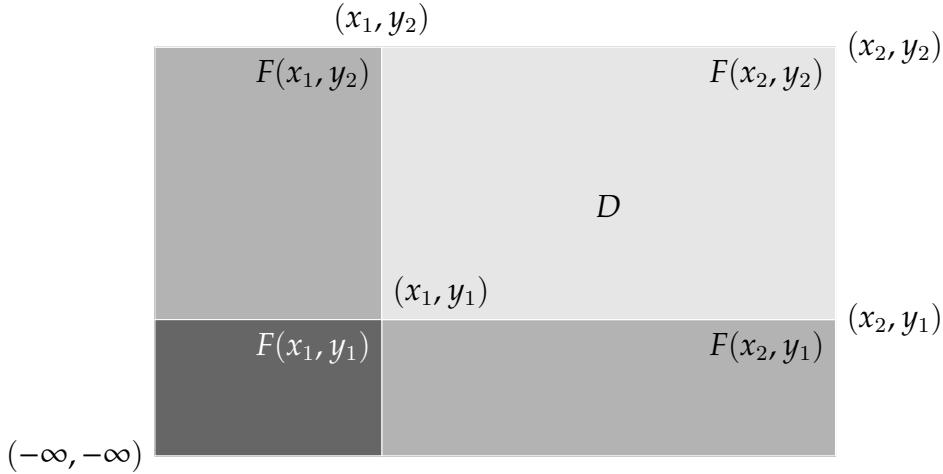


图 2.15: 公式 (2.20) 的直观解释: $P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = P(X \leq x_2, Y \leq y_2) - P(X \leq x_1, Y \leq y_2) - P(X \leq x_2, Y \leq y_1) + P(X \leq x_1, Y \leq y_1)$ 。

\rightsquigarrow 定理 2.11. 若非负函数 $F(x, y)$ 满足性质 2.16 中的 ①②, 则 $F(x, y)$ 是某个二维随机向量的分布函数当且仅当对于任意的 $x_1 < x_2$ 和 $y_1 < y_2$ 皆有

$$F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \geq 0 \quad (2.21)$$

例 2.20. 条件式 (2.21) 是必要的, 下面构造一个二元函数 $F(x, y)$ 满足性质 2.16 中的 ①②, 但不满足条件式 (2.21), $F(x, y)$ 不是分布函数。

$$F(x, y) = \begin{cases} 0 & \text{如果 } x + y < 0 \\ 1 & \text{如果 } x + y \geq 0 \end{cases}$$

如果 $F(x, y)$ 是一个分布函数, 由式 (2.20) 则有 $P(-1 < X \leq 3, -1 < Y \leq 3) = F(3, 3) - F(3, -1) - F(-1, 3) + F(-1, -1) = -1$, 矛盾!

定理 2.12. 类似于性质 2.16, n 维随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的分布函数 $F_{\mathbf{X}}(\mathbf{x})$ 具有下述性质。

- ① 对每个变量 $x_j, j = 1, 2, \dots, n$ 来说, 函数 $F(x_1, \dots, x_j, \dots, x_n)$ 是非减的、右连续的, 并且有变量趋向于 $-\infty$ 时, 函数 F 趋向于 0。

$$\lim_{x_j \rightarrow -\infty} F(x_1, \dots, x_j, \dots, x_n) = 0$$

$$\lim_{x_1 \rightarrow +\infty} F(x_1, \dots, x_j, \dots, x_n) = 1$$

$$\vdots \\ x_n \rightarrow +\infty$$

② 对任意的 $a_j < b_j, j = 1, 2, \dots, n$, 总有

$$\begin{aligned} P\{a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, \dots, a_n < X_n \leq b_n\} &= F(b_1, b_2, \dots, b_n) - \\ &[F(a_1, b_2, \dots, b_n) + F(b_1, a_2, \dots, b_n) + \dots + F(b_1, b_2, \dots, a_n)] + \\ &[F(a_1, a_2, b_3, \dots, b_n) + F(a_1, b_2, a_3, b_4, \dots, b_n) + \\ &\dots + F(b_1, b_2, \dots, b_{n-2}, a_{n-1}, a_n)] + \dots + (-1)^n F(a_1, \dots, a_n) \geq 0 \end{aligned}$$

反之, 类似定理 2.11, 若非负 n 元函数 $F(x_1, x_2, \dots, x_n)$ 满足上述条件 ①②, 则它必是某 n 维随机向量的分布函数。

定义 2.17. 已知随机向量 $(X, Y)^\top$ 具有分布函数 $F(x, y)$, 与随机变量类似, 也可以定义离散型和连续型。

离散型: 如果 $(X, Y)^\top$ 所有可能的取值是至多可数的, 概率函数为 $P(X = x_i, Y = y_j) = p_{ij}$, 它满足

$$\sum_{i,j=1}^{\infty} p_{ij} = 1 \text{ 并且 } F(x, y) = \sum_{\substack{x_i \leq x \\ y_j \leq y}} p_{ij}$$

连续型: 如果存在非负函数 $f(x, y)$ 使得 $\forall (x, y) \in \mathbb{R}^2$ 皆有

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(s, t) dt ds$$

此处 $f(x, y)$ 称为 $(X, Y)^\top$ 的(联合)密度函数, 满足如下性质。

$$\begin{aligned} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dy dx &= F(+\infty, +\infty) = 1 \\ \text{并且 } \frac{\partial^2 F(x, y)}{\partial x \partial y} &= f(x, y) \end{aligned}$$

练习 2.11. 令 n 是某自然数且正实数 p, q 满足 $0 < p + q < 1$, 请验证如下定义的函数

$$P(X = i, Y = j, Z = k) = \frac{n!}{i! j! k!} p^i q^j (1 - p - q)^k$$

其中 i, j, k 都是非负整数且 $n = i + j + k$, 保证了 $(X, Y, Z)^\top$ 为一个离散型随机向量。
提示: $1 = [p + q + (1 - p - q)]^n = \sum_{i,j,k=0}^n P(X = i, Y = j, Z = k)$ 。

例 2.21. 令 $m(\Omega)$ 表示区域 $\Omega \subset \mathbb{R}^2$ 的面积, 我们称连续型随机向量 $(X, Y)^\top$ 服从区

域 Ω 上的均匀分布, 记作 $(X, Y)^\top \sim U(\Omega)$, 如果它的密度函数为

$$f(x, y) = \begin{cases} \frac{1}{m(\Omega)} & \text{如果 } (X, Y)^\top \in \Omega \subset \mathbb{R}^2 \\ 0 & \text{其他} \end{cases}$$

例 2.22 (二元正态分布). 如果随机向量 $(X, Y)^\top$ 的密度函数如下, 则称 $(X, Y)^\top$ 服从参数为 $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的二元正态分布, 记作 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 其中参数 $-1 < \rho < 1$ 是随机变量 X 与 Y 的相关系数, 第 204 页的例 2.84 将给出详细论证。

$$\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{g(x, y)}{2(1-\rho^2)}\right\} \quad (2.22)$$

$$\begin{aligned} \text{其中, } g(x, y) &= \frac{(x - \mu_X)^2}{\sigma_X^2} - \frac{2\rho(x - \mu_X)(y - \mu_Y)}{\sigma_X\sigma_Y} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} \\ &= \left(\frac{y - \mu_Y}{\sigma_Y} - \rho\frac{x - \mu_X}{\sigma_X}\right)^2 + (1 - \rho^2)\frac{(x - \mu_X)^2}{\sigma_X^2} \\ &= \left(\frac{x - \mu_X}{\sigma_X} - \rho\frac{y - \mu_Y}{\sigma_Y}\right)^2 + (1 - \rho^2)\frac{(y - \mu_Y)^2}{\sigma_Y^2} \end{aligned}$$

二元正态分布 $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的密度函数曲面呈钟形, 等高线是椭圆。不难发现, $|\rho|$ 越接近 1 而其他参数不变, 等高线就越“扁”。

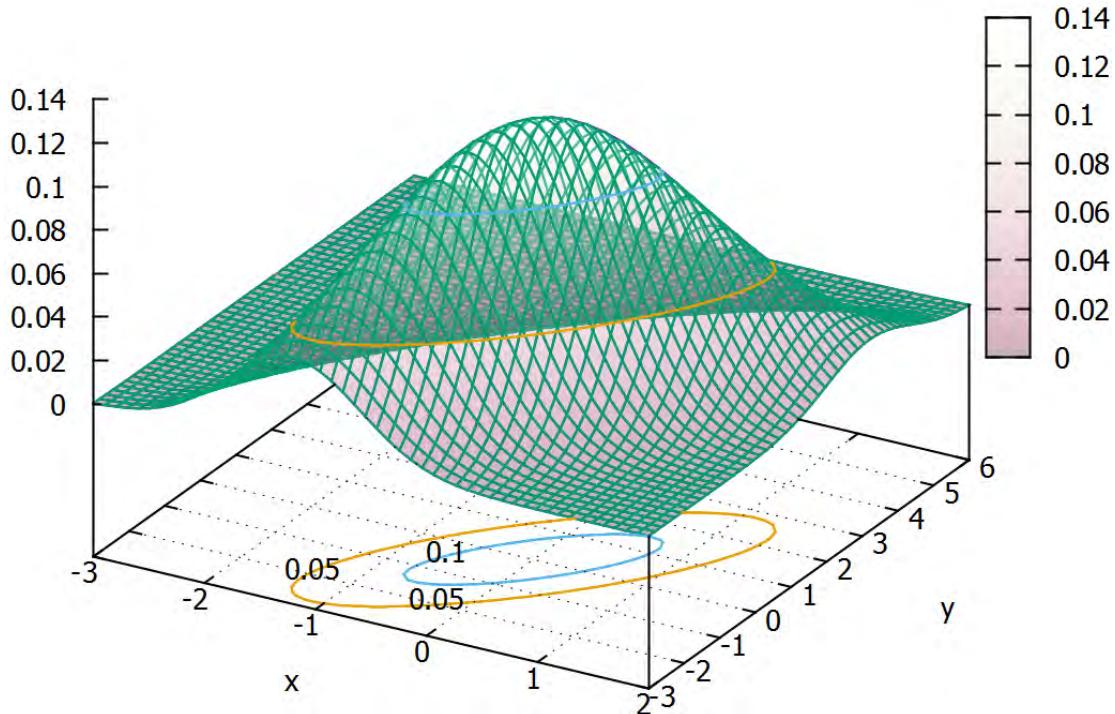


图 2.16: 二元正态分布的密度函数曲面 $z = \phi(x, y | 0, 0, 1, 4, 0.8)$ 及其等高线。

练习 2.12. 请读者验证 (2.22) 是概率密度函数, 即

$$\iint \phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) dx dy = 1$$

二元正态分布的密度函数的表达式 (2.22) 比较复杂, §4.3.4 将介绍 n 元正态分布。作为特款, 二元正态分布有下面的密度函数。

$$\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = \frac{1}{2\pi \sqrt{|\Sigma|}} \exp \left\{ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right\}$$

其中, $\mathbf{z} = \begin{pmatrix} x \\ y \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$, $\Sigma = \begin{pmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}$

$$\Sigma^{-1} = \frac{1}{(1 - \rho^2)\sigma_X^2 \sigma_Y^2} \begin{pmatrix} \sigma_Y^2 & -\rho \sigma_X \sigma_Y \\ -\rho \sigma_X \sigma_Y & \sigma_X^2 \end{pmatrix}$$

练习 2.13. 令 $\rho \neq 0$, 二元正态分布的密度函数曲面 (2.22) 的等高线是椭圆, 请写出其一般方程, 以及长轴所在的直线。答案:

$$\frac{y - \mu_Y}{x - \mu_X} = \tan \theta, \text{ 其中 } \theta = \begin{cases} \frac{1}{2} \arctan \frac{2\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2} & \text{若 } \sigma_X > \sigma_Y \\ \frac{\pi}{2} + \frac{1}{2} \arctan \frac{2\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2} & \text{若 } \sigma_X < \sigma_Y \end{cases}$$

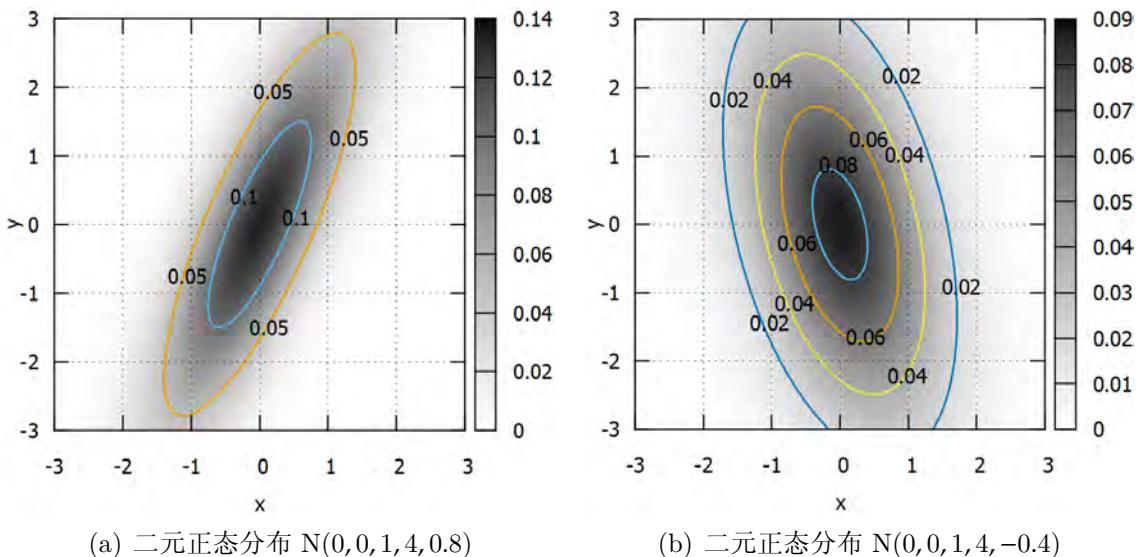


图 2.17: 二元正态分布 $N(0,0,1,4,0.8)$ 和 $N(0,0,1,4,-0.4)$ 密度函数的等高图。通过对比发现 $|\rho|$ 越接近 1, 等高线就显得越“扁”。

性质 2.17. 已知 n 维连续型随机向量 \mathbf{X} 的密度函数为 $f(\mathbf{x})$, 则随机变量 $Z = g(\mathbf{X})$

的分布函数为

$$F_Z(z) = P(Z \leq z) \xrightarrow{A=\{x \in \mathbb{R}^n : g(x) \leq z\}} P(\mathbf{X} \in A) = \int_A f(\mathbf{x}) d\mathbf{x}$$

※例 2.23. 接着例 2.22, 将密度函数 $\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 简记为 $f(x, y)$ 。令 $Z = X + Y$, 求随机变量 Z 的分布。

解. 记 $A = \{(X, Y)^\top \in \mathbb{R}^2 : x + y \leq z\}$, 则 Z 的分布函数 $F_Z(z)$ 为

$$\begin{aligned} F_Z(z) &= \iint_A f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{z-x} f(x, y) dy \right] dx, \text{ 变量替换 } s = x + y \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^z f(x, s-x) ds \right] dx \\ &= \int_{-\infty}^z \left[\int_{-\infty}^{+\infty} f(x, s-x) dx \right] ds \end{aligned}$$

因为 $f_Z(z) = F'_Z(z)$, 于是

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f(x, z-x) dx, \text{ 其中} \\ f(x, z-x) &\propto \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(z-x-\mu_Y)}{\sigma_X \sigma_Y} + \frac{(z-x-\mu_Y)^2}{\sigma_Y^2} \right] \right\} \\ &\propto \exp \left\{ -\frac{(x-\mu_*)^2}{2\sigma_*^2} \right\} \exp \left\{ -\frac{(z-\mu_Z)^2}{2\sigma_Z^2} \right\} \end{aligned}$$

此处, $\mu_Z = \mu_X + \mu_Y$, $\sigma_Z^2 = \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2$, 而 μ_* 是某不含 x 的复杂表达式, σ_*^2 是某不含 x, z 的复杂表达式。有趣的是, 读者不必细究 μ_*, σ_*^2 的具体形式, 因为它们作为某正态分布的密度函数在积分 $\int_{-\infty}^{+\infty} f(x, z-x) dx$ 中被归一了! 所以,

$$\begin{aligned} f_Z(z) &\propto \phi(z | \mu_Z, \sigma_Z^2), \text{ 即} \\ Z &= X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2) \end{aligned}$$

例如, 人类的父代身高 Y 和子代身高 X 的联合分布是正态的 (见第 199 页的图 2.44), 两代的身高之和也是正态的。进而, 两代的平均身高也是正态的。

$$\frac{X+Y}{2} \sim N\left(\frac{\mu_X + \mu_Y}{2}, \frac{\sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2}{4}\right)$$

本节内容

由随机向量的联合分布可以诱导出边缘分布和条件分布，第一小节举例说明边缘分布和条件分布是从“独特视角”看待联合分布。第二、三小节利用边缘分布和条件分布刻画了随机变量之间的一种极端关系——独立性和条件独立性，这种简单关系是很多后续研究的基础。第四小节重点讨论如何由已知的随机向量通过变换构造出新的随机向量或随机变量，并计算它们的分布。此外，还重温了两个函数的卷积，它将用于第3章说明为何要定义特征函数。

关键知识

(1) 随机向量的联合分布、边缘分布和条件分布；(2) 随机变量之间的独立性、条件独立性；(3) 随机向量经过变换后的分布，特别是独立随机变量之和的分布。

2.2.1 边缘分布和条件分布

本小节介绍如何从联合分布构造出两类新的分布——边缘分布和条件分布，它们仅仅承载了联合分布的部分信息。

定义 2.18 (边缘分布). 已知随机变量 X 和 Y 的联合分布，则 X 或 Y 的分布可从该联合分布导出，称为边缘分布 (marginal distribution)。“边缘”一词本来多余，用它是为了强调边缘分布是从联合分布推导出来的。

□ 如果 $(X, Y)^\top$ 是离散型的随机向量，满足 $P(X = x_i, Y = y_j) = p_{ij}$ ，定义 X 的边缘分布的概率函数为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij} = p_{i\cdot}$$

即“矩阵” (p_{ij}) 逐行求和“抹掉”了 Y 的信息而得到的分布列。类似地， Y 的边缘分布定义为 $P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij} = p_{\cdot j}$ ，它是“矩阵” (p_{ij}) 逐列求和，“抹掉”了 X 的信息而得到的分布列。显然，

$$\sum_{i=1}^{\infty} p_{i\cdot} = \sum_{j=1}^{\infty} p_{\cdot j} = \sum_{i,j=1}^{\infty} p_{ij} = 1$$

表 2.2: 常用下面的分布列描述二维离散型随机向量。有时候，也用 $(X, Y)^\top \sim p_{11}\langle x_1, y_1 \rangle + p_{12}\langle x_1, y_2 \rangle + \dots + p_{ij}\langle x_i, y_j \rangle + \dots$ 来表示。

X	Y	y_1	y_2	\cdots	y_j	\cdots	X 的边缘分布
x_1		p_{11}	p_{12}	\cdots	p_{1j}	\cdots	$p_{1\cdot}$
x_2		p_{21}	p_{22}	\cdots	p_{2j}	\cdots	$p_{2\cdot}$
\vdots		\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_j		p_{i1}	p_{i2}	\cdots	p_{ij}	\cdots	$p_{i\cdot}$
\vdots		\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Y 的边缘分布		$p_{\cdot 1}$	$p_{\cdot 2}$	\cdots	$p_{\cdot j}$	\cdots	1

□ 如果 $(X, Y)^\top$ 是连续型的随机向量，密度函数为 $f(x, y)$ ，定义随机变量 X 的边缘分布的密度函数（简称边缘密度） $f_X(x)$ 如下：

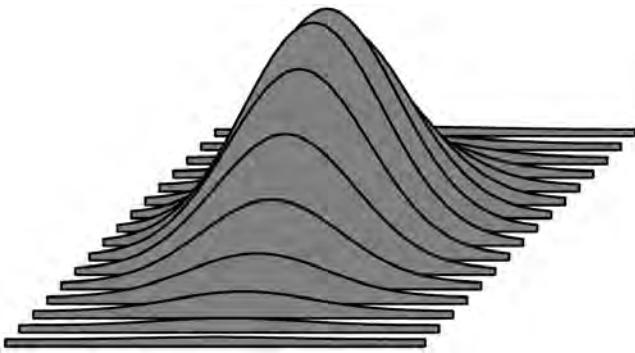
$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad (2.23)$$

边缘密度 $f_X(x)$ 的几何意义见图 2.18。类似地， Y 的边缘分布的密度函数定义

如下，其意义不再赘述。

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

图 2.18: 随机变量 X 的边缘密度 $f_X(x)$ 就是，暂时固定 x ，曲线 $z = f(x, y)$ 与 $z = 0$ 所围成切片的面积。显然，随机变量 Y 的信息被积分 (2.23) “抹掉”了。边缘分布的信息全部来自联合分布，从联合分布求边缘分布就像是从“精细”信息中整合出“粗糙”信息，但反之行不通，即从边缘分布不能重构联合分布。



例 2.24. 假设 $\{1, 2, \dots, 21\}$ 中每个整数被选中的机会等同，考虑所选整数被 2 或 3 整除的概率。令随机变量 X 服从 0-1 分布，表示被 3 整除与否（“1”表示“是”，“0”表示“否”）。类似地，令随机变量 Y 表示被 2 整除与否。随机变量 X 和 Y 的联合分布和边缘分布如下表描述。

表 2.3: 自然数 $1, 2, \dots, 21$ 中被 2 整除和被 3 整除的情况。

$X \ Y$	被 2 整除	不能被 2 整除	X 的边缘分布
被 3 整除	$p_{11} = 3/21$	$p_{12} = 4/21$	$p_{1 \cdot} = 7/21$
不能被 3 整除	$p_{21} = 7/21$	$p_{22} = 7/21$	$p_{2 \cdot} = 14/21$
Y 的边缘分布	$p_{\cdot 1} = 10/21$	$p_{\cdot 2} = 11/21$	1

练习 2.14. 请读者构造一个与例 2.24 不同的联合分布，但边缘分布与例 2.24 中的相同。提示： $(X, Y)^T \sim \frac{4}{21}\langle x_1, y_1 \rangle + \frac{3}{21}\langle x_1, y_2 \rangle + \frac{6}{21}\langle x_2, y_1 \rangle + \frac{8}{21}\langle x_2, y_2 \rangle$ 。

例 2.25. 已知二维随机向量 $(X, Y)^T \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ，则 $X \sim N(\mu_X, \sigma_X^2)$ 且 $Y \sim N(\mu_Y, \sigma_Y^2)$ 。事实上，从式 (2.22) 可得

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} \phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) dy \\ &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{\left(\frac{y-\mu_Y}{\sigma_Y} - \rho\frac{x-\mu_X}{\sigma_X}\right)^2}{2(1-\rho^2)} - \frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} dy \\ &= \phi(x | \mu_X, \sigma_X^2) \int_{-\infty}^{+\infty} \phi(y | \mu_Y, (1-\rho^2)\sigma_Y^2) dy, \text{ 其中 } \mu \text{ 是不含 } y \text{ 的项} \\ &= \phi(x | \mu_X, \sigma_X^2) \end{aligned}$$

该结果表明，二元正态分布的边缘分布依然是正态分布。由 $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ 无法得到 $(X, Y)^\top$ 的分布，因为参数 ρ 的信息丢失了。

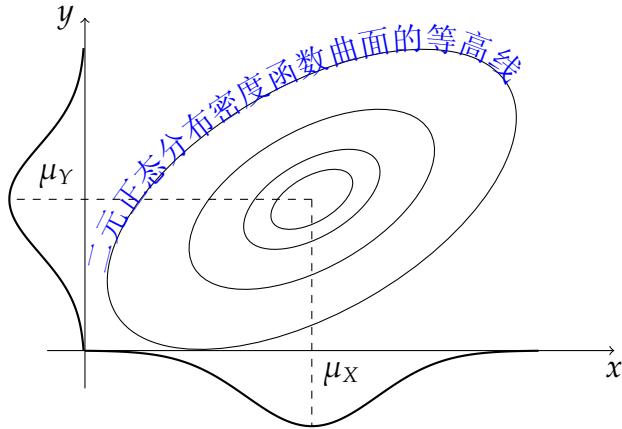


图 2.19: 例 2.25 的直观图示：二元正态分布的边缘分布依然是正态分布。

由联合分布和边缘分布，还能构造出一类新的分布——条件分布。条件分布函数也是分布函数，定语“条件”无非强调它们是由联合分布函数和边缘分布函数诱导出来的。下面，分离散型和连续型两种情形分别定义它。

定义 2.19 (条件概率函数). 已知离散型随机向量 $(X, Y)^\top$ 的概率函数 $P(X = x_i, Y = y_j) = p_{ij}$ ，在给定 $X = x_i$ 的条件下 $Y = y_j$ 的概率为

$$P(Y = y_j | X = x_i) = \frac{p_{ij}}{p_{\cdot i}} \quad (2.24)$$

我们称 (2.24) 为“在给定 $X = x_i$ 的条件下 Y 的概率函数（或分布列）”，简称 $Y|X = x_i$ 的条件概率函数（或条件分布列）。

类似地，可定义 $X|Y = y_j$ 的条件概率函数（即在给定 $Y = y_j$ 的条件下 X 的概率函数）为

$$P(X = x_i | Y = y_j) = \frac{p_{ij}}{p_{\cdot j}}$$

性质 2.18. 对于离散型随机向量 $(X, Y)^\top$ ，由定义 2.19，显然有

$$\begin{aligned} \sum_{j=1}^{\infty} P(Y = y_j | X = x_i) &= 1, \text{ 并且} \\ \sum_{i=1}^{\infty} P(Y = y_j | X = x_i) p_{\cdot i} &= p_{\cdot j} \end{aligned} \quad (2.25)$$

读者不难看出，上式就是古典概率的全概率公式 $P(B) = \sum_{j=1}^{\infty} P(A_j)P(B|A_j)$ 的“离散型随机变量版”，即

$$P(Y = y_j) = \sum_{i=1}^{\infty} P(X = x_i)P(Y = y_j|X = x_i)$$

例 2.26. 接着**例 2.24**， $X|Y = 0$ 和 $X|Y = 1$ 的条件分布列分别为

$$\begin{aligned} P(X = 1|Y = 1) &= \frac{p_{11}}{p_{\cdot 1}} = \frac{3}{10}, & P(X = 0|Y = 1) &= \frac{p_{21}}{p_{\cdot 1}} = \frac{7}{10} \\ P(X = 1|Y = 0) &= \frac{p_{12}}{p_{\cdot 2}} = \frac{4}{11}, & P(X = 0|Y = 0) &= \frac{p_{22}}{p_{\cdot 2}} = \frac{7}{11} \end{aligned}$$

$P(X = 1|Y = 1) = 3/10$ 的含义是，从 $\{1, 2, \dots, 21\}$ 里随机地选取一个整数，已知它能被 2 整除，则它也能被 3 整除的概率是 $3/10$ 。

定义 2.20 (条件分布函数). 已知连续型的随机向量 $(X, Y)^T$ 的密度函数为 $f(x, y)$ ，在给定 $X = x$ 的条件下 Y 的分布函数定义为

$$F_{Y|X}(y|x) = \lim_{\Delta x \rightarrow 0} P(Y \leq y | x < X \leq x + \Delta x) \quad (2.26)$$

该分布函数通常简称为 $Y|X = x$ 的条件分布函数 (conditional distribution)。继续简化式 (2.26)，不难得得到 $F_{Y|X}(y|x)$ 的表达式。

$$\begin{aligned} F_{Y|X}(y|x) &= \lim_{\Delta x \rightarrow 0} \frac{P(Y \leq y, x < X \leq x + \Delta x)}{P(x < X \leq x + \Delta x)} \\ &= \lim_{\Delta x \rightarrow 0} \frac{\frac{1}{\Delta x} \int_x^{x+\Delta x} \int_{-\infty}^y f(s, t) dt ds}{\frac{1}{\Delta x} \int_x^{x+\Delta x} \int_{-\infty}^{+\infty} f(s, y) dy ds} \\ &= \frac{\int_{-\infty}^y f(x, t) dt}{\int_{-\infty}^{+\infty} f(x, y) dy} \\ &= \frac{\int_{-\infty}^y f(x, t) dt}{f_X(x)} \end{aligned} \quad (2.27)$$

式 (2.27) 两边对 y 求导进而得到 $Y|X = x$ 的条件密度函数

$$f_{Y|X}(y|x) = \frac{f(x, y)}{\int_{-\infty}^{+\infty} f(x, y) dy} = \frac{f(x, y)}{f_X(x)} \quad (2.28)$$

在不引起混淆的前提下， $F_{Y|X}(y|x)$ 和 $f_{Y|X}(y|x)$ 通常简记作 $F(y|x)$ 和 $f(y|x)$ 。类

似地, 可定义 $X|Y = y$ 的条件分布函数 $F_{X|Y}(x|y)$, 并得出条件密度函数的表达式

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

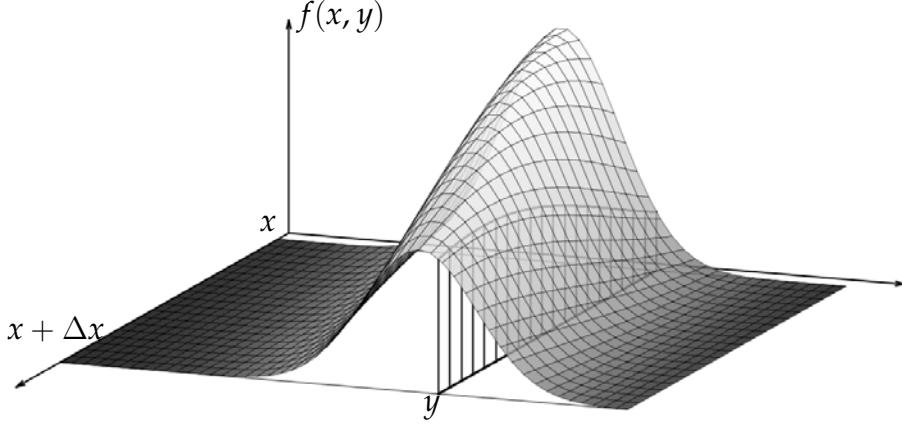


图 2.20: 条件分布函数 $F_{Y|X}(y|x)$ 的含义是随机向量 $(X, Y)^\top$ 落于 xoy 平面上的子区域 $(x, x + \Delta x] \times (-\infty, y]$ 的概率和落于 $(x, x + \Delta x] \times \mathbb{R}$ 的概率之比, 即曲面 $z = f(x, y)$ 在这两个区域上所围成的体积之比, 在 $\Delta x \rightarrow 0$ 时的极限。

\curvearrowleft **性质 2.19 (全概率公式).** 从式 (2.27) 和式 (2.28), 可直接得到全概率公式 (1.32) 的“连续型随机变量版”如下,

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{+\infty} f_X(x)F(y|x)dx \\ f_Y(y) &= \int_{-\infty}^{+\infty} f_X(x)f(y|x)dx \end{aligned}$$

例 2.27. 已知随机向量 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 试证明:

$$Y|X = x \sim N\left(\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right) \quad (2.29)$$

证明. 由第 141 页的**例 2.25** 知, 边缘分布 $X \sim N(\mu_X, \sigma_X^2)$ 。根据式 (2.22) 求得 $Y|X = x$ 的条件密度函数 $f(y|x)$ 如下。

$$\begin{aligned} f(y|x) &= \frac{\phi(x, y|\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)}{\phi(x|\mu_X, \sigma_X^2)} \\ &\propto \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_Y}{\sigma_Y} - \rho \frac{x-\mu_X}{\sigma_X}\right)^2\right\} \\ &\propto \phi\left(y \left| \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right.\right) \end{aligned} \quad \square$$

例如, 根据**例 2.27** 的结果, 从 $(X, Y)^\top \sim N(0, 0, 1, 4, 0.8)$ 得到条件密度函数

$f(y|x) = \phi(y|1.6x, 1.44)$ 。即，对于每个固定的 x , $f(y|x)$ 都是一个正态分布，均值是 $1.6x$ ，方差是 1.44。该条件分布的几何意义见下图所示的 $f(y|x)$ 的曲面。

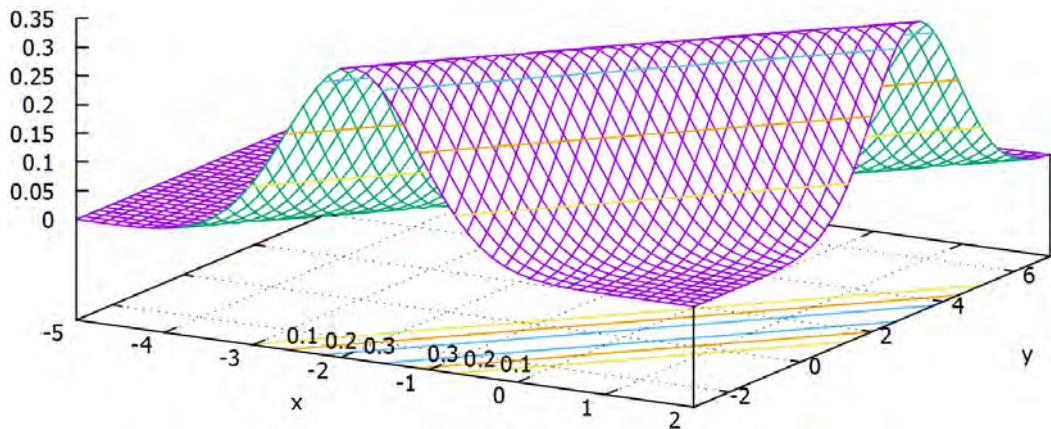


图 2.21: 条件密度函数 $f(y|x) = \phi(y|1.6x, 1.44)$ 的曲面。用 $x = x_0$ 平面截曲面 $f(y|x)$, 剖面曲线为 $\phi(y|1.6x_0, 1.44)$ 。用 $y = y_0$ 能截出什么曲线?

2.2.2 随机变量间的独立性

§1.3.3 和 §1.3.4 曾讨论过随机事件间的独立性和条件独立性，基于此，本节将定义随机变量间的独立性与条件独立性。

定义 2.21 (独立性). 定义在同一个概率空间 (Ω, \mathcal{S}, P) 上的随机变量 X_1, X_2, \dots, X_n 是（相互）独立的 (independent)，记作 $\perp\!\!\!\perp \{X_1, X_2, \dots, X_n\}$ ，当且仅当对于任意 Borel 集 $B_1, B_2, \dots, B_n \in \mathfrak{B}_1$ 皆有

$$P\left(\bigcap_{j=1}^n \{X_j \in B_j\}\right) = \prod_{j=1}^n P\{X_j \in B_j\} \quad (2.30)$$

练习 2.15. 如果 $\perp\!\!\!\perp \{X_1, X_2, \dots, X_n\}$ ，则其中一部分随机变量 $X_{k_1}, X_{k_2}, \dots, X_{k_m}$ 也是独立的。特别地，任意两个随机变量之间是相互独立的。

提示：令式 (2.30) 中某些 $B_j = \mathbb{R}$ 。

性质 2.20. 随机变量 X_1, X_2, \dots, X_n 是独立的当且仅当 X_1, X_2, \dots, X_n 的联合分布函数 $F(x_1, x_2, \dots, x_n)$ 能分解为边缘分布函数之积，即

$$F(x_1, x_2, \dots, x_n) = \prod_{j=1}^n F_{X_j}(x_j), \text{ 其中 } F_{X_j}(x_j) \text{ 是 } X_j \text{ 的分布函数} \quad (2.31)$$

证明. 往证 “ \Rightarrow ”：取 $B_j = (-\infty, x_j]$ ，由式 (2.30) 轻易证得结果 (2.31)。“ \Leftarrow ” 的证明本书不作要求，感兴趣的读者可参阅 Loève 的《概率论》[107] 第五章第十六节。□

 正是因为性质 2.20 揭示了式 (2.30) 与式 (2.31) 的等价性，在很多教科书中，式 (2.31) 经常作为随机变量间独立性的定义。

性质 2.21. 若 X, Y 是相互独立的离散型随机变量，则

$$\begin{aligned} p_{ij} &= P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) = p_i p_{\cdot j} \\ p_{\cdot i} &= P(X = x_i | Y = y_j) \text{ 且 } P(Y = y_j | X = x_i) = p_{\cdot j} \end{aligned}$$

若 X, Y 是相互独立的连续型随机变量，则

- 联合密度与边缘密度： $f(x, y) = f_X(x)f_Y(y)$ 且反之亦然。
- 条件分布与边缘分布： $F(y|x) = F_Y(y)$ 且 $F(x|y) = F_X(x)$ 。

例 2.28. 随机变量 X, Y 之间的独立性是蕴含在联合分布函数中的，即便 X, Y 有相同的分布，它们之间也有可能不是独立的。例如，

$X \ Y$	0	1	X 的边缘分布
0	1/21	4/21	5/21
1	4/21	12/21	16/21
Y 的边缘分布	5/21	16/21	1

定义 2.22. 如果独立的随机变量 X_1, X_2, \dots, X_n 服从相同的分布 F , 则称 X_1, X_2, \dots, X_n 独立同分布 (independent and identically distributed, iid) 于 F , 记作 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} F$ 。此时, X_1, X_2, \dots, X_n 的联合分布是

$$F(x_1, x_2, \dots, x_n) = \prod_{j=1}^n F(x_j)$$

譬如, $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ 的联合分布是 $\prod_{j=1}^n \phi(x_j | \mu, \sigma^2)$ 。

定义 2.23. 随机变量间独立性的定义可自然推广到随机向量上, 譬如 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 与 $\mathbf{Y} = (Y_1, \dots, Y_m)^\top$ 相互独立当且仅当对任意的 Borel 集 $B_1 \in \mathfrak{B}_n, B_2 \in \mathfrak{B}_m$ 皆有

$$P(\mathbf{X} \in B_1, \mathbf{Y} \in B_2) = P(\mathbf{X} \in B_1)P(\mathbf{Y} \in B_2)$$

请读者注意, \mathbf{X} 与 \mathbf{Y} 相互独立并不能推出 X_1, \dots, X_n 是独立的。

定理 2.13. 令 F_X 和 F_Y 分别是 \mathbf{X} 和 \mathbf{Y} 的分布函数。与定义 2.23 等价的定义是随机向量 $\mathbf{Z} = (X_1, \dots, X_n, Y_1, \dots, Y_m)^\top$ 的分布函数有如下的分解,

$$F_Z(x_1, \dots, x_n, y_1, \dots, y_m) = F_X(x_1, \dots, x_n)F_Y(y_1, \dots, y_m)$$

练习 2.16. 随机变量 $X \sim N(\mu_X, \sigma_X^2)$ 和 $Y \sim N(\mu_Y, \sigma_Y^2)$ 相互独立当且仅当随机向量 $(X, Y)^\top$ 的密度函数为 $\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, 0)$ 。

定理 2.14. 若随机向量 \mathbf{X}, \mathbf{Y} 相互独立, 令 g_1, g_2 是 Borel 可测函数, 则随机向量 $g_1(\mathbf{X})$ 与 $g_2(\mathbf{Y})$ 也相互独立。

证明. 由定理 2.10 知 $g_1(\mathbf{X}), g_2(\mathbf{Y})$ 也是随机向量, 根据定义 2.23 有

$$\begin{aligned} P\{g_1(\mathbf{X}) \leq \mathbf{x}, g_2(\mathbf{Y}) \leq \mathbf{y}\} &= P\{\mathbf{X} \in g_1^{-1}(-\infty, \mathbf{x}], \mathbf{Y} \in g_2^{-1}(-\infty, \mathbf{y]}\} \\ &= P\{\mathbf{X} \in g_1^{-1}(-\infty, \mathbf{x}]\}P\{\mathbf{Y} \in g_2^{-1}(-\infty, \mathbf{y]}\} \\ &= P\{g_1(\mathbf{X}) \leq \mathbf{x}\}P\{g_2(\mathbf{Y}) \leq \mathbf{y}\} \end{aligned} \quad \square$$

练习 2.17. 若随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 与 $\mathbf{Y} = (Y_1, \dots, Y_m)^\top$ 相互独立, 则 $(X_{n_1}, \dots, X_{n_t})^\top$ 与 $(Y_{m_1}, \dots, Y_{m_s})^\top$ 也相互独立, 其中指标 $\{n_1, \dots, n_t\}$ 是 $\{1, \dots, n\}$ 的非空子集, 指标 $\{m_1, \dots, m_s\}$ 是 $\{1, \dots, m\}$ 的非空子集。

例 2.29. 已知 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 现代统计学奠基人之一、英国著名数学家 Ronald Aylmer Fisher (1890-1962) 证得了下面的结果。

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j \perp\!\!\!\perp (X_1 - \bar{X}, \dots, X_n - \bar{X})^\top$$

该结果被称为 Fisher 定理, 即本书第 245 页的定理 3.16 (它的证明需要用到特征函数这一工具, 暂且不表)。利用定理 2.14, 由 Fisher 定理立即可证得下面的结论, 即正态总体的样本均值和样本方差相互独立。

$$\bar{X} \perp\!\!\!\perp S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2$$

Fisher 定理的证明见第 3 章, 而第 7 章将具体论证随机变量 \bar{X} 和 S^2 服从的分布, 这些都是后话。

例 2.30. 随机变量 X_1, X_2, \dots, X_n 中任意两个随机变量都是独立的 (简称两两独立, pairwise independent) 推不出 X_1, X_2, \dots, X_n 是独立的。

例如, 已知 $X_1, X_2 \stackrel{\text{iid}}{\sim} \frac{1}{2}\langle -1 \rangle + \frac{1}{2}\langle 1 \rangle$, 则 $X_1, X_2, X_1 X_2$ 两两独立但 $X_1, X_2, X_1 X_2$ 并不是相互独立的 (留给读者验证)。

定义 2.24 (可交换性). 随机变量间的可交换性是一个比独立性稍弱一些的概念。随机变量 X_1, X_2, \dots, X_n 称为可交换的 (exchangable), 如果对于 $1, 2, \dots, n$ 上的任意置换 σ 皆有 X_1, X_2, \dots, X_n 的联合分布等同于 $X_{\sigma(1)}, X_{\sigma(2)}, \dots, X_{\sigma(n)}$ 的联合分布, 即

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = F_{X_{\sigma(1)}, X_{\sigma(2)}, \dots, X_{\sigma(n)}}(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)})$$

由定义容易看出, 函数 $F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ 关于变量是对称的。

显然, 如果随机变量 X_1, X_2, \dots, X_n 是独立同分布的, 则它们一定是可交换的。但反之不成立, 见下面的例子。

例 2.31. 如果随机向量 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 则随机变量 X, Y 是可交换的, 但当 $\rho \neq 0$ 时, 它们是不独立的。

2.2.3 条件独立性

随机变量间的独立性可自然推广至条件独立性，二者本质上是一样的，定义如下。条件独立性也有类似性质 2.21 的结果，不再赘述。

定义 2.25 (条件独立性). 在给定 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)^\top$ 的条件下，随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的分布函数 $F_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ 若有如下的分解，其中 $\mathbf{x} = (X_1, X_2, \dots, X_n)^\top$ ，

$$F_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^n F_{X_i|\mathbf{Y}}(x_i|\mathbf{y})$$

其中 $F_{X_i|\mathbf{Y}}(x_i|\mathbf{y})$ 是 $X_i|\mathbf{Y}$ 的条件分布函数，则称 X_1, X_2, \dots, X_n 关于 \mathbf{Y} 条件独立 (conditionally independent)，记作 $\perp\!\!\!\perp_{\mathbf{Y}} \{X_1, X_2, \dots, X_n\}$ ，或者 $\perp\!\!\!\perp \{X_1, X_2, \dots, X_n\}|\mathbf{Y}$ 。

例 2.32. 在自然语言处理中，一个文档被抽象为词的序列 w_1, w_2, \dots, w_n 。一元生成模型假设文档是这样产生的：首先按照主题（譬如，体育、政治、经济、学术、娱乐等）的分布 $p(z)$ 随机地产生一个主题 z ，然后按照该主题之下词的条件分布 $p(w|z)$ 独立地产生 n 个词。进而，我们得到这个文档的产生概率是

$$p(w_1, w_2, \dots, w_n) = \sum_z p(z) \prod_{j=1}^n p(w_j|z)$$

简而言之，一元生成模型假设在给定主题的条件下，词的产生是条件独立的。显然，该模型不考虑词的出现次序。

例 2.33 (独立性与条件独立性的关系). 已知随机变量 X, Y 关于 Z 条件独立，能否推导出 $X \perp\!\!\!\perp Y$? 反之，由 $X \perp\!\!\!\perp Y$ 能否推导出 $X \perp\!\!\!\perp_Z Y$?

解. 两个回答都是“不能”，构造反例如下。

□ 令随机变量 $Z \sim 0.5\langle 1 \rangle + 0.5\langle 0 \rangle$ ，并且假设 $X|Z$ 和 $Y|Z$ 的条件分布满足

$$\begin{aligned} X|Z = 1, Y|Z = 1 &\stackrel{\text{iid}}{\sim} 0.9\langle 1 \rangle + 0.1\langle 0 \rangle \\ X|Z = 0, Y|Z = 0 &\stackrel{\text{iid}}{\sim} 0.1\langle 1 \rangle + 0.9\langle 0 \rangle \end{aligned}$$

不难算得 $P(X = 0) = 0.5 < P(X = 0|Y = 0) = 0.82$ 。即， X, Y 不独立。

□ 设 $X, Y \stackrel{\text{iid}}{\sim} 0.9\langle 1 \rangle + 0.1\langle 0 \rangle$ 且 $Z = X+Y$ ，则 $P(X = 0|Z = 1) = P(Y = 0|Z = 1) = \frac{1}{2}$ ，
 $P(X = 0, Y = 0|Z = 1) = 0$ 。即， $X \perp\!\!\!\perp_Z Y$ 不成立。

性质 2.22. 已知随机变量 X, Y, Z, W , 条件独立性具有以下常见的性质。

$$X \perp\!\!\!\perp (Y, Z)^\top \Leftrightarrow X \perp\!\!\!\perp Z \text{ 且 } X \perp\!\!\!\perp_Z Y \quad (2.32)$$

$$X \perp\!\!\!\perp_W (Y, Z)^\top \Leftrightarrow X \perp\!\!\!\perp_W Z \text{ 且 } X \perp\!\!\!\perp_{W,Z} Y \quad (2.33)$$

证明. 往证式 (2.32) 的 “ \Rightarrow ”: 由 $X \perp\!\!\!\perp (Y, Z)^\top$ 可得 $X \perp\!\!\!\perp Z$, 并且

$$\begin{aligned} \pi(x, y, z) &= \pi_X(x)\pi(y, z) = \pi_X(x)\pi_Z(z)\pi(y|z) \\ \text{于是, } \pi(x, y|z) &= \frac{\pi(x, y, z)}{\pi_Z(z)} = \pi_X(x)\pi(y|z) = \pi(x|z)\pi(y|z) \end{aligned}$$

往证式 (2.32) 的 “ \Leftarrow ”: 由已知条件 $\pi(x, y|z) = \pi_X(x)\pi(y|z)$, 易得

$$\pi(x, y, z) = \pi_Z(z)\pi(x, y|z) = \pi_X(x)\pi_Z(z)\pi(y|z) = \pi_X(x)\pi(y, z)$$

结果 (2.33) 是结果 (2.32) 的推论, 请读者补全证明。 \square

为什么需要条件独立性这个概念? 一个重要的原因是简化模型。譬如, 若 $X_2 \perp\!\!\!\perp_{X_1} X_3$, 则 X_1, X_2, X_3 的联合密度函数为

$$\begin{aligned} \pi(x_1, x_2, x_3) &= \pi_1(x_1)\pi_2(x_2|x_1)\pi_3(x_3|x_1, x_2) \\ &= \pi_1(x_1)\pi_2(x_2|x_1)\pi_3(x_3|x_1) \end{aligned}$$

一般地, 假设随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的密度函数为 $\pi(\mathbf{x})$, 其中 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 。条件独立性有助于简化 $\pi(\mathbf{x})$ 的分解, 从而简化模型和概率计算。一旦给定了 $\pi(\mathbf{x})$ 的某个分解, 我们就可以利用以 X_1, X_2, \dots, X_n 为节点的有向无圈图 (directed acyclic graph, DAG) 来形象地描述该分解, 反之亦然。另外, 根据该图的某些结构特点, 可以直观地看出随机变量之间的条件独立性或者独立性。

定义 2.26. 我们把条件分布 $\pi_i(x_i|x_{i_1}, \dots, x_{i_k})$ 中的每个条件 x_{i_1}, \dots, x_{i_k} 称为 x_i 的父辈节点 (parent), 把集合 $\{x_{i_1}, \dots, x_{i_k}\}$ 简记作 $\text{pa}(x_i)$ 。在不引起歧义的时候, 也称 X_{i_1}, \dots, X_{i_k} 为 X_i 的父辈节点。没有父辈节点的节点称为根节点。

描述联合密度函数的有向无圈图被称为贝叶斯网络 (Bayesian network), 或者信念网络 (belief network), 它是一类概率图模型 (probabilistic graphical model)。

贝叶斯网络允许有多个根节点, 非根节点可以有一个或多个父辈节点。例如, 图 2.22 中, X_4 的父辈节点是 X_1, X_3 , 而 X_1 又是 X_3 的父辈节点。图 2.23 中的 (d), X_1, X_2 是根节点, $\text{pa}(X_3) = \{X_1, X_2\}$ 。

性质 2.23. 联合密度函数 $\pi(\mathbf{x})$ 的任意分解方式都能抽象地表示为

$$\pi(\mathbf{x}) = \prod_{i=1}^n \pi_i(x_i|\text{pa}(x_i)) \quad (2.34)$$

给定了贝叶斯网络的结构及式 (2.34) 中每个 $\pi_i(x_i|\text{pa}(x_i))$ 的信息，对于任何有关这几个变量的条件概率问题，有多种算法可实现数值计算 [92]。然而，如何自动获得贝叶斯网络的结构仍然是机器学习领域一个悬而未决的难题。

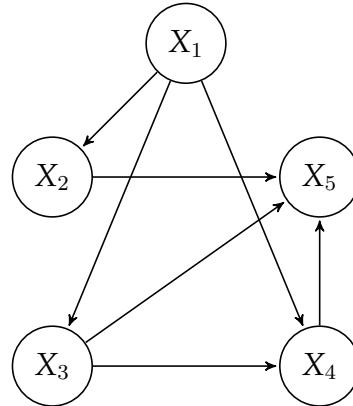


图 2.22: X_1, \dots, X_5 的联合分布: $\pi_1(x_1)\pi_2(x_2|x_1)\pi_3(x_3|x_1)\pi_4(x_4|x_1, x_3)\pi(x_5|x_2, x_3, x_4)$ 。

不计较变量名，三个节点的连通有向无圈图的拓扑结构不外乎图 2.23 所示的四种情况。除了 (a)，其他三种情况都蕴含条件独立性或者独立性。

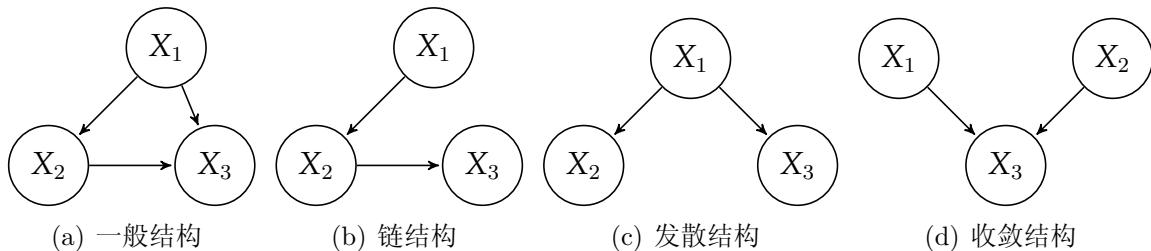


图 2.23: 结构 (b)、(c)、(d) 是 (a) 删除一条边后得到的，蕴含着（条件）独立性。

- (a) 对应着分解 $\pi(x_1, x_2, x_3) = \pi_1(x_1)\pi_2(x_2|x_1)\pi_3(x_3|x_2, x_1)$ 。由该分解不能推导出随机变量 X_1, X_2, X_3 之间有独立性或者条件独立性，但并不意味着没有。
- (b) 对应着分解 $\pi(x_1, x_2, x_3) = \pi_1(x_1)\pi_2(x_2|x_1)\pi_3(x_3|x_2)$ ，进而 $X_1 \perp\!\!\!\perp_{X_2} X_3$ ，因为

$$\pi(x_1, x_3|x_2) = \frac{\pi(x_1, x_2, x_3)}{\pi(x_2)} = \frac{\pi(x_1, x_2)\pi_3(x_3|x_2)}{\pi(x_2)} = \pi(x_1|x_2)\pi(x_3|x_2)$$

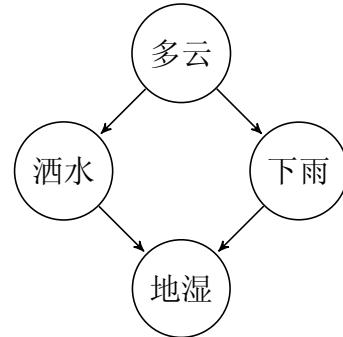
或者, 由 $\pi_3(x_3|x_2, x_1) = \pi_3(x_3|x_2)$ 直接得到 $X_1 \perp\!\!\!\perp_{X_2} X_3$ 。

- (c) 对应着 $\pi(x_1, x_2, x_3) = \pi_1(x_1)\pi_2(x_2|x_1)\pi_3(x_3|x_1)$, 进而 $X_2 \perp\!\!\!\perp_{X_1} X_3$ (留作练习)。
 (d) 对应着 $\pi(x_1, x_2, x_3) = \pi_1(x_1)\pi_2(x_2)\pi_3(x_3|x_1, x_2)$, 显然 $X_1 \perp\!\!\!\perp X_2$, 但一般没有 $X_1 \perp\!\!\!\perp_{X_3} X_2$ 。

 贝叶斯网络描绘的并非变量间的因果关系, 而是条件独立性。譬如, $X_1 \rightarrow X_2 \rightarrow X_3$ 和 $X_1 \leftarrow X_2 \leftarrow X_3$ 是等价的, 描述的都是 $X_1 \perp\!\!\!\perp_{X_2} X_3$ 。另外, 从给定的 DAG 推知两个节点的条件独立性, 但不能判定二者间的非独立性。譬如, 图 2.22 蕴含 $X_2 \perp\!\!\!\perp_{X_1} X_3$, 但是否有 $X_1 \perp\!\!\!\perp_{X_3} X_2$? 从该图中看不出来。

※例 2.34. 令服从 0-1 分布的离散型随机变量 C, S, R, W 分别表示是否“多云”、“洒水”、“下雨”、“地湿”, 其联合概率函数 $P(C = c, S = s, R = r, W = w)$ 具有右图所示的分解, 其中 $c, s, r, w \in \{0, 1\}$, 即

$$\begin{aligned} P(C = c, S = s, R = r, W = w) \\ = P(C = c)P(S = s|C = c)P(R = r|C = c)P(W = w|S = s, R = r) \end{aligned}$$



上式右边的每一项都是已知的, 换句话说, 联合分布是已知的。不妨设, 随机变量 $C \sim 0.5\langle 1 \rangle + 0.5\langle 0 \rangle$; 给定条件 C , 随机变量 S 和 R 的条件分布分别为

C	$P(S = 1 C)$	$P(S = 0 C)$	$P(R = 1 C)$	$P(R = 0 C)$
1	0.1	0.9	0.8	0.2
0	0.5	0.5	0.2	0.8

条件概率函数 $P(W = w|S = s, R = r)$ 如下所示, 其中 $P(W = 1|S = 0, R = 1) = 0.9$ 表示在已知未洒水并且下雨的条件下, 地湿的概率为 0.9。

S	R	$P(W = 1 S, R)$	$P(W = 0 S, R)$
1	1	0.99	0.01
1	0	0.9	0.1
0	1	0.9	0.1
0	0	0.0	1.0

基于上述信息, 在任何给定的条件下, 每个随机变量的条件概率都可以计算出来。例如, 观察到地湿的概率为

$$P(W = 1) = \sum_{c,r,s \in \{0,1\}} P(C = c, S = s, R = r, W = 1) \approx 0.6471$$

在观察到地湿的条件下，洒水的概率为

$$\begin{aligned} P(S = 1|W = 1) &= \frac{P(S = 1, W = 1)}{P(W = 1)} \\ &= \frac{\sum_{c,r \in \{0,1\}} P(C = c, S = 1, R = r, W = 1)}{P(W = 1)} \\ &\approx 0.43 \end{aligned}$$

※练习 2.18. 接着例 2.34, 若观察到地湿, 问下雨和洒水哪个可能性大些?

提示: $P(S = 1|W = 1) < P(R = 1|W = 1) \approx 0.708$ 。

2.2.4 随机向量的函数

为了由已知随机向量构造出新的随机向量，我们需要使用随机向量的函数。现在我们面临这样一个问题：已知 n 维随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的密度函数为 $f_{\mathbf{X}}(\mathbf{x})$ ，并且一一映射 $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 是连续的，具体定义为

$$\mathbf{y} = g(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_i(\mathbf{x}), \dots, g_n(\mathbf{x}))^\top$$

向量 $g(\mathbf{x})$ 中的每个分量 $g_i(\mathbf{x})$ 都具有连续的偏导数 $\partial g_i(\mathbf{x}) / \partial x_j$ 。问新构造的随机向量 $\mathbf{Y} = g(\mathbf{X})$ 的密度函数 $f_{\mathbf{Y}}(\mathbf{y})$ ？

\curvearrowleft 定理 2.15. 如果 \mathbb{R}^n 到自身的连续映射 g 的雅可比矩阵（见第 771 页的定义 E.10） $J_g = \partial \mathbf{y} / \partial \mathbf{x}$ 非奇异，则存在 g 的逆映射 $\mathbf{x} = h(\mathbf{y})$ 使得

$$\begin{aligned} P(\mathbf{X} \in D) &= \int_D f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \int_S |\det(J_h)| \cdot f_{\mathbf{X}}(h(\mathbf{y})) d\mathbf{y} \end{aligned}$$

其中区域 $S = g(D)$ 并且 $J_h = \partial \mathbf{x} / \partial \mathbf{y}$ 为逆变换 h 的雅可比矩阵。 $\det(J_h)$ 表示雅可比矩阵 J_h 的行列式，简称为雅可比行列式。随机向量 $\mathbf{Y} = g(\mathbf{X})$ 的密度函数为

$$f_{\mathbf{Y}}(\mathbf{y}) = |\det(J_h)| \cdot f_{\mathbf{X}}(h(\mathbf{y})) = \left| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right| \cdot f_{\mathbf{X}}(h(\mathbf{y}))$$

例 2.35. 已知随机向量 $(X, Y)^\top \in \mathbb{R}^2$ 的密度函数 $f(x, y)$ ，由 X, Y 构造新的随机变量 Z ，分别求 $Z = X + Y, X - Y$ 和 X/Y 的密度函数。

解. 考虑新的随机向量 $(\frac{X}{Z})$ 或者 $(\frac{Y}{Z})$ 的联合分布，然后求得 Z 的边缘分布。

□ 容易算得变换 $\begin{cases} X = X \\ Z = X + Y \end{cases}$ 的逆变换 $\begin{cases} X = X \\ Y = Z - X \end{cases}$ 的雅可比行列式为 1，由定理 2.15 知随机向量 $(X, Z)^\top$ 的密度函数为 $f(x, z - x)$ ，进而求得随机变量 $Z = X + Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z - x) dx$$

□ 类似地，随机变量 $Z = X - Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, x - z) dx$$

□ 变换 $\begin{cases} Y = Y \\ Z = X/Y \end{cases}$ 的逆变换 $\begin{cases} Y = Y \\ X = YZ \end{cases}$ 的雅可比行列式为 y , 由定理 2.15, $(Y, Z)^\top$ 的密度函数为 $|y| \cdot f(yz, y)$, 进而求得 $Z = X/Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} |y| \cdot f(yz, y) dy$$

性质 2.24. 已知随机变量 X, Y 相互独立, 密度函数分别是 $f_X(x), f_Y(y)$, 则随机变量 $Z = X + Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z - x) dx = \int_{-\infty}^{+\infty} f_Y(y) f_X(z - y) dy \quad (2.35)$$

令 $F_X(x), F_Y(y), F_Z(z)$ 分别是随机变量 X, Y, Z 的分布函数, 则

$$F_Z(z) = \int_{-\infty}^{+\infty} F_Y(z - x) dF_X(x) = \int_{-\infty}^{+\infty} F_X(z - y) dF_Y(y) \quad (2.36)$$

由于独立随机变量之和经常被研究, 如中心极限定理, 从式 (2.35) 可以提炼出非线性映射 $(f_X, f_Y) \mapsto f_Z$, 即卷积运算, 定义如下。

■ **定义 2.27** (函数的卷积). 实数域 \mathbb{R} 上的两个可积函数 $f(x)$ 与 $g(x)$ 的卷积 (convolution) 是一个积分运算, 把 f, g 映为可积函数 h , 记作 $h = f * g$, 具体定义为

$$h(z) = f * g = \int_{-\infty}^{+\infty} f(x) g(z - x) dx$$

这里, 我们用 z 作卷积后的变量名, 有时候也喜欢用 t 。选哪个变量名无所谓, 只要不与 f, g 的变量重名引起误解就行。

卷积的物理意义可参阅本书的附录 B, 卷积的几何意义见图 2.24。结果 (2.35) 也可表述为两个独立随机变量之和的密度函数是这两个随机变量密度函数的卷积。

练习 2.19. 请验证卷积满足交换律、结合律、对加法的分配律, 即

$$\begin{aligned} f * g &= g * f \\ (f * g) * h &= f * (g * h) \\ (r_1 f_1 + r_2 f_2) * g &= r_1 f_1 * g + r_2 f_2 * g, \text{ 其中 } r_1, r_2 \in \mathbb{R} \end{aligned}$$

练习 2.20. 证明两个密度函数的卷积依然是密度函数。提示: 利用定理 2.6。

例 2.36. 令 $f(x)$ 为非负判定函数 $J(x)$, 定义见 (2.10)。令 $g(x) = \frac{J(x)}{2} \exp(-\frac{x}{2})$, 即

$$g(x) = \begin{cases} \frac{1}{2} \exp(-\frac{x}{2}) & \text{若 } x \geq 0 \\ 0 & \text{若 } x < 0 \end{cases}$$

$$f * g = \int_{-\infty}^{+\infty} f(x)g(t-x)dx = \begin{cases} 1 - \exp(-\frac{t}{2}) & \text{若 } t \geq 0 \\ 0 & \text{若 } t < 0 \end{cases}$$

函数 $g(t-x)$ 的几何解释是 $g(x)$ 关于 $x=0$ 的镜像 $g(-x)$ 沿水平方向平移 t 。我们把卷积 $f * g$ 比喻成一种更广泛的滑动平均 (moving average)^{*}: 对暂时固定的 t , $\int_{-\infty}^{+\infty} f(x)g(t-x)dx$ 是对 $f(x)$ 在整个实轴上的加权平均, 权重是 $g(t-x)$ 。

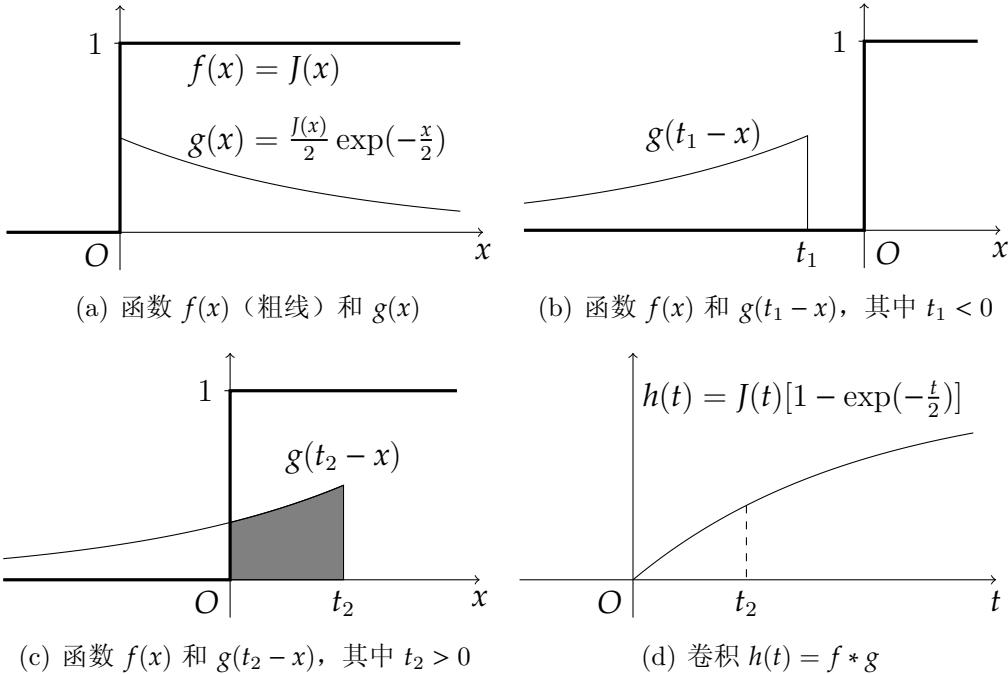


图 2.24: 卷积的几何意义: 在例 2.36 中, 若 $f(x) = J(x)$, 卷积 $h(t) = f * g$ 在点 t_2 的函数值 $h(t_2)$ 即是子图 (c) 中阴影部分的面积。

练习 2.21. 令 $f(x)$ 为非负判定函数 $J(x)$, 求卷积 $f * f$ 并解释它的几何意义。

定义 2.28 (数列的卷积). 两个数列 $\{a_i : i = 0, 1, 2, \dots\}$ 和 $\{b_j : j = 0, 1, 2, \dots\}$ 的卷积是一个新的数列 $\{c_k : k = 0, 1, 2, \dots\}$, 记作 $\{c_k\} = \{a_i\} * \{b_j\}$, 其中 c_k 如下定义,

$$c_k = a_0 b_k + a_1 b_{k-1} + \dots + a_{k-1} b_1 + a_k b_0 = \sum_{i+j=k} a_i b_j, \quad \text{其中 } k = 0, 1, 2, \dots \quad (2.37)$$

^{*}给定数列 $a_0, a_1, \dots, a_t, \dots$, 其简单滑动平均 (simple moving average, SMA) 定义为一个新的数列 $b_0, b_1, \dots, b_t, \dots$, 其中 b_t 是 $a_t, a_{t-1}, \dots, a_{\max(0, t-n+1)}$ 的算术平均, n 为一给定自然数, 称为窗宽。

这里, c_k 可以看作是 a_0, a_1, \dots, a_k 的加权平均, 权重为 b_k, b_{k-1}, \dots, b_0 , 不必归一。例如, $(1, 2, 3) * (1, 2) = (1, 4, 7, 6)$, $(1, 2, 3) * (1, 2, 3) = (1, 4, 10, 12, 9)$ 。

例 2.37. 已知离散型随机变量 X 与 Y 相互独立, 它们的分布列分别为 $P(X = i) = a_i$, $P(Y = j) = b_j$, 其中 $i, j = 0, 1, 2, \dots$ 。不难得到离散型随机变量 $Z = X + Y$ 的分布列 $P(Z = k) = c_k$, 其中 $\{c_k\} = \{a_i\} * \{b_j\}$, $k = 0, 1, 2, \dots$ 。

例 2.38. 已知随机变量 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + q\langle 0 \rangle$, 其中 $q = 1 - p$, 试证明: 随机变量 $X = X_1 + \dots + X_n \sim B(n, p)$ 。

证明. 用归纳法往证 $p_k = P(X = k) = C_n^k p^k q^{n-k}$ 。当 $n = 2$ 时, 易证 $Z = X_1 + X_2 \sim B(2, p)$ 。设 $Y = X_1 + \dots + X_{n-1} \sim B(n-1, p)$, 则

$$\{p_k\} = \{C_{n-1}^0 p^0 q^{n-1}, \dots, C_{n-1}^{n-1} p^{n-1} q^0\} * \{p, q\}$$

$$\text{其中, } p_k = C_{n-1}^{k-1} p^{k-1} q^{n-k} p + C_{n-1}^k p^k q^{n-1-k} q = C_n^k p^k q^{n-k}, k = 0, 1, \dots, n \quad \square$$

例 2.39. 令 $u(x)$ 是 $X \sim U[0, 1]$ 的密度函数, 已知 $X, Y \stackrel{\text{iid}}{\sim} U[0, 1]$, 由**性质 2.24** 不难得 $Z = X + Y$ 的密度函数为

$$f_Z(z) = u * u = \begin{cases} z & \text{当 } 0 < z \leq 1 \\ 2-z & \text{当 } 1 < z \leq 2 \\ 0 & \text{其他} \end{cases} \quad (2.38)$$

求解 $f_Z(z) = u * u$ 的过程如图 2.25 所示。

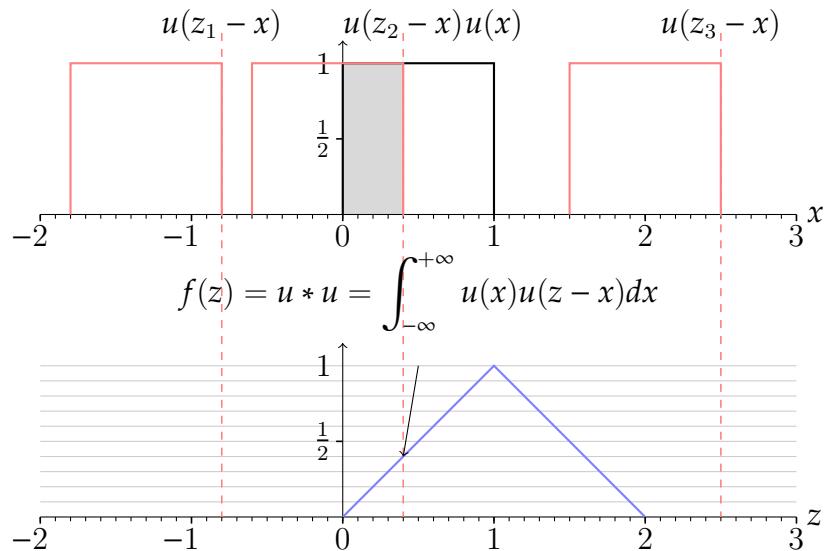


图 2.25: 对每个暂时固定的 z , 密度函数 $u(x)$ 经过翻转和平移得到 $u(z-x)$, 卷积 $f(z) = u * u$ 解释为 $u(x)u(z-x)$ 在整个 x 轴上的累积, 即阴影部分的面积。

解法二. 利用式 (2.36) 和 $Y \sim U[0, 1]$ 的分布函数 $F_Y(y)$ 可以得到 Z 的分布函数 $F_Z(z)$, 即下图中阴影部分的面积。请读者给出 $f_Z(z) = F'_Z(z)$ 。

$$F_Z(z) = \int_0^1 F_Y(z-x)dx = \begin{cases} 0 & \text{当 } z \leq 0 \\ z^2/2 & \text{当 } 0 < z \leq 1 \\ -z^2/2 + 2z - 1 & \text{当 } 1 < z \leq 2 \\ 1 & \text{当 } z > 2 \end{cases}$$

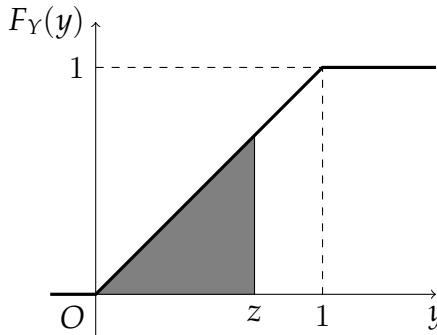
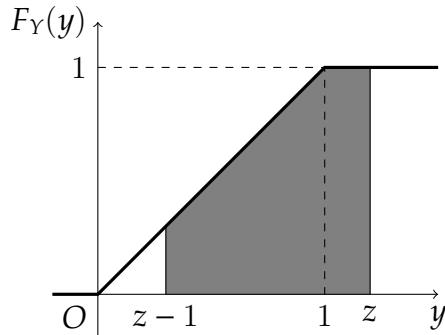
(a) $0 < z \leq 1$ (b) $1 < z \leq 2$

图 2.26: 由式 (2.36) 可知, 例 2.39 所求的 Z 的分布函数 $F_Z(z)$ 等于 $F_Y(y)$ 在区间 $[z-1, z]$ 上的积分, 即图中阴影部分的面积。

解法三. 由已知条件知, 随机向量 $(X, Y)^\top \sim U([0, 1] \times [0, 1])$ 。随机变量 Z 的分布函数为 $F_Z(z) = P(X + Y \leq z)$, 即下图中阴影部分的面积。

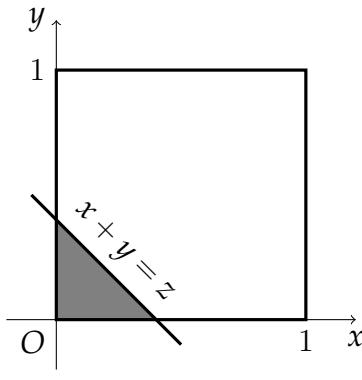
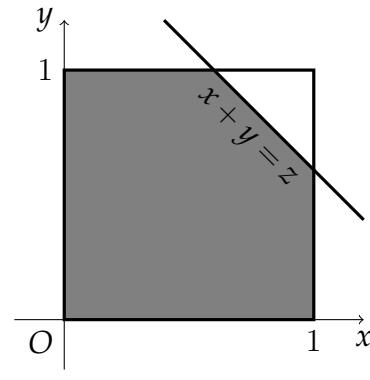
(a) $0 < z \leq 1$ (b) $1 < z \leq 2$

图 2.27: 已知随机向量 $(X, Y)^\top \sim U([0, 1] \times [0, 1])$, 随机变量 $Z = X + Y$ 的分布函数 $P(Z \leq z)$ 即 $(X, Y)^\top$ 落于图中阴影部分的概率, 即该阴影部分的面积。

例 2.39 还有解法四, 见第 242 页的例 3.30, 用到了特征函数这一新工具, 这是后话。具有式 (2.38) 这样密度函数的随机变量被称为服从 $[0, 2]$ 上的等腰三角形分布, 记作 $\Delta[0, 2]$ 。等腰三角形分布是三角形分布的一个特款, 见第 4 章。

例 2.40. 已知随机变量 $X \sim U[0, 1]$ 与 $Y \sim N(0, 1)$ 相互独立, 由结果 (2.35), 随机变量 $Z = X + Y$ 的密度函数为

$$f_Z(z) = \int_0^1 \phi(z-x)dx = \Phi(z) - \Phi(z-1)$$

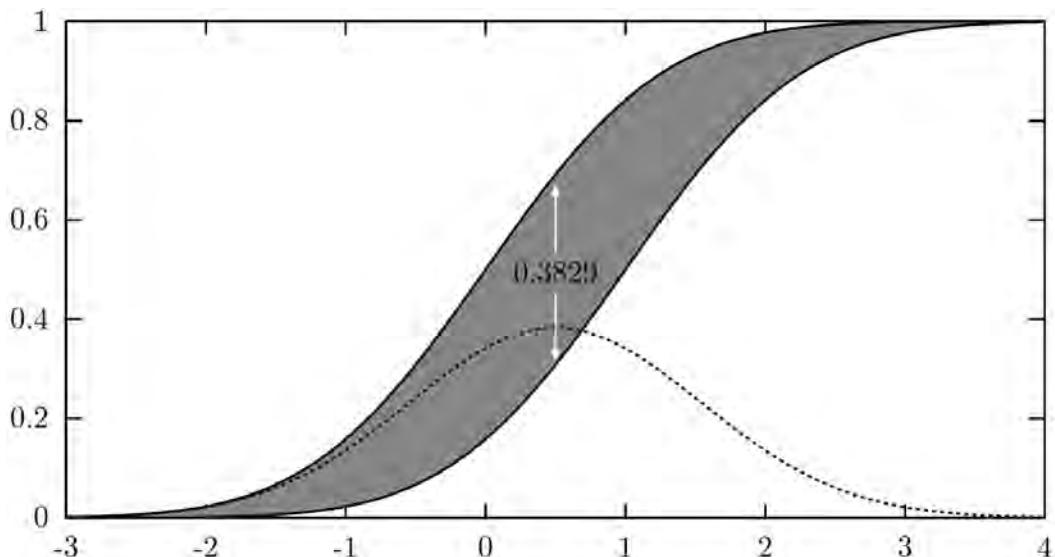


图 2.28: 阴影部分是由曲线 $\Phi(z)$ 和 $\Phi(z-1)$ 围成的, 虚线是密度函数 $f_Z(z) = \Phi(z) - \Phi(z-1)$, 但不是正态的。 $\Phi(z)$ 与 $\Phi(z-1)$ 的 Kolmogorov 距离大约是 0.3829。

曲线 $f_Z(z)$ 关于 $z = 1/2$ 对称, 也呈现钟形, 但 Z 不是正态分布——如若不然, 必存在某 $\sigma > 0$ 使得 $\Phi(z) - \Phi(z-1) = \phi(z|1/2, \sigma^2)$ 。下面往证这是不可能的。

- 当 $z = 1/2$ 时, 利用式 (2.15), 得到方程 $2\Phi(1/2) - 1 = (\sigma \sqrt{2\pi})^{-1}$, 求解得到 $\sigma \approx 1.041828977196953$, 精度为 10^{-15} 。
- 当 $z = 1$ 时, $\Phi(1) - \Phi(0) \approx 0.34134474606854$, 然而 $\phi(z|1/2, \sigma^2) \approx 0.34127030187026$ 。上述结果的精度至少为 10^{-12} , 于是 $\Phi(z) - \Phi(z-1) \neq \phi(z|1/2, \sigma^2)$ 。

例 2.41. 已知 $X, Y \stackrel{\text{iid}}{\sim} U[-1/2, 1/2]$, 求 $Z = XY$ 的分布函数 $F_Z(z)$ 和密度函数 $f_Z(z)$ 。

解. 由已知条件知, 随机向量 $(X, Y)^\top \sim U([-1/2, 1/2] \times [-1/2, 1/2])$ 。随机变量 Z 的分布函数为 $F_Z(z) = P(XY \leq z)$, 即 $(X, Y)^\top$ 落于图 2.29 中阴影部分的概率, 即该阴影部分的面积。下面, 就 $z < 0$ 和 $z > 0$ 这两种情形分别求 $F_Z(z)$ 。

$$F_Z(z) = \begin{cases} 2 \int_{-2z}^{1/2} dy \int_{-1/2}^{z/y} dx = \frac{1}{2} + 2z - 2z \ln(-4z) & \text{当 } z < 0 \\ 1 - 2 \int_{2z}^{1/2} dy \int_{z/y}^{1/2} dx = \frac{1}{2} + 2z - 2z \ln(4z) & \text{当 } z > 0 \end{cases}$$

由随机变量 $Z = XY$ 的分布函数 $F_Z(z)$ 得到其密度函数 $f_Z(z)$ 为

$$f_Z(z) = \begin{cases} -2 \ln(4|z|) & \text{当 } |z| \leq 1/4 \\ 0 & \text{当 } |z| > 1/4 \end{cases}$$

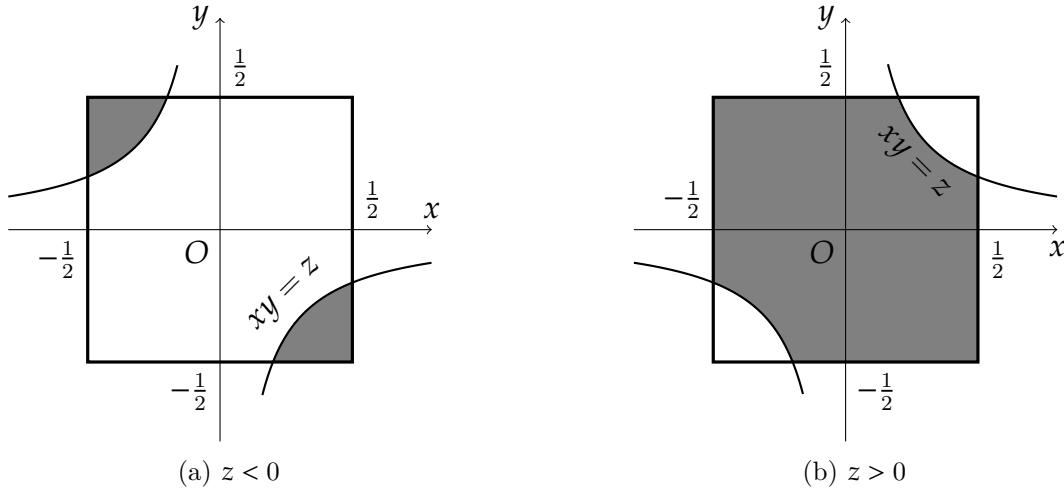


图 2.29: 已知 $(X, Y)^T \sim U([-1/2, 1/2] \times [-1/2, 1/2])$, 随机变量 $Z = XY$ 的分布函数 $P(Z \leq z)$ 即 $(X, Y)^T$ 落于阴影部分的概率, 即该阴影部分的面积。

例 2.42. 已知随机变量 X 和 Y 独立同分布, 其密度函数都是

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{当 } x > 0 \\ 0 & \text{当 } x \leq 0 \end{cases}$$

令 $U = \max(X, Y), V = \min(X, Y)$, 求随机向量 $(U, V)^T$ 的密度函数。

解. 显然, 随机向量 $(X, Y)^T$ 的密度函数为 $f(x, y) = p(x)p(y)$, 并且 $P(X = Y) = 0$ 。而随机向量 $(U, V)^T$ 的分布函数为

$$\begin{aligned} F(u, v) &= P(U \leq u, V \leq v) = 2P(X > Y, X \leq u, Y \leq v) \\ &= 2 \int_0^v \left[\int_y^u p(x)p(y)dx \right] dy \end{aligned}$$

因此, 随机向量 $(U, V)^T$ 的密度函数为

$$f(u, v) = \frac{\partial^2 F(u, v)}{\partial u \partial v} = \begin{cases} 2\lambda^2 e^{-\lambda(u+v)} & \text{当 } u > v > 0 \\ 0 & \text{其他} \end{cases}$$

例 2.43. 若随机变量 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$ 相互独立, 求 $Z = X + Y$ 的分布。

解. 该例是第 138 页的例 2.23 的特款, 因为 $(X, Y)^\top$ 的密度函数是

$$\phi(x|\mu_X, \sigma_X^2)\phi(y|\mu_Y, \sigma_Y^2) = \phi(x, y|\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, 0)$$

由例 2.23 的结论可知,

$$Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

另一解法. 设随机变量 Z 的密度函数为 $f(z)$, 由性质 2.24 的结果, 我们有

$$\begin{aligned} f(z) &= \phi(x|\mu_1, \sigma_1^2) * \phi(x|\mu_2, \sigma_2^2) \\ &= \int_{-\infty}^{+\infty} \phi(x|\mu_1, \sigma_1^2)\phi(z-x|\mu_2, \sigma_2^2)dx \end{aligned} \quad (2.39)$$

仿照第 66 页的例 1.50, 利用配方法, 令 $\rho = \sigma_1^{-2} + \sigma_2^{-2}$,

$$\begin{aligned} \phi(x|\mu_1, \sigma_1^2)\phi(z-x|\mu_2, \sigma_2^2) &\propto \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(z-x-\mu_2)^2}{\sigma_2^2}\right]\right\} \\ &= \exp\left\{-\frac{\rho}{2}\left[x - \frac{1}{\rho}\left(\frac{\mu_1}{\sigma_1^2} + \frac{z}{\sigma_2^2}\right)\right]^2 - \frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right\} \\ &\propto \phi\left(x \left| \frac{1}{\rho}\left(\frac{\mu_1}{\sigma_1^2} + \frac{z}{\sigma_2^2}\right), \frac{1}{\rho}\right. \right) \phi(z|\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \end{aligned}$$

将之带入到 (2.39) 右边, 积分归一了密度函数 $\phi(x|\cdot, \cdot)$, 因此,

$$f(z) \propto \phi(z|\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

由于 $f(z)$ 是一个密度函数, 必有 $f(z) = \phi(z|\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ 。这个解法稍微笨了些, 简单的解法见定理 4.6 的证明, 要用到第 3 章的特征函数工具。

例 2.44. 考虑二元正态分布 $(X, Y)^\top \sim N(0, 0, \sigma_X^2, \sigma_Y^2, \rho)$, 其密度函数 (见例 2.22) 为 $f(x, y) = \phi(x, y|0, 0, \sigma_X^2, \sigma_Y^2, \rho)$, 即

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right]\right\}$$

试求经过正交变换 (即, 坐标旋转变换) 后所得的随机向量 $(U, V)^\top$ 的密度函数 $p(u, v)$, 其中 $U = X \cos \alpha + Y \sin \alpha, V = -X \sin \alpha + Y \cos \alpha$ 。

解. 由定理 2.15 得到随机向量 $(U, V)^\top$ 的密度函数为

$$\begin{aligned} p(u, v) &= f(u \cos \alpha - v \sin \alpha, u \sin \alpha + v \cos \alpha) \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{Au^2 - 2Buv + Cv^2}{2(1-\rho^2)}\right\} \end{aligned}$$

其中, $A = \frac{\cos^2 \alpha}{\sigma_X^2} - 2\rho \frac{\cos \alpha \sin \alpha}{\sigma_X\sigma_Y} + \frac{\sin^2 \alpha}{\sigma_Y^2}$
 $B = \frac{\cos \alpha \sin \alpha}{\sigma_X^2} - \rho \frac{\sin^2 \alpha - \cos^2 \alpha}{\sigma_X\sigma_Y} - \frac{\cos \alpha \sin \alpha}{\sigma_Y^2}$
 $C = \frac{\sin^2 \alpha}{\sigma_X^2} + 2\rho \frac{\cos \alpha \sin \alpha}{\sigma_X\sigma_Y} + \frac{\cos^2 \alpha}{\sigma_Y^2}$

令 α 满足 $\tan \alpha = 2\rho\sigma_X\sigma_Y/(\sigma_X^2 - \sigma_Y^2)$, 则系数 $B = 0$ 。由练习 2.16, 此时随机变量 U, V 相互独立。也就是说, 二元正态分布经过坐标旋转变换后依然为正态分布, 总能选择合适的角度使得各分量间是独立的。

练习 2.22. 根据例 2.44 的结果设计算法来产生 $(X, Y)^\top \sim N(0, 0, \sigma_X^2, \sigma_Y^2, \rho)$ 的随机数。

※例 2.45. 已知随机变量 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1)$, 求随机变量 $Z = \sqrt{Y}$ 和 $Y = \sum_{j=1}^n X_j^2$ 的密度函数 $f_Z(z)$ 和 $f_Y(y)$ 。

解. 随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的密度函数为 $f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^n \phi(x_j)$ 。

□ 先求 Z 的分布函数 $F_Z(z)$ 和密度函数 $f_Z(z)$ 。

$$\begin{aligned} F_Z(z) &= \int_D f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \text{ 其中 } D = \left\{ \mathbf{x} : \sum_{j=1}^n x_j^2 \leq z^2 \right\} \\ &= \int_D \cdots \int_D \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp \left\{ -\sum_{j=1}^n \frac{x_j^2}{2} \right\} dx_1 dx_2 \cdots dx_n \end{aligned}$$

利用球面坐标变换, 令上式中变量 x_1, x_2, \dots, x_n 分别替换为

$$\begin{aligned} x_1 &= \rho \cos \alpha_1 \cos \alpha_2 \cdots \cos \alpha_{n-1} \\ x_2 &= \rho \cos \alpha_1 \cos \alpha_2 \cdots \sin \alpha_{n-1} \\ &\vdots \\ x_n &= \rho \sin \alpha_1 \end{aligned}$$

舍弃常数因子，利用正比关系将 $F_Z(z)$ 整理为

$$\begin{aligned} F_Z(z) &\propto \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} \left[\int_0^z e^{-\rho^2/2} \rho^{n-1} d\rho \right] \Psi(\alpha_1, \dots, \alpha_{n-1}) d\alpha_{n-1} \cdots d\alpha_1 \\ &\propto \int_0^z e^{-\rho^2/2} \rho^{n-1} d\rho, \text{ 其中 } \Psi(\alpha_1, \dots, \alpha_{n-1}) \text{ 是有关 } \alpha_1, \dots, \alpha_{n-1} \text{ 的某函数} \\ F_Z(z) &= C_n \int_0^z e^{-\rho^2/2} \rho^{n-1} d\rho, \text{ 进而 } f_Z(z) = \begin{cases} C_n e^{-z^2/2} z^{n-1} & \text{当 } z > 0 \\ 0 & \text{当 } z \leq 0 \end{cases} \end{aligned}$$

由 Gamma 函数的定义 $\Gamma(s) = \int_0^{+\infty} e^{-x} x^{s-1} dx$, 求得归一因子 C_n 为

$$C_n = \left[\int_0^{+\infty} e^{-\rho^2/2} \rho^{n-1} d\rho \right]^{-1} = \left[2^{\frac{n}{2}-1} \Gamma\left(\frac{n}{2}\right) \right]^{-1}$$

$n = 2$ 时, Z 的分布称为 Rayleigh 分布 (详见第 313 页的定义 4.29); $n = 3$ 时, Z 的分布称为 Maxwell 分布 (详见定义 4.30)。Rayleigh 分布和 Maxwell 分布是物理学中常见的分布, 第 4 章将详细讨论它们。

□ 利用第 131 页的例 2.17 的结果, 容易得到 Y 的密度函数为

$$f_Y(y) = \begin{cases} 0 & \text{当 } y \leq 0 \\ \frac{y^{n/2-1} e^{-y/2}}{2^{n/2} \Gamma(n/2)} & \text{当 } y > 0 \end{cases} \quad (2.40)$$

随机变量 Y 的分布在统计学中占有重要地位, 称作自由度为 n 的 χ^2 分布 (国内的有些文献里把它译为“卡方”分布, 详见第 296 页的定义 4.17), 记作 $Y \sim \chi_n^2$, 其中 n 表示自由度。

※例 2.46. 已知随机变量 $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} \frac{1}{2}\langle 0 \rangle + \frac{1}{2}\langle 2 \rangle$, 定义新的随机变量如下,

$$X = \sum_{j=1}^{\infty} \frac{1}{3^j} X_j$$

该随机变量取值范围是 Cantor 集, 它正是例 2.11 定义的奇异型随机变量。

2.3 随机变量的数字特征

分布的重要数字特征包括：分位数、众数、期望 (expectation)、方差 (variance)、矩 (moment) 等。如同素描，通过这些数字特征可以寥寥数笔勾勒出分布的“大致轮廓”。下面，先介绍由分布函数 $F(x)$ 直接定义的分位数概念。

定义 2.29 (分位数). 已知随机变量 X 的分布函数为 $F(x)$ ，对于 $0 \leq \alpha \leq 1$ ，实数 $q_\alpha = \inf\{x : \alpha \leq F(x)\}$ 称为下侧 α -分位数 (lower α -th quantile)^{*}，简称 α -分位数，它总是存在的。特别地，分位数 $q_{1/2}$ 称为中位数 (median)，记作 $M(X)$ 或 MX 或 m_X 。通过若干分位数，人们可以大致了解分布函数的形态。

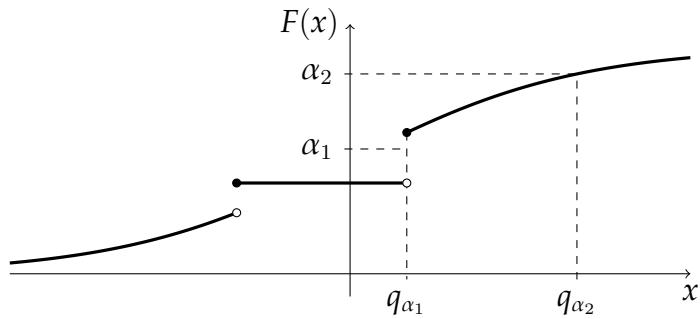


图 2.30: 当 $F(x)$ 严格单调增时， α -分位数 q_α 就是方程 $F(x) = \alpha$ 的解。

性质 2.25. 已知分布函数 $F(x)$ ，则 α -分位数 q_α 满足性质

$$F(q_\alpha - 0) \leq \alpha \leq F(q_\alpha)$$

证明. 由**定义 2.29**， $\alpha \leq F(q_\alpha)$ 是显然的。同时，对于任意的 $h > 0$ 皆有 $F(q_\alpha - h) < \alpha$ ，所以 $F(q_\alpha - 0) \leq \alpha$ 。 \square

例 2.47. 已知分布函数 $F(x) = 1 - \exp(-\beta x)$ ，其中 $\beta > 0$ ，则它的 α -分位数是

$$q_\alpha = -\beta^{-1} \ln(1 - \alpha)$$

练习 2.23. 在第 126 页的**图 2.12** 中，直观地解释 $N(0, 1)$ 的 α -分位数，并近似地求解它。提示：利用 Pólya 近似式 (2.16)。

练习 2.24. 若分布函数 $F(x)$ 有跳跃点 $x = h$ ，则 $\forall \alpha \in (F(h - 0), F(h)]$ 皆有 $q_\alpha = h$ 。

^{*}本书所采用的分位数缺省地是下侧 α -分位数。有的教科书采用上侧 α -分位数 (upper α -th quantile) $q'_\alpha = \sup\{x : F(x) < \alpha\}$ 。读者在阅读文献的时候，注意上下文中对分位数的约定。

性质 2.26. 若分布函数 $F(x)$ 连续，则 $F(q_\alpha) = \alpha$ 。并且，对于 $0 \leq \alpha_1 < \alpha_2 \leq 1$ 有

$$P(q_{\alpha_1} < X \leq q_{\alpha_2}) = \alpha_2 - \alpha_1$$

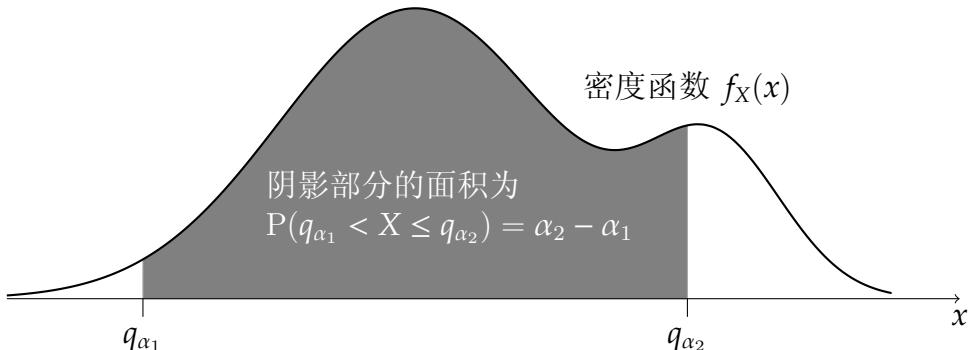


图 2.31: 分位数的直观含义: 随机变量 $X \in (-\infty, q_\alpha]$ 的概率是 α 。

性质 2.27. 如果随机变量 X 的密度函数 $f_X(x)$ 关于 $x = 0$ 对称，则有

$$q_\alpha = -q_{1-\alpha}, \text{ 其中 } 0 < \alpha < 1$$

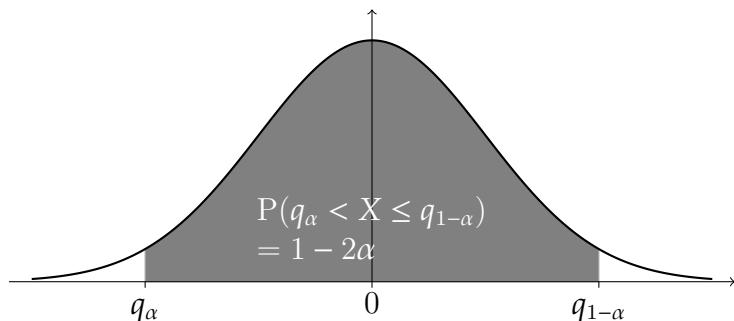


图 2.32: 如果随机变量 X 的密度函数 $f_X(x)$ 关于 $x = 0$ 对称，则 $q_\alpha = -q_{1-\alpha}$ 。

例 2.48. 标准正态分布 $Z \sim N(0, 1)$ 的 α -分位数 z_α 的数值计算:

定义 2.30. 众数 (mode) 是随机变量的数字特征，它可以不存在，也可以不唯一。

- 一个连续型随机向量 $\mathbf{X} \sim f(\mathbf{x})$ 的众数定义为密度函数 $f(\mathbf{x})$ 的极大值点。
- 对于离散型随机向量 $\mathbf{X} \sim p_1\langle \mathbf{x}_1 \rangle + p_2\langle \mathbf{x}_2 \rangle + \cdots + p_j\langle \mathbf{x}_j \rangle + \cdots$, 若 $p_j = \max(p_1, p_2, \dots)$, 则称 \mathbf{x}_j 为众数。

我们把有一个和多个众数的分布分别称作单峰分布 (unimodal distribution) 和多峰分布 (multimodal distribution)。

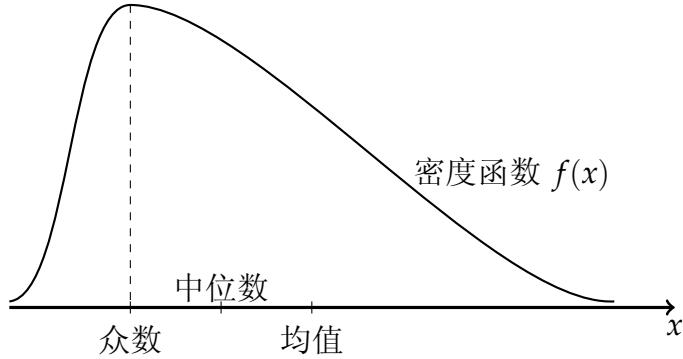


图 2.33: 单峰右长尾的密度函数的均值、众数、中位数的直观表示。

例 2.49. 分布 $N(\mu, \sigma^2)$ 的众数是 μ , 而 $U(0, 1)$ 的众数不唯一。例 2.41 中, 随机变量 Z 的众数不存在。

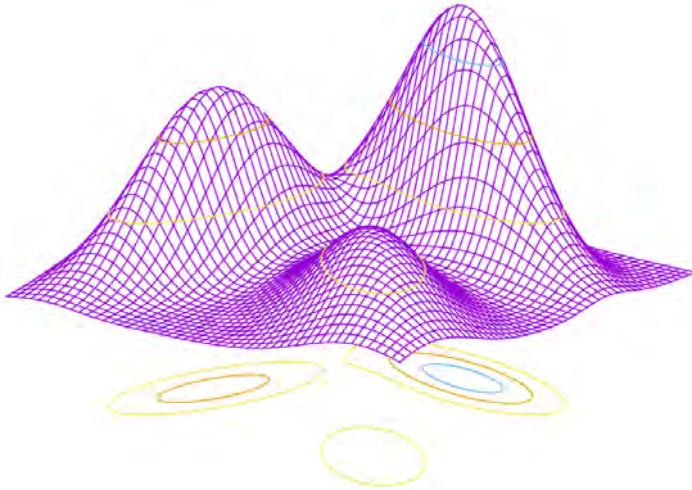


图 2.34: 多峰分布的密度函数曲面：众数就是局部极大值点。

定理 2.16 (Khinchin, 1938). 随机变量 X 服从众数为 0 的单峰分布当且仅当存在独立的随机变量 $U \sim U[0, 1]$ 和 Y 使得 $X = YU$ 。

例 2.50. 设随机变量 $Y \sim N(0, 1)$ 与 $U \sim U[\epsilon, 1]$ 相互独立, 其中 $0 < \epsilon < 1$, 仿照例 2.35, 求得随机变量 $X = YU$ 的密度函数为

$$\begin{aligned} f_X(x, \epsilon) &= \int_{\epsilon}^1 f_Y(x/u) f_U(u) \frac{1}{u} du \\ &= \int_{\epsilon}^1 \frac{\phi(x/u)}{u(1-\epsilon)} du \end{aligned}$$

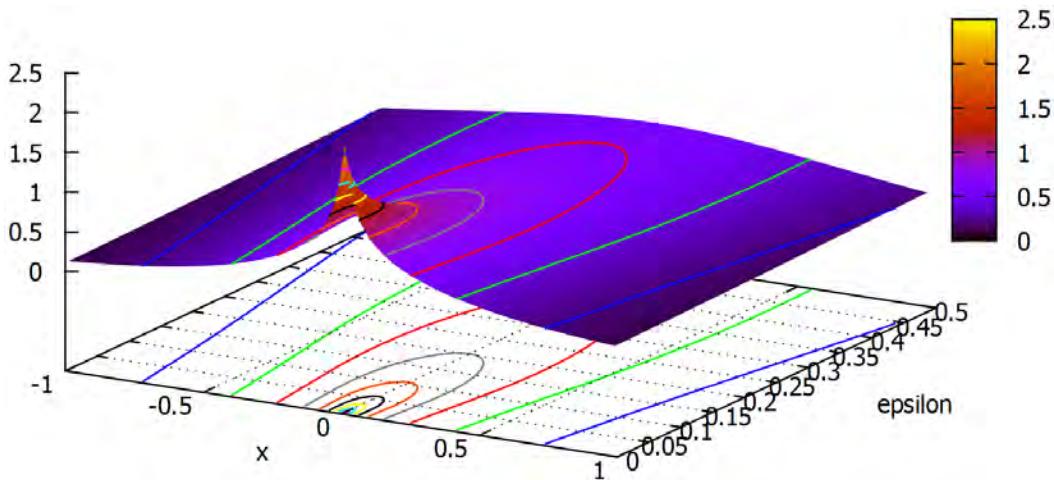


图 2.35: 例 2.50 中, 函数 $f_X(x, \epsilon)$ 的曲面, 对于每个固定的 ϵ 在 $x = 0$ 处取得最大值。

本节的内容颇多, 核心概念是“期望”, 它是随机变量 X 最重要的数字特征, 其他重要概念, 如“方差”、“矩”、“熵”等都是由 X 构造的某个随机变量的期望。

除了能勾勒出分布的“大致轮廓”, 数字特征还能揭示随机变量的内在规律, 而不必计较具体的分布。其中不乏一些经典的结果, 如双期望定理, 以及以 Markov、Chebyshev、Kolmogorov、Lévy、Bernstein、Hoeffding 等数学家命名的若干不等式, 它们在概率统计中有着广泛的应用。

本节内容

第一小节介绍了随机变量最重要的数字特征——期望, 它刻画的是随机变量的“平均取值”。第二小节定义了条件期望, 并得到颇有意思的“双期望定理”。第三小节介绍了随机变量另一重要的数字特征——方差, 它刻画的是随机变量的取值在期望周围的分散程度。优美且应用广泛的 Markov 不等式、Chebyshev 不等式、Kolmogorov 不等式等结果揭示了期望和方差的关系, 它们是第四小节的主要内容。还有一些数字特征, 我们放在第五小节讨论, 如原点矩、中心矩、绝对矩、偏度系数、峰度系数、变异系数、协方差、相关系数等。其中, 相关系数是由方差和协方差定义的一个数字特征, 它衡量了两个随机变量之间的线性相关程度。最后, 第六小节重点讨论最小二乘法和回归, 用来研究两个非独立变量之间的函数关系, 当然也包括最简单的线性关系。

关键知识

(1) 期望、方差、协方差、相关系数等; (2) 概率不等式; (3) 最小二乘法法。

2.3.1 数学期望

随机变量 X 最常用的一个数字特征就是它的期望，也称作均值 (mean)、期望值 (expected value)、数学期望等，常记为 $E(X)$ 或 EX ，有时也简记为 μ_X 或 μ 。

期望这一重要概念的产生与下面的“赌资分配问题”有关，该问题十五世纪末就已提出，但在很长时间里悬而未决。de Méré 曾请教过 Pascal 赌资分配问题，在此之前 Pascal 和 Fermat 就已经多次通信讨论过概率问题，而这一次他们不仅联手解决了赌资分配问题，还提出了期望的概念。

例 2.51 (赌资分配问题). 甲乙二人玩赌博游戏，每局输赢机会等同，先赢够 6 局的人得到全部赌资 64 个金币。由于某原因游戏未决出胜负就中止了，目前的状态是甲赢了 5 局，乙赢了 2 局。问：甲乙二人如何公平地分配赌资？

当时，很多人认为甲乙二人所得应该是 5 : 2，但也有人认为甲距离赢得所有金币只有一步之遥，所得赌资应该更多些，众说纷纭，莫衷一是。赌资分配问题在 1654 年 Pascal 和 Fermat 的多次通信中终于得到解决，二人所用方法不同，但殊途同归，Fermat 的方法更接近现代的解法。而期望的概念则是这两位天才讨论所得的副产品。

在赌资分配问题中，甲只需再赢一次便可获胜，而乙还需要再赢四次才能获胜。让我们想像赌博继续了下去：

至多再赌四局一切结果都分明了，甲获胜的可能是 $15/16$ ，超过四局的“虚拟”赌博甲获胜的可能依然是 $15/16$ 。所以，甲乙所得应该按照 $15 : 1$ 来分配赌资。在这场虚拟赌博中，甲的所得是随机变量

$$X \sim \frac{15}{16}\langle 64 \rangle + \frac{1}{16}\langle 0 \rangle$$

因此，在现实中分配给甲的赌资（金币数目）应是 X 的均值，即下面的加权平均值。

$$64 \times \frac{15}{16} + 0 \times \frac{1}{16} = 60$$

定义 2.31 (期望损失). 受随机因素或缺失信息的影响，人们经常遇到这样的决策问题：有若干可选的行为 a_1, \dots, a_n ，假设每一行为都将产生几个可能的结果，如何选出最优行为呢？关键是给出行为的评价标准，理性的方法是列出行为 a 可能导致的所有结果和相应的概率 $p_i (i = 1, 2, \dots)$ ，并给出相应的损失 l_i （譬如用金钱来计量），利用期望损失，即加权平均值 $\sum_{i=1}^{\infty} p_i l_i$ 来评价行为



a , 期望损失^{*}最小的行为就是该决策问题的解。



Pascal 在他的遗作《思想录》(1670) 第三编《必须打赌》里利用期望损失来论述应该“赌上帝存在”, 因为“假如你赢了, 你就赢得了一切; 假如你输了, 你却一无所失。”在哲学史上, Pascal 是把概率论用于解决传统形而上学问题的第一人。

例 2.52. 现有 10000 元人民币, 若存银行可稳赚利息 200 元; 若用于投资, 有 10% 的可能血本无归, 也有 90% 的可能赢得 2000 元。试问这笔钱该用于投资还是存银行?

行为	投资成功 (90%)	投资失败 (10%)	期望损失
投资	-2000	10000	-800
存银行	-200	-200	-200

解. “投资”行为有两个不确定的结果, 它的期望损失是 $-2000 \times 0.9 + 10000 \times 0.1 = -800$, 而“存银行”行为的期望损失是 -200, 理性的决策会选择期望损失小的行为[†], 即“投资”。

通俗地讲, 随机变量的期望就是它的平均取值。对离散型随机变量而言, 期望的意义显得更直观一些, 就是加权平均。

定义 2.32 (离散型随机变量的期望). 已知 X 是离散型随机变量, 其概率函数 $P(X = x_j) = p_j, j = 1, 2, \dots$ 满足下面的绝对收敛条件

$$\sum_{j=1}^{\infty} |x_j| p_j < \infty \quad (2.41)$$

则称以下级数为离散型随机变量 X 的期望或均值, 记作 $E(X)$ 或 EX 。

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j) = \sum_{j=1}^{\infty} x_j p_j \quad (2.42)$$

条件 (2.41) 保证了级数 (2.42) 是绝对收敛的, 于是式 (2.42) 的值与求和次序无关, 这样 $E(X)$ 才是有意义的。

^{*}统计学家是悲观主义者, 习惯用损失; 而经济学家是乐观主义者, 习惯用收益。损失即是负的收益, 所以期望损失最小等价于期望收益最大。

[†]把例 2.52 中的货币单位“元”改为“亿元”后情况会怎样呢? 相信多数人会选“存银行”。统计决策最终要牵扯到效用 (utility), 本书不作深入介绍, 感兴趣的读者可参阅 J. O. Berger 的《统计决策论及贝叶斯分析》[9] 和 M. H. DeGroot 的《最优统计决策》[32]。

离散型随机变量的期望可看作是对算术平均的推广：将式 (2.42) 理解为 x_1, x_2, \dots 的加权平均，权重分别为 p_1, p_2, \dots 。而算术平均值 $\frac{1}{n} \sum_{j=1}^n y_j$ （假设 y_1, y_2, \dots, y_n 两两不等）恰是均匀分布 $Y \sim \frac{1}{n}\langle y_1 \rangle + \dots + \frac{1}{n}\langle y_n \rangle$ 的期望。

例 2.53. 单点分布 $X \sim \langle c \rangle$ 的期望 $E(X) = c$ 。两点分布 $Y \sim p\langle a \rangle + (1-p)\langle b \rangle$ 的期望 $E(Y) = ap + b(1-p) = b + (a-b)p$ 。

例 2.54. 二项分布 $X \sim B(n, p)$ 的期望是

$$\begin{aligned} E(X) &= \sum_{k=0}^n k \cdot P(X=k) = \sum_{k=0}^n k C_n^k p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^n C_{n-1}^{k-1} p^{k-1} (1-p)^{n-k} = np[p + (1-p)]^{n-1} = np \end{aligned}$$

当 n 很大时， np 就是 $P(X=k)$ 的对称位置，见第 23 页的图 1.7。

例 2.55 (Gini 混乱度). 已知随机变量 X 属于 n 个类，满足 $X \sim p_1\langle 1 \rangle + \dots + p_n\langle n \rangle$ 。
 X 被错分的概率是

$$I_G(X) = \sum_{j=1}^n p_j(1-p_j) \quad (2.43)$$

(2.43) 被称为 Gini 混乱度 (impurity)，由意大利统计学家、人口学家和社会学家 Corrado Gini (1884-1965) 命名。Gini 混乱度被用于机器学习中著名的分类与回归树 (Classification And Regression Tree, CART) 算法。



定义 2.33 (连续型随机变量的期望). 若 X 是连续型随机变量，其密度函数 $f(x)$ 满足下面的绝对可积条件

$$\int_{-\infty}^{+\infty} |x| f(x) dx < \infty \quad (2.44)$$

则称如下积分为连续型随机变量 X 的期望或均值，记作 $E(X)$ 或 EX 。

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad (2.45)$$

例 2.56. 均匀分布 $X \sim U[a, b]$ （其中 $-\infty < a < b < +\infty$ ）的期望是

$$E(X) = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{b+a}{2}$$

例 2.57. 正态分布 $X \sim N(\mu, \sigma^2)$ 的期望就是密度函数的对称位置, 即

$$E(X) = \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = \mu$$

 若条件式 (2.41) 或式 (2.44) 不成立, 即便式 (2.42) 或式 (2.45) 的结果为有限值, 按定义随机变量的期望也不存在。如下面两例,

例 2.58. 考察离散型随机变量 X , 设它的概率质量函数为

$$P\left\{X = (-1)^k \frac{2^k}{k}\right\} = \frac{1}{2^k}, \text{ 其中 } k = 1, 2, \dots$$

按照条件 (2.41), 随机变量 X 的期望不存在, 这是因为

$$\sum_{k=1}^{\infty} |x_k| p_k = \sum_{k=1}^{\infty} \frac{1}{k} = \infty$$

此时, $\sum_{k=1}^{\infty} x_k p_k = \sum_{k=1}^{\infty} (-1)^k / k = -\ln 2$ 是没有任何意义的。

练习 2.25. 请验证如下定义的离散型随机变量的期望不存在。

$$X \sim \frac{6}{\pi^2} \langle 1 \rangle + \frac{6}{4\pi^2} \langle 2 \rangle + \cdots + \frac{6}{k^2\pi^2} \langle k \rangle + \cdots$$

例 2.59 (Cauchy 分布). 若连续型随机变量 X 具有如下密度函数, 则称它服从参数为 (μ, λ) 的 Cauchy 分布, 记作 $X \sim \text{Cauchy}(\mu, \lambda)$, 其中 μ 为位置参数, λ 为尺度参数。

$$f(x|\mu, \lambda) = \frac{\lambda}{\pi[(x-\mu)^2 + \lambda^2]}, \text{ 其中 } \lambda > 0, -\infty < x, \mu < \infty \quad (2.46)$$

尽管 $X \sim \text{Cauchy}(\mu, \lambda)$ 的密度函数关于 $x = \mu$ 对称, 但因条件 (2.44) 不成立, 所以 Cauchy 分布的期望不存在。

Cauchy 分布的密度函数 $f(x|\mu, \lambda)$

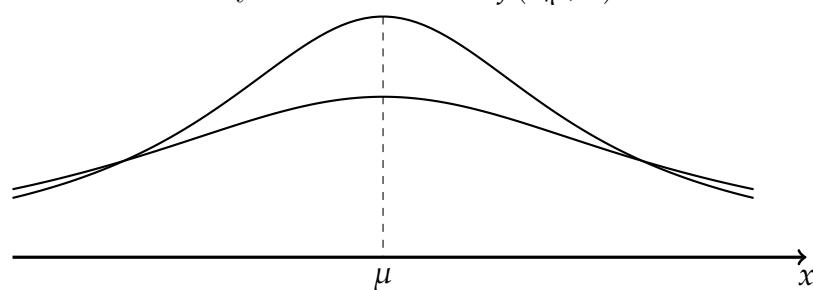


图 2.36: 尺度参数 λ 越小, 分布 $\text{Cauchy}(\mu, \lambda)$ 的密度函数曲线越“高瘦”。

 我们约定, 当使用符号 $E(X)$ 时, 都缺省地表示随机变量 X 的期望存在。不论随机变量 X 是离散型的还是连续型的, 它的期望可表示成 Riemann-Stieltjes 积分或 Lebesgue-Stieltjes 积分^{*} (见第 761 页的定理 D.3)。

$$E(X) = \int_{-\infty}^{+\infty} x dF(x), \text{ 其中 } F(x) \text{ 是 } X \text{ 的分布函数} \quad (2.47)$$

当 X 是离散型随机变量时, 式 (2.47) 的含义见第 758 页的图 C.2。另外, Stieltjes 积分也能带来书写和讨论的便捷, 譬如

$$P(X \in A) = \int_A dF(x), \text{ 其中 } A \subseteq \mathbb{R}$$

如果读者不熟悉 Riemann-Stieltjes 积分 (或 Lebesgue-Stieltjes 积分), 也可以仅把它视作约定的符号记法, 指代式 (2.42) 和式 (2.45)。

 定理 2.17. 设 h 是一个 Borel 可测函数, 若 $Y = h(X)$ 的期望存在, 则

$$E(Y) = \int_{-\infty}^{+\infty} h(x) dF(x) = \begin{cases} \sum_{j=1}^{\infty} h(x_j) p_j & \text{离散型} \\ \int_{-\infty}^{+\infty} h(x) f(x) dx & \text{连续型} \end{cases}$$

※证明. 见 Michel Loève (1907-1979) 的《概率论》第三章第十节。 □

例 2.60. 若随机变量 $X \sim N(0, \tau^2)$, 求随机变量 $Y = \Phi(X/\sigma)$ 的期望。

解. 由第 128 页的例 2.14 的结果, 我们有

$$E\{\Phi(X/\sigma)\} = \int_{-\infty}^{+\infty} \phi(x|0, \tau^2) \Phi(x|0, \sigma^2) dx = \frac{1}{2}$$

定义 2.34. 已知二维随机向量 $(X, Y)^\top$ 的联合分布函数为 $F(x, y)$, g 是 \mathbb{R}^2 上的 Borel 可测函数。定义 $g(X, Y)$ 的期望为 Lebesgue-Stieltjes 积分

$$E[g(X, Y)] = \int_{\mathbb{R}^2} g(x, y) dF(x, y)$$

离散型: 若 $\sum_{i,j=1}^{\infty} |g(x_i, y_j)| p_{ij} < \infty$, 则定义 $g(X, Y)$ 的期望为

$$E[g(X, Y)] = \sum_{i,j=1}^{\infty} g(x_i, y_j) p_{ij}$$

*见附录 C 和附录 D 的简介或 W. Rudin 的名著《数学分析原理》[140] 第六章。

连续型: 若 $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |g(x, y)| f(x, y) dx dy < \infty$, 则定义 $g(X, Y)$ 的期望为

$$E[g(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy$$

性质 2.28. 如果随机变量 X, Y 的期望都存在, 则 $\forall a, b \in \mathbb{R}$ 有

$$E(b) = b \quad (2.48)$$

$$E(aX + b) = aE(X) + b, \text{ 特别地 } E[X - E(X)] = 0 \quad (2.49)$$

$$E(X + Y) = E(X) + E(Y) \quad (2.50)$$

证明. 由单点分布 $P(X = b) = 1$ 易证式 (2.48)。式 (2.49) 由式 (2.47) 和 Riemann-Stieltjes 积分的性质 (见附录 C) 导出。作为练习, 读者也可以从**定义 2.32** 和**定义 2.33** 出发给出式 (2.49) 和式 (2.50) 的证明, 就像下面证明式 (2.50) 的离散情形。

$$\begin{aligned} E(X + Y) &= \sum_{i,j} p_{ij}(x_i + y_j) = \sum_i x_i \sum_j p_{ij} + \sum_j y_j \sum_i p_{ij} \\ &= \sum_i x_i p_{i\cdot} + \sum_j y_j p_{\cdot j} = E(X) + E(Y) \end{aligned} \quad \square$$

推论 2.1. 由结果 (2.48), 总有 $E[X - E(X)] = 0$ 。

$E[X - E(X)] = 0$ 是直观的: 把 $X - E(X)$ 解释为随机变量 X 偏离均值 $E(X)$ 的误差, 这误差可正可负, 但其均值为零。

性质 2.29. 如果随机变量 X, Y 相互独立且期望都存在, 则 XY 的期望存在且

$$E(XY) = E(X)E(Y)$$

证明. 下面给出的是离散情形的证明, 请读者补证连续的情形。

$$E(XY) = \sum_{i,j} p_{ij} x_i y_j = \sum_{i,j} p_{i\cdot} p_{\cdot j} x_i y_j = \sum_i x_i p_{i\cdot} \sum_j y_j p_{\cdot j} = E(X)E(Y) \quad \square$$

下面的性质非常重要, 常用来判定一个事件以概率 1 发生。该性质的证明要用到即将介绍的 Chebyshev 不等式, 以后再说。

性质 2.30. 随机变量 X 满足 $E(X^2) = 0 \Leftrightarrow X$ 服从单点分布 $P(X = 0) = 1$ 。

2.3.2 条件期望与双期望定理

很自然地，我们可以把期望的定义用于条件分布，得到条件期望的定义。

定义 2.35 (条件期望). 已知随机向量 $(X, Y)^\top$ 的概率函数或密度函数，在给定 $X = x$ 的条件下 Y 的期望，简称 $Y|X = x$ 的条件期望，定义为

$$E(Y|X = x) = \int_{-\infty}^{+\infty} y dF(y|x), \text{ 其中 } F(y|x) \text{ 为 } Y|X = x \text{ 的条件分布函数} \quad (2.51)$$

显然， $E(Y|X = x)$ 是一个有关 x 的函数。分为离散型和连续型两种情形，条件期望 $E(Y|X = x)$ 具体定义为

$$\begin{aligned} E(Y|X = x_i) &= \sum_j y_j P(Y = y_j|X = x_i) = \sum_j y_j \frac{p_{ij}}{p_i} && \text{离散型} \\ E(Y|X = x) &= \int_{-\infty}^{+\infty} y f(y|x) dy = \int_{-\infty}^{+\infty} y \frac{f(x, y)}{f_X(x)} dy && \text{连续型} \end{aligned}$$

类似地，在给定 $Y = y$ 的条件下 X 的期望（简称 $X|Y = y$ 的条件期望）定义为

$$E(X|Y = y) = \int_{-\infty}^{+\infty} x dF(x|y)$$

例 2.61. 已知 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ，其中 $|\rho| < 1$ 。根据**例 2.27** 和**例 2.57** 的结果，求得 $Y|X = x$ 的条件期望 $E(Y|X = x)$ 。

$$E(Y|X = x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X) \quad (2.52)$$

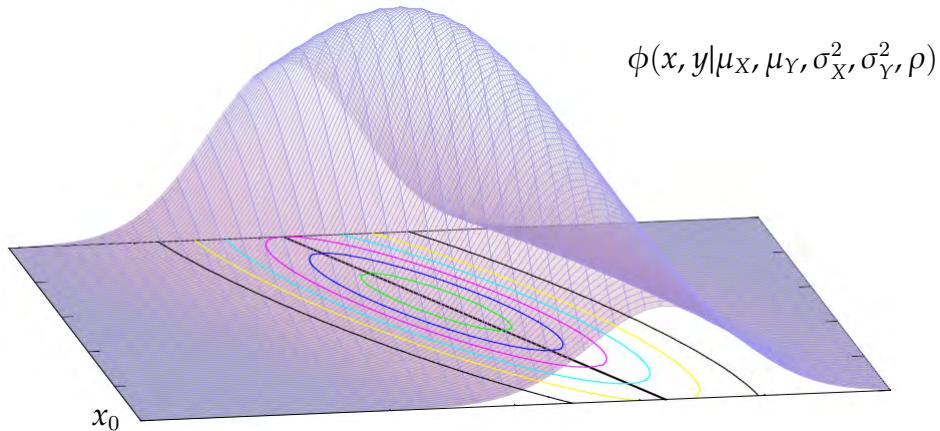


图 2.37：图中 xoy 平面上的直线 $l(x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$ 就是 $E(Y|X = x)$ ，它不是等高椭圆的长轴。用平面 $x = x_0$ 截曲面 $\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ，剖面曲线正比于正态分布 $N(l(x_0), (1 - \rho^2)\sigma_Y^2)$ 的密度函数，只相差一个常数因子 $\phi(x_0 | \mu_X, \sigma_X^2)$ 。

性质 2.31. 已知 h 是 X 的函数, 则 $E(h(X)Y|X) = h(X)E(Y|X)$ 。

证明. $\forall x, E(h(X)Y|X = x) = E(h(x)Y|X = x) = h(x)E(Y|X = x)$ 。 \square

 对比式 (2.47) 和式 (2.51), 读者不难发现二者在形式上是一致的。在给定 $X = x$ 的条件下 Y 的数字特征, 即 $Y|X = x$ 的数字特征, 都是条件分布函数 $F(y|x)$ 的深加工产品, 深加工的手法与从 $F_X(x)$ 定义 X 的数字特征的手法是一致的。所以, 在下文中不再针对条件分布定义方差、原点矩、中心矩等数字特征。

 **定理 2.18 (双期望).** 按照条件期望的定义, $E(Y|X)$ 是关于 X 的函数, 也是一个随机变量, 它的期望是 $E(Y)$ 。即

$$E[E(Y|X)] = E(Y) \quad (2.53)$$

证明. 我们证离散的情形, 连续的情形类似, 留作练习。

$$\begin{aligned} E[E(Y|X)] &= \sum_i P(X = x_i)E(Y|X = x_i) \\ &= \sum_i p_i \sum_j y_j \frac{p_{ij}}{p_i} = \sum_i \sum_j y_j p_{ij} = \sum_j y_j p_{\cdot j} = E(Y) \end{aligned} \quad \square$$

 双期望定理的直观含义是, 在不同条件下考虑 Y 的均值, 然后再算这些均值的均值, 等价于直接计算 Y 的均值。譬如, 令 X, Y 分别表示公务员的受教育程度 (单位: 年) 和年收入, 平均年收入 $E(Y)$ 可以这样分两步求得: (1) 对每个受教育程度 $X = x_i$, 计算出这类公务员的平均年收入 $E(Y|X = x_i)$; (2) 按照 X 的分布, 计算 $E(Y|X = x_i), i = 1, 2, \dots, n$ 的均值。

推论 2.2. 已知 h 是 X 的函数, 则

$$\textcircled{1} \quad E(h(X)Y) = E(h(X)E(Y|X))$$

$$\textcircled{2} \quad E[h(X)(Y - E(Y|X))] = 0$$

证明. 由定理 2.18, $E(h(X)Y) = E[E(h(X)Y|X)]$, 再利用性质 2.31 便证得 ①。② 可由性质 2.31 以及 $E[Y - E(Y|X)] = 0$ 推出。 \square

例 2.62. 接着例 2.61, 验证 $E[E(Y|X)] = E(Y)$ 如下。

$$E[E(Y|X)] = E\left[\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(X - \mu_1)\right] = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}[E(X) - \mu_1] = \mu_2 = E(Y)$$

譬如, 若 $(X, Y)^T \sim N(0, 0, 1, 4, 0.8)$, 则 $Y - E(Y|X) = Y - 1.6X$ 。由推论 2.2 的 ② 可知 $E[X(Y - 1.6X)] = 0$, 而 $(X, Y - 1.6X)^T$ 的散点图见第 207 页的图 2.47 中的右图。

例 2.63. 已知随机变量 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 和 $Z \sim q\langle 1 \rangle + (1-q)\langle 0 \rangle$, 则 $Y = XZ \sim pq\langle 1 \rangle + (1-pq)\langle 0 \rangle$ 。下面验证 $E(E(Y|X)) = E(Y)$, 事实上,

$$E(Y|X=x) = E(XZ|X=x) = qx$$

$$E(E(Y|X)) = E(qX) = pq = E(Y)$$

定理 2.19. 对于随机变量 X 和 Y , 若函数 g 满足 $E[Y - g(X)]^2 < \infty$, 则

$$E[Y - g(X)]^2 = E[Y - E(Y|X)]^2 + E[g(X) - E(Y|X)]^2 \quad (2.54)$$

证明. 因为 $g(X) - E(Y|X)$ 是 X 的函数, 利用推论 2.2 的结果 ②, 有 $E[(Y - E(Y|X))(g(X) - E(Y|X))] = 0$ 。因此,

$$\begin{aligned} E[Y - g(X)]^2 &= E[Y - E(Y|X) + E(Y|X) - g(X)]^2 \\ &= E[Y - E(Y|X)]^2 + E[g(X) - E(Y|X)]^2 \end{aligned} \quad \square$$

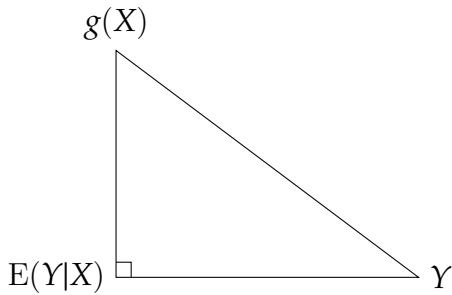


图 2.38: 公式 (2.54) 可以想象成“勾股定理”, $E[Y - g(X)]^2$ 就是斜边边长的平方, 能够分解为两条直角边边长的平方和, 即 $E[Y - E(Y|X)]^2 + E[g(X) - E(Y|X)]^2$ 。

定义 2.36. 作为 Y 的估计, $g(X)$ 的均方误差 (mean squared error, MSE) 和均方根误差 (root mean squared error, RMSE) 分别定义为

$$MSE = E(Y - g(X))^2 \quad RMSE = \sqrt{E(Y - g(X))^2}$$

练习 2.26. 在定理 2.19 的条件下, 总有 $E[Y - E(Y|X)]^2 \leq E[Y - g(X)]^2$, 其中等号成立当且仅当 $P\{g(X) = E(Y|X)\} = 1$ 。提示: 利用定理 2.19 和性质 2.30。

 从练习 2.26 不难看出, 如果用均方误差 $E(Y - g(X))^2$ 来度量随机变量 Y 和 $g(X)$ 间的差距, 在由 X 构造的所有随机变量中, $E(Y|X)$ 是唯一最接近 Y 的。就如同用 $E(Y - c)^2$ 来度量 Y 与任意实数 c 间的差距, $E(Y)$ 是唯一最接近 Y 的实数。

2.3.3 方差与条件方差

对于一个随机变量 X , 我们可以用 $[X - E(X)]^2$ 来刻画随机变量 X 离开均值 $E(X)$ 的幅度*, 这幅度是一个非负的随机变量, 其均值在直观上可视为 X 对点 $E(X)$ 的平均平方差异。由此我们得到了随机变量的一个新的数字特征——方差 (variance)。

定义 2.37 (方差). 设随机变量 X 的分布函数为 $F(x)$, 如果 $E(X)$ 和 $E(X^2)$ 皆存在, 则 X 的方差[†] $V(X)$ 定义为

$$V(X) = E[X - E(X)]^2 = \int_{-\infty}^{+\infty} (x - E(X))^2 dF(x) \quad (2.55)$$

将 X 分为离散型和连续型两种情形, 式 (2.55) 具体为

$$V(X) = \begin{cases} \sum_j (x_j - E(X))^2 p_j & \text{离散型} \\ \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx & \text{连续型} \end{cases}$$

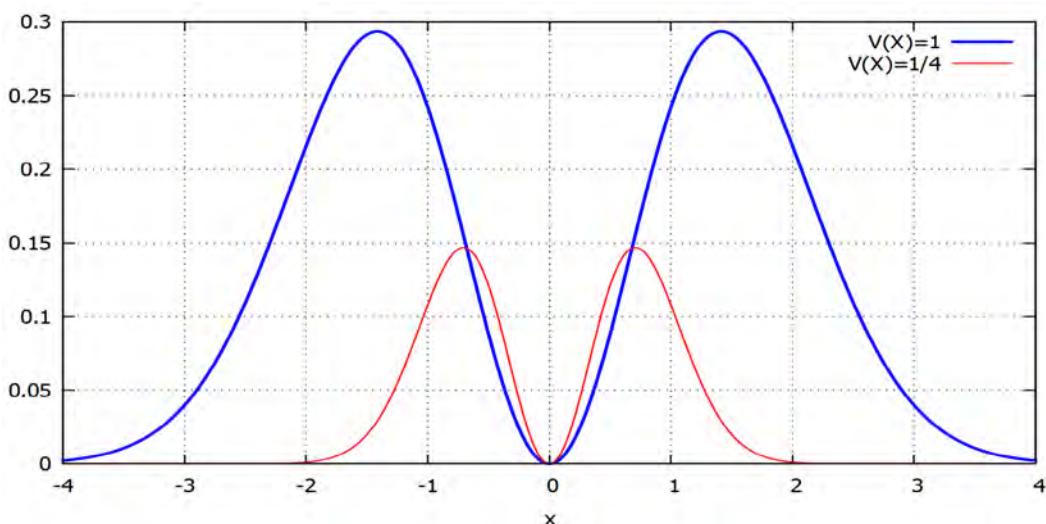


图 2.39: 曲线 $x^2\phi(x)$ 和 $x^2\phi(x|0, 1/4)$: $(x - \mu)^2 f(x)$ 与 x 轴围成的面积就是方差。

式 (2.55) 可进一步简化为

$$V(X) = E\{X^2 - 2X \cdot E(X) + [E(X)]^2\} = E(X^2) - [E(X)]^2 \quad (2.56)$$

式 (2.56) 常作为方差的等价定义用于实际计算。有时 X 的方差 $V(X)$ 也记作 σ_X^2

*当然, 我们也可以用平均差异 $|X - E(X)|$ 来衡量随机变量 X 的取值在 $E(X)$ 周围的分散程度, 但出于计算上的考虑, 方差更便利一些。

[†]有些文献中把 X 的方差记作 $\text{var}(X)$ 、 $D(X)$ 、 σ_X^2 等, 本书采用最流行的记号 $V(X)$ 。

或者 σ^2 。方差的算术平方根 $\sigma_X = \sqrt{V(X)}$ 称为 X 的标准差 (standard deviation)。直观上，方差或标准差可作为计量随机变量在其均值周围分散程度的一个尺度。

 **性质 2.32.** 若随机变量 X 的方差 $V(X)$ 存在，则有 $V(aX + b) = a^2V(X)$ ，其中 $\forall a, b \in \mathbb{R}$ 。另外， $V(X) \leq E(X^2)$ ，更一般地有

$$V(X) \leq E(X - c)^2, \text{ 其中 } \forall c \in \mathbb{R} \quad (2.57)$$

证明. $V(aX + b) = E[aX + b - E(aX + b)]^2 = a^2E[X - E(X)]^2 = a^2V(X)$ 。

$$\begin{aligned} E(X - c)^2 &= E[X - E(X) + E(X) - c]^2 \\ &= E[X - E(X)]^2 + 2E[X - E(X)] \cdot [E(X) - c] + [E(X) - c]^2 \\ &= V(X) + [E(X) - c]^2 \geq V(X) \end{aligned} \quad \square$$

 从 $V(X + b) = V(X)$ 不难看出，平移一个随机变量，丝毫不能改变其取值的分散程度，即方差是一个平移不变量。不等式 (2.57) 意味着函数 $f(c) = E(X - c)^2$ 在 $c = E(X)$ 处取最小值 $V(X)$ 。类似地，若已知 X, Y 的联合分布，在给定 $X = x$ 的条件下， $E(Y - c)^2$ 在 $c = E(Y|X = x)$ 处取得最小值，该最小值称为 $Y|X = x$ 的条件方差，即

$$\begin{aligned} V(Y|X = x) &= E([Y - E(Y|X = x)]^2|X = x) \\ &= \int_{-\infty}^{+\infty} (y - E(Y|X = x))^2 dF(y|x) \end{aligned}$$

 **定理 2.20.** 如果随机变量 Y 的方差存在，则对任意随机变量 X 皆有

$$V(Y|X) = E(Y^2|X) - [E(Y|X)]^2 \quad (2.58)$$

$$V(Y) = E\{V(Y|X)\} + V\{E(Y|X)\} \quad (2.59)$$

$$V\{E(Y|X)\} \leq V(Y), \text{ 其中等号成立当且仅当 } P\{Y = E(Y|X)\} = 1 \quad (2.60)$$

证明. $V(Y|X)$ 和 $E(Y|X)$ 都是由 X 定义的随机变量，与 (2.56) 的证明类似，我们有

$$\begin{aligned} V(Y|X) &= E\{[Y - E(Y|X)]^2|X\} \\ &= E\{Y^2 - 2YE(Y|X) + [E(Y|X)]^2|X\} \\ &= E(Y^2|X) - 2[E(Y|X)]^2 + [E(Y|X)]^2 \\ &= E(Y^2|X) - [E(Y|X)]^2 \end{aligned}$$

利用定理 2.19, 令 $g(X) = E(Y)$, 将之代入式 (2.54) 得到,

$$\begin{aligned} V(Y) &= E(Y - E(Y|X))^2 + E[E(Y|X) - E(Y)]^2 \\ &= E(E([Y - E(Y|X)]^2|X)) + E[E(Y|X) - E(E(Y|X))]^2 \\ &= E\{V(Y|X)\} + V\{E(Y|X)\} \end{aligned}$$

不等式 (2.60) 由式 (2.59) 直接推出。等号成立当且仅当

$$E\{V(Y|X)\} = E(Y - E(Y|X))^2 = 0$$

根据性质 2.30, $V(Y) = V\{E(Y|X)\} \Leftrightarrow P\{Y = E(Y|X)\} = 1$ 得证。 \square

例 2.64. 接着例 2.63, 验证 $V(Y) = E\{V(Y|X)\} + V\{E(Y|X)\} = pq(1 - pq)$ 。事实上,

$$\begin{aligned} V\{E(Y|X)\} &= V(qX) = q^2V(X) = q^2p(1 - p) \\ V(Y|X = x) &= E([XZ - qx]^2|X = x) = x^2E[(Z - q)^2] = x^2V(Z) = x^2q(1 - q) \\ E\{V(Y|X)\} &= E\{q(1 - q)X^2\} = pq(1 - q) \end{aligned}$$

1853 年, 法国统计学家 Irénée-Jules Bienaym  (1796-1878) 证明了下面的结果 (2.61), 称为 Bienaym  公式。

~定理 2.21 (Bienaym , 1853). 若随机变量 X_1, X_2, \dots, X_n 相互独立, 并且它们的方差都存在, 则随机变量之和的方差等于各自方差之和。即,

$$V\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n V(X_j) \quad (2.61)$$

证明. 只证 $n = 2$ 的情形: 由式 (2.56) 和性质 2.29,

$$\begin{aligned} V(X_1 + X_2) &= E[(X_1 + X_2)^2] - [E(X_1 + X_2)]^2 \\ &= E(X_1^2) + E(X_2^2) + 2E(X_1)E(X_2) - [E(X_1)]^2 - [E(X_2)]^2 - 2E(X_1)E(X_2) \\ &= V(X_1) + V(X_2) \end{aligned} \quad \square$$

~性质 2.33. $V(X) = 0$ 当且仅当 X 服从单点分布 $P\{X = E(X)\} = 1$, 即 X 几乎必然等于常数 $E(X)$ 。

证明. 令 $Y = X - E(X)$, 利用性质 2.30 可证。 \square

例 2.65. 求服从二项分布的随机变量 $X \sim B(n, p)$ 的方差。

解. 利用式 (2.56), 首先计算 $E(X^2)$,

$$\begin{aligned} E(X^2) &= \sum_{k=0}^n k^2 C_n^k p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^n k C_{n-1}^{k-1} p^{k-1} (1-p)^{n-k} \\ &= np \left[1 + \sum_{k=1}^n (k-1) C_{n-1}^{k-1} p^{k-1} (1-p)^{n-k} \right], \text{ 根据例 2.54 的结果} \\ &= np[1 + (n-1)p] = np(1-p+pn) \end{aligned}$$

于是, $V(X) = E(X^2) - [E(X)]^2 = np(1-p+pn) - n^2 p^2 = np(1-p)$ 。

例 2.66. 求均匀分布 $X \sim U[a, b]$ 的方差。

解. 由例 2.56 知 $E(X) = (b+a)/2$,

$$E(X^2) = \int_0^1 \frac{x^2}{b-a} dx = \frac{b^2 + ab + a^2}{3}$$

于是, $V(X) = E(X^2) - [E(X)]^2 = (b-a)^2/12$ 。

例 2.67. 试证明: 正态分布 $X \sim N(\mu, \sigma^2)$ 的方差就是尺度参数 σ^2 , 并且 σ^2 越小, 概率 $P\{X \in (\mu-\epsilon, \mu+\epsilon)\}$ 越大, 其中 ϵ 是一个给定的正数。

证明. 根据下面的结果不难算得 $E(X^2) = \sigma^2 + \mu^2$, 进而 $V(X) = \sigma^2$ 。

$$\int_{-\infty}^{+\infty} x^2 \exp\left\{-\frac{x^2}{\sigma^2}\right\} dx = \frac{\sigma^3 \sqrt{\pi}}{2} \quad (2.62)$$

另外, 从 $P\{X \in (\mu-\epsilon, \mu+\epsilon)\} = \Phi(\epsilon/\sigma) - \Phi(-\epsilon/\sigma) = 2\Phi(\epsilon/\sigma) - 1$ 不难看出 σ^2 越小 $P\{X \in (\mu-\epsilon, \mu+\epsilon)\}$ 越大。 \square

练习 2.27. 已知 $(X, Y)^\top \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, 其中 $|\rho| < 1$, 求 $Y|X=x$ 的条件方差 $V(Y|X=x)$ 并与 $V(Y) = \sigma_2^2$ 做比较。

提示: 根据例 2.61 的结果, $V(Y|X=x) = (1-\rho^2)\sigma_2^2$ 。显然, 在得知 $X=x$ 的信息之后, Y 的条件方差比 σ_2^2 要小些, 这得益于已知信息。

练习 2.28. 已知连续型随机变量 X 的密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } x < 0 \\ \frac{x^m}{m!} e^{-x} & \text{当 } x \geq 0, \text{ 其中 } m \text{ 为自然数} \end{cases}$$

求 $E(X)$ 和 $V(X)$ 。答案: $E(X) = V(X) = m+1$ 。

定义 2.38. 如果随机变量 Y 满足 $E(Y) = 0$ 且 $V(Y) = 1$, 则称 Y 为标准化的 (normalized) 随机变量。

例 2.68. 性质 2.12 中, $Y = (X - \mu)/\sigma \sim N(0, 1)$ 就是标准化的随机变量。

练习 2.29. 已知随机变量 X 的期望和标准差分别为 $\mu = E(X)$ 和 $\sigma = \sqrt{V(X)}$, 则 $Y = (X - \mu)/\sigma$ 是 X 经过标准化得到的随机变量。

2.3.4 熵、互信息和 Kullback-Leibler 散度

设信号为一个离散型随机变量 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \cdots + p_n\langle x_n \rangle + \cdots$, 为了提高通信效率, 信号 x_j 出现的概率 p_j 越大, 在通信中对 x_j 的编码长度就应该越短。例如, 在自然语言中, 一般来说越常用的词语其音节也越短, 例如, 我、I、je、ich 等。为了便于交流, 有时甚至人为地制造缩略语来减少音节, 例如“彩电”是“彩色电视机”的缩略语, MIDI 是 Musical Instrument Digital Interface 的缩略语。

为刻画平均编码长度, 1948 年, 信息论之父、美国电子工程师 Claude Elwood Shannon (1916-2001) 将热力学中熵^{*}的概念引入信息论, 在《通信的数学理论》[143] 一文中定义了离散型随机变量的熵。熵对信息进行了量化, 熵越大表示信息的不确定性程度越高, 平均编码长度越长。用于数据压缩技术中的 Huffman 编码就是基于熵的这一特点(见 Cormen 等人的《算法导论》[27] 中 Huffman 编码的贪心算法)。



定义 2.39 (Shannon 熵). 离散型随机变量 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \cdots + p_n\langle x_n \rangle + \cdots$ 的 Shannon 熵(或离散熵), 简称熵(entropy), 定义为 $-\ln X$ 的期望, 即

$$H(X) = E(-\ln X) = - \sum_{j=1}^{\infty} p_j \ln p_j \quad (2.63)$$

在式 (2.63) 中, $-\ln p_j$ 是信号 x_j 的编码长度(与 $-\log_k p_j$ 相差一个常数因子, 其中 $k \geq 2$), 所以 $H(X)$ 也可看作是编码长度的期望, 即平均编码长度(为方便讨论, 我们约定 $0 \ln 0 = 0$), 总是非负的。

编码长度的单位是比特(bit), 它是信息量的最小单位, 即二进制中的一位。例如, 二进制数 0101 是四比特。一个字节(byte)是八比特(换算成十进制就是 0 至 255 的自然数)。ASCII 码中, 一个英文字母占一个字节的空间, 一个汉字占两个字节的空间。在 Unicode 编码中, 中英文字符都是两个字节, 可以表示 $2^{16} = 65536$ 个字符。例如, “熵”的 Unicode 编码是 71B5。

抛一枚均匀的硬币, 其结果要么是 0 要么是 1, 熵为 $-\log_2 \frac{1}{2} = 1$ 比特。抛这枚硬币 n 次, 其结果是一个长度为 n 的 0-1 串, 熵为 n 比特。

练习 2.30. 对随机变量 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \cdots + p_n\langle x_n \rangle$, 试证明: $H(X) \leq \ln n$, 并且当 $p_1 = p_2 = \cdots = p_n = 1/n$ 时, 熵 $H(X)$ 最大。提示: 利用 Jensen 不等式(F.1)

^{*}在热力学中, 系统的状态越有序, 熵越小; 越无序熵越大。热力学第二定律说, 在隔离或绝热的条件下, 系统总是自发地从有序走向无序, 直至达到熵最大的平衡状态。

可得

$$\sum_{j=1}^n \alpha_j \ln \frac{1}{x_j} \leq \ln \sum_{j=1}^n \frac{\alpha_j}{x_j}, \text{ 其中 } \alpha_j > 0 \text{ 满足 } \sum_{j=1}^n \alpha_j = 1$$

例 2.69 (连续熵). 作为 Shannon 熵的推广, 密度函数为 $f(x)$ 的连续型随机变量 X 的连续熵或微分熵 (differential entropy) 定义为

$$H(X) = E(-\ln f(X)) = - \int_{-\infty}^{+\infty} f(x) \ln f(x) dx \quad (2.64)$$

离散熵 (即 Shannon 熵) 满足非负性, 而连续熵 (2.64) 则不满足, 例子见练习 2.31。

例 2.70. 利用公式 (2.62), 求得连续型随机变量 $X \sim N(\mu, \sigma^2)$ 的连续熵是

$$H(X) = \frac{1}{2} \ln(2\pi e \sigma^2)$$

正态分布的熵与位置参数 μ 无关, 只与方差 σ^2 有关。显然, σ^2 越大, 密度函数 $\phi(x|\mu, \sigma^2)$ 的曲线越扁平, X 的不确定性越高, 熵越大。

练习 2.31. 求随机变量 $X \sim U[0, 1/2]$ 的连续熵。答案: $H(X) = -\ln 2$ 。

定义 2.40 (联合熵). 设离散型随机向量 $(X, Y)^\top$ 的分布列为 $P(X = x_i, Y = y_j) = p_{ij}$, 其中 $i = 1, 2, \dots, m, \dots, j = 1, 2, \dots, n, \dots$ 。仿照式 (2.63), 随机向量 $(X, Y)^\top$ 的熵, 也称之为 X 与 Y 的联合熵, 定义如下:

$$H(X, Y) = - \sum_{i,j=1}^{\infty} p_{ij} \ln p_{ij}$$

类似地, 若 $f(x, y)$ 是连续型随机向量 $(X, Y)^\top$ 的密度函数, 则联合熵定义为

$$H(X, Y) = - \iint_{\mathbb{R}^2} f(x, y) \ln f(x, y) dx dy$$

例 2.71. 随机向量 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的熵与位置参数 μ_X, μ_Y 无关, 为

$$H(X, Y) = \ln[2\pi e \sqrt{1 - \rho^2} \sigma_X \sigma_Y]$$

定义 2.41 (条件熵). 已知随机变量 Y 的条件密度函数 $f(y|x)$, 定义 Y 在给定条件

$X = x$ 之下的条件熵 (conditional entropy) 为

$$H(Y|X=x) = - \int_{-\infty}^{+\infty} f(y|x) \ln f(y|x) dy$$

显然, $H(Y|X=x)$ 是一个有关 x 的函数 (也可为常数)。我们定义 Y 在条件 X 之下的条件熵为所有 $H(Y|X=x)$ 的加权平均, 即

$$\begin{aligned} H(Y|X) &= \int_{-\infty}^{+\infty} f_X(x) H(Y|X=x) dx \\ &= - \iint_{\mathbb{R}^2} f_X(x) f(y|x) \ln \frac{f(x,y)}{f_X(x)} dx dy \\ &= - \iint_{\mathbb{R}^2} f(x,y) \ln f(x,y) dx dy + \int_{-\infty}^{+\infty} \ln f_X(x) \left[\int_{-\infty}^{+\infty} f(x,y) dy \right] dx \\ &= H(X, Y) - H(X) \end{aligned} \quad (2.65)$$

显然, 若 $H(Y|X=x)$ 是个与 x 无关的常数, 则 $H(Y|X) = H(Y|X=x)$ 。离散型条件熵的定义是类似的, 不再赘述, 请读者模仿上述定义给出。

例 2.72. 已知随机向量 $(X, Y)^T \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 根据例 2.27 和例 2.70 的结果,

$$H(Y|X=x) = \frac{1}{2} \ln[2\pi e(1-\rho^2)\sigma_Y^2]$$

此处, $H(Y|X=x)$ 是个常数, 于是

$$H(Y|X) = H(Y|X=x) = \frac{1}{2} \ln[2\pi e(1-\rho^2)\sigma_Y^2]$$

这个结果也可以根据 例 2.70 和 例 2.71 的结果, 通过 (2.65) 得到。读者不难发现, $H(Y|X) \leq H(Y)$ 。参考第 145 页的图 2.21 所示条件分布 $Y|X=x$ 的几何意义, 上述结果更容易理解。

性质 2.34. 不难证明条件熵具有以下性质:

$$H(Y|X) = H(X|Y) - H(X) + H(Y)$$

$$H(Y|X) \leq H(Y)$$

$$H(Y|X) = H(Y), \text{ 如果 } X, Y \text{ 相互独立}$$

定义 2.42 (互信息). 为了刻画离散型随机变量 X 与 Y 相互关联的程度, Shannon 还

定义了 X 与 Y 的互信息 (mutual information, MI), 记作 $I(X, Y)$ 。

$$I(X, Y) = \sum_{i,j=1}^{\infty} p_{ij} \ln \frac{p_{ij}}{p_i p_{\cdot j}}$$

这个定义可以自然地推广到连续型随机变量, 即

$$I(X, Y) = \iint_{\mathbb{R}^2} f(x, y) \ln \frac{f(x, y)}{f_X(x)f_Y(y)} dx dy$$

性质 2.35. 互信息常用测量两个随机变量 X, Y 之间共享了多少信息,

① 互信息是对称的, 即 $I(X, Y) = I(Y, X)$ 。另外, 互信息可等价地定义为

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \quad (2.66)$$

② 随机变量 X 与 Y 的互信息非负, 即 $I(X, Y) \geq 0$, 其中 $I(X, Y) = 0$ 当且仅当 X, Y 相互独立。

证明. 若 X, Y 是离散型随机变量, 则

$$\begin{aligned} I(X, Y) &= \sum_{i,j=1}^{\infty} p_{ij} \ln p_{ij} - \sum_{i=1}^{\infty} p_{i\cdot} \ln p_{i\cdot} - \sum_{j=1}^{\infty} p_{\cdot j} \ln p_{\cdot j} \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

若 X, Y 是连续型随机变量, 证明是类似的。结论 ② 是稍后即将介绍的定理 2.22 的推论, $I(X, Y) \geq 0$ 并且等号成立当且仅当 $p_{ij} = p_{i\cdot} p_{\cdot j}$ 或者 $f(x, y) = f_X(x)f_Y(y)$ 。请读者补全。 \square

练习 2.32. 接着第 141 页的例 2.24, 计算 X 和 Y 的熵, 以及联合熵和互信息。答案: $H(X) \approx 0.6365, H(Y) \approx 0.6920, H(X, Y) \approx 1.3262, I(X, Y) \approx 0.0023$ 。

练习 2.33. 试证明: 随机变量 X, Y 相互独立当且仅当 $H(X, Y) = H(X) + H(Y)$ 。并且,

$$H(Y) = H(Y|X) + I(X, Y)$$

提示: 利用公式 (2.65) 和 (2.66)。

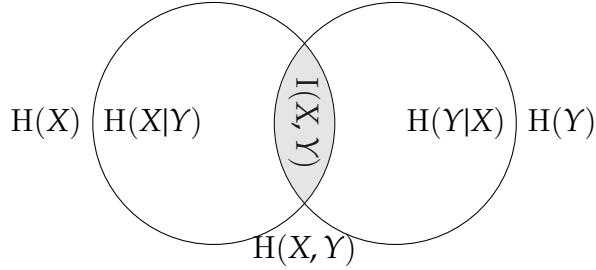


图 2.40: 熵、条件熵、互信息和联合熵的关系。

为刻画两个密度函数 $f(\mathbf{x})$ 和 $g(\mathbf{x})$ 之间的相似程度, 1968 年美国密码专家兼数学家 Solomon Kullback (1907-1994, 左照片) 和 Richard A. Leibler (1914-2003, 右照片) 定义了相对熵 (relative entropy) 或 Kullback-Leibler 信息散度 (information divergence), 简称 Kullback-Leibler 散度如下。



定义 2.43 (Kullback-Leibler 信息散度). 已知两个密度函数 $f(\mathbf{x})$ 和 $g(\mathbf{x})$, 其中 $\mathbf{x} \in \mathbb{R}^d$, 则 g 到 f 的 Kullback-Leibler 信息散度定义为

$$K(f/g) = E_f \left\{ \ln \frac{f(\mathbf{X})}{g(\mathbf{X})} \right\} = \int_{\mathbb{R}^d} f(\mathbf{x}) \ln \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} \quad (2.67)$$

这个定义可以推广到离散型随机向量。譬如, 若 X, Y 是离散型随机变量, 分布列分别为 $X \sim p_1\langle 1 \rangle + \cdots + p_j\langle j \rangle + \cdots$, $Y \sim q_1\langle 1 \rangle + \cdots + q_j\langle j \rangle + \cdots$, 则定义 Y 到 X 的 Kullback-Leibler 信息散度为

$$K(X/Y) = \sum_{j=1}^{\infty} p_j \ln \frac{p_j}{q_j}$$

例 2.73. 令 f, g 分别是分布 $N(0, 1)$ 和 $\text{Laplace}(0, 1)$ 的密度函数, 由**定义 2.43**,

$$\begin{aligned} K(f/g) &= \ln \sqrt{\frac{2}{\pi}} + \sqrt{\frac{2}{\pi}} - \frac{1}{2} \approx 0.0721 \\ K(g/f) &= \ln \sqrt{\frac{\pi}{2}} \approx 0.2258 \end{aligned}$$

 有的文献中将 Kullback-Leibler 散度称为“Kullback-Leibler 距离”或“信息距离”, 但它不是真正意义上的“距离”。距离是一个满足非负性、对称性和三角不等式的二元关系, 而 Kullback-Leibler 散度不满足对称性, 即 $K(f/g) \neq K(g/f)$ (见上例)。从**定义 2.42** 不难看出, $I(X, Y)$ 即是概率函数 $P(X = x_i)P(Y = y_j)$ 到

$P(X = x_i, Y = y_j)$, 或者密度函数 $f_X(x)f_Y(y)$ 到 $f(x, y)$ 的 Kullback-Leibler 信息散度。

定理 2.22. Kullback-Leibler 散度非负, 即 $K(f/g) \geq 0$ 。等号成立当且仅当 $f(\mathbf{x}) = g(\mathbf{x})$ 。

证明. 根据 $-\ln x$ 是凸函数的事实以及 Jensen 不等式 (见第 777 页的定理 F.2),

$$K(f/g) = - \int_{\mathbb{R}^d} f(\mathbf{x}) \ln \frac{g(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x} \geq -\ln \int_{\mathbb{R}^d} f(\mathbf{x}) \frac{g(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x} = -\ln 1 = 0 \quad \square$$

1961 年, 匈牙利数学家 Alfréd Rényi (1921-1970) 提出一类更广泛的熵 [133], Shannon 熵是它的特例。

定义 2.44 (Rényi 熵). 离散型随机变量 $X \sim p_1\langle x_1 \rangle + \cdots + p_n\langle x_n \rangle$ 的 α 阶 Rényi 熵定义为

$$H_\alpha(X) = \frac{1}{1-\alpha} \ln \left(\sum_{j=1}^n p_j^\alpha \right), \text{ 其中 } \alpha \geq 0 \text{ 且 } \alpha \neq 1$$

性质 2.36. $\lim_{\alpha \rightarrow 1} H_\alpha(X) = H(X)$, 记作 $H_1(X)$ 。另外,

$$\begin{aligned} H_0(X) &= \ln n \geq H_1(X) \\ &\geq H_2(X) = -\ln \left(\sum_{j=1}^n p_j^2 \right) \\ &\geq H_\infty(X) = \min_j (-\ln p_j) = -\ln \left(\max_j p_j \right) \end{aligned}$$

证明. 当 $\alpha \rightarrow 1$ 时, $H_\alpha(X)$ 是 $\frac{0}{0}$ 型, 利用 L'Hôpital 法则可得 $H_1(X) = H(X)$ 。其他结果, 利用 Jensen 不等式即可证得。 \square

2.3.5 原点矩、中心矩和绝对矩

矩 (moment) 也是概率分布重要的数字特征，它是期望和方差的自然推广。本节讨论三类矩：原点矩、中心矩和绝对矩。

定义 2.45 (矩). 令 $k \in \mathbb{N}$, 随机变量 X 的 k 阶原点矩 (简称矩) m_k , k 阶中心矩 (central moment) μ_k 和 k 阶绝对矩 β_k 分别定义为

$$\text{矩: } m_k = E(X^k) = \int_{-\infty}^{+\infty} x^k dF_X(x) = \begin{cases} \sum_{j=1}^{\infty} x_j^k p_j & \text{离散型} \\ \int_{-\infty}^{+\infty} x^k f_X(x) dx & \text{连续型} \end{cases}$$

中心矩: $\mu_k = E[X - E(X)]^k$, 显然 $\mu_1 = 0, \mu_2 = V(X)$

绝对矩: $\beta_k = E(|X|^k)$

练习 2.34. 由**定义 2.45**, 验证以下中心矩和矩的关系。

$$\mu_2 = m_2 - m_1^2$$

$$\mu_3 = m_3 - 3m_1 m_2 + 2m_1^3$$

$$\mu_4 = m_4 - 4m_1 m_3 + 6m_1^2 m_2 - 3m_1^4$$

 存在不同的分布函数, 其各阶矩都相等, 即矩不足以确定分布。譬如, W. Feller 在 [46] 中指出对数正态分布 (见本书第 4 章) 不被它的各阶矩唯一决定。

定义 2.46. 令随机变量 X 的期望和标准差分别为 μ 和 $\sigma > 0$, 三阶、四阶中心矩为 μ_3, μ_4 。随机变量 X 的以下数字特征也是常用的:

 偏度系数 (coefficient of skewness) 刻画了随机变量关于期望的对称程度。

$$c_s = \frac{\mu_3}{\sigma^3}$$

显然, 若 X 的密度函数或概率函数关于 $E(X)$ 对称, 则 $c_s = 0$ 。统计学之父 K. Pearson 曾建议用 $3[E(X) - M(X)]/\sqrt{V(X)}$ 定偏度系数, 但未流行。

 峰度系数 (coefficient of kurtosis) 常用来衡量单峰的密度函数曲线顶部与正态密度函数曲线顶部的相对陡峭程度。

$$c_k = \frac{\mu_4}{\sigma^4} - 3$$

□ 变异系数 (coefficient of variation) 是随机变量取值分散程度的（无量纲的）相对度量。

$$c_v = \frac{\sigma}{\mu}$$

例 2.74. 参考第 128 页的练习 2.9 可知, $f_X(x|\alpha, \eta) = 2\alpha\phi(\alpha x)\Phi(\eta x)$ 是一个密度函数, 其中参数 $\alpha > 0, \eta > 0$ 。利用符号计算工具验证下面的结果。

$$\begin{aligned} m_1 &= \frac{\eta}{\alpha} \sqrt{\frac{2}{\pi(\alpha^2 + \eta^2)}} & \beta_1 &= \frac{1}{\alpha} \sqrt{\frac{2}{\pi}} \\ m_2 = \beta_2 &= \frac{1}{\alpha^2} & \beta_3 &= \frac{2}{\alpha^3} \sqrt{\frac{2}{\pi}} \\ m_3 &= \frac{\eta(3\alpha^2 + 2\eta^2)}{\alpha^3(\alpha^2 + \eta^2)^{3/2}} \sqrt{\frac{2}{\pi}} & m_4 = \beta_4 &= \frac{3}{\alpha^4} \end{aligned}$$

练习 2.35. 验证正态分布 $N(\mu, \sigma^2)$ 的偏度系数和峰度系数分别为 $c_s = 0$ 和 $c_k = 0$ 。更多分布的偏度系数和峰度系数见第 4 章。

性质 2.37. 如果随机变量 X 的 k 阶绝对矩 β_k 存在, 则原点矩 m_1, \dots, m_k 都存在, 并且绝对矩 $\beta_1, \dots, \beta_{k-1}$ 也存在。

证明. 由 $|x|^{k-1} < |x|^k + 1$ 可得到

$$\int_{-\infty}^{+\infty} |x|^{k-1} dF_X(x) = \int_{-\infty}^{+\infty} (|x|^k + 1) dF_X(x) < \infty$$

于是, β_{k-1} 存在。类似地, $\beta_1, \dots, \beta_{k-2}$ 也都存在。原点矩 m_1, \dots, m_k 的存在性留给读者证明。 \square

定理 2.23 (Lyapunov 不等式). 假设随机变量 X 的 n 阶绝对矩 $E(|X|^n)$ 存在, 则对 $k = 1, 2, \dots, n-1$ 有不等式 $\sqrt[k]{\beta_k} \leq \sqrt[k+1]{\beta_{k+1}}$ 成立, 即

$$\beta_1 \leq \sqrt[2]{\beta_2} \leq \sqrt[3]{\beta_3} \leq \dots \leq \sqrt[n]{\beta_n}$$

证明. 令 r 是任意实数, 则有不等式

$$\int_{-\infty}^{+\infty} \left(r|x|^{\frac{k-1}{2}} + |x|^{\frac{k+1}{2}} \right)^2 dF_X(x) = \beta_{k-1}r^2 + 2r\beta_k + \beta_{k+1} \geq 0$$

于是, 根的判别式 $4\beta_k^2 - 4\beta_{k-1}\beta_{k+1} \leq 0$, 即 $\beta_k^2 \leq \beta_{k-1}\beta_{k+1}$, 进而得到 $\beta_k^{2k} \leq$

$\beta_{k-1}^k \beta_{k+1}^k, k = 1, 2, \dots, n-1$, 其中 $\beta_0 = 1$ 。即

$$\begin{aligned}\beta_1^2 &\leq \beta_0^1 \beta_2^1 \\ \beta_2^4 &\leq \beta_1^2 \beta_3^2 \\ &\vdots \\ \beta_{n-1}^{2(n-1)} &\leq \beta_{n-2}^{n-1} \beta_n^{n-1}\end{aligned}$$

将前 k 个不等式相乘便得到 $\beta_{k-1}^k \leq \beta_k^{k-1}$, 即 $\sqrt[k]{\beta_k} \leq \sqrt[k+1]{\beta_{k+1}}$ 。 \square

2.3.6 概率不等式

不管随机变量 X, Y 独立与否, 由期望的定义, 不难发现

$$|E(XY)| = \left| \int_{\mathbb{R}^2} xy dF(x, y) \right| \leq E|XY| = \int_{\mathbb{R}^2} |xy| dF(x, y)$$

下面的定理实际上是 Cauchy-Schwarz 不等式 $|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$ 的期望版, 其中 $\langle \cdot, \cdot \rangle$ 是定义在线性空间 V 上的内积*, $x, y \in V$ 。

 定理 2.24 (Cauchy-Schwarz 不等式). 对于任意随机变量 X 和 Y , 皆有

$$E|XY| \leq \sqrt{E(X^2)E(Y^2)} \quad (2.68)$$

证明. 若 $E(Y^2) = 0$, 由性质 2.30 易证得等号成立。不妨设 $E(Y^2) > 0$,

$$E(|X| + t|Y|)^2 = E(X^2) + 2tE|XY| + t^2E(Y^2) \geq 0$$

上式是关于 t 的二次多项式, 因为它是非负的, 所以必有根的判别式 $\Delta = 4(E|XY|)^2 - 4E(X^2)E(Y^2) \leq 0$, 得证。 \square

 式 (2.68) 之所以称为 Cauchy-Schwarz 不等式的期望版, 是因为 $\langle X, Y \rangle = E(XY)$ 定义了一个内积, 只要满足 $\langle X, X \rangle = 0$ 当且仅当 $X \stackrel{a.s.}{=} 0$, 即 $P(X = 0) = 1$ 。站得更高一点看, 式 (2.68) 是下述结果的推论。

 定理 2.25 (Hölder 不等式). 已知实数 $r, s > 1$ 满足 $1/r + 1/s = 1$, 则

$$E|XY| \leq \{E|X|^r\}^{1/r} \{E|Y|^s\}^{1/s} \quad (2.69)$$

证明. 利用不等式 (2.7), 用 $X\{E|X|^r\}^{-1/r}$ 替换 X , 用 $Y\{E|Y|^s\}^{-1/s}$ 替换 Y , 则

$$|XY| \leq \frac{1}{r}|X|^r\{E|X|^r\}^{1/r-1}\{E|Y|^s\}^{1/s} + \frac{1}{s}|Y|^s\{E|Y|^s\}^{1/s-1}\{E|X|^r\}^{1/r}$$

上式两边求期望便证得结果 (2.69)。 \square

显然, Cauchy-Schwarz 不等式 (2.68) 是 Hölder 不等式 (2.69) 在 $r = s = 2$ 时的特例。我们称 $\{E|X|^r\}^{1/r}$ 为随机变量 X 的 L_r 范数, 记作 $\|X\|_r$ 。因此, Hölder 不等式 (2.69) 也常表示为

$$\|XY\|_1 \leq \|X\|_r \|Y\|_s$$

*定义在线性空间 V 上的内积 $\langle \cdot, \cdot \rangle$ 就是满足下述条件的二元函数 $V \times V \rightarrow \mathbb{R}$: (1) 对称性 $\langle x, y \rangle = \langle y, x \rangle$; (2) 非负性 $\langle x, x \rangle \geq 0$, 等号仅当 $x = 0$ 时成立; (3) 线性 $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$, 并且 $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$, 其中 $\alpha \in \mathbb{R}$ 。

推论 2.3 (Minkowski 不等式^{*}). 已知实数 $r \geq 1$ 且 $\|X\|_r, \|Y\|_r$ 皆存在, 则

$$\|X + Y\|_r \leq \|X\|_r + \|Y\|_r \quad (2.70)$$

证明. 当 $r = 1$ 时, $|X + Y| \leq |X| + |Y|$ 是显然的。设 $r > 1$, 则

$$|X + Y|^r \leq |X| \cdot |X + Y|^{r-1} + |Y| \cdot |X + Y|^{r-1}$$

令 s 满足 $1/r + 1/s = 1$, 即 $(r - 1)s = r$ 。根据 Hölder 不等式 (2.69), 有

$$\begin{aligned} E|X + Y|^r &\leq \{E|X|^r\}^{1/r} \{E|X + Y|^{(r-1)s}\}^{1/s} + \{E|Y|^r\}^{1/r} \{E|X + Y|^{(r-1)s}\}^{1/s} \\ &= [\{E|X|^r\}^{1/r} + \{E|Y|^r\}^{1/r}] \{E|X + Y|^r\}^{1/s} \end{aligned}$$

无论 $E|X + Y|^r$ 是否为零, 都可以得到 $\{E|X + Y|^r\}^{1/r} \leq \{E|X|^r\}^{1/r} + \{E|Y|^r\}^{1/r}$, 即 Minkowski 不等式 (2.70) 总是成立的。□

定理 2.26 (Jensen 不等式的数学期望版). 已知 $g(\mathbf{x})$ 是 $S \subset \mathbb{R}^d$ 上的凸函数, d 维随机向量 \mathbf{X} 有有限期望 $E\mathbf{X}$ 且 $P(\mathbf{X} \in S) = 1$, 则 $E\mathbf{X} \in S$, 并且 $g(\mathbf{X})$ 的期望存在, 满足

$$g(E\mathbf{X}) \leq E\{g(\mathbf{X})\} \quad (2.71)$$

等号成立当且仅当存在 $c \in \mathbb{R}$ 和 $\mathbf{w} \in \mathbb{R}^d$ 使得 $P\{g(\mathbf{X}) = \mathbf{w}^\top \mathbf{X} + c\} = 1$ 。

证明. 见附录 F 中的定理 F.1 和定理 F.2。□

例 2.75. 因为 $g(x) = x^2$ 在 \mathbb{R} 上是凸函数, 所以 $(E\mathbf{X})^2 \leq E(X^2)$ 。

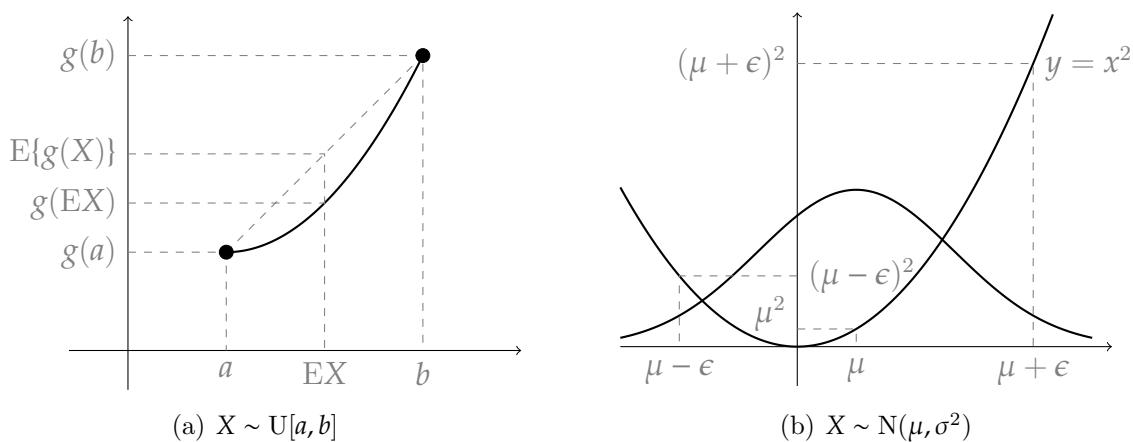


图 2.41: 定理 2.26 的几何解释: 若 g 是凸函数, 则 $g(E\mathbf{X})$ 不会超过 $E\{g(\mathbf{X})\}$ 。图 (b) 解释了简单性质 $(E\mathbf{X})^2 \leq E(X^2)$ 。

^{*}Hermann Minkowski (1864-1909), 德国数学家, Hilbert 的挚友, Einstein 的老师。

推论 2.4. 如果非常数的随机变量 $X > 0$ 具有有限期望, 则

$$[E(X)]^{-1} \leq E(X^{-1})$$

$$E(\ln X) \leq \ln[E(X)]$$

俄国数学家、机械学家 P. L. Chebyshev (1821-1894) 是圣彼得堡数学学派的创始人, 在解析数论、概率论、函数逼近理论、变分法等方面颇多建树。Chebyshev 对概率论的贡献包括: (i) 1866 年发表论文《论均值》, 利用 Chebyshev 不等式证明了 Chebyshev 弱大数律。(ii) 1867 年建立了有关各阶绝对矩一致有界的独立随机变量序列的中心极限定理, 但其证明欠妥, 1898 年经他的学生 A. A. Markov 进一步完善成为中心极限定理的第一个严格证明。1900-1901 年, Chebyshev 的另一个学生 A. M. Lyapunov 利用特征函数给出了更简单的严密证明, 实现了极限定理研究方法的突破。圣彼得堡学派在这个关键问题上的传承极大地推动了概率论的发展。



Markov 发现的下述不等式虽然晚于 Chebyshev 不等式, 但从它可以推导出 Chebyshev 不等式, 为方便陈述, 先简介 Markov 不等式。

引理 2.1 (Markov 不等式). 令非负随机变量 Y 有期望 $E(Y)$, 则 $\forall k > 0$,

$$P(Y \geq k) \leq \frac{E(Y)}{k}, \text{ 或者等价地, } P(Y < k) \geq 1 - \frac{E(Y)}{k} \quad (2.72)$$

证明. 设 Y 的分布函数为 $F(y)$, 则

$$E(Y) = \int_0^{+\infty} y dF(y) \geq \int_k^{+\infty} y dF(y) \geq k \int_k^{+\infty} dF(y) \geq k P(Y \geq k) \quad \square$$

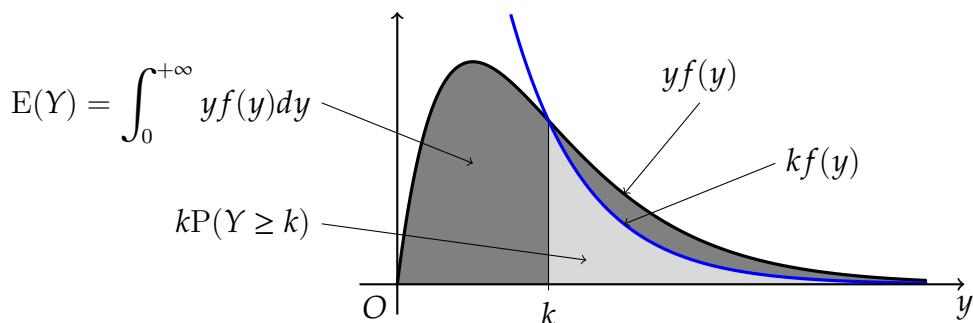


图 2.42: Markov 不等式 (2.72) 的直观解释: 显然, 函数 $yf(y)$ 与 $kf(y)$ 相交于 $y = k$ 处。进而, $kP(Y \geq k) \leq E(Y)$ 。

因为 $P(Y \leq k) \geq P(Y < k)$, 式 (2.72) 在不影响应用的时候也可以减弱为下面的不等式, 同样称为 Markov 不等式。

$$P(Y \leq k) \geq 1 - \frac{E(Y)}{k}$$

推论 2.5 (推广了的 Markov 不等式). 令 $h(X)$ 是随机变量 X 的非负的 Borel 可测函数, 假设 $E\{h(X)\}$ 存在, 则 $\forall k > 0$ 有

$$P\{h(X) \geq k\} \leq \frac{E\{h(X)\}}{k}$$

例 2.76. 假设随机变量 X 的期望 $E(X) = 0$, 方差 $V(X) = \sigma^2$, 试证明:

$$P(X \geq x) \begin{cases} \leq \frac{\sigma^2}{x^2 + \sigma^2} & \text{如果 } x > 0 \\ \geq \frac{x^2}{x^2 + \sigma^2} & \text{如果 } x < 0 \end{cases}$$

证明. 往证第一个不等式: $h(X) = (X+c)^2$ 是一个非负的 Borel 可测函数, 其中 $c > 0$ 。对于 $X \geq x$ 而言总有 $h(X) \geq (x+c)^2$ 。

$$\begin{aligned} P\{X \geq x\} &\leq P\{h(X) \geq (x+c)^2\} \\ &\leq \frac{E(X+c)^2}{(x+c)^2} = \frac{\sigma^2 + c^2}{(x+c)^2} \end{aligned}$$

上式右端在 $c = \sigma^2/x$ 时达到最小值 $\sigma^2/(x^2 + \sigma^2)$ 。第二个不等式可由第一个不等式推出, 请读者尝试用其他方法给出第二个不等式的证明 (留作习题)。 □

定理 2.27 (Chebyshev 不等式*, 1866). 若随机变量 X 的期望 $E(X)$ 和方差 $V(X)$ 都存在, 则 $\forall \epsilon > 0$, 下面的不等式成立并且相互等价。

$$P\{|X - E(X)| \geq \epsilon \sqrt{V(X)}\} \leq \frac{1}{\epsilon^2} \quad (2.73)$$

$$P\{|X - E(X)| \geq \epsilon\} \leq \frac{V(X)}{\epsilon^2} \quad (2.74)$$

$$P\{|X - E(X)| < \epsilon\} \geq 1 - \frac{V(X)}{\epsilon^2} \quad (2.75)$$

证明. 令 $Y = [X - E(X)]^2$, 它是非负随机变量。将它和 $k = \epsilon^2 V(X)$ 代入 Markov 不等式 (2.72) 便可证得 Chebyshev 不等式 (2.73)。Chebyshev 不等式的另两个等价形式 (2.74) 和 (2.75) 也是显然的, 请读者给出证明。 □

*该不等式由法国统计学家 I. Bienaym 于 1853 年首次发现, 由 Chebyshev 于 1866 年再次独立发现并用于证明“Chebyshev 弱大数律”, 习惯上我们把这个不等式称作 Chebyshev 不等式。

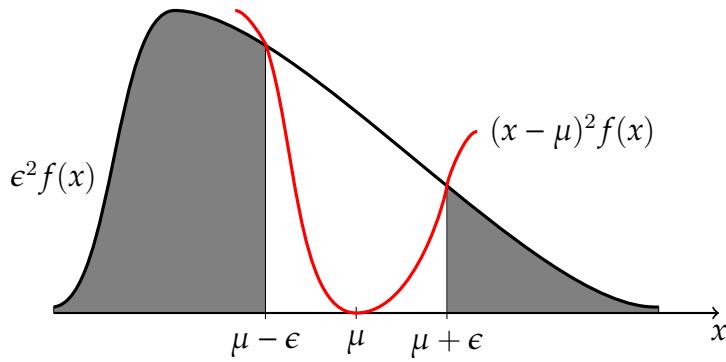


图 2.43: Chebyshev 不等式 (2.74) 的直观解释: 令 $\mu = E(X)$, 显然, 函数 $(x - \mu)^2 f(x)$ 与 $\epsilon^2 f(x)$ 相交于 $x = \mu \pm \epsilon$ 处。在 $(\mu - \epsilon, \mu + \epsilon)$ 之外, $\epsilon^2 f(x) \leq (x - \mu)^2 f(x)$ 。进而, $\epsilon^2 P\{|X - \mu| \geq \epsilon\} \leq V(X)$ 。即阴影部分的面积不超过 $(x - \mu)^2 f(x)$ 与 x 轴围成的面积。

类似于正态分布的 3σ 原则, Chebyshev 不等式 (2.73) 蕴含了任意随机变量 X 距离它的期望 $E(X)$ 不超过 3 个标准差的概率大于 $8/9$ 。Chebyshev 不等式 (2.74) 揭示了越远离期望的值取到的概率越小, 粗略地形容为“中间高两边低”。Chebyshev 不等式 (2.75) 说明了方差越小, X 落于 $(EX - \epsilon, EX + \epsilon)$ 的概率就越大, 至少为 $1 - V(X)/\epsilon^2$ 。

简洁而普适的 Chebyshev 不等式是概率论中的一件常用工具, 它特别适用于随机变量之和的研究 (如证明 Chebyshev 弱大数律, 见定理 5.5), 而不适用于概率的精确估计。即便无法精确估计 $P\{|X - E(X)| < \epsilon\}$, 也不影响 Chebyshev 不等式有像例 2.77 这样有趣的应用。



例 2.77. 假设随机事件 A 在一次 Bernoulli 试验中发生的概率为 $p = 1/2$, 需要独立重复多少次该试验, 才能使得“ $|A$ 出现的频率 $- p| < 0.01$ ”发生的概率至少为 95%?

解. 令随机变量 $X \sim B(n, p)$ 表示事件 A 在 n 重 Bernoulli 试验中出现的次数, 则 A 出现的频率为 X/n , 由 Chebyshev 不等式 (2.75),

$$\begin{aligned} P\left\{\left|\frac{X}{n} - p\right| < \epsilon\right\} &= P\{|X - np| < n\epsilon\} \\ &\geq 1 - \frac{np(1-p)}{(n\epsilon)^2} \geq 0.95 \end{aligned}$$

将 $\epsilon = 0.01, p = 1/2$ 代入, 解此不等式得到 $n \geq 5 \times 10^4$ 。请读者注意, 此解并非精确解, 而只是保证题目要求的结果成立。

Chebyshev 不等式的一个非平凡推广是下面的 Kolmogorov 不等式 (在第 5 章将被用于证明 Kolmogorov 强大数律, 见定理 5.11 的证明), 而 Lévy 不等式 (定

理 2.29) 又是对 Kolmogorov 不等式的推广。

\nwarrow 定理 2.28 (Kolmogorov 不等式, 1928-1929). 如果随机变量 X_1, X_2, \dots, X_n 相互独立且都有有限方差, 则对任意 $\epsilon > 0$ 有以下不等式成立。

$$P\left(\bigcup_{i=1}^n \left\{\left|\sum_{j=1}^i X_j - EX_j\right| \geq \epsilon\right\}\right) \leq \frac{\sum_{i=1}^n V(X_i)}{\epsilon^2} \quad (2.76)$$

如果 X_j 是有界的, 不妨设 $|X_j| \leq c$, 其中 $j = 1, 2, \dots, n$, 则还有

$$1 - \frac{(\epsilon + 2c)^2}{\sum_{i=1}^n V(X_i)} \leq P\left(\bigcup_{i=1}^n \left\{\left|\sum_{j=1}^i X_j - EX_j\right| \geq \epsilon\right\}\right) \quad (2.77)$$

\nwarrow 证明. 令 $Y_j = X_j - EX_j$ 且 $Z_i = \sum_{j=1}^i Y_j$, 令 A_0 表示事件 “ $\forall k \leq n, |Z_k| < \epsilon$ ”。令 A_i 表示事件 “ $\forall k \leq i-1, |Z_k| < \epsilon$ 且 $|Z_i| \geq \epsilon$ ”, 其中 $i = 1, 2, \dots, n$ 。显然, 事件 A_1, A_2, \dots, A_n 是两两互斥的。下面的事件是等价的:

$$\begin{aligned} &\text{至少有一个 } i \ (1 \leq i \leq n) \ \text{使得 } |Z_i| \geq \epsilon \\ \Leftrightarrow & \max_{1 \leq i \leq n} |Z_i| \geq \epsilon \iff \sum_{i=1}^n A_i \end{aligned}$$

因 $V(Z_n) = \sum_{i=1}^n V(X_i)$, 故往证 (2.76) 即往证 $\sum_{i=1}^n P(A_i) \leq V(Z_n)/\epsilon^2$ 。另外, $V(Z_n)$ 具有如下的分解。

$$V(Z_n) = \sum_{i=0}^n P(A_i)E(Z_n^2|A_i) \geq \sum_{i=1}^n P(A_i)E(Z_n^2|A_i)$$

若能证得 $E(Z_n^2|A_i) \geq \epsilon^2$ 便万事大吉。下面验证 ϵ^2 的确是 $E(Z_n^2|A_i)$ 的下界: 将 Z_n 分解为 $Z_n = Z_i + (Y_{i+1} + \dots + Y_n)$, 进而

$$\begin{aligned} Z_n^2 &= Z_i^2 + (Y_{i+1} + \dots + Y_n)^2 + 2Z_i(Y_{i+1} + \dots + Y_n) \\ E(Z_n^2|A_i) &= E\left(Z_i^2 + \sum_{j>i} Y_j^2 + 2 \sum_{j>i} Z_i Y_j + 2 \sum_{k>j>i} Y_k Y_j \middle| A_i\right) \\ &\geq E\left(Z_i^2 + 2 \sum_{j>i} Z_i Y_j + 2 \sum_{k>j>i} Y_k Y_j \middle| A_i\right) \geq \epsilon^2 \end{aligned}$$

最后一步因为 $E(Z_i Y_j|A_i) = E(Z_i|A_i)E(Y_j|A_i) = 0$ 且 $E(Y_k Y_j|A_i) = 0$ 。不等式 (2.77) 的证明见 M. Loève 的《概率论》[107] 第 248 页。 \square

练习 2.36. 验证 Kolmogorov 不等式 (2.76) 还等价于

$$P\left(\max_{1 \leq i \leq n} \left| \sum_{j=1}^i X_j - EX_j \right| \geq \epsilon\right) \leq \frac{\sum_{i=1}^n V(X_i)}{\epsilon^2} \quad (2.78)$$

$$P\left(\bigcap_{i=1}^n \left\{ \left| \sum_{j=1}^i X_j - EX_j \right| < \epsilon \right\}\right) \geq 1 - \frac{\sum_{i=1}^n V(X_i)}{\epsilon^2} \quad (2.79)$$

请读者说明 Kolmogorov 不等式是 Chebyshev 不等式的推广。

推论 2.6. 设 X_1, X_2, \dots, X_n 是期望为 0 且具有有限方差的独立随机变量, 令 $S_k = \sum_{j=1}^k X_j, k = 1, 2, \dots, n$, 则 $\forall \epsilon > 0$ 有以下不等式成立。

$$P\left(\max_{1 \leq k \leq n} |S_k| \geq \epsilon\right) \leq \frac{E(S_n^2)}{\epsilon^2}$$

下面不加证明地补充几个有关随机变量之和的经典不等式, 对它们的证明感兴趣的读者可参阅 M. Loève 的《概率论》[107] 或 W. Feller 的《概率论及其应用》下卷 [46]。

定理 2.29 (Lévy 不等式, 1937). 如果随机变量 X_1, X_2, \dots, X_n 相互独立, 令 $S_k = \sum_{j=1}^k X_j$, 则对任意 $x \in \mathbb{R}$ 有以下不等式成立。

$$\begin{aligned} P\left(\max_{1 \leq k \leq n} [S_k - M(S_k - S_n)] \geq x\right) &\leq 2P\{|S_n| \geq x\} \\ P\left(\max_{1 \leq k \leq n} |S_k - M(S_k - S_n)| \geq x\right) &\leq 2P\{|S_n| \geq x\} \end{aligned}$$

定理 2.30. 接着定理 2.29 的条件, 如果 $g(x) \geq 0$ 是凸的单调函数且 $Eg(|S_n|) < \infty$, 则

$$P\left(\max_{1 \leq k \leq n} |S_k| \geq x\right) \leq \frac{Eg(|S_n|)}{g(x)}$$

定理 2.31 (Bernstein-Kolmogorov 不等式, 1911, 1929). 已知随机变量 X_1, X_2, \dots, X_n 的期望都为 0 且都有界 (不妨设 $|X_j| \leq c$, 其中 $c > 0$ 为常数)。令 $\sigma^2 = V(X_1 + X_2 + \dots + X_n)$, 则 $\forall \epsilon > 0$ 下面的不等式成立。

$$P\{|X_1 + X_2 + \dots + X_n| \geq \epsilon\} \leq 2 \exp\left\{\frac{-\epsilon^2}{2(\sigma^2 + c\epsilon/3)}\right\}$$

1963 年, 美国统计学家 Wassily Hoeffding (1914-1991) 发现了有关独立随机变量之和的下述不等式结果, 常用于有界的随机变量。

定理 2.32 (Hoeffding 不等式 [74], 1963). 已知随机变量 X_1, X_2, \dots, X_n 独立且 $P\{X_j -$

$E(X_j) \in [a_j, b_j]$ } = 1, 其中 $j = 1, 2, \dots, n$ 。令 $S_n = X_1 + X_2 + \dots + X_n$, 则 $\forall \epsilon > 0$ 下面的不等式成立。

$$P\{S_n - E(S_n) \geq \epsilon\} \leq \exp\left\{\frac{-2\epsilon^2}{\sum_{j=1}^n (b_j - a_i)^2}\right\} \quad (2.80)$$

$$P\{|S_n - E(S_n)| \geq \epsilon\} \leq 2 \exp\left\{\frac{-2\epsilon^2}{\sum_{j=1}^n (b_j - a_i)^2}\right\} \quad (2.81)$$

例 2.78. 利用 Hoeffding 不等式 (2.81) 证明 Bernoulli 弱大数律 (第 62 页的定理 1.1)。

证明. 设 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 则 $S_n = X_1 + X_2 + \dots + X_n \sim B(n, p)$ 且 $P\{X_1 - E(X_1) \in [-1, 1]\} = 1$, 利用 Hoeffding 不等式 (2.81) 可证得当 $n \rightarrow \infty$ 时,

$$P\left\{\left|\frac{S_n}{n} - p\right| \leq \epsilon\right\} = P\{|S_n - np| \leq n\epsilon\} \geq 1 - 2 \exp\left\{-\frac{n\epsilon^2}{2}\right\} \rightarrow 1 \quad \square$$

2.4 随机变量之间的关系

不论参数 ρ 取何值, 第 174 页的例 2.61 揭示了二元正态分布的两个随机变量之间存在像结果 (2.52) 那样的确定关系。这个事实首先被英国人类学家、统计学家、地理学家、气象学家、心理学家、遗传学家 Francis Galton (1822-1911) 发现。

1865 年, Galton 受其表兄、《物种起源》的作者 Charles Robert Darwin 的影响转而研究遗传学。1885 年, Galton 考察了 205 对夫妇以及他们的 928 个成年子女的身高, 发现了父代身高 Y 和子代身高 X 呈现出一定的规律。Galton 发现虽然高(矮)的父代产生的子代平均也高(矮), 但有向父代均值退化的趋势, Galton 称之为“回归到平常”——设父代的身高均值为 μ_Y , 身高为 $h > \mu_Y$ 的父代, 其子代身高的均值小于 h 而往 μ_Y 方向回归。1886 年, Galton 发表了著名论文《遗传结构中向中心的回归》, 他论证了 $(X, Y)^\top$ 呈二元正态分布, 并得到了回归直线 (2.52)。

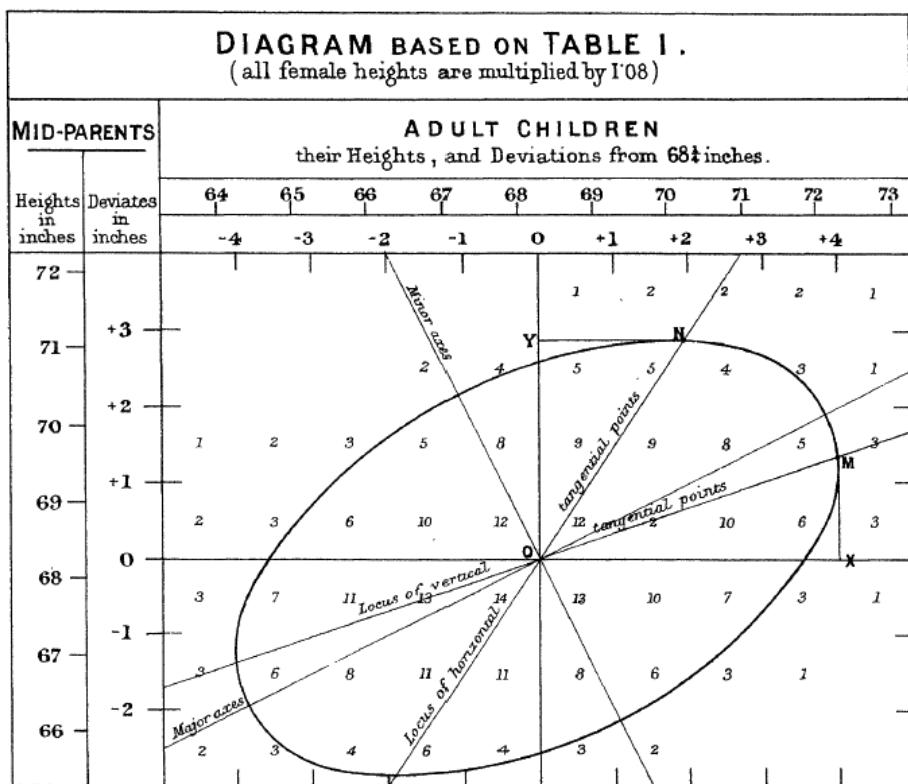


图 2.44: 因为女性的身高一般低于男性, Galton 利用男女平均身高之比把女性的身高乘以 1.08 换算成男性身高。Galton 定义了中亲 (mid-parents) 身高 = $\frac{1}{2}(\text{父亲的身高} + 1.08 \times \text{母亲的身高})$ 来刻画父代的身高。此图来自 Galton 的学生、统计学之父 K. Pearson 的文章《对相关性的历史注记》[119]。

例 2.79. 在两次考试中，第一次成绩最低的学生群体在第二次考试中成绩有升有降，平均成绩将有所提升；类似地，第一次成绩最高的学生群体在第二次考试中平均成绩将有所下降。这就是所谓的“回归效应”，即向均值退化。

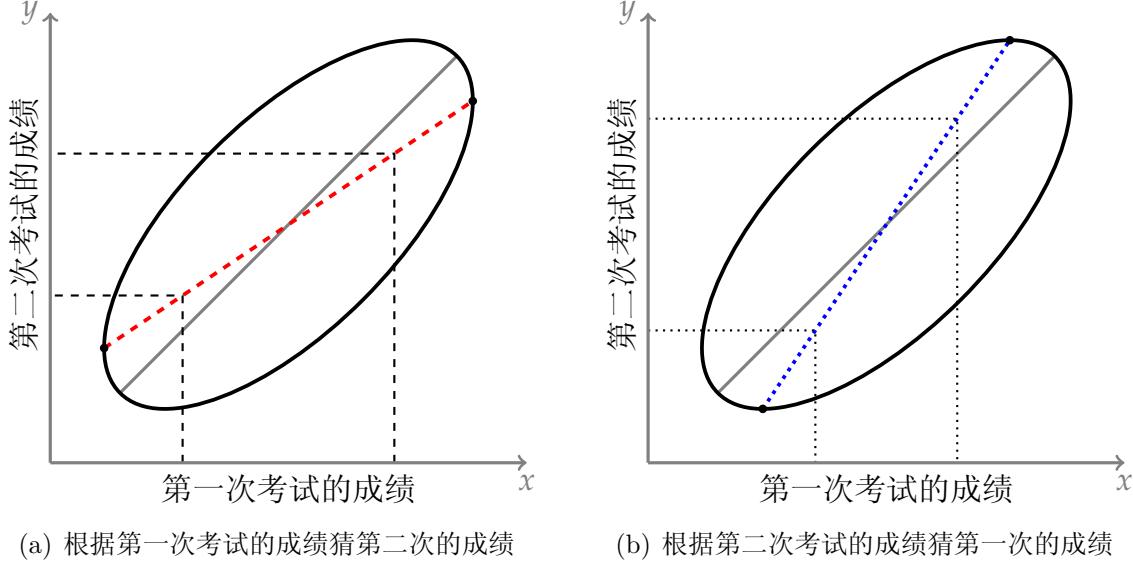


图 2.45：若两次考试的成绩 X 和 Y 的联合分布是正态的，椭圆是联合密度函数的某个非退化等高线。虚线是根据某次考试的成绩猜出另一次考试的成绩。

随机变量 X, Y 的关系早就蕴含在它们的联合分布之中。除了期望和方差，刻画两个随机变量之间关系的还有一个极为重要的数字特征——协方差，定义如下。

定义 2.47 (协方差). 随机向量 $(X, Y)^\top$ 的 (s, t) 阶矩定义为

$$m_{s,t} = E(X^s Y^t), \text{ 其中 } s, t \text{ 为自然数}$$

类似地， (s, t) 阶中心矩定义为

$$\mu_{s,t} = E[(X - E(X))^s (Y - E(Y))^t]$$

例如， $m_{1,0} = E(X), m_{2,0} = E(X^2), \mu_{1,0} = 0, \mu_{2,0} = V(X)$ 。特别地， $\mu_{1,1} = E[(X - EX)(Y - EY)] = E(XY) - E(X)E(Y)$ 被称为 X, Y 的协方差 (covariance)，记作 $Cov(X, Y)$ 。

例 2.80. 已知 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ，计算协方差 $Cov(X, Y)$ 。

解. 二元正态分布 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的详情见例 2.22，其密度函数

$\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 见式 (2.22)。按照定义 2.47,

$$\text{Cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y) \phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) dx dy$$

令 $u = (x - \mu_X)/\sigma_X, v = (y - \mu_Y)/\sigma_Y$, 代入上式作变量替换得

$$\begin{aligned} \text{Cov}(X, Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{uv \sigma_X \sigma_Y}{2\pi \sqrt{1-\rho^2}} \exp \left\{ -\frac{(u-\rho v)^2}{2(1-\rho^2)} - \frac{v^2}{2} \right\} du dv \\ &= \int_{-\infty}^{+\infty} \frac{v \sigma_X \sigma_Y \exp(-v^2/2)}{2\pi \sqrt{1-\rho^2}} \left[\int_{-\infty}^{+\infty} u \exp \left\{ -\frac{(u-\rho v)^2}{2(1-\rho^2)} \right\} du \right] dv \\ &= \frac{\rho \sigma_X \sigma_Y}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} v^2 \exp(-v^2/2) dv, \text{ 根据式 (2.62) 可得} \\ &= \rho \sigma_X \sigma_Y \end{aligned}$$

性质 2.38. 根据式 (2.56) 很容易证得下面的结果,

$$\text{V}(X + Y) = \text{V}(X) + \text{V}(Y) + 2\text{Cov}(X, Y) \quad (2.82)$$

推论 2.7. 由定理 2.21, 若随机变量 X, Y 独立, 则 $\text{Cov}(X, Y) = 0$ 。

定义 2.48. 如果随机变量 X, Y 满足 $\text{Cov}(X, Y) = 0$, 则称它们是不相关的 (uncorrelated)。显然, 若 X, Y 独立, 则它们也是不相关的。但反之不成立, 请看下例。

例 2.81. 已知 $\theta \sim U[-\pi, \pi]$, 往证 $X = \sin \theta, Y = \cos \theta$ 是不相关的。

$$\left. \begin{array}{l} \text{E}(X) = 0 \\ \text{E}(XY) = 0 \end{array} \right\} \Rightarrow \text{Cov}(X, Y) = 0 \Rightarrow \rho(X, Y) = 0$$

然而, 由 $X^2 + Y^2 = 1$ 可知 X, Y 不独立, 这说明“不相关”比“独立”要弱些。

例 2.82. 第 175 页的推论 2.2 的结果 ② 揭示, $h(X)$ 与 $Y - E(Y|X)$ 是不相关的。

练习 2.37. 令随机变量 U, V 具有相同的期望和方差, 请验证

$$\text{Cov}(U + V, U - V) = 0$$

对随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 而言, 最常用的数字特征是如下定义的期望和协方差矩阵, 它们是随机变量的期望和方差的自然推广。

定义 2.49 (随机向量的期望与协方差矩阵). 随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的期望定义为 $E\mathbf{X} = (EX_1, \dots, EX_n)^\top$, 它的方差-协方差矩阵 (variance-covariance matrix) 记作

$\text{Cov}(\mathbf{X}, \mathbf{X})$ 或者 $\Sigma_{\mathbf{XX}}$, 定义为

$$\text{Cov}(\mathbf{X}, \mathbf{X}) = E[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^\top] = (\text{Cov}(X_i, X_j))_{n \times n}$$

更一般地, 两个随机向量 $\mathbf{Y} = (Y_1, \dots, Y_m)^\top$ 与 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的方差-协方差矩阵记作 $\text{Cov}(\mathbf{Y}, \mathbf{X})$ 或者 $\Sigma_{\mathbf{YX}}$, 定义为

$$\text{Cov}(\mathbf{Y}, \mathbf{X}) = E[(\mathbf{Y} - E\mathbf{Y})(\mathbf{X} - E\mathbf{X})^\top] = (\text{Cov}(Y_i, X_j))_{m \times n}$$

显然, $\Sigma_{\mathbf{XY}} = \Sigma_{\mathbf{YX}}^\top$ 。为方便起见, 方差-协方差矩阵通常简称为协方差矩阵。

例 2.83. 由例 2.80 知, $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的协方差矩阵是 $\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$

性质 2.39. 已知 \mathbf{X} 是一个 n 维随机向量, 则对于任意矩阵 $A_{m \times n}$, 有

$$E(A\mathbf{X}) = AE(\mathbf{X}) \quad (2.83)$$

$$V(A\mathbf{X}) = ACov(\mathbf{X}, \mathbf{X})A^\top \quad (2.84)$$

$$\text{Cov}(A_{s \times n}\mathbf{X}, B_{t \times m}\mathbf{Y}) = ACov(\mathbf{X}, \mathbf{Y})B^\top \quad (2.85)$$

特别地, 结果 (2.83) 和 (2.84) 蕴涵 $\forall \mathbf{a} \in \mathbb{R}^n$, 皆有

$$E(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top E(\mathbf{X})$$

$$V(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \text{Cov}(\mathbf{X}, \mathbf{X}) \mathbf{a}$$

证明. 结果 (2.84) 是 (2.85) 的推论。下面往证 (2.85):

$$\text{Cov}(AX, BY) = E\{A[\mathbf{X} - E(\mathbf{X})] \cdot [Y^\top - E(Y^\top)]B^\top\} = ACov(\mathbf{X}, \mathbf{Y})B^\top \quad \square$$

性质 2.40. 已知 \mathbf{X}, \mathbf{Y} 分别是 n 维和 m 维随机向量, 设 $\Sigma_{\mathbf{XX}}, \Sigma_{\mathbf{YY}}$ 都非退化。

$$\text{若 } \mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}, \text{ 则有 } \text{Cov}(\mathbf{Z}, \mathbf{Z}) = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix} \quad (2.86)$$

$$\text{若 } \mathbf{W} = \begin{pmatrix} \Sigma_{\mathbf{XX}}^{-\frac{1}{2}} \mathbf{X} \\ \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{Y} \end{pmatrix}, \text{ 则有 } \text{Cov}(\mathbf{W}, \mathbf{W}) = \begin{pmatrix} I_n & \Sigma_{\mathbf{XX}}^{-\frac{1}{2}} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \\ \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-\frac{1}{2}} & I_m \end{pmatrix} \quad (2.87)$$

证明. 结果 (2.86) 直接由定义可得。下面，利用 (2.85) 来证明 (2.87)。

$$\begin{aligned} \mathbf{W} &= \begin{pmatrix} \Sigma_{\mathbf{XX}}^{-\frac{1}{2}} & O_{n \times m} \\ O_{m \times n} & \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = A\mathbf{Z}, \text{ 其中 } A = \begin{pmatrix} \Sigma_{\mathbf{XX}}^{-\frac{1}{2}} & O_{n \times m} \\ O_{m \times n} & \Sigma_{\mathbf{YY}}^{-\frac{1}{2}} \end{pmatrix} \text{ 是对称矩阵} \\ &\Downarrow \\ \text{Cov}(\mathbf{W}, \mathbf{W}) &= \text{Cov}(A\mathbf{Z}, A\mathbf{Z}) = A\text{Cov}(\mathbf{Z}, \mathbf{Z})A^{\top} \end{aligned}$$
□

练习 2.38. 利用式 (2.84) 将结果 (2.82) 推广为：

$$\text{V}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^n \text{V}(X_i) + 2 \sum_{1 \leq j < k \leq n} \text{Cov}(X_j, X_k) \quad (2.88)$$

定理 2.33. 方阵 $\Sigma_{n \times n}$ 是一个 n 维随机向量的协方差矩阵当且仅当 Σ 对称且半正定（有关半正定矩阵的性质见附录 E 中的定理 E.3）。

证明. 往证 “ \Rightarrow ”：若方阵 $\Sigma = (\sigma_{ij})_{n \times n}$ 是随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^{\top}$ 的协方差矩阵，则 $\sigma_{ij} = \sigma_{ji} = \text{Cov}(X_i, X_j)$ ，并且 $\forall \mathbf{x} \in \mathbb{R}^n$ 皆有

$$\mathbf{x}^{\top} \Sigma \mathbf{x} = \mathbf{x}^{\top} \text{E}[(\mathbf{X} - \text{E}\mathbf{X})(\mathbf{X} - \text{E}\mathbf{X})^{\top}] \mathbf{x} = \text{E}[\mathbf{x}^{\top} (\mathbf{X} - \text{E}\mathbf{X})]^2 \geq 0$$

往证 “ \Leftarrow ”：因为 $\Sigma_{n \times n}$ 对称且半正定，则存在 $n \times k$ 矩阵 A 使得 $\Sigma = AA^{\top}$ ，其中 $1 \leq k \leq n$ 。令 $X_1, X_2, \dots, X_k \stackrel{\text{iid}}{\sim} N(0, 1)$ 且 $\mathbf{X} = (X_1, X_2, \dots, X_k)^{\top}$ 。随机向量 $\mathbf{Y} = A\mathbf{X}$ 的协方差矩阵为

$$\text{E}(\mathbf{Y}\mathbf{Y}^{\top}) = \text{E}[(A\mathbf{X})(A\mathbf{X})^{\top}] = A\text{E}(\mathbf{X}\mathbf{x}^{\top})A^{\top} = AIA^{\top} = AA^{\top} = \Sigma \quad \square$$

练习 2.39. 验证例 2.83 里的协方差矩阵 Σ 是半正定的。

2.4.1 相关系数

定义 2.50 (相关系数). 已知随机向量 $(X, Y)^\top$, $\sigma_X > 0$ 和 $\sigma_Y > 0$ 分别是 X 和 Y 的标准差。随机变量 X 与 Y 之间的相关系数 (correlation coefficient, CC) $\rho(X, Y)$ (也记作 $\rho_{X,Y}$ 或 ρ) 定义为

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\mu_{1,1}}{\sigma_X\sigma_Y} \quad (2.89)$$

例 2.84. 已知 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 计算相关系数 $\rho(X, Y)$ 。

解. 由**例 2.80**, $\text{Cov}(X, Y) = \rho\sigma_X\sigma_Y$ 。按照**定义 2.50**, 于是 $\rho(X, Y) = \rho$ 。

性质 2.41. 随机变量 X, Y 之间的相关系数 $\rho(X, Y)$ 满足 $-1 \leq \rho(X, Y) \leq 1$ 。

证明. 下面往证 $|\rho(X, Y)| \leq 1$ 。由 Cauchy-Schwarz 不等式 (2.68), 不难得到

$$\begin{aligned} |\text{Cov}(X, Y)| &= |\mathbb{E}[(X - EX)(Y - EY)]| \\ &\leq \sqrt{\mathbb{E}[(X - EX)^2]\mathbb{E}[(Y - EY)^2]} \\ &= \sqrt{V(X)V(Y)} \end{aligned}$$

□

性质 2.42. 对于非零实数 a, c 和任意实数 b, d 皆有

$$|\rho(aX + b, cY + d)| = |\rho(X, Y)|$$

即, 线性变换不改变两个随机变量之间相关系数的绝对值。或者说, 相关系数的绝对值是线性变换的不变量。

例 2.85. 已知 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 则随机变量 X, Y 相互独立当且仅当相关系数 $\rho = 0$, 即独立性等价于不相关性。

证明. 不相关性导出独立性几乎是显然的: 将 $\rho = 0$ 代入密度函数 (2.22), 有

$$\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, 0) = \phi(x | \mu_X, \sigma_X^2)\phi(y | \mu_Y, \sigma_Y^2)$$

□

定理 2.34 (线性相关的判定). 随机变量 X, Y 的相关系数 $\rho(X, Y)$ 的绝对值等于 1 当且仅当 X, Y 之间几乎必然存在线性关系, 即

$$\rho^2(X, Y) = 1 \Leftrightarrow \exists (a, b) \in \mathbb{R}^2 \text{ 使得 } P(Y = aX + b) = 1$$

证明. (1) 首先往证 “ \Leftarrow ”: 由式 (2.53), $E(Y) = P(Y = aX + b)E(Y|Y = aX + b) + P(Y \neq$

$aX + b)E(Y|Y \neq aX + b) = aE(X) + b$, 于是

$$\begin{aligned}\sigma_Y^2 &= E(Y - EY)^2 = E[aX - aE(X)]^2 = a^2\sigma_X^2 \\ \mu_{1,1} &= E[(X - EX)(aX - aEX)] = a\sigma_X^2\end{aligned}$$

由**定义 2.50** 得出结论 $\rho^2(X, Y) = 1$, 于是 “ \Leftarrow ” 得证。

(2) 下面往证 “ \Rightarrow ”: 由 $\mu_{1,1}^2 - \sigma_X^2\sigma_Y^2 = 0$ 和**性质 2.41** 的证明细节以及事实 $\sigma_X\sigma_Y > 0$ 得知, 存在唯一的非零实数 r_0 使得

$$E[r_0(X - EX) + (Y - EY)]^2 = 0$$

由**性质 2.30** 可得 $P\{r_0(X - EX) + (Y - EY) = 0\} = 1$ 。 □

 参见第 137 页的**图 2.17**, 相关系数 $|\rho|$ 越接近 1, 分布 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的密度函数曲面就越窄, 随机变量 X, Y 之间的线性关系就越明显。可以说, 相关系数 $\rho(X, Y)$ 刻画的是随机变量 X, Y 之间的线性相关情况: 当 $\rho = \pm 1$ 时, X 与 Y 之间存在线性关系的概率是 1。对于非线性关系的描述, 如 $Y = aX^2 + b$ 等, 相关系数 $\rho(X, Y)$ 是无能为力的, 不过我们可以尝试 $\rho(X^2, Y)$ 。

2.4.2 最小二乘法和回归

如果随机变量 X, Y 不独立, 如“身高”和“体重”, 它们之间存在着某个未知的关系, 不妨将之抽象为函数 $y = g(x)$, 如何找出这个函数呢? 我们必须依靠“最小二乘法”这件工具。“二乘”一词来自日文, 即“平方”的意思。

法国数学家 Adrien-Marie Legendre (1752-1833) 在著作《计算彗星轨道的新方法》(1805) 的附录中明确提出了最小二乘法。虽然 Gauss 对最小二乘法的研究更早些, 但 Legendre 发表结果在先, 数学史把 Legendre 也列作最小二乘法的创立者之一。Legendre 在数论 (素数定理、二次互反律等)、数学分析 (椭圆积分)、天体力学上也有诸多贡献。他为人处事低调, 竟没有正式的肖像留世, 右边的水彩画是 Legendre 唯一的肖像。



1794-1795 年, 伟大的天才数学家 C. F. Gauss 系统研究了最小二乘法 (method of least squares), 是年十八岁, 并以它为工具计算出了谷神星的运动轨迹, Gauss 于 1809 年在《天体运动论》中详尽地著述了这一成果。

按照最小二乘原则, 关系函数 g 的选取要使得 $E[Y - g(X)]^2$ 达到最小, 这里假设 EY^2 和 $E[g(X)]^2$ 都存在。这是一个最优化问题, 以连续型的随机向量 $(X, Y)^\top$ 为例,

$$\begin{aligned} E[Y - g(X)]^2 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [y - g(x)]^2 f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} f_X(x) \left\{ \int_{-\infty}^{+\infty} [y - g(x)]^2 f_{Y|X}(y|x) dy \right\} dx \end{aligned}$$

根据练习 2.26, 当 $g(x) = E(Y|X = x)$ 时上式花括号内的积分达到最小, 进而 $E[Y - g(X)]^2$ 达到最小。离散型的情形也是类似的。

定义 2.51 (回归). 函数 $y = E(Y|X = x)$ 称为 Y 关于 X 的回归 (regression), 其中 $E(Y|X = x)$ 是固定 X 的取值 x 后 Y 的均值 (参见第 141 页的图 2.18)。我们把如下定义的曲线 $l_{Y|X}$ 称为 Y 关于 X 的回归曲线。

$$l_{Y|X} = \{(x, E(Y|X = x))\}$$

一旦给定了 X, Y 的联合分布, 回归曲线则由该联合分布完全确定。类似地, 函数 $x = E(X|Y = y)$ 称为 X 关于 Y 的回归, 且 X 关于 Y 的回归曲线是

$$l_{X|Y} = \{(E(X|Y = y), y)\}$$

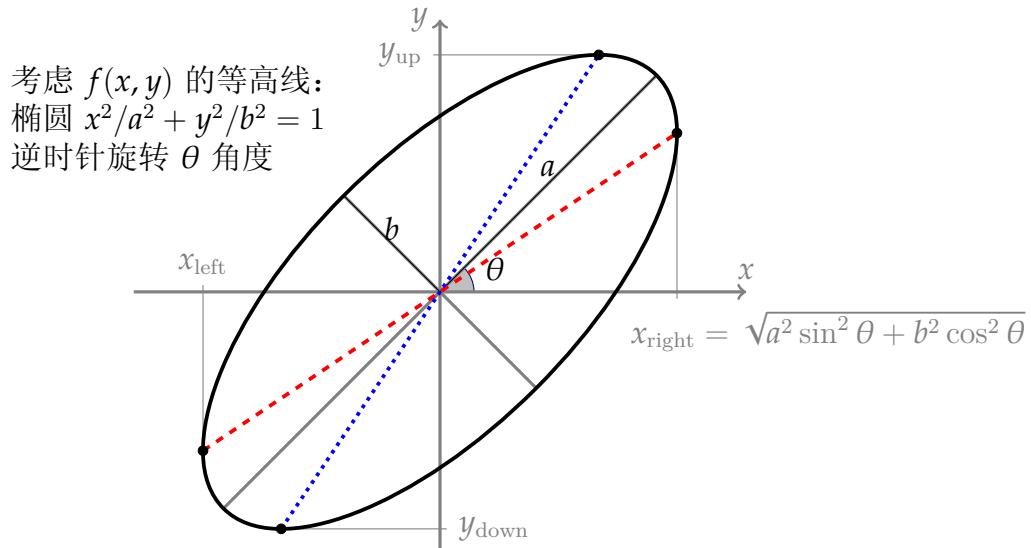


图 2.46: 回归曲线 $l_{Y|X}$ (红色虚线) 的几何含义 (请参考图 2.44): 不妨设联合密度函数 $f(x, y)$ 的等高线是椭圆 (即联合分布是一个二维正态分布, 见第 137 页的练习 2.13), 给定 $X = x$, 与椭圆所交线段的中点即是 $Y|X = x$ 的均值。将这些中点连接成线就是回归曲线。 $x_{\text{right}}, x_{\text{left}}$ 分别是该椭圆曲线在 x 轴投影的最大值和最小值。回归曲线 $l_{X|Y}$ (蓝色点线) 的几何含义是类似的。

两个随机变量 X, Y 之间不存在严格的函数关系, 弄清楚以谁为视角来观察谁很重要。一般地, $l_{Y|X} \neq l_{X|Y}$, 即这两条回归曲线不是简单的反函数的关系。请看下例。

例 2.86. 接着第 144 页的例 2.27 考虑二元正态分布 $(X, Y)^T \sim N(0, 0, 1, 4, 0.8)$ 。 Y 关于 X 的回归曲线 (见图 2.47 中的粗实线) 是

$$y = \int_{-\infty}^{+\infty} y f(y|x) dy = 1.6x$$

上述结果请参考第 145 页的图 2.21。类似地, X 关于 Y 的回归曲线是

$$x = \int_{-\infty}^{+\infty} x f(x|y) dx = 0.4y$$

对于例 2.86 和图 2.46 所揭示的正态联合分布情形下的回归曲线, 结合着第 174 页的图 2.37 来看更清楚。在图 2.37 中, 每个 $X = x$ 处截断面的曲线都正比于正态密度函数, 其均值就是 $E(Y|X = x)$ 。回归曲线 $l_{Y|X}$ 估计或预测 x 相应的 y 值为 $E(Y|X = x)$, 可谓“中庸之道”。

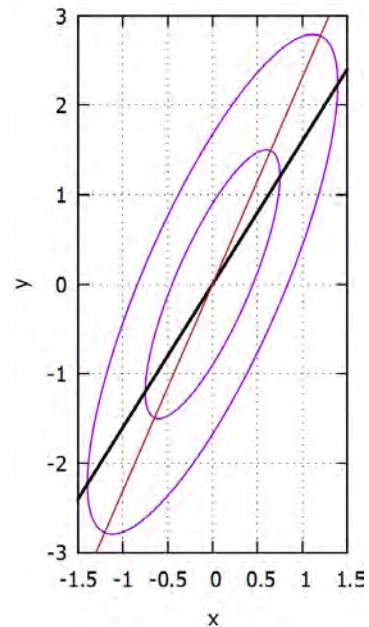


图 2.47: 对于 $(X, Y)^T \sim N(0, 0, 1, 4, 0.8)$, 给定 $X = x$, Y 的平均取值是 $1.6x$ 。

练习 2.40. 请计算图 2.46 中的点 x_{right} 对应的 y 值, 点 y_{up} 及其对应的 x 值。

例 2.86 中回归曲线为直线。一般地, 当已知 Y 关于 X 的回归曲线为直线 $y = \alpha x + \beta$ 时, 其中参数 α, β 未知, 使得下面的损失函数达到最小。

$$\begin{aligned} L(\alpha, \beta) &= E(Y - \alpha X - \beta)^2 \\ &= EY^2 + \alpha^2 EX^2 + \beta^2 - 2\alpha E(XY) + 2\alpha\beta EX - 2\beta EY \end{aligned}$$

为使 $L(\alpha, \beta)$ 达到最小, 求解下面有关 α, β 的方程组。

$$\begin{aligned} \frac{\partial L}{\partial \alpha} &= 2\alpha EX^2 - 2E(XY) + 2\beta EX = 0 \\ \frac{\partial L}{\partial \beta} &= 2\beta + 2\alpha EX - 2EY = 0 \end{aligned}$$

性质 2.43. 若 $V(X) \neq 0$, Y 关于 X 的回归直线为 $y = \alpha x + \beta$, 其中

$$\alpha = \frac{\text{Cov}(X, Y)}{V(X)}, \text{ 且 } \beta = EY - \alpha EX \quad (2.90)$$

性质 2.44. 由式 (2.90) 得到的 α, β 满足 $E(Y - \alpha X - \beta) = 0$ 。另外

$$\begin{aligned} E[Y - (\alpha X + \beta)]^2 &= (1 - \rho^2)V(Y) \\ &= \min_{c_1, c_2} E(Y - c_1 X - c_2)^2 \end{aligned} \quad (2.91)$$

$$E[(X - EX)(Y - \alpha X - \beta)] = 0 \quad (2.92)$$

 式 (2.91) 说明 ρ^2 越接近 1, Y 偏离 $\alpha X + \beta$ 的平均程度就越小, 直至 $\rho^2 = 1$ 时达到极致 (见定理 2.34)。作为 Y 的近似, $\alpha X + \beta$ 的 $\text{MSE} = E[Y - (\alpha X + \beta)]^2$ 只是 $V(Y)$ 的 $1 - \rho^2$ 倍 (或者 $\text{RMSE} = \sigma_Y \sqrt{1 - \rho^2}$), 显然 $\text{MSE} \leq V(Y)$ (或者 $\text{RMSE} \leq \sigma_Y$)。换句话说, 得知与 Y 相关的 X 的信息去猜 Y 的值为 $\alpha X + \beta$ 比直接猜 Y 取值 $E(Y)$ 要精准些。而式 (2.92) 意味着 X 与误差 $Y - \alpha X - \beta$ 是不相关的 (该结论也可由第 175 页的推论 2.2 的结果 ② 直接推得), 二者之间不存在线性关系。

定理 2.35. 根据 (2.90) 所示的推导, Y 关于 X 的回归直线 $l_{Y|X}$ 的方程为

$$y - EY = \frac{\text{Cov}(X, Y)}{V(X)}(x - EX) \quad (2.93)$$

练习 2.41. 请读者验证, X 关于 Y 的回归直线 $l_{X|Y}$ 的方程为

$$x - EX = \frac{\text{Cov}(X, Y)}{V(Y)}(y - EY) \quad (2.94)$$

定义 2.52. 我们把式 (2.93) 中的斜率 $\gamma_{Y|X} = \text{Cov}(X, Y)/V(X)$ 称为 Y 关于 X 的回归系数, 把 $\gamma_{X|Y} = \text{Cov}(X, Y)/V(Y)$ 称为 X 关于 Y 的回归系数。显然,

$$\rho^2(X, Y) = \gamma_{Y|X}\gamma_{X|Y} \quad (2.95)$$

练习 2.42. 回归直线 $l_{X|Y}$ 和 $l_{Y|X}$ 在什么条件下重合? 答案: 参考式 (2.95), 当 $\rho^2 = 1$ 时两条回归直线重合。

例 2.87. 若 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 利用定理 2.35 和性质 2.44, 也可以得到第 144 页的例 2.27 的结果 (2.29), 即

$$Y|X = x \sim N\left(\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right)$$

该结果简记作 $Y|X \sim N(l_{Y|X}, \text{MSE})$ 或者 $Y|X \sim N(l_{Y|X}, \text{RMSE}^2)$ 。

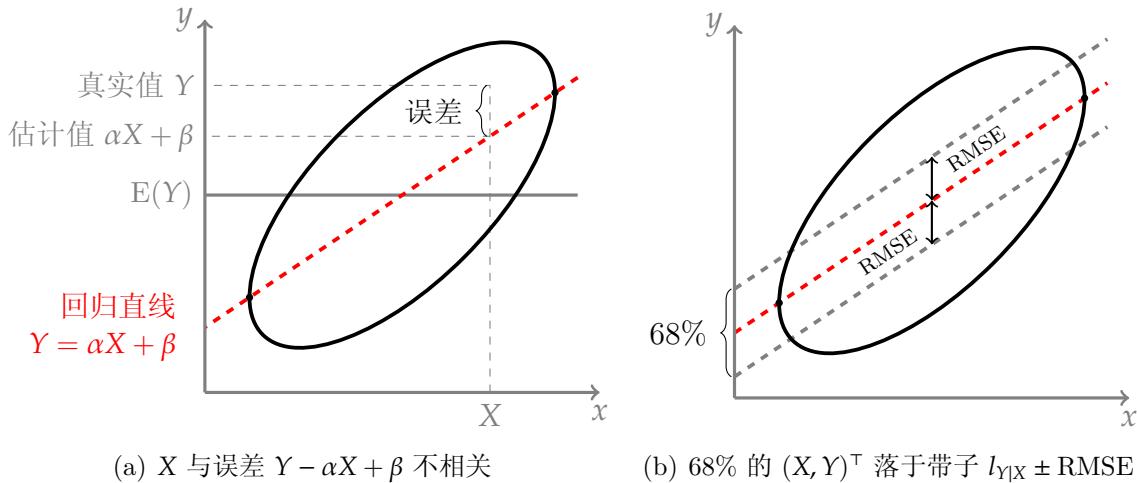


图 2.48: 性质 2.44 的直观解释: 回归直线的均方根误差 $\text{RMSE} = \sigma_Y \sqrt{1 - \rho^2}$, 其中 ρ 是相关系数。在所有直线当中, 回归直线的 RMSE 达到最小。

练习 2.43. 请说明 95% 的 $(X, Y)^\top$ 落于带子 $l_{Y|X} \pm 2\sigma_Y \sqrt{1 - \rho^2}$ (简记作 $l_{Y|X} \pm 2\text{RMSE}$)。

2.4.3 随机向量的主成分

受第 161 页的例 2.44 的启发, 对于随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top$, 是否存在正交变换使得 \mathbf{X} 经过此变换后协方差矩阵为一个对角阵? 换句话说, 经过该正交变换所得的新随机向量 $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ 的各个分量之间的相关系数为零。

Galton 和他的得意门生、继承人 Karl Pearson (1857-1936) 最早研究过这个问题。1933 年, 美国统计学家 Harold Hotelling (1895-1973, 照片见右) 彻底解决该问题, 明确地提出了主成分的概念和多元统计的主成分分析方法 [77]。主成分分析是数据表示的重要方法之一, 它能抓住数据在空间分布的重要特征。



定理 2.36. 设随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的协方差矩阵为 $\Sigma_{n \times n}$, 根据第 767 页的定理 E.4, 存在谱分解 $\Sigma = U\Lambda U^\top$, 其中, $U = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ 是一个正交矩阵, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ 是 Σ 的特征根 $\lambda_1 \geq \dots \geq \lambda_n \geq 0$ 构成的对角阵, 并且 $\mathbf{Y} = U^\top \mathbf{X}$ 的协方差矩阵为 Λ 。

证明. 根据公式 (2.85), \mathbf{Y} 的协方差矩阵是

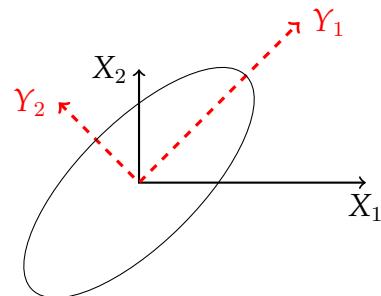
$$\text{Cov}(\mathbf{Y}, \mathbf{Y}) = \text{Cov}(U^\top \mathbf{X}, U^\top \mathbf{X}) = U^\top \text{Cov}(\mathbf{X}, \mathbf{X}) U = U^\top \Sigma U = U^\top U \Lambda U^\top U = \Lambda \quad \square$$

在定理 2.36 中, 如果协方差矩阵 Σ 有重复的特征根, 则正交矩阵 U 不唯一 (但这并不影响结果)。

定义 2.53. 接着定理 2.36, \mathbf{Y} 的分量 $Y_j = \mathbf{u}_j^\top \mathbf{X}, j = 1, \dots, n$ 被称为 \mathbf{X} 的第 j 主成分 (principal component)。显然, 第一主成分的方差 λ_1 最大, 最能描述 \mathbf{X} 的散落情况; 第二主成分次之, ……。

随机向量 \mathbf{X} 的主成分就是换个“角度”看 \mathbf{X} 的分量, 使它们之间的关系简化为无关。对于正态分布的随机向量, 分量之间的无关性和独立性是等价的, 主成分的意义就更直观了。

图 2.49: 接着第 137 页的练习 2.13, 对于二元正态分布 $\mathbf{X} \sim N(0, 0, \sigma_1^2, \sigma_2^2, \rho)$, 随机向量在长 (短) 轴所在直线上的投影就是第一 (二) 主成分。



练习 2.44. 在定理 2.36 中, \mathbf{Y} 的协方差矩阵 Λ 的迹和行列式与 Σ 的相同。

协方差矩阵 Σ 的迹是随机向量 \mathbf{X} 各分量的方差之和, 练习 2.44 揭示它等同于 \mathbf{X} 的主成分的方差之和。下面的概念量化了各个主成分的重要地位。

定义 2.54. 在定理 2.36 中, 我们把下面的量称为第 j 主成分的贡献率。

$$c_j = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_n}, j = 1, 2, \dots, n$$

显然, $0 \leq c_n \leq \cdots \leq c_1 \leq 1$ 且 $c_1 + \cdots + c_n = 1$ 。如果前 k 个主成分的贡献率 $c_1 + \cdots + c_k \geq 85\%$, 则 $\mathbf{Y}' = (Y_1, \dots, Y_k, 0, \dots, 0)^\top$ 将是对 \mathbf{Y} 的一个很好的近似。

例 2.88. 二元正态分布 $\mathbf{X} \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ 的第一主成分的贡献率是

$$c_1 = \frac{1}{2} + \frac{\sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\rho^2\sigma_1^2\sigma_2^2}}{2(\sigma_1^2 + \sigma_2^2)}$$

显然, $|\rho|$ 越大, 第一主成分的贡献率就越大。例如, 第 137 页的图 2.17 所示的二元正态分布 $\mathbf{X} \sim N(0, 0, 1, 4, 0.8)$, 其协方差矩阵的谱分解是

$$\begin{pmatrix} 1 & 1.6 \\ 1.6 & 4 \end{pmatrix} = \begin{pmatrix} 0.3975 & -0.9176 \\ 0.9176 & 0.3975 \end{pmatrix} \begin{pmatrix} 4.6932 & 0 \\ 0 & 0.3068 \end{pmatrix} \begin{pmatrix} 0.3975 & 0.9176 \\ -0.9176 & 0.3975 \end{pmatrix}$$

第一主成分 $0.3975X_1 + 0.9176X_2$ 的贡献率是 93.86%, 第二主成分 $-0.9176X_1 + 0.3975X_2$ 的贡献率是 6.14%。

练习 2.45. 计算图 2.17 所示的 $\mathbf{X} \sim N(0, 0, 1, 4, -0.4)$ 的主成分及其贡献率。

答案: 仿照上例, 第一主成分 $-0.2425X_1 + 0.9701X_2$ 的贡献率是 84%, 第二主成分 $-0.9701X_1 - 0.2425X_2$ 的贡献率是 16%。

2.5 习题

- 2.1. 令映射 $p_j : \mathbb{R}^n \rightarrow \mathbb{R}$ 是 n 维向量 $\mathbf{x} = (x_1, \dots, x_j, \dots, x_n)^\top$ 到其第 j 个分量的投射, 即 $p_j(\mathbf{x}) = x_j$ 。已知定义在可测空间 (Ω, \mathcal{S}) 上的向量值函数 $h : \Omega \rightarrow \mathbb{R}^n$, 令 $h_j = p_j \circ h, j = 1, 2, \dots, n$, 则 h 是 Borel 可测的当且仅当 h_1, h_2, \dots, h_n 都是 Borel 可测的。
 - 2.2. 某狙击手射中目标的概率为 p , 连续向同一目标射击直至击中目标为止, 请给出射击次数 X 的分布列。
 - 2.3. 考虑第 113 页的例 2.4, 请给出 X 的分布函数。
 - 2.4. 接着例 2.3, 定义随机变量 $X_1(k) = k$ 和 $X_2(k) = 7 - k$, 其中 $k = 1, 2, \dots, 6$ 。如果骰子是均匀的, 试说明 X_1, X_2 具有相同的分布。
 - 2.5. 将 3 个球随机地放入标号为 1, 2, 3, 4 的 4 个盒子中, 请给出有球的盒子的最大标号 X 的分布列。
 - 2.6. 盒中装有 8 个球, 其中 4 个白球, 4 个黑球。一次一个不放回地抽取直至取得 1 个白球, 求所抽取的球的个数 X 的分布列。
 - 2.7. 对于二项分布 $X \sim B(n, p)$, 问 k 取何值时 $P\{X = k\}$ 最大?
 - 2.8. 设随机变量 $X \sim B(2, p), Y \sim B(3, p)$ 。若 $P\{X \geq 1\} = 5/9$, 求 $P\{Y \geq 1\}$ 。
 - 2.9. 设随机变量 X 的分布列为 $P(X = k) = \lambda^k e^{-\lambda} / k!$, 其中 $k = 0, 1, 2, \dots$ 。(1) 求 X 取偶数的概率; (2) 若 $P(X = 2) = P(X = 3)$, 求 X 取偶数的概率。
 - 2.10. 设 D 是曲线 $y = 1 - x^2$ 与 x 轴围成的区域, 在 D 内任取一点, 该点到 x 轴的距离为 X , 求 X 的分布函数。
 - 2.11. 已知 $F(x)$ 是一个分布函数, 证明: 对任意的 $h > 0$, 函数 $G(x) = \frac{1}{h} \int_x^{x+h} F(y) dy$ 也是一个分布函数。
 - 2.12. 往闭区间 $[0, 1]$ 上随机地投钉, 以 X 表示落点的坐标, 并设该点落在 $[0, 1]$ 的任意子区间内的概率与这个子区间的长度成正比。试求 X 的分布。
 - ☆ 2.13. 令 X, Y 是两个随机变量, 对于任意实数 a 和 $\epsilon > 0$, 皆有
- $$P(Y \leq a) \leq P(X \leq a + \epsilon) + P(|Y - X| > \epsilon) \quad (2.96)$$
- 2.14. 已知连续型随机变量 X 的概率密度函数为 $f_X(x) = ae^{-|x|}$, 其中 $x \in \mathbb{R}$ 。求: (1) 常数 a ; (2) X 的分布函数 $F_X(x)$; (3) $P\{-1 < X < 2\}$ 。

2.15. 已知随机变量 X 的分布列为 $P\{X = k\} = \frac{1}{ck!}$, 其中 $k = 1, 2, \dots$, 求正常数 c 。

2.16. 设随机变量 X 的概率密度函数为 $f_X(x) = \begin{cases} 2|x|/5 & \text{当 } -2 < x < 1 \\ 0 & \text{其它} \end{cases}$

求随机变量 $Y = 2X + 1$ 的概率密度函数 $f_Y(y)$ 。

2.17. 设 $X \sim N(0, 1)$, 求下面新随机变量的密度函数: (1) $Y = e^X$; (2) $Z = \sqrt{|X|}$ 。

2.18. 已知随机变量 X 的分布函数为 $F_X(x) = \begin{cases} 1 - e^{-2x} & \text{当 } x > 0 \\ 0 & \text{当 } x \leq 0 \end{cases}$

试证明: $Y = 1 - e^{-2X} \sim U(0, 1)$ 。

☆ 2.19. 若 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} U[0, a]$, 其中 $a > 0$, 求 $Y = \max(X_1, \dots, X_n)$ 的密度函数。

2.20. 如果随机变量 $X \sim U[-r, r]$, 其中 $r > 0$, 并且方程 $4t^2 + 4Xt + X + 2 = 0$ 有实根的概率概率为 $1/4$, 试求 X 的概率分布。

2.21. 已知 $X, Y \stackrel{\text{iid}}{\sim} \frac{1}{k+1}\langle 0 \rangle + \frac{1}{k+1}\langle 1 \rangle + \dots + \frac{1}{k+1}\langle k \rangle$, 求 $Z = X + Y$ 的分布列。

2.22. 设二维随机向量 $(X, Y)^\top$ 在矩形区域 $D = \{(x, y) : 1 \leq x \leq 3, 1 \leq y \leq 3\}$ 上服从均匀分布, 求 $Z = |X - Y|$ 的概率密度。

2.23. 已知随机变量 $U \sim U[0, 1], V \sim U[-1, 1]$ 相互独立, 令 $a > 0$, 则 $X = aV/U + b$ 的密度函数是

$$f_X(x) = \begin{cases} \frac{1}{4a} & \text{若 } b - a \leq x \leq b + a \\ \frac{a}{4(x - b)^2} & \text{否则} \end{cases}$$

2.24. 已知离散型随机向量 $(X, Y)^\top$ 的分布列为
$$\begin{array}{c|ccc} X & Y & -1 & 1 & 2 \\ \hline -1 & 1/10 & 2/10 & 3/10 \\ 2 & -2/10 & 1/10 & 1/10 \end{array}$$

试求下面这些随机变量的分布列: (1) $Z = X + Y$; (2) $Z = XY$; (3) $Z = X/Y$; (4) $Z = \max(X, Y)$ 。

2.25. 设随机变量 X 与 Y 都服从正态分布 $N(0, \sigma^2)$, 且 $P\{X \leq 0, Y \geq 0\} = 1/3$, 求 $P\{X > 0, Y < 0\}$ 。

2.26. 设二维随机向量 $(X, Y)^\top$ 的分布函数 $F(x, y) = (a + b \arctan x)(a + b \arctan y)[1 + 1/2(a - b \arctan x)(a - b \arctan y)]$, 试求: (1) 常数 a, b ; (2) $P\{X \geq 0, Y \geq 0\}$ 。

2.27. 设随机向量 $(X, Y)^\top$ 具有如下密度函数, 计算 $Z = XY$ 的密度函数。

$$f(x, y) = \begin{cases} 24xy(1 - x^2) & \text{当 } 0 < x < 1, 0 < y < 1 \\ 0 & \text{其他} \end{cases}$$

2.28. 已知随机变量 $X \sim U(1, 2)$ 与 $Y \sim U(3, 4)$ 相互独立, 计算 $Z = XY$ 的密度函数。

☆ 2.29. 接着定理 2.16, 令 $F_X(x), F_Y(y)$ 分别是 X, Y 的分布函数, 试证明:

$$F_X(x) = \int_0^1 F_Y\left(\frac{x}{u}\right) du$$

2.30. 设随机向量 $(X, Y)^\top$ 的概率密度为 $f(x, y) = \begin{cases} e^{-y} & \text{当 } 0 < x < 1, y > 0 \\ 0 & \text{其他} \end{cases}$

(1) 判断 X 和 Y 是否独立; (2) 求 $Z = X + Y$ 的分布函数 $F_Z(z)$; (3) 求 $P\{Z > 3\}$ 。

☆ 2.31. 设随机向量 $(X, Y)^\top \sim N(0, 0, \sigma_X^2, \sigma_Y^2, \rho)$, 求随机变量 $Z = X/Y$ 的密度函数 $f_Z(z)$ 。若 X, Y 相互独立, X/Y 是怎样的分布?

☆ 2.32. 设随机变量 $X \sim N(0, 1)$ 与 $Y \sim \chi_n^2$ 相互独立, 其中 Y 的密度函数见式 (2.40), 求 $T = \frac{X}{\sqrt{Y/n}}$ 的密度函数 $f_T(t)$ (该分布就是著名的 t 分布, 记作 $T \sim t_n$, 它是英国统计学家 W. S. Gosset 发现的, 详见第 305 页的定义 4.21)。

☆ 2.33. 如果随机变量 $X \sim \chi_m^2$ 与 $Y \sim \chi_n^2$ 相互独立, 求 $Z = \frac{X/m}{Y/n}$ 的密度函数 $f_Z(z)$ (该分布是统计学中著名的 F 分布, 记作 $Z \sim F_{m,n}$, 详见第 306 页的定义 4.22)。

2.34. 设随机向量 $(X, Y)^\top$ 服从区域 D 上的均匀分布, D 是由直线 $x = 0, y = 0, x + y = 1$ 围成的闭区域, (1) 判断 X 与 Y 是否独立; (2) 求 $Z = X + Y$ 的分布函数 $F_Z(z)$ 。

2.35. 设随机向量 $(X, Y)^\top$ 服从圆盘 $D = \{(X, Y)^\top \in \mathbb{R}^2 : x^2 + y^2 \leq r^2\}$ 上的均匀分布, 试求: (1) 边缘密度 $f_X(x)$; (2) 条件密度 $f_{X|Y}(x|y)$ 。

2.36. 设随机变量 X 只取 $[0, 1]$ 上的值, 试证明 $V(X) \leq 1/4$ 并指出何时取等号。

2.37. 盒子里有 n 个球, 标号依次是 $1, 2, \dots, n$ 。独立有放回地均匀抽取 n' 次, 若 n 很大, 求所抽结果中不同球的期望个数?

☆ 2.38. 设连续型随机变量 X 的取值范围是 \mathbb{R} , 若 $\lambda > 0$ 是一个常数, 令 $Y = e^{\lambda X}$, 试证明: $P\{X \geq a\} \leq e^{-\lambda a} E(Y)$, 其中 a 为任意实数。

2.39. 已知随机变量 $X \sim N(\mu, \sigma^2)$, 试证明 $E(|X - \mu|) = \sigma \sqrt{2/\pi}$ 。

- ☆ 2.40. 设随机向量 $(X, Y)^\top \sim N(0, 0, 1, 1, \rho)$, 试求 $E[\max(X, Y)] = ?$
- 2.41. 若 $X \sim p_1\langle x_1 \rangle + \cdots + p_n\langle x_n \rangle$, 试证明 $p_1 = \cdots = p_n = 1/n$ 时, 熵 $H(X)$ 最大。

☆ 2.42. 试证明第 173 页的性质 2.30。

2.43. 接着练习 2.28, 试证明: $P\{0 < X < 2(m+1)\} \geq m/(m+1)$ 。

2.44. 试证明: 若 $E(\exp\{X^2\})$ 存在, 则 $P(|X| \geq \epsilon) \leq E(\exp\{X^2\})/\epsilon^2$ 。

2.45. 如果 $g(x)$ 是正的单调增函数, 而且 $E(g(X))$ 存在, 试证明:

$$P(|X| > t) \leq \frac{E(g(X))}{g(t)}$$

☆ 2.46. 试证明 Paley-Zygmund 不等式 (1932): 如果非负随机变量 Z 的方差有限, 则

$$P(Z \geq \lambda EZ) \geq (1 - \lambda)^2 \frac{(EZ)^2}{EZ^2}, \text{ 其中 } 0 < \lambda < 1$$

☆ 2.47. 试用其他方法证明第 194 页的例 2.76 中的第二个不等式。

2.48. 若 $X \sim N(1, 2)$ 和 $Y \sim N(0, 1)$ 独立, 求 $Z = 2X - Y + 3$ 的分布。

2.49. 已知 $X, Y \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, 令 $Z_1 = aX + bY$ 且 $Z_2 = aX - bY$, (1) 求 Z_1, Z_2 的相关系数; (2) 问 Z_1, Z_2 是否相关? 是否独立?

☆ 2.50. 若矩阵 $A_{n \times n} = (a_{ij})$ 正定, 试证明下面的 $f(x_1, \dots, x_n)$ 是概率密度函数。

$$f(x_1, \dots, x_n) = \sqrt{\frac{|A|}{\pi^n}} \exp \left\{ - \sum_{i,j=1}^n a_{ij}(x_i - \mu_i)(x_j - \mu_j) \right\} \quad (2.97)$$

其中, $|A|$ 是 A 的行列式, $\mu_1, \dots, \mu_n \in \mathbb{R}$ 。

2.51. 设随机向量 $(X, Y)^\top$ 在圆盘 $x^2 + y^2 \leq r^2$ 上服从均匀分布, (1) 求 X 与 Y 的相关系数 ρ ; (2) 问 X 与 Y 是否独立?

☆ 2.52. 设随机变量 X 与 Y 相互独立, 密度函数分别为 $f_X(x) = f(x) = \begin{cases} e^{-x} & \text{当 } x > 0 \\ 0 & \text{当 } x \leq 0 \end{cases}$ 和 $f_Y(y) = f(y)$ 。试证明: 随机变量 $U = X + Y$ 与 $V = X/Y$ 也是相互独立的。

☆ 2.53. 已知随机变量 $\{X_n : n = 1, 2, \dots\}$ 相互独立, 且 X_n 的概率函数为 $P(X_n = \pm \sqrt{n+1}) = 1/(n+1)$, $P(X_n = 0) = 1 - 2/(n+1)$ 。试证明: $\forall \epsilon > 0$,

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{j=1}^n X_j \right| < \epsilon \right\} = 1$$

☆ 2.54. 求证: 随机变量 X 的中位数 $M(X)$ 满足 $|M(X) - E(X)| \leq \sqrt{2V(X)}$ 。

☆ 2.55. 若 $\|X\|_p < \infty$, 则 $\forall q \in (0, p)$, 皆有 $\|X\|_q \leq \|X\|_p$ 。

2.56. 已知随机变量 X 的 k 阶绝对矩存在, 试证明: $\forall \epsilon > 0$, $P\{|X| \geq \epsilon\} \leq E|X|^k/\epsilon^k$ 。

2.57. 设随机向量 $(X, Y)^\top$ 的密度函数为 $f(x, y) = \begin{cases} e^{-y} & \text{当 } 0 < x < y < \infty \\ 0 & \text{其他} \end{cases}$

试求: 相关系数 $\rho(X, Y)$ 和回归系数 $\gamma_{Y|X}, \gamma_{X|Y}$ 。

2.58. 接着例 1.54, 令 Y 表示不动点个数, 求 $E(Y)$ 和 $V(Y)$ 。

☆ 2.59. 已知随机变量 X, Y 的期望和方差分别为 $E(X) = E(Y) = 0$, $V(X) = V(Y) = 1$ 且 $Cov(X, Y) = \rho$ 。试证明: $E[\max(X^2, Y^2)] \leq 1 + \sqrt{1 - \rho^2}$ 。

2.60. 已知随机向量 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 验证 Y 关于 X 的回归曲线 $l_{Y|X} = \{(x, E(Y|X=x))\}$ 是直线 $y = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ 。

2.61. 随机变量 X, Y 的联合密度函数为 $f(x, y) = \begin{cases} 2 & \text{当 } 0 < x < y < 1 \\ 0 & \text{其他} \end{cases}$

试求: (1) Y 关于 X 的回归; (2) X 关于 Y 的回归。

★ 2.62. 设连续型随机变量 X 的密度函数为 $f_\theta(x)$, 其中 $\theta \in \Theta$ 是未知参数, 试证明:

$$E_\theta \left[\frac{\partial \ln f_\theta(X)}{\partial \theta} \right] = 0 \text{ 且 } E_\theta \left[\frac{\partial \ln f_\theta(X)}{\partial \theta} \right]^2 = -E_\theta \left[\frac{\partial^2 \ln f_\theta(X)}{\partial \theta^2} \right] \quad (2.98)$$

第三章

特征函数

半亩方塘一鉴开，天光云影共徘徊。问渠哪得清如许，为有源头活水来。

朱熹《观书有感》

为了方便推理或运算，人们常把复杂的原问题“翻译”成另外一种形式，以便能够简单地进行处理。例如，对数变换可把乘积运算转化为加法运算，Fourier 变换可把卷积运算转化为乘积运算 [142]。

1807 年，法国数学家、物理学家 Joseph Fourier (1768-1830) 在研究热传导问题时提出了 Fourier 级数。Fourier 变换是复 Fourier 级数的一般化，是 Fourier 分析的主要研究内容之一，现已发展成为数值计算、振动分析、声学、光学、量子力学、信号处理、图像处理、电子工程、计量经济学等领域的常用工具。在一定条件下，Fourier 变换把实值函数 $g(x)$ 变为复值函数 $\varphi(t)$ 。然而，在不同的领域，Fourier 变换的定义有些许的不同。在概率论中，Fourier 变换主要用于定义特征函数，即针对概率质量函数或者概率密度函数进行变换。



定义 3.1 (Fourier 变换). 在概率论中，Fourier 变换定义为

$$\begin{aligned}\mathcal{F}(g) &= \int_{-\infty}^{+\infty} e^{itx} g(x) dx \\ &= \int_{-\infty}^{+\infty} \cos(tx) g(x) dx + i \int_{-\infty}^{+\infty} \sin(tx) g(x) dx\end{aligned}\tag{3.1}$$

其中， $i = \sqrt{-1}$ 是虚数单位。

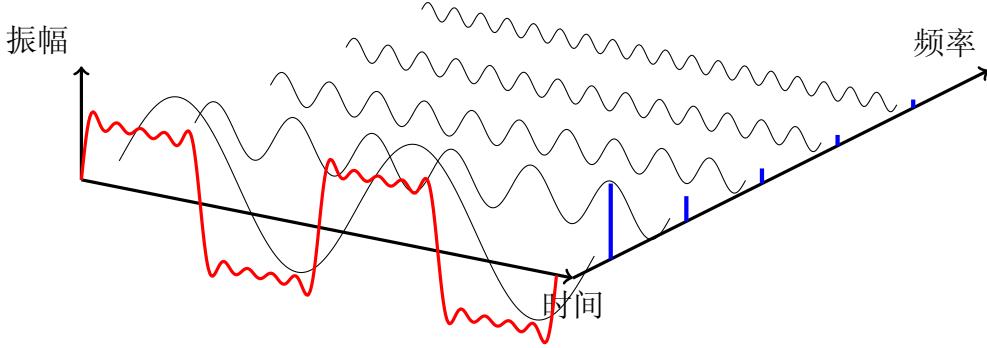
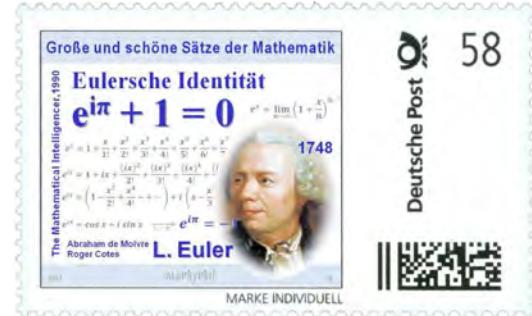


图 3.1: Fourier 变换是复 Fourier 级数的一般化。在信号处理中, Fourier 变换的直观含义是把原始信号(粗线)转化为频率的复值函数。

式 (3.1) 用到了下面著名的 Euler 公式 (1743), 它堪称最伟大的数学公式之一, 在数学、物理学、工程科学中随处可见。

$$e^{i\theta} = \cos \theta + i \sin \theta, \text{ 其中 } \theta \in \mathbb{R}$$

该公式的推导过程见右边的邮票, Euler 恒等式是它的推论。



例 3.1. 考虑均匀分布 $U[-\frac{1}{2}, \frac{1}{2}]$ 的密度函数 $I_{[-\frac{1}{2}, \frac{1}{2}]}(x)$ 的 Fourier 变换,

$$\begin{aligned}\varphi(t) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \cos(tx) dx + i \int_{-\frac{1}{2}}^{\frac{1}{2}} \sin(tx) dx \\ &= \frac{2}{t} \sin\left(\frac{t}{2}\right)\end{aligned}$$

例 3.2. 考虑高斯函数 $g(x) = \exp(-\alpha x^2)$ (其中 $\alpha > 0$) 的 Fourier 变换, 读者不难发现它依然是高斯函数。

$$\begin{aligned}\varphi(t) &= \int_{-\infty}^{+\infty} \cos(tx) \exp(-\alpha x^2) dx + i \int_{-\infty}^{+\infty} \sin(tx) \exp(-\alpha x^2) dx \\ &= \sqrt{\frac{\pi}{\alpha}} \exp\left(\frac{-t^2}{4\alpha}\right)\end{aligned}$$

定义 3.2 (Fourier 逆变换). 若实值函数 $\varphi(t)$ 在 \mathbb{R} 上可积, 则如下定义的变换被称为

$\varphi(t)$ 的 Fourier 逆变换。

$$\mathcal{F}^{-1}(\varphi) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi(t) dt \quad (3.2)$$

 (3.1) 定义的 Fourier 变换常用于概率论。而在信号处理中, Fourier 变换把时间的实值函数(即信号)转化为频率的复值函数, 其定义如下。

$$\begin{aligned} F(\omega) &= \int_{-\infty}^{+\infty} e^{-i\omega t} f(t) dt, \text{ 或者} \\ F(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i\omega t} f(t) dt \end{aligned}$$

在信号处理中, Fourier 逆变换定义为

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{i\omega t} F(\omega) d\omega, \text{ 或者} \\ f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{i\omega t} F(\omega) d\omega \end{aligned}$$

不管 Fourier 变换 \mathcal{F} 采用哪种定义, 总之它要和 Fourier 逆变换成对出现, 以确保

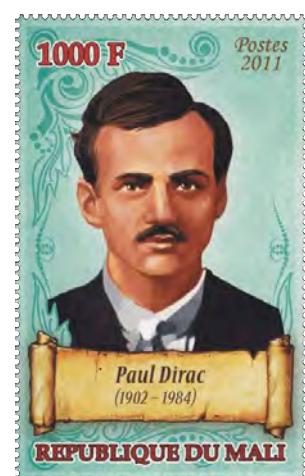
$$\mathcal{F}^{-1}[\mathcal{F}(g)] = g$$

※例 3.3. 英国理论物理学家 Paul Dirac (1902-1984) 在其著作《量子力学原理》(1927)里提出了一个广义函数——Dirac delta 函数, 简称 delta 函数, 定义如下。

$$\delta(x) = \mathcal{F}^{-1}(1) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} dt = \lim_{n \rightarrow \infty} \frac{\sin(nx)}{\pi x} \quad (3.3)$$

在信号处理中, delta 函数常用作单位脉冲函数。显然, $\mathcal{F}(\delta(x)) = 1$ 。也有文献这样定义 Dirac delta 函数:

$$\begin{aligned} \delta(x) &= \lim_{\sigma \rightarrow 0} \phi(x|0, \sigma^2), \text{ 或者} \\ \delta(x) &= \begin{cases} +\infty & \text{当 } x = 0 \\ 0 & \text{当 } x \neq 0 \end{cases} \\ \text{使得 } \int_{-\infty}^{+\infty} \delta(x-c) h(x) dx &= h(c) \end{aligned} \quad (3.4)$$



其中 $h(x)$ 是任意连续函数, c 为常数。特别地, delta 函数满足

$$\int_{-\infty}^{+\infty} \delta(x) dx = 1$$

$\delta(x)$ 的 k 阶导数 $\delta^{(k)}(x)$ 类似地定义为

$$\int_{-\infty}^{+\infty} \delta^{(k)}(x-c) h(x) dx = (-1)^k h^{(k)}(c)$$

※练习 3.1. 请读者证明 delta 函数的下述性质。

$$\begin{aligned}\delta(-x) &= \delta(x) && \text{对于常数 } c \neq 0, \quad \delta(cx) = \frac{1}{|c|} \delta(x) \\ x\delta(x) &= 0 && \delta(x) + x\delta'(x) = 0\end{aligned}$$

※定理 3.1 (卷积定理). \mathbb{R} 上可积函数 g_1 和 g_2 的卷积 $g_1 * g_2$ 在 Fourier 变换之下转化为乘积运算, 即

$$\mathcal{F}(g_1 * g_2) = \mathcal{F}(g_1)\mathcal{F}(g_2)$$

卷积定理有一个直接的应用: 给定 n 个独立的随机变量 X_i , 设其密度函数为 $f_i(x), i = 1, 2, \dots, n$, 则随机变量 $X = X_1 + X_2 + \dots + X_n$ 的密度函数为 $(f_1 * f_2 * \dots * f_n)(x)$ 。然而, 做多次卷积运算是相当麻烦的, 出于简化计算的目的, 需在卷积定理的基础上提炼出特征函数的概念。

※定义 3.3. 已知随机变量 X 的分布函数为 $F(x)$, 具有分布列 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \dots$, 或者密度函数 $f(x)$, X 的特征函数 (characteristic function) $\varphi_X(t)$ 定义为

$$\varphi_X(t) = \mathbb{E}e^{itX} = \int_{-\infty}^{+\infty} e^{itx} dF(x) = \begin{cases} \sum_{n=1}^{\infty} e^{itx_n} p_n & \text{离散型} \\ \int_{-\infty}^{+\infty} e^{itx} f(x) dx & \text{连续型} \end{cases} \quad (3.5)$$

积分变换 (3.5) 也称为分布函数 $F(x)$ 的 Fourier-Stieltjes 变换。显然, 对于连续型随机变量 X , 特征函数即是 $\varphi_X(t) = \mathcal{F}(f(x))$ 。由于 $e^{itx} = \cos(tx) + i \sin(tx)$ 的模长^{*}等于 1, 所以级数 $\sum_{n=1}^{\infty} e^{itx_n} p_n$ 绝对收敛, 函数 $e^{itx} f(x)$ 在 \mathbb{R} 上绝对可积, 因此特征函数 $\varphi_X(t)$ 总是存在的。在不引起歧义时, 也记作 $\varphi(t)$ 。

^{*}复数 $a + bi$ 的模长是 $|a + bi| = \sqrt{a^2 + b^2}$, 即复平面里原点到点 (a, b) 的距离。

此章节所讨论的特征函数几乎未用到复分析 (complex analysis) 的知识, Fourier 变换 (3.2) 靠实积分来实现。与特征函数类似的工具还有如下定义的矩母函数 (moment-generating function) $M_X(s)$, 然而矩母函数并不像特征函数那样总是存在的 [137], 所以本书对矩母函数不作深入介绍。

$$M_X(t) = \mathbb{E}(e^{tX}) \quad (3.6)$$

例 3.4. 两点分布 $X \sim p\langle a \rangle + (1-p)\langle b \rangle$ 的特征函数为 $pe^{ita} + (1-p)e^{itb}$ 。

练习 3.2. 试证明: 二项分布 $X \sim \text{B}(n, p)$ 的特征函数为 $[1 + p(e^{it} - 1)]^n$ 。

提示: $\varphi_X(t) = \sum_{k=0}^n e^{itk} C_n^k p^k q^{n-k} = (q + pe^{it})^n$, 其中 $q = 1 - p$ 。

例 3.5. 均匀分布 $X \sim U[a, b]$ 的特征函数为

$$\begin{aligned} \varphi_X(t) &= \frac{1}{b-a} \left[\int_a^b \cos(tx) dx + i \int_a^b \sin(tx) dx \right] \\ &= \frac{e^{itb} - e^{ita}}{it(b-a)} \end{aligned}$$

例 3.6. 求正态分布 $X \sim N(\mu, \sigma^2)$ 的特征函数 $\varphi_X(t)$ 。

解. 利用**例 3.2** 的结果, 不难得到

$$\begin{aligned} \varphi_X(t) &= \mathcal{F}(\phi(x|\mu, \sigma^2)) \\ &= e^{it\mu} \int_{-\infty}^{+\infty} e^{it(x-\mu)} \phi(x|\mu, \sigma^2) dx \\ &= \exp \left\{ it\mu - \frac{\sigma^2 t^2}{2} \right\} \end{aligned}$$

练习 3.3. 请利用**例 3.6** 的结果证明:

$$\int_{-\infty}^{+\infty} \phi(x|\mu + it, \sigma^2) dx = 1, \text{ 其中 } \mu, t \in \mathbb{R} \quad (3.7)$$

定义 3.4 (随机向量的特征函数). 已知 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 为一个 n 维随机向量, 我们称下面的函数 $\varphi_{\mathbf{X}} : \mathbb{R}^n \rightarrow \mathbb{C}$ 为随机向量 \mathbf{X} 的特征函数,

$$\begin{aligned} \varphi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E} \left\{ e^{i\mathbf{t}^\top \mathbf{X}} \right\} \\ &= \mathbb{E} \left\{ e^{i(t_1 X_1 + t_2 X_2 + \dots + t_n X_n)} \right\} \end{aligned}$$

其中, 系数 $\mathbf{t} = (t_1, t_2, \dots, t_n)^\top \in \mathbb{R}^n$ 。

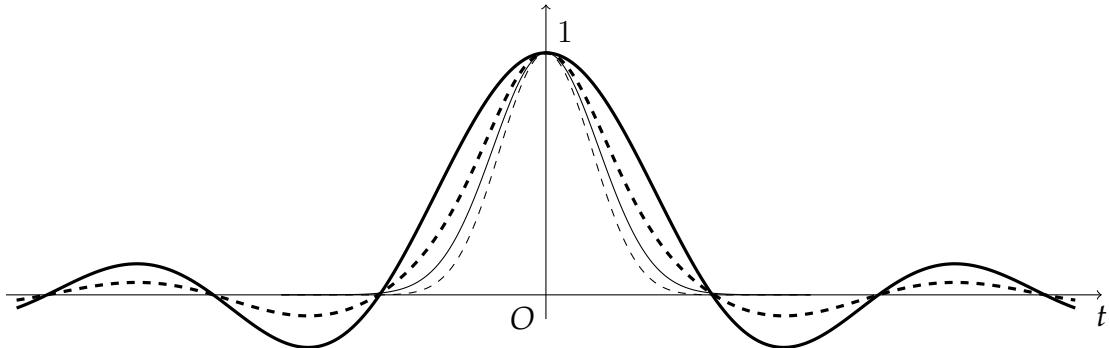


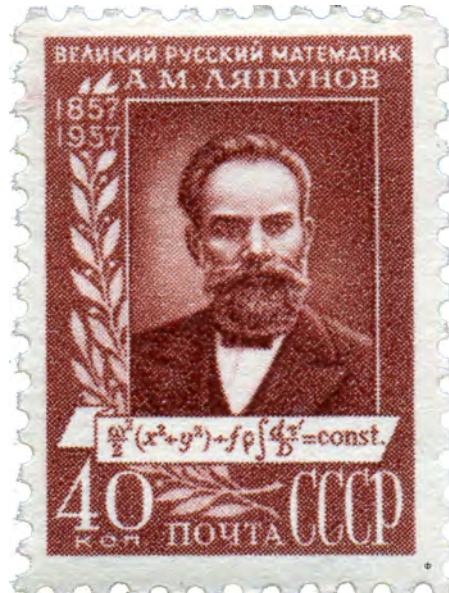
图 3.2: 均匀分布 $X \sim U(-1, 1)$ 和标准正态分布 $Y \sim N(0, 1)$ 的特征函数都是实值函数, 分别是 $\varphi_X(t) = \frac{1}{t} \sin t$ (粗实线) 和 $\varphi_Y(t) = \exp(-t^2/2)$ (细实线)。虚线分别是 $0.4\varphi_X(t) + 0.6\varphi_Y(t)$ (粗) 和 $\varphi_X(t)\varphi_Y(t)$ (细), 它们是不是特征函数?

例 3.7. 二元正态分布 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的特征函数为

$$\begin{aligned}\varphi(s, t) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp(isx + ity) \phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) dx dy \\ &= \exp \left\{ is\mu_X + it\mu_Y - \frac{1}{2} (\sigma_X^2 s^2 + 2\rho\sigma_X\sigma_Y st + \sigma_Y^2 t^2) \right\}\end{aligned}$$

1900-1901 年, 伟大的俄国数学家、物理学家 A. M. Lyapunov (1857-1918) 首次用特征函数的方法来证明中心极限定理而使其大放异彩, 后来成为概率论常用的分析方法之一。Lyapunov 的这件工作意义深远, 因为经过法国数学家 P. Lévy 的发展, 特征函数已成为概率论研究中一件强大的分析工具。本章主要介绍特征函数的以下内容: (1) 独立随机变量之和的特征函数; (2) 利用特征函数计算随机变量的各阶原点矩; (3) 揭示分布函数与特征函数之间关系的反演 (inversion) 公式*和 Lévy 连续性定理。

Lyapunov 在概率论、微分方程、动力系统、位势论等领域建树颇丰。Lyapunov 和他的好友 Andrey Markov (1856-1922) 都师承 Chebyshev (1821-1894), 也都是俄国圣彼得堡学派的代表人物。1892 年, Lyapunov 的博士毕业论文《运动稳定性的一般问题》开创了微分方程稳定性理论, 百年来影响非凡, 堪称传世名著。1902 年, Lyapunov 接替 Chebyshev 在圣彼得堡大学的教席, 专心致力于数学研究。1918 年 10 月 30 日, 因妻子病逝, Lyapunov 开枪殉情, 三天后不治身亡。谨以此章纪念这位卓越的俄罗斯学者在特征函数及其应用上的开创性的工作。



*国内的文献也有将之译为“逆转公式”、“反转公式”等。

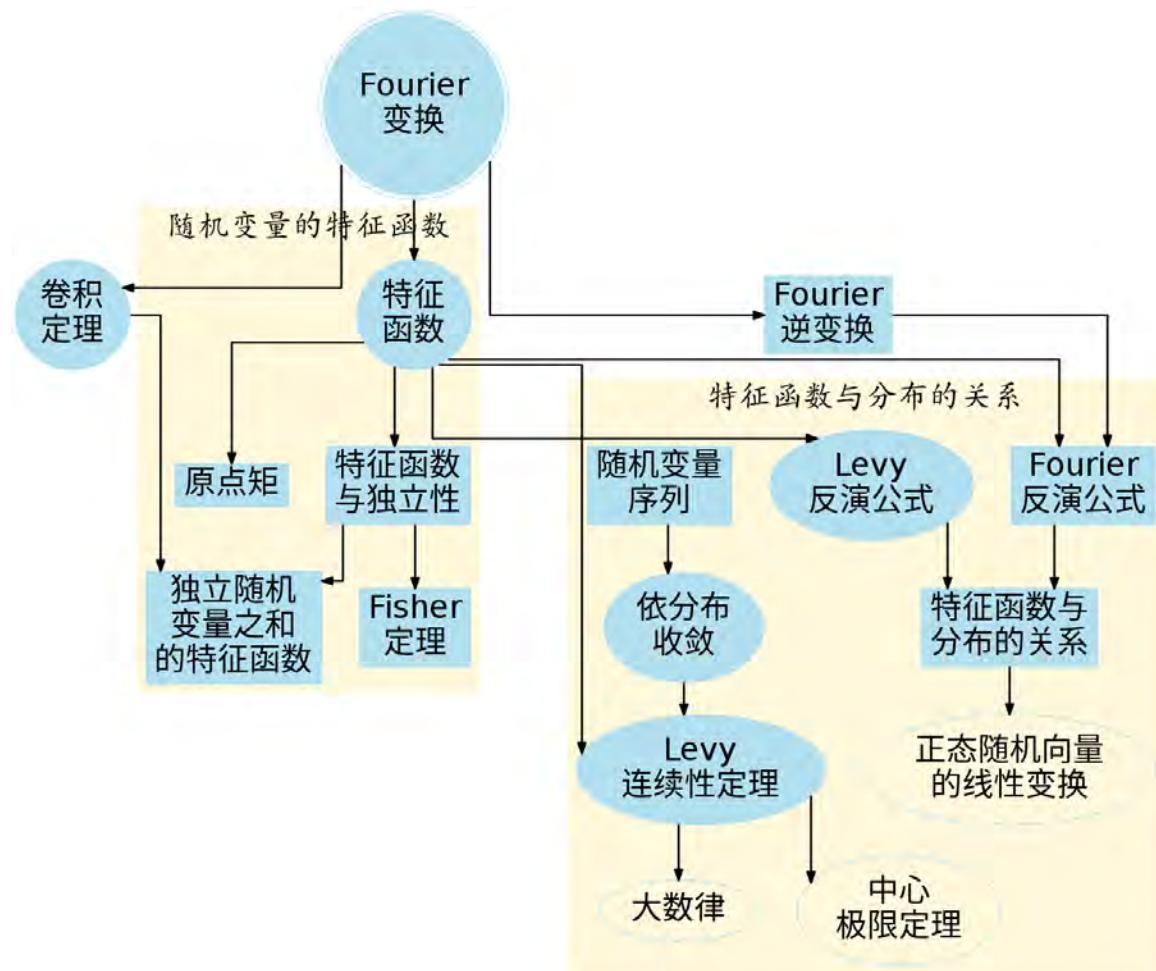
完成一件精美的艺术品，少不了他山之石。数学研究需要“引而伸之，触类而长之，天下之能事毕矣也”。Fourier 变换就是打磨概率论的工具。

鹤鸣于九皋，声闻于野。
鱼潜在渊，或在于渚。
乐彼之园，爰有树檀，其下维萚。
他山之石，可以为错。

鹤鸣于九皋，声闻于天。
鱼在于渚，或潜于渊。
乐彼之园，爰有树檀，其下维穀。
它山之石，可以攻玉。

《诗经·小雅·鹤鸣》

第三章的主要内容及其关系



3.1 特征函数的基本性质

由式(3.5), 特征函数是概率函数的离散 Fourier 变换或密度函数的 Fourier 变换的结果, 它的首个优点就是: 对任意随机变量 X 而言, 其特征函数 $\varphi_X(t)$ 总是存在的。与特征函数类似的工具是矩母函数 $M_X(t) = E(e^{tX})$, 虽然矩母函数比特征函数在计算上要略微简单些, 但对某些分布(如 Cauchy 分布)来说其矩母函数并不存在, 这一缺憾让我们在本书中还是选择了特征函数作为随机变量的分析工具。本节的开始主要介绍特征函数常见的性质, 包括特征函数的一些判定准则等。

性质 3.1. 随机变量 X 的矩母函数 $M_X(t)$ 如果存在, 则 $M_X(t) = \varphi_X(-it)$ 。

例 3.8. 二项分布、均匀分布、正态分布的矩母函数和特征函数如下。

分布	矩母函数	特征函数
$B(n, p)$	$[1 + p(e^t - 1)]^n$	$[1 + p(e^{-it} - 1)]^n$
$U[a, b]$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\frac{e^{itb} - e^{ita}}{it(b-a)}$
$N(\mu, \sigma^2)$	$\exp(t\mu + \sigma^2 t^2/2)$	$\exp(it\mu - \sigma^2 t^2/2)$

定理 3.2. 特征函数 $\varphi(t)$ 在 \mathbb{R} 上是一致连续的*, 并且

$$\varphi(0) = 1 \text{ 且 } |\varphi(t)| \leq 1$$

证明. 由特征函数的定义, $\varphi(0) = \int_{-\infty}^{+\infty} dF(x) = 1$ 且

$$|\varphi(t)| = \left| \int_{-\infty}^{+\infty} e^{itx} dF(x) \right| \leq \int_{-\infty}^{+\infty} |e^{itx}| dF(x) = 1$$

下面往证一致连续性: 首先对于任意 $t, h \in \mathbb{R}$ 皆有

$$|\varphi(t+h) - \varphi(t)| = \left| \int_{-\infty}^{+\infty} e^{itx} (e^{ihx} - 1) dF(x) \right| \leq \int_{-\infty}^{+\infty} |e^{ihx} - 1| dF(x)$$

要证得函数 $\varphi(t)$ 的一致连续性, 只需证明 $\forall \epsilon > 0$, 存在 h 使得对任意的 t 皆有 $|\varphi(t+h) - \varphi(t)| < \epsilon$ 。为此, 先选一个足够大的 $a > 0$ 使得

$$\int_{|x| \geq a} dF(x) < \epsilon/4$$

*即 $\forall \epsilon > 0$, 总存在 $\delta > 0$ 使得 $|t_1 - t_2| < \delta$ 时就有 $|\varphi(t_1) - \varphi(t_2)| < \epsilon$ 。一致连续性比连续性要强: 连续性是局部性质, 一致连续性则以同一尺度全局保证了只要两自变量充分地接近, 它们的因变量也充分地接近。细节请参阅 R. Courant 的名著《微积分和数学分析引论》第一卷 [28]。

接着, 再选足够小的 h 使得

$$|e^{ihx} - 1| = 2 \left| \sin \frac{hx}{2} \right| < \frac{\epsilon}{2}, \text{ 其中 } \forall x \in [-a, a]$$

于是, 得到 $|\varphi(t+h) - \varphi(t)| < \epsilon$, 具体推导过程是

$$\begin{aligned} |\varphi(t+h) - \varphi(t)| &\leq \int_{-a}^a |e^{ithx} - 1| dF(x) + 2 \int_{|x| \geq a} dF(x) \\ &< \frac{\epsilon}{2} \int_{-\infty}^{+\infty} dF(x) + \frac{\epsilon}{2} = \epsilon \end{aligned} \quad \square$$

例 3.9. 均匀分布 $X \sim U[a, b]$ 的特征函数 $\varphi_X(t)$ 见**例 3.5**, 显然

$$|\varphi_X(t)| = \frac{\sqrt{2 - 2 \cos[(b-a)t]}}{(b-a)|t|} \leq 1$$

练习 3.4. 已知 $\varphi_X(t)$ 是随机变量 X 的特征函数, 则

$$\varphi_X(-t) = \overline{\varphi_X(t)} \quad (3.8)$$

并且, 随机变量 $aX + b$ 的特征函数为

$$\varphi_{aX+b}(t) = e^{itb} \varphi_X(at) \quad (3.9)$$

练习 3.5. 试证明: 分布 $Y \sim \text{Cauchy}(\mu, \lambda)$ 的特征函数为 $\exp\{it\mu - \lambda|t|\}$ 。

提示: 先由 $\mathcal{F}((x^2+1)^{-1}) = \pi \exp(-|t|)$ 求得 $X \sim \text{Cauchy}(0, 1)$ 的特征函数为 $\exp(-|t|)$, 再根据式 (3.9) 求 $Y = \lambda X + \mu \sim \text{Cauchy}(\mu, \lambda)$ 的特征函数。

练习 3.6. 二维随机向量 $(X, Y)^\top$ 的特征函数 $\varphi(s, t)$ 具有性质:

$$\begin{aligned} |\varphi(s, t)| &\leq 1 \text{ 且 } \varphi(0, 0) = 1 \\ \varphi(-s, -t) &= \overline{\varphi(s, t)} \\ \varphi(s, 0) &= \varphi_X(s) \text{ 且 } \varphi(0, t) = \varphi_Y(t) \end{aligned}$$

定理 3.3. 下面不加证明地列举一些特征函数的判定准则, 对其证明感兴趣的读者可参阅 W. Feller 的《概率论及其应用》第二卷。

- ① 由随机向量 $(X, Y)^\top$ 的特征函数 $\varphi(s, t)$ 可推得 X 和 Y 的特征函数分别为 $\varphi_X(s) = \varphi(s, 0)$ 和 $\varphi_Y(t) = \varphi(0, t)$ 。
- ② 至多可数个特征函数的凸线性组合 $\sum_{n=1}^{\infty} \alpha_n \varphi_n(t)$ 依然是一个特征函数, 其中 $\alpha_n \geq 0$ 且 $\sum_{n=1}^{\infty} \alpha_n = 1$ 。

- ③ 至多可数个特征函数的积 $\prod_{n=1}^{\infty} \varphi_n(t)$ 依然是一个特征函数。
- ④ 若 $\varphi(t)$ 是一个特征函数, α 是一个常数, 则 $\overline{\varphi(t)}$ 、 $\varphi(\alpha t)$ 、 $\Re(\varphi(t))$ 、 $|\varphi(t)|^2$ 也都是特征函数, 其中 $\Re(\varphi(t))$ 表示 $\varphi(t)$ 的实部。
- ⑤ Bochner 准则: 函数 $\varphi : \mathbb{R}^n \rightarrow \mathbb{C}$ 是某随机向量的特征函数当且仅当 φ 半正定 (见定义 E.5), 且在原点连续并取值为 1。
- ⑥ Khinchin 准则: 复值函数 $\varphi(t)$ 满足 $\varphi(0) = 1$, 它是众数为 0 的单峰分布的特征函数当且仅当存在特征函数 $\tilde{\varphi}(u)$ 使得

$$\varphi(t) = \frac{1}{t} \int_0^t \tilde{\varphi}(u) du$$

- ⑦ Pólya 准则: 如果实值连续函数 $\varphi(t)$ 满足如下条件, 则 φ 是某个绝对连续对称分布的特征函数。
- (a) $\varphi(0) = 1$ 且 $\varphi(\infty) = 0$ 。
- (b) $\varphi(t)$ 是偶函数, 且当 $t > 0$ 时为凸函数。

例 3.10. 在图 3.2 中, 函数 $0.4\varphi_X(t) + 0.6\varphi_Y(t)$ 和 $\varphi_X(t)\varphi_Y(t)$ 都是特征函数。

例 3.11. 根据 Pólya 准则, $\tilde{\varphi}(t) = \exp(-t^2/2)$ 是一个特征函数。由例 3.6 我们知道, $\tilde{\varphi}(t)$ 是标准正态分布 $N(0, 1)$ 的特征函数。根据 Khinchin 准则, 下述函数 $\varphi(t)$ 是众数为 0 的某一单峰分布的特征函数。

$$\varphi(t) = \frac{1}{t} \int_0^t \tilde{\varphi}(u) du = \frac{2\Phi(t) - 1}{t} \sqrt{\frac{\pi}{2}}$$

本节内容

第一小节证明了“几个独立随机变量之和的特征函数等于这些随机变量的特征函数的乘积”, 同时举例子说明从“几个随机变量之和的特征函数等于这些随机变量的特征函数的乘积”不能判定这些随机变量是独立的。第二小节给出了用特征函数计算随机变量的各阶原点矩的方法, 以及如何用原点矩表示特征函数。

关键知识

(1) 特征函数的定义和常见性质; (2) 独立随机变量之和的特征函数; (3) 利用随机变量的特征函数计算该随机变量的各阶原点矩。

3.1.1 独立随机变量之和的特征函数

为何要引入特征函数这一工具？很重要的缘由是对独立随机变量之和的研究都要基于下面这个简单却不乏重要性的结果。

\nwarrow 定理 3.4. 已知随机变量 X_1, X_2, \dots, X_n 是独立的，特征函数分别为 $\varphi_1(t), \varphi_2(t), \dots, \varphi_n(t)$ ，则随机变量 $Y = \sum_{k=1}^n X_k$ 的特征函数为

$$\varphi_Y(t) = \prod_{k=1}^n \varphi_k(t)$$

证明. 由 $\varphi_Y(t) = Ee^{it(X_1+X_2+\dots+X_n)} = \prod_{k=1}^n Ee^{itX_k} = \prod_{k=1}^n \varphi_k(t)$ ，得证。 \square

例 3.12. 已知随机变量 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ ，由例 3.4 知 $\varphi_{X_1}(t) = 1 + p(e^{it} - 1)$ 。根据定理 3.4，随机变量 $Y = X_1 + X_2 + \dots + X_n$ 的特征函数为

$$\varphi_Y(t) = [\varphi_{X_1}(t)]^n = [1 + p(e^{it} - 1)]^n$$

由练习 3.2，该特征函数也是二项分布 $B(n, p)$ 的特征函数。要严格证明 $Y \sim B(n, p)$ ，需要用到定理 3.14，该定理揭示随机变量由特征函数唯一确定，这是后话。

事实上，二项分布 $B(n, p)$ 可以由 n 个独立同分布于 $p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 的随机变量之和来定义，它的概率意义也是直观的。

练习 3.7. 设 $\varphi_{X_1}(t), \varphi_{X_2}(t), \dots, \varphi_{X_n}(t)$ 分别是独立的随机变量 X_1, X_2, \dots, X_n 的特征函数，求 $Y = c + \sum_{k=1}^n \alpha_k X_k$ 的特征函数。提示：利用 (3.9)，求得 $e^{itc} \prod_{k=1}^n \varphi_{X_k}(\alpha_k t)$ 。

例 3.13. 已知随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的特征函数 $\varphi_{\mathbf{X}}(t)$ ，其中 $t = (t_1, t_2, \dots, t_n)^\top$ ，求随机变量 $Y = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n$ 的特征函数，其中 $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}$ 。

解. 随机变量 $Y = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n$ 的特征函数为

$$\begin{aligned} \varphi_Y(t) &= E \left\{ e^{it(\alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n)} \right\} \\ &= \varphi_{\mathbf{X}}(\alpha_1 t, \alpha_2 t, \dots, \alpha_n t) \end{aligned}$$

譬如，例 3.7 中 $Z = X + Y$ 的特征函数为 $\varphi_Z(t) = \exp\{it(\mu_1 + \mu_2) - \frac{1}{2}(\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2)t^2\}$ ，正是正态分布 $N(\mu_1 + \mu_2, \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2)$ 的特征函数。请读者参考第 138 页的例 2.23，显然 $Z = X + Y$ 的特征函数也可以从例 2.23 的结论直接导出。

例 3.14. 已知两个独立的随机变量 X_1, X_2 的概率函数分别为 $P(X_1 = k) = \lambda_1^k e^{-\lambda_1}/k!$ 和 $P(X_2 = k) = \lambda_2^k e^{-\lambda_2}/k!$ ，其中 $k = 0, 1, 2, \dots$ 。求 $Y = X_1 - X_2$ 的特征函数。

解. 离散型随机变量 X_1 的特征函数为

$$\varphi_{X_1}(t) = \sum_{k=0}^{\infty} e^{itk} \frac{\lambda_1^k e^{-\lambda_1}}{k!} = e^{-\lambda_1} \sum_{k=0}^{\infty} \frac{(\lambda_1 e^{it})^k}{k!} = e^{-\lambda_1} \exp\{\lambda_1 e^{it}\} = \exp\{\lambda_1(e^{it} - 1)\}$$

根据结果 (3.9), $-X_2$ 的特征函数为 $\exp\{\lambda_2(e^{-it} - 1)\}$, 由定理 3.4 进而求得 Y 的特征函数 $\varphi_Y(t) = \exp\{\lambda_1 e^{it} + \lambda_2 e^{-it} - \lambda_1 - \lambda_2\}$ 。

 人们很自然地要问定理 3.4 的逆命题是否成立, 即如果 $Y = X_1 + X_2 + \dots + X_n$ 的特征函数为 $\varphi(t) = \prod_{i=1}^n \varphi_i(t)$, 能否判定随机变量 X_1, X_2, \dots, X_n 独立? 答案是“不能”, 构造两个反例如下。

例 3.15. 由练习 3.5, 随机变量 $X \sim \text{Cauchy}(0, 1)$ 的特征函数为 $\varphi_X(t) = \exp(-|t|)$ 。设 $Y = cX$, 其中 $c > 0$ 为常数。显然, X, Y 不独立。随机变量 Y 的特征函数为

$$\varphi_Y(t) = \exp(-c|t|), \text{ 并且 } \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

※例 3.16. 已知随机向量 $(X, Y)^\top$ 的密度函数为

$$f(x, y) = \begin{cases} \frac{1}{4}[1 + xy(x^2 - y^2)] & \text{如果 } |x| \leq 1 \text{ 且 } |y| \leq 1 \\ 0 & \text{否则} \end{cases} \quad (3.10)$$

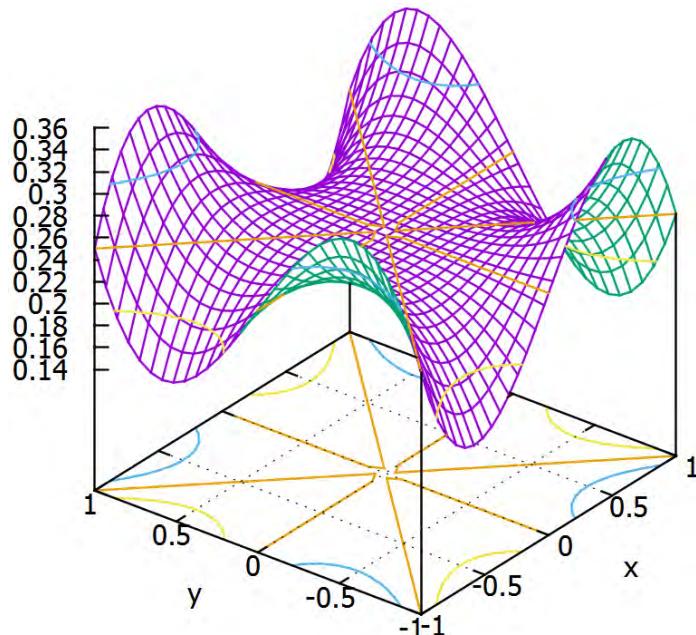


图 3.3: 由式 (3.10) 定义的密度函数曲面。求得 X 和 Y 的边缘分布都是 $U[-1, 1]$, 显然 $f(x, y) \neq f_X(x)f_Y(y)$, 这意味着随机变量 X, Y 并不独立。

根据第 154 页的例 2.35 的结果, 随机变量 $Z = X + Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z-x) dx$$

$$= \begin{cases} \int_{-1}^{z+1} f(x, z-x) dx = \frac{1}{4}(2+z) & \text{当 } -2 \leq z \leq 0 \\ \int_{z-1}^1 f(x, z-x) dx = \frac{1}{4}(2-z) & \text{当 } 0 < z \leq 2 \\ 0 & \text{当 } |z| > 2 \end{cases}$$

请读者验证随机变量 X, Y 的边缘分布都是 $U[-1, 1]$, 进而求得随机变量 X, Y, Z 的特征函数满足 $\varphi_Z(t) = \varphi_X(t)\varphi_Y(t)$, 但 X, Y 不独立。事实上,

$$\varphi_X(t) = \varphi_Y(t) = \frac{1}{2} \int_{-1}^1 e^{itx} dx = \frac{\sin t}{t}$$

$$\varphi_Z(t) = \frac{1}{4} \int_{-2}^0 (2+z)e^{itz} dz + \frac{1}{4} \int_0^2 (2-z)e^{itz} dz = \frac{\sin^2 t}{t^2}$$

3.1.2 特征函数与矩

定理 3.5. 如果随机变量 X 有 n 阶绝对矩, 则 X 的特征函数 $\varphi(t)$ 是 n 次可微的, 并且当 $k \leq n$ 时

$$\frac{d^k}{dt^k} \varphi(t) = E(i^k X^k e^{itX}) = \begin{cases} \sum_{n=1}^{\infty} i^k x_n^k e^{itx_n} p_n & \text{离散型} \\ \int_{-\infty}^{+\infty} i^k x^k e^{itx} f(x) dx & \text{连续型} \end{cases}$$

特别地, 随机变量 X 的 k 阶矩

$$m_k = \frac{\varphi^{(k)}(0)}{i^k} \quad (3.11)$$

证明. 因为随机变量 X 有 n 阶绝对矩, 所以对于任意自然数 $k \leq n$, 其 k 阶绝对矩也存在, 皆有 $\int_{-\infty}^{+\infty} |x|^k dF(x) < \infty$, 进而

$$\left| \int_{-\infty}^{+\infty} x^k e^{itx} dF(x) \right| \leq \int_{-\infty}^{+\infty} |x|^k dF(x) < \infty$$

因此, $\varphi^{(k)}(t)$ 存在。而结果 (3.11) 是因为 $\varphi^{(k)}(0) = i^k E(X^k) = i^k m_k$ 。

□

 定理 3.5 的逆命题不成立, 下面的例子说明特征函数在零点 k 阶可导, 但 k 阶矩不存在。所以, 结果 (3.11) 必须在保证 k 阶矩 m_k 存在的前提下才可以使用。

例 3.17. 设随机变量 X 的密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } |x| \leq 2 \\ \frac{c}{x^2 \ln|x|} & \text{当 } |x| > 2, \text{ 其中常数 } c \text{ 是使得 } \int_{-\infty}^{+\infty} f(x) dx = 1 \text{ 的归一因子} \end{cases}$$

经过计算, X 的特征函数 $\varphi(t)$ 满足 $\varphi'(0) = 0$ 。然而, $\int_2^s |x| f(x) dx = c(\ln \ln s - \ln \ln 2)$ 随着 $s \rightarrow \infty$ 而趋向无穷, 即 X 的期望不存在。

例 3.18. 对于标准正态分布 $X \sim N(0, 1)$, 其特征函数为 $\varphi(t) = \exp\{-t^2/2\}$ 。由式 (3.11) 易得 $\forall k \in \mathbb{N}$, $m_{2k-1} = 0$ 且 $m_{2k} = 1 \cdot 3 \cdot 5 \cdots (2k-1) = (2k-1)!!$ 。

推论 3.1. 若随机变量 X 的 $k = 1, 2, \dots, n$ 阶矩 m_k 都存在, 则当 $t \rightarrow 0$ 时, 特征函

数 $\varphi_X(t)$ 可表示为

$$\begin{aligned}\varphi_X(t) &= \sum_{k=0}^n \frac{\varphi_X^{(k)}(0)}{k!} t^k + o(t^n) \\ &= \sum_{k=0}^n m_k \frac{(it)^k}{k!} + o(t^n)\end{aligned}$$

\curvearrowleft 定理 3.6. 若随机变量 X 的任意 k 阶绝对矩 $\beta_k = E|X|^k$ 都存在，并且

$$\overline{\lim}_{k \rightarrow \infty} \frac{\sqrt[k]{\beta_k}}{k} = \frac{1}{eR} < \infty$$

则原点矩 $m_k, k = 1, 2, \dots$ 唯一决定 $F(x)$ ，并且对于 $|t| < R$ 总有

$$\varphi_X(t) = \sum_{k=0}^{\infty} m_k \frac{(it)^k}{k!} \quad (3.12)$$

\diamond 证明. 由第 189 页的性质 2.37 可知任意阶原点矩都存在。令 $t_0 \in (0, R)$ ，由 Stirling 公式我们有

$$\overline{\lim}_{k \rightarrow \infty} \frac{\sqrt[k]{\beta_k}}{k} < \frac{1}{et_0} \Rightarrow \overline{\lim}_{k \rightarrow \infty} \frac{\sqrt[k]{\beta_k t_0^k}}{k} < \frac{1}{e} \Rightarrow \lim_{k \rightarrow \infty} \sqrt[k]{\frac{\beta_k t_0^k}{k!}} < 1$$

根据正项级数的 Cauchy 收敛准则，级数 $\sum_{k=0}^{\infty} \beta_k t_0^k / k!$ 收敛，于是对于 $|t| < t_0$ ，级数 $\sum_{k=0}^{\infty} m_k (it)^k / k!$ 收敛。由推论 3.1 证得结果 (3.12)。原点矩 $m_k, k = 1, 2, \dots$ 唯一决定 $F(x)$ 的证明见 [145] 的第 295 页。

 在定理 3.6 的条件之下，在 $|t| < R$ 内特征函数被原点矩 $m_k, k = 1, 2, \dots$ 唯一决定。一般地，分布函数不能由原点矩 $m_k, k = 1, 2, \dots$ 唯一决定。

定义 3.5 (半不变量). 若随机变量 X 的各阶矩都存在，定义一个新函数 $\psi_X(t) = \ln \varphi_X(t)$ 。在开圆盘 $|z| < 1$ 上，解析函数 $\ln(1+z)$ 有如下的级数展开。

$$\ln(1+z) = \frac{z}{1} - \frac{z^2}{2} + \frac{z^3}{3} - \dots, \text{ 其中 } z \in \mathbb{C} \quad (3.13)$$

根据推论 3.1，若令 $z = \sum_{k=1}^{\infty} m_k (it)^k / k!$ ，则有

$$\psi_X(t) = \ln \varphi_X(t) = \ln(1+z) = \frac{z}{1} - \frac{z^2}{2} + \frac{z^3}{3} - \dots = \sum_{k=1}^{\infty} \varkappa_k \frac{(it)^k}{k!} \quad (3.14)$$

式 (3.14) 中，系数 \varkappa_k 被称作 k 阶半不变量。

练习 3.8. 接着[定义 3.5](#), 请读者验证半不变量的如下性质。

$$\begin{aligned}\kappa_k &= i^{-k} \psi_X^{(k)}(0) \\ E(X) &= -i\psi'_X(0) = \kappa_1 \\ V(X) &= -\psi''_X(0) = \kappa_2\end{aligned}$$

定理 3.7. 接着[定理 3.4](#), 随机变量 $Y = X_1 + X_2 + \dots + X_n$ 的 k 阶半不变量等于 X_1, X_2, \dots, X_n 各自的 k 阶半不变量之和。

例 3.19. 如果随机变量 X, Y 具有关系 $Y = X + b$, 其中 $b \neq 0$, 一般地 X 的 k 阶矩不等于 Y 的 k 阶矩。由 $\ln \varphi_Y(t) = b t + \ln \varphi_X(t)$ 易见, $\kappa_2, \kappa_3, \dots$ 却是变换 $Y = X + b$ 之下的不变量。

※例 3.20. 若随机变量 X 的各阶矩都存在, 利用 e^z 在 $z = 0$ 处的幂级数展开 $e^z = 1 + z + z^2/2! + \dots + z^n/n! + \dots$ 寻找矩与半不变量之间的关系。

$$\begin{aligned}\varphi(t) &= 1 + \sum_{k=1}^{\infty} m_k \frac{(it)^k}{k!} \\ &= \exp \left\{ \sum_{k=1}^{\infty} \kappa_k \frac{(it)^k}{k!} \right\} \\ &= 1 + \sum_{k=1}^{\infty} \kappa_k \frac{(it)^k}{k!} + \frac{1}{2!} \left[\sum_{k=1}^{\infty} \kappa_k \frac{(it)^k}{k!} \right]^2 + \frac{1}{3!} \left[\sum_{k=1}^{\infty} \kappa_k \frac{(it)^k}{k!} \right]^3 + \dots\end{aligned}$$

通过对比系数, 容易得到

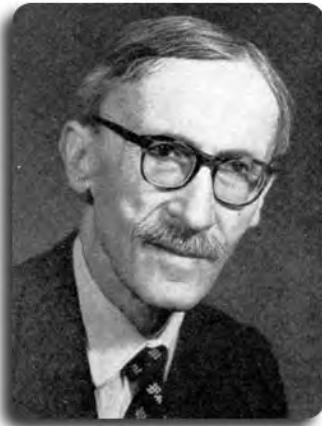
$$\begin{cases} m_1 = \kappa_1 \\ m_2 = \kappa_2 + \kappa_1^2 \\ m_3 = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3 \\ \dots \end{cases} \quad \text{或者} \quad \begin{cases} \kappa_1 = m_1 \\ \kappa_2 = m_2 - m_1^2 \\ \kappa_3 = m_3 - 3m_1m_2 + 2m_1^3 \\ \dots \end{cases}$$

3.2 特征函数与分布函数的关系

特征函数与分布函数之间是一一对应的，因此对分布函数序列的收敛性的研究可以转移到特征函数序列上。现代概率论的开拓者之一、法国数学家 Paul Pierre Lévy (1886-1971, 照片见下) 继 A. M. Lyapunov 之后于 1919-1925 年系统地建立了特征函数理论，其中反演公式和连续性定理最为重要，它们揭示了分布函数与特征函数之间的内在联系。

自 Lévy 的工作起，特征函数成为概率分析的重要工具，用于中心极限定理的证明（详见第 5 章）和独立增量过程的研究。在正式介绍 Lévy 反演公式和连续性定理这两个重要结果之前，有一些预备知识和概念需要交代清楚。首先，我们定义随机变量序列及其最弱的收敛方式依分布收敛^{*}，该收敛方式在实践中很常用，尤其是对中心极限定理而言。

定义 3.6 (随机变量序列). 若 $X_1, X_2, \dots, X_n, \dots$ 是定义在同一概率空间 (Ω, \mathcal{S}, P) 上的随机变量，则称 $X_1, X_2, \dots, X_n, \dots$ 为一个随机变量序列，简记为 $\{X_n\}_{n=1}^{\infty}$ 或 $\{X_n\}$ 。研究随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 的收敛性直观上就是看 n 很大时 X_n 近似地服从怎样的分布。



定义 3.7. 一个随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 称为独立的当且仅当对任意 $n = 2, 3, \dots$ 皆有 X_1, X_2, \dots, X_n 相互独立。

例 3.21. 随机试验中事件 A 发生的概率 $P(A)$ 的直观含义是：在相同条件下的多次重复试验中，事件 A 发生的频率之稳定值（见例 1.48）。为了更明确地表述概率的频率解释，现定义随机变量 X_j 如下，

$$X_j = \begin{cases} 1 & \text{在第 } j \text{ 次试验中, 事件 } A \text{ 发生} \\ 0 & \text{在第 } j \text{ 次试验中, 事件 } A \text{ 未发生} \end{cases}$$

显然，随机变量 X_1, X_2, \dots, X_n 相互独立，事件 A 在 n 次重复试验中发生的频率为 $Y_n = \frac{1}{n} \sum_{j=1}^n X_j$ 。对于随机变量序列 $\{Y_n\}_{n=1}^{\infty}$ ，概率的频率解释意味着当 n 很大时，随机变量 Y_n 近似地服从单点分布 $\langle P(A) \rangle$ 。

^{*}第 5 章还将介绍几乎必然收敛、依概率收敛，并讨论这些收敛方式间的关系。

例 3.22. 已知随机变量 X_n 的分布函数是

$$F_n(x) = \begin{cases} \left(\frac{x}{x+1}\right)^n & \text{若 } x \geq 0 \\ 0 & \text{若 } x < 0 \end{cases}$$

显然, $\lim_{n \rightarrow \infty} F_n(x) = 0$ 不满足右连续性, 因此并不是分布函数 (见第 119 页的定理 2.3)。

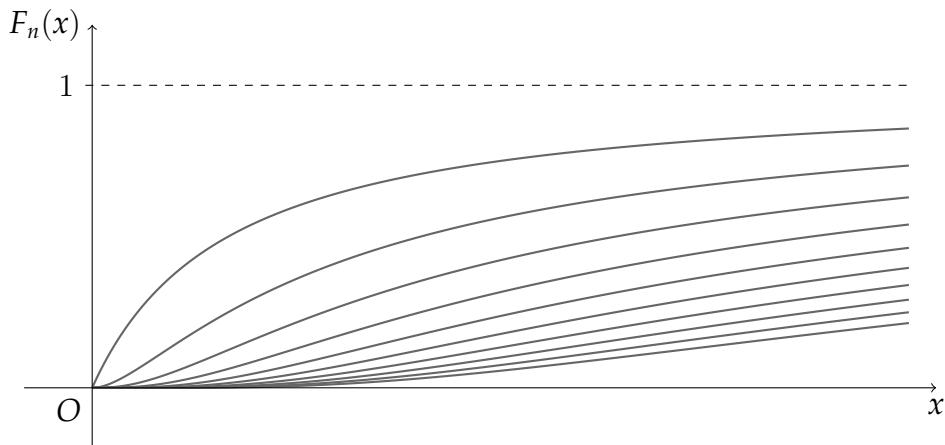


图 3.4: 例 3.22 中的分布函数序列 $F_n(x)$ 的极限不是分布函数。

以奥地利数学家 Eduard Helly (1884-1943, 照片见右) 命名的下述两个定理在研究分布函数序列的极限时非常有用, 其证明详见 Gnedenko 的《概率论教程》[58] 第七章。Helly 第一定理断言从任一分布函数序列中总能“沙里淘金”地找到一个收敛的子序列, 其极限“差不多”是一个分布函数。



定理 3.8 (Helly 第一定理, 1923). 对于任意给定的分布函数的序列 $\{F_n(x)\}$, 总存在一个子序列 $\{F_{n_k}(x)\}$ 和一个非减的、右连续函数 $F(x)$ 使得对于 $F(x)$ 的任意连续点 x 皆有

$$F(x) = \lim_{k \rightarrow \infty} F_{n_k}(x)$$

例 3.23. 在定理 3.8 中, $F(x)$ 不一定是分布函数。例如, 构造分布函数

$$F_n(x) = aI_{[n, +\infty)} + bI_{[-n, +\infty)} + cG(x)$$

其中, $G(x)$ 是一个分布函数且常数 $a, b, c \in (0, 1)$ 满足 $a + b + c = 1$ 。显然,

$$F_n(x) \rightarrow F(x) = b + cG(x)$$

然而, $F(x)$ 不是一个分布函数, 原因是 $F(-\infty) = b, F(+\infty) = 1 - a$ 。

定理 3.9 (Helly 第二定理, 1923). 如果分布函数的序列 $\{F_n(x)\}$ 在非减函数 $F(x)$ 的连续点上收敛于 $F(x)$, 并且 $F(-\infty) = 0, F(\infty) = 1$, 则对于 \mathbb{R} 上的任意连续函数 $g(x)$ 皆有

$$\lim_{n \rightarrow \infty} \int_{-\infty}^{+\infty} g(x) dF_n(x) = \int_{-\infty}^{+\infty} g(x) dF(x) \quad (3.15)$$

若能保证分布函数的序列 $\{F_n(x)\}$ 在分布函数 $F(x)$ 的连续点上收敛于 $F(x)$, 根据定理 3.9 就能拥有式 (3.15) 这样好的性质——求极限和求积分交换次序, 下面的定义正是为了达到该目的而提炼出“依分布收敛”这一概念。

定义 3.8 (依分布收敛). 已知 $\{F_n(x)\}$ 是随机变量序列 $\{X_n\}$ 对应着的分布函数序列, 如果对于分布函数 $F_X(x)$ 的任意连续点 x 皆有

$$\lim_{n \rightarrow \infty} F_n(x) = F_X(x)$$

则称随机变量序列 $\{X_n\}$ 依分布收敛 (converge in law/distribution) 于随机变量 X , 记作 $X_n \xrightarrow{L} X$ 。有时候, 也称分布函数序列 $\{F_n(x)\}$ 弱收敛于分布函数 $F_X(x)$ 。

 **例 3.22** 说明, 分布函数序列的极限未必是分布函数。所以, 在**定义 3.8** 中, 极限函数要求必须是分布函数。在上下文中, 只要不引起歧义, 我们常把“随机变量序列 $\{X_n\}$ 依分布收敛到单点分布 $X \sim \langle \mu \rangle$ ”简记作 $X_n \xrightarrow{L} \langle \mu \rangle$, 或者 $X_n \xrightarrow{L} \mu$ 。

例 3.24. 若随机变量序列 $X_n \sim N(\mu_n, \sigma_n^2), n = 1, 2, \dots$ 满足 $\lim_{n \rightarrow \infty} \mu_n = \mu$ 和 $\lim_{n \rightarrow \infty} \sigma_n^2 = \sigma^2$, 则 $X_n \xrightarrow{L} X$, 其中随机变量 $X \sim N(\mu, \sigma^2)$ 。

例 3.25. 随机变量序列 $X_n \sim \langle 1/n \rangle, n = 1, 2, \dots$ 依分布收敛于单点分布的随机变量 $X \sim \langle 0 \rangle$ 。事实上, X_n 和 X 的分布函数具有关系

$$F_n(x) = \begin{cases} 0 & \text{当 } x < 1/n \\ 1 & \text{当 } x \geq 1/n \end{cases} \quad \text{弱收敛于 } F(x) = \begin{cases} 0 & \text{当 } x < 0 \\ 1 & \text{当 } x \geq 0 \end{cases}$$

在 $F(x)$ 的不连续点 $x = 0$ 上,

$$\lim_{n \rightarrow \infty} F_n(0) = \lim_{n \rightarrow \infty} 0 \neq 1 = F(0)$$

而在 $F(x)$ 的任意连续点 $x = x_0 \neq 0$ 上, 皆有

$$\lim_{n \rightarrow \infty} F_n(x_0) = F(x_0)$$

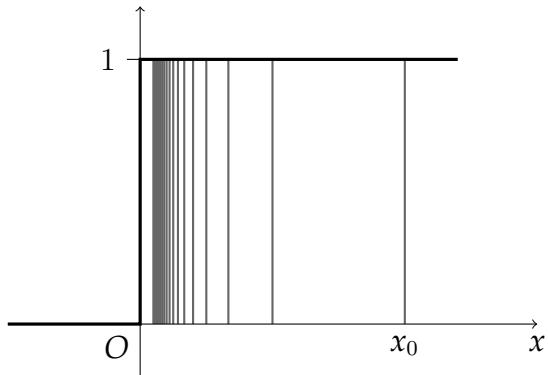


图 3.5: 随机变量序列 $X_n \sim \langle 1/n \rangle, n = 1, 2, \dots$ 对应的分布函数序列 $F_n(x)$ 弱收敛于单点分布 $\langle 0 \rangle$ 的分布函数, 即 $X_n \xrightarrow{L} 0$ 。

例 3.26. 已知取值范围为 $[0, 1]$ 的随机变量 X_n 的分布函数是

$$F_n(x) = x - \frac{\sin(2n\pi x)}{2n\pi}, \text{ 若 } 0 \leq x \leq 1$$

显然, 在闭区间 $[0, 1]$ 上, $\lim_{n \rightarrow \infty} F_n(x) = x$ 是均匀分布 $U[0, 1]$ 的分布函数, 所以 $X_n \xrightarrow{L} X$, 其中 $X \sim U[0, 1]$ 。

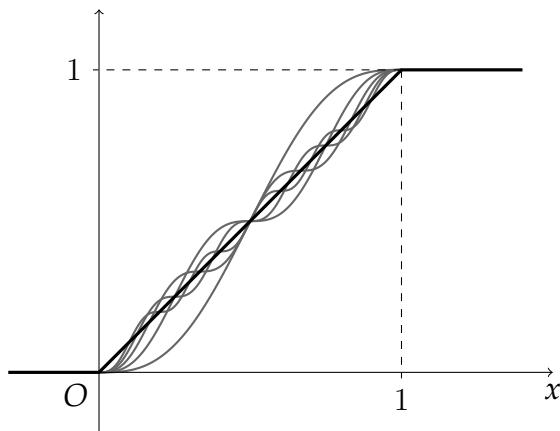


图 3.6: 例 3.26 定义的分布函数序列 $\{F_n(x)\}$ 弱收敛于 $U[0, 1]$ 的分布函数。

连续型随机变量序列可以依分布收敛到离散型随机变量, 离散型随机变量序列也可以依分布收敛到连续型随机变量, 请看下面的两个例子。

例 3.27. 随机变量序列 $Y_n \sim N(\mu, \sigma^2/n), n = 1, 2, \dots$ 依分布收敛于 $Y \sim \langle \mu \rangle$, 即在

$F_Y(y)$ 的任意连续点 $y \neq \mu$ 上, $\lim_{n \rightarrow \infty} F_{Y_n}(y) = F_Y(y)$ 。事实上,

$$\lim_{n \rightarrow \infty} F_{Y_n}(y) = \lim_{n \rightarrow \infty} \Phi(y|\mu, \sigma^2/n) = \begin{cases} 1 & \text{若 } y > \mu \\ 0 & \text{若 } y < \mu \end{cases}$$

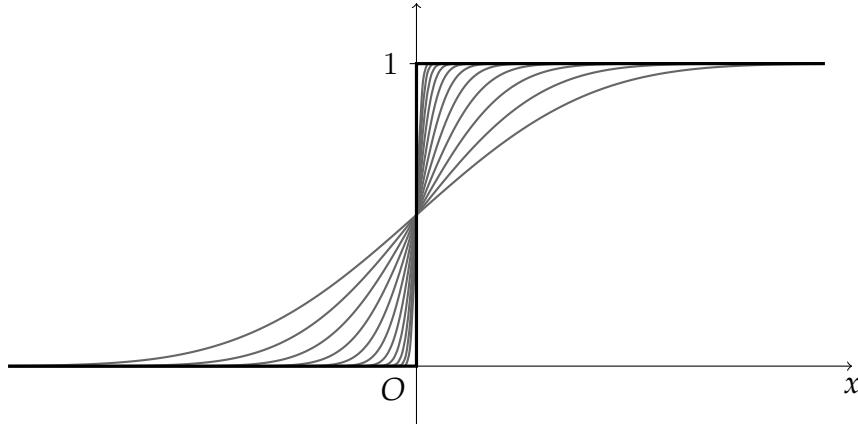


图 3.7: 随机变量序列 $Y_n \sim N(0, 2^{-n})$, $n = 0, 1, 2, \dots$ 对应的分布函数序列 $F_n(y) = \Phi(y|0, 2^{-n})$ 弱收敛于单点分布 $\langle 0 \rangle$ 的分布函数, 即 $Y_n \xrightarrow{L} 0$ 。

练习 3.9. 若随机变量 X_n 的分布函数 $F_n(x) = 1 - (1-x)^n$, 其中 $0 \leq x \leq 1$, 试证明 $X_n \xrightarrow{L} 0$ 。提示: 仿照例 3.25 和例 3.27。

例 3.28. 设离散型随机变量 X_n 等概率地取值 $\frac{1}{n}, \frac{2}{n}, \dots, 1$, 则 $X_n \xrightarrow{L} X \sim U[0, 1]$ 。

证明. X_n 的分布函数是

$$F_{X_n}(x) = \begin{cases} 0 & \text{若 } x \leq 0 \\ \frac{\lfloor nx \rfloor}{n} & \text{若 } 0 \leq x \leq 1 \\ 1 & \text{若 } x \geq 1 \end{cases}$$

根据 $nx - 1 \leq \lfloor nx \rfloor \leq nx$ 可知 $n \rightarrow \infty$ 时, $F_{X_n}(x) \rightarrow x$, 其中 $x \in [0, 1]$ 。 \square

在上例中, 对于所有的 $n \in \mathbb{N}$ 皆有 $P(X_n \in \mathbb{Q}) = 1$, 而 $P(X \in \mathbb{Q}) = 0$ 。依分布收敛并不能把 X_n 的所有属性带给它的极限 X 。

\nwarrow **性质 3.2.** 依分布收敛还有一些等价的定义: $X_n \xrightarrow{L} X$ 当且仅当

- ① 对于任意有界连续函数 g 皆有 $Eg(X_n) \rightarrow Eg(X)$ 。
- ② 对于任意非负连续函数 g 皆有 $\liminf_{n \rightarrow \infty} Eg(X_n) \geq Eg(X)$ 。
- ③ 对于任意开集 G 皆有 $\liminf_{n \rightarrow \infty} P(X_n \in G) \geq P(X \in G)$ 。

④ 对于任意闭集 F 皆有 $\limsup_{n \rightarrow \infty} P(X_n \in F) \leq P(X \in F)$ 。

※证明. 详见 A. W. van der Vaart 的《渐近统计学》[153] 的第二章的引理 2.2, 或者 P. Billingsley 的《概率测度的收敛性》[14] 的第一章第三节, 本书不作要求。□

本节内容

介绍特征函数的两个重要结果: (1) 由 Lévy 反演公式推导出的唯一性定理揭示了分布函数与特征函数之间的一一对应关系; (2) Lévy 连续性定理保证了对随机变量序列依分布收敛的研究可以转嫁到考察其特征函数序列的收敛性上, 反之亦然。

关键知识

(1) Lévy 反演公式; (2) Lévy 连续性定理; (3) 利用特征函数判定独立性, 如 Fisher 定理 3.16 的证明中所使用的方法。

3.2.1 Lévy 反演公式

已知随机变量 X 的特征函数 $\varphi(t)$, 若 $x_1 < x_2$ 是 $F(x)$ 的两个连续点, 反演公式揭示 $P(x_1 < X \leq x_2)$ 可通过 $\varphi(t)$ 来计算。唯一性定理说明特征函数承载了随机变量的所有信息, 它是分布函数的替代品。

\nwarrow 定理 3.10 (Lévy 反演公式). 已知随机变量 X 的分布函数和特征函数分别为 $F(x)$ 和 $\varphi(t)$, 假定 $F(x)$ 在 $x_0 \pm h$ 上连续 ($h > 0$), 则

$$F(x_0 + h) - F(x_0 - h) = \lim_{s \rightarrow \infty} \frac{1}{\pi} \int_{-s}^s \frac{\sin(ht)}{t} e^{-itx_0} \varphi(t) dt \quad (3.16)$$

※证明. 在计算式 (3.16) 右边的极限之前, 先对它做一些整理。

$$\begin{aligned} J_s &= \frac{1}{\pi} \int_{-s}^s \frac{\sin(ht)}{t} e^{-itx_0} \varphi(t) dt \\ &= \frac{1}{\pi} \int_{-s}^s \left\{ \int_{-\infty}^{+\infty} \frac{\sin(ht)}{t} e^{-itx_0} e^{itx} dF(x) \right\} dt \\ &= \frac{1}{\pi} \int_{-\infty}^{+\infty} \left\{ \int_{-s}^s \frac{\sin(ht)}{t} e^{itx-itx_0} dF(x) \right\} dt \\ &= \int_{-\infty}^{+\infty} \left\{ \frac{2}{\pi} \int_0^s \frac{\sin(ht)}{t} \cos[(x-x_0)t] dt \right\} dF(x) \\ &= \int_{-\infty}^{+\infty} g_s(x) dF(x), \text{ 其中 } g_s(x) = \frac{2}{\pi} \int_0^s \frac{\sin(ht)}{t} \cos[(x-x_0)t] dt \end{aligned}$$

在上面第三步的推导中, 利用了数学分析中的 Fubini 定理*, 两个积分之所以可以交换次序是因为对 t 的积分是有限的, 并且

$$\int_{-\infty}^{+\infty} \left| \frac{\sin(ht)}{t} e^{itx-itx_0} \right| dF(x) \leq h$$

■ 接着对 $g_s(x)$ 进一步做整理得

$$g_s(x) = \frac{1}{\pi} \int_0^s \left\{ \frac{\sin[(x-x_0+h)t]}{t} - \frac{\sin[(x-x_0-h)t]}{t} \right\} dt$$

由数学分析的知识, 积分 $\int_0^s \frac{\sin x}{x} dx$ 对所有的 $s > 0$ 皆是有界的, 于是 $|g_s(x)|$ 有

*意大利数学家 Guido Fubini (1879-1943) 发现: 如果实函数 $f(\mathbf{x}, \mathbf{y})$ 在 $A \times B \subseteq \mathbb{R}^m \times \mathbb{R}^n$ 上绝对可积, 则 $f(\mathbf{x}, \mathbf{y})$ 在 $A \times B$ 上的多重积分和累次积分的结果相同, 即

$$\int_{A \times B} f(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}) = \int_A d\mathbf{x} \int_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \int_B d\mathbf{y} \int_A f(\mathbf{x}, \mathbf{y}) d\mathbf{x}$$

界，并且有下面的结果（读者也可以用 Maxima 验证之），

$$\lim_{s \rightarrow \infty} \int_0^s \frac{\sin x}{x} dx = \frac{\pi}{2}$$

进而， $\lim_{s \rightarrow \infty} \frac{1}{\pi} \int_0^s \frac{\sin(\beta t)}{t} dt = \begin{cases} \frac{1}{2} & \text{当 } \beta > 0 \\ 0 & \text{当 } \beta = 0 \\ -\frac{1}{2} & \text{当 } \beta < 0 \end{cases}$

$$\lim_{s \rightarrow \infty} g_s(x) = \begin{cases} 0 & \text{当 } |x - x_0| > h \\ \frac{1}{2} & \text{当 } |x - x_0| = h \\ 1 & \text{当 } |x - x_0| < h \end{cases}$$

□ 由第 763 页的 Lebesgue 控制收敛定理 D.4, $\lim_{s \rightarrow \infty} \int_{-\infty}^{+\infty} g_s(x) dF(x)$ 中的求极限与求积分可交换次序，于是

$$\lim_{s \rightarrow \infty} J_s = \int_{-\infty}^{+\infty} \lim_{s \rightarrow \infty} g_s(x) dF(x) = \int_{x_0-h}^{x_0+h} dF(x) = F(x_0 + h) - F(x_0 - h)$$

最后一步是因为 $F(x)$ 在 $x_0 \pm h$ 处连续，结果 (3.16) 得证。 □

推论 3.2. 定理 3.10 所揭示的反演公式 (3.16) 有下面两个等价形式，

□ 若 x_1, x_2 是分布函数 $F(x)$ 的两个连续点，不妨设 $x_1 < x_2$ ，则

$$F(x_2) - F(x_1) = \lim_{s \rightarrow \infty} \frac{1}{2\pi} \int_{-s}^s \frac{e^{-itx_1} - e^{-itx_2}}{it} \varphi(t) dt$$

□ 令 x 是 $F(x)$ 的连续点， y 沿着 $F(x)$ 的连续点趋向 $-\infty$ ，则

$$F(x) = \lim_{y \rightarrow -\infty} \lim_{s \rightarrow \infty} \frac{1}{2\pi} \int_{-s}^s \frac{e^{-ity} - e^{-itx}}{it} \varphi(t) dt \quad (3.17)$$

下面的结论揭示了密度函数与特征函数之间的关系。

~**推论 3.3** (Fourier 反演). 已知 $\varphi(t)$ 是某个连续型随机变量 X 的特征函数且在 \mathbb{R} 上可积，则 X 具有有界连续密度函数 $f(x) = \mathcal{F}^{-1}(\varphi)$ ，即

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi(t) dt \quad (3.18)$$

证明. 由于 $\varphi(t)$ 在 \mathbb{R} 上绝对可积，Lévy 反演公式 (3.16) 进



一步简化为

$$\frac{F(x + \Delta x) - F(x)}{\Delta x} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\sin(t\Delta x/2)}{t\Delta x/2} e^{-it(x+\Delta x/2)} \varphi(t) dt$$

令 $\Delta x \rightarrow 0$, 因为 $\varphi(t)$ 绝对可积, 求极限可与求积分交换次序,

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \lim_{\Delta x \rightarrow 0} \frac{\sin(t\Delta x/2)}{t\Delta x/2} e^{-it(x+\Delta x/2)} \varphi(t) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi(t) dt \end{aligned} \quad \square$$

例 3.29. 由**例 3.6** 知, $X \sim N(\mu, \sigma^2)$ 的特征函数为 $\varphi(t) = \exp(it\mu - \sigma^2 t^2/2)$, 下面利用配方法和结果 (3.7) 来验证 Fourier 反演公式 (3.18)。

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi(t) dt &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp \left\{ -it(x - \mu) - \frac{\sigma^2 t^2}{2} \right\} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp \left\{ -\frac{\left[t + i \left(\frac{x-\mu}{\sigma^2} \right) \right]^2}{2\sigma^2} - \frac{(x-\mu)^2}{2\sigma^2} \right\} dt \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \end{aligned}$$

练习 3.10. 利用 Fourier 反演公式 (3.18), 验证特征函数 $\varphi(t)$ 所对应的密度函数。

特征函数	\leftrightarrow	密度函数
$\varphi(t) = \begin{cases} 1 - t & \text{当 } t \leq 1 \\ 0 & \text{当 } t > 1 \end{cases}$	\leftrightarrow	$f(x) = \frac{1 - \cos x}{\pi x^2}$
$\varphi(t) = \exp\{- t \}$	\leftrightarrow	$f(x) = \frac{1}{\pi(x^2 + 1)}$

例 3.30. 第 157 页的**例 2.39** 也可以通过特征函数的方法求解。由**例 3.5**, $U[0, 1]$ 的特征函数为 $\varphi(t) = (e^{it} - 1)/(it)$, 于是 $Z = X + Y$ 的特征函数为

$$\varphi_Z(t) = [\varphi(t)]^2 = -\frac{(e^{it} - 1)^2}{t^2}$$

进而, 由 Fourier 反演公式 (3.18) 得到 Z 的密度函数。

$$f_Z(z) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itz} \varphi_Z(t) dt = -\frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itz} \frac{(e^{it} - 1)^2}{t^2} dt = \begin{cases} z & \text{当 } 0 < z \leq 1 \\ 2 - z & \text{当 } 1 < z \leq 2 \\ 0 & \text{其他} \end{cases}$$

作为定理 3.10 的补充, 下面不加证明地给出刻画特征函数和分布函数关系的另外两个结果。

定理 3.11. 已知随机变量 X 的分布函数为 $F(x)$,

□ 若 $F(x)$ 在 $x = x_0$ 处不连续, 则

$$F(x_0) - F(x_0-) = \lim_{s \rightarrow \infty} \frac{1}{2s} \int_{-s}^s e^{-itx_0} \varphi(t) dt$$

□ J. Gil-Pelaez 于 1951 年证得: 若 $F(x)$ 在 $x = x_0$ 处连续, 则

$$F(x_0) = \frac{1}{2} - \frac{1}{\pi} \int_0^{+\infty} \frac{\Im[e^{-itx_0} \varphi(t)]}{t} dt$$

其中, $\Im(z)$ 表示复数 $z \in \mathbb{C}$ 的虚部。

△ 定理 3.12 (分布列的反演公式). 对于离散型随机变量 X , 不妨设其取值范围是 \mathbb{Z} , 令 $p_k = P(X = k)$, 其中 $k \in \mathbb{Z}$, 则 X 的特征函数 $\varphi(t) = \sum_{k=-\infty}^{+\infty} p_k e^{itk}$ 的反演公式如下。

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \varphi(t) dt \quad (3.19)$$

证明. 如果 $l \neq k$, 则

$$\int_{-\pi}^{\pi} e^{-it(k-l)} dt = \int_{-\pi}^{\pi} \{\cos[t(k-l)] - i \sin[t(k-l)]\} dt = 0$$

结果 (3.19) 由下面的事实立得,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \varphi(t) dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ p_k + \sum_{l \neq k} p_l e^{-it(k-l)} \right\} dt = p_k \quad \square$$

练习 3.11. 由例 3.4 知, 0-1 分布 $p|1\rangle + (1-p)|0\rangle$ 的特征函数为 $\varphi(t) = 1 + p(e^{it} - 1)$ 。请用该特征函数验证分布列的反演公式 (3.19)。

定理 3.13 (随机向量的 Lévy 反演公式). 类似推论 3.2, 设 $\varphi(t_1, t_2, \dots, t_n)$ 是随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的特征函数, 并且 \mathbf{X} 落在长方体 $(a_1, b_1] \times (a_2, b_2] \times \dots \times (a_n, b_n]$ 边界上的概率为零, 则

$$\begin{aligned} & P\{a_k < X_k \leq b_k, k = 1, 2, \dots, n\} \\ &= \lim_{s \rightarrow \infty} \frac{1}{(2\pi)^n} \int_{-s}^s \cdots \int_{-s}^s \prod_{k=1}^n \frac{e^{-it_k a_k} - e^{-it_k b_k}}{it_k} \varphi(t_1, t_2, \dots, t_n) dt_1 dt_2 \cdots dt_n \end{aligned} \quad (3.20)$$

\nwarrow 定理 3.14 (唯一性定理). 两个分布函数 $F_1(x)$ 和 $F_2(x)$ 恒等当且仅当它们的特征函数 $\varphi_1(t)$ 和 $\varphi_2(t)$ 相同。即，分布函数 $F_X(x)$ 与特征函数 $\varphi_X(t)$ 相互唯一决定。对于 n 维随机向量 \mathbf{X} 也有类似的结果。

证明. “ \Rightarrow ” 是显然的。现在往证 “ \Leftarrow ”: 若 $\forall t \in \mathbb{R}, \varphi_1(t) = \varphi_2(t)$, 记 A 为 $F_1(x), F_2(x)$ 的不连续点集, 它至多可数。(1) 对于 $x \notin A$, 由式 (3.17) 得 $F_1(x) = F_2(x)$ 。(2) 对于 $y \in A$, 取一列 $x_n \notin A$ 满足 $x_n \downarrow y$, 由分布函数的右连续性知,

$$F_1(y) = \lim_{n \rightarrow \infty} F_1(x_n) = \lim_{n \rightarrow \infty} F_2(x_n) = F_2(y)$$

例 3.31. 按照定理 3.3 中的 Pólya 准则, $\varphi_1(t) = \exp\{-|t|\}$ 和 $\varphi_2(t) = \max\{\varphi_1(t), 2/[e(1+|t|)]\}$ 都是特征函数。这两个特征函数仅在区间 $[-1, 1]$ 上重叠, 它们对应着不同的密度函数。所以, 特征函数的局部性质无法决定密度函数。

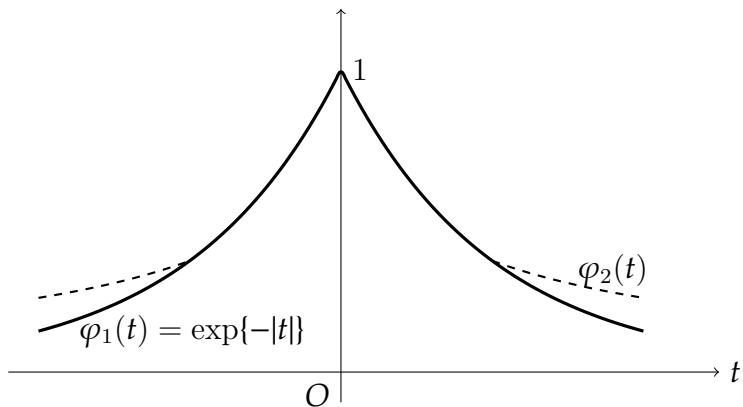


图 3.8: 特征函数 $\varphi_1(t) = \exp\{-|t|\}$ 和 $\varphi_2(t) = \max\{\varphi_1(t), 2/[e(1+|t|)]\}$ 。

\nwarrow 定理 3.15 (用特征函数判定独立性). 随机变量 X, Y 相互独立当且仅当 $(X, Y)^\top$ 的特征函数 $\varphi(s, t) = \varphi_X(s)\varphi_Y(t)$ 。此结论可自然推广至随机向量。

证明. 往证 “ \Rightarrow ”: $\varphi(s, t) = \mathbb{E}e^{isX+itY} = \mathbb{E}e^{isX} \cdot \mathbb{E}e^{itY} = \varphi_X(s)\varphi_Y(t)$ 。往证 “ \Leftarrow ”: 只证连续的情形, 离散的情形留作练习。

$$\begin{aligned} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{isx+ity} f(x, y) dx dy &= \int_{-\infty}^{+\infty} e^{isx} f_X(x) dx \int_{-\infty}^{+\infty} e^{ity} f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{isx+ity} f_X(x) f_Y(y) dx dy \end{aligned}$$

由唯一性定理 3.14 知 $f(x, y) = f_X(x)f_Y(y)$, 即 X, Y 相互独立。 \square

利用定理 3.15 和定理 3.14 来验证随机变量 X, Y 相互独立, 本质上与验证 $f(x, y) = f_X(x)f_Y(y)$ 是等价的。但有时需要借特征函数的道儿绕开复杂的联合分布, 譬如下面的重要结果 (在第 148 页的例 2.29 中已经介绍过了, 但还欠一个证明)。

\nwarrow 定理 3.16 (Fisher, 1925). 已知 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 则随机变量 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ 与随机向量 $(X_1 - \bar{X}, \dots, X_n - \bar{X})^\top$ 相互独立。

证明. 设随机向量 $(\bar{X}, X_1 - \bar{X}, \dots, X_n - \bar{X})^\top$ 的特征函数为 $\varphi(t, t_1, \dots, t_n)$, 由定义 3.4, 我们先写出该特征函数, 再根据条件 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ 加以整理。

$$\begin{aligned}\varphi(t, t_1, \dots, t_n) &= E \exp\{it\bar{X} + it_1(X_1 - \bar{X}) + \dots + it_n(X_n - \bar{X})\} \\ &= E \exp\left\{i \sum_{k=1}^n X_k \left(t_k - \frac{t_1 + \dots + t_n - t}{n}\right)\right\} \\ &= \prod_{k=1}^n E \exp\left\{\frac{iX_k[t + n(t_k - \bar{t})]}{n}\right\}, \text{ 其中 } \bar{t} = \frac{t_1 + t_2 + \dots + t_n}{n}\end{aligned}$$

由定理 3.3 中的第一个结果, \bar{X} 和 $(X_1 - \bar{X}, \dots, X_n - \bar{X})^\top$ 的特征函数分别为 $\varphi(t, 0, \dots, 0)$ 和 $\varphi(0, t_1, \dots, t_n)$ 。由于 $X_k \sim N(\mu, \sigma^2)$, 其特征函数为 $\varphi(t) = \exp\{it\mu - \sigma^2 t^2/2\}$, 所以

$$\begin{aligned}\varphi(t, t_1, \dots, t_n) &= \prod_{k=1}^n \exp\left\{\frac{i[t + n(t_k - \bar{t})]\mu}{n} - \frac{\sigma^2[t + n(t_k - \bar{t})]^2}{2n^2}\right\} \\ &= \exp\left\{it\mu - \frac{\sigma^2 t^2}{2n}\right\} \exp\left\{-\frac{\sigma^2 \sum_{k=1}^n (t_k - \bar{t})^2}{2}\right\} \\ &= \varphi(t, 0, \dots, 0) \varphi(0, t_1, \dots, t_n) \\ &= \varphi_{\bar{X}}(t) \varphi_{X_1 - \bar{X}, \dots, X_n - \bar{X}}(t_1, \dots, t_n)\end{aligned}$$

由定理 3.15 知, \bar{X} 与 $(X_1 - \bar{X}, \dots, X_n - \bar{X})^\top$ 相互独立。 \square

3.2.2 Lévy 连续性定理

\nwarrow 定理 3.17. 已知随机变量序列 $\{X_n\}$ 对应着的分布函数和特征函数序列分别是 $\{F_n(x)\}$ 和 $\{\varphi_n(t)\}$ 。下面两个结果合称为 Lévy 连续性定理 (Lévy's continuity theorem)。

① 正极限定理: 若 $X_n \xrightarrow{L} X$, 则 $\{\varphi_n(t)\}$ 收敛于 X 的特征函数 $\varphi(t)$, 并且在 t 的任一有限区间上收敛是一致的 (函数序列 $\{\varphi_n(t)\}$ 一致收敛的定义见第 758 页)。

② 逆极限定理: 设特征函数序列 $\{\varphi_n(t)\}$ 收敛于一个在 $t = 0$ 处连续的函数 $\varphi(t)$, 则 $\varphi(t)$ 是某随机变量 X 特征函数, 而且 $X_n \xrightarrow{L} X$ 。

粗略地说, Lévy 连续性定理保证了下列图表交换: 研究 $X_n \xrightarrow{L} X$ 可转嫁到 $X_n \Leftrightarrow \varphi_n(t) \rightarrow \varphi(t) \Leftrightarrow X$ 上, 这是数学里的“迂回战术”。

$$\begin{array}{ccc} X_n & \xrightarrow{L} & X \\ \Updownarrow & & \Updownarrow \\ \varphi_n(t) & \rightarrow & \varphi(t) \end{array}$$

证明. 往证正极限定理: 令 $a < 0$ 和 $b > 0$ 是 $F(x)$ 的两个连续点。

$$\begin{aligned} \varphi_n(t) &= \int_{-\infty}^a e^{itx} dF_n(x) + \int_a^b e^{itx} dF_n(x) + \int_b^{+\infty} e^{itx} dF_n(x) = C_{n1} + C_{n2} + C_{n3} \\ \varphi(t) &= \int_{-\infty}^a e^{itx} dF(x) + \int_a^b e^{itx} dF(x) + \int_b^{+\infty} e^{itx} dF(x) = C_1 + C_2 + C_3 \end{aligned}$$

对于任意的 $\epsilon > 0$, 总可以令 $|a|$ 足够地大, 使得

$$|C_{n1} - C_1| \leq \int_{-\infty}^a dF_n(x) + \int_{-\infty}^a dF(x) = F_n(a) + F(a) < \frac{\epsilon}{6} + \frac{\epsilon}{6} = \frac{\epsilon}{3}$$

类似地, 令 $|b|$ 足够地大, 使得 $|C_{n3} - C_3| \leq \epsilon/3$ 。对于 t 的任意一个有限区域, 总存在 $N \in \mathbb{N}$ 使得当 $n > N$ 时,

$$|C_{n2} - C_2| \leq |F_n(b) - F(b)| + |F_n(a) - F(a)| + |t| \int_a^b |F_n(x) - F(x)| dx < \frac{\epsilon}{9} + \frac{\epsilon}{9} + \frac{\epsilon}{9} = \frac{\epsilon}{3}$$

于是, 在 t 的一个有限区域上, $\{\varphi_n(t)\}$ 一致收敛到 $\varphi(t)$ 。 □

※证明. 往证逆极限定理: 显然 $0 \leq F(x) \leq 1$ 。为了证明 $F(+\infty) - F(-\infty) = 1$, 下面用归谬法, 令 $a = F(+\infty) - F(-\infty) < 1$ 。因为 $\varphi(0) = \lim_{n \rightarrow \infty} \varphi_n(0) = 1$ 且 $\varphi(t)$ 连续, 于

是 $\forall \epsilon \in (0, 1 - a)$, 存在 $T > 0$ 使得下式成立。

$$\frac{1}{2T} \left| \int_{-T}^T \varphi(t) dt \right| > 1 - \frac{\epsilon}{2} > a + \frac{\epsilon}{2}$$

由于 $\lim_{k \rightarrow \infty} \varphi_{n_k}(t) = \varphi(t)$, 对足够大的 k 有

$$\frac{1}{2T} \left| \int_{-T}^T \varphi_{n_k}(t) dt \right| > a + \frac{\epsilon}{2} \quad (3.21)$$

令 $b > 4/(T\epsilon)$, 对足够大的 k 有 $a_k = F_{n_k}(b) - F_{n_k}(-b) < a + \epsilon/4$ 。

$$\begin{aligned} \frac{1}{2T} \left| \int_{-T}^T \varphi_{n_k}(t) dt \right| &= \frac{1}{2T} \left| \int_{-T}^T \left\{ \int_{-\infty}^{+\infty} e^{itx} dF_{n_k}(x) \right\} dt \right| \\ &= \frac{1}{2T} \left| \int_{-\infty}^{+\infty} \left\{ \int_{-T}^T e^{itx} dt \right\} dF_{n_k}(x) \right| \\ &\leq \frac{1}{2T} \int_{|x| \leq b} \left| \int_{-T}^T e^{itx} dt \right| dF_{n_k}(x) + \frac{1}{2T} \int_{|x| > b} \left| \int_{-T}^T e^{itx} dt \right| dF_{n_k}(x) \\ &\leq \int_{|x| \leq b} dF_{n_k}(x) + \frac{1}{2T} \int_{|x| > b} \frac{2}{|x|} dF_{n_k}(x) \\ &\leq a_k + \frac{1}{bT} \leq a + \frac{\epsilon}{2} \end{aligned} \quad (3.22)$$

(3.22) 的推导中用到了以下事实,

$$\left| \int_{-T}^T e^{itx} dt \right| = \frac{2|\sin(Tx)|}{|x|} \leq \frac{2}{|x|}$$

首先, (3.21)+(3.22) \Rightarrow 矛盾! 于是, $F(+\infty) - F(-\infty) = 1$, 即 $F(x)$ 是一个分布函数。其次, 若 $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ 不真, 则存在子序列 $\{F_{n_j}(x)\}$ 使得 $\lim_{j \rightarrow \infty} F_{n_j}(x) = \tilde{F}(x) \neq F(x)$, 二者却有相同的特征函数, 矛盾!

□

例 3.32. 在逆极限定理中, 条件“在 $t = 0$ 处连续”必不可少, 否则将导致特征函数序列的极限不一定是特征函数。例如,

$$\begin{aligned} \text{特征函数 } \varphi_n(t) &= \frac{\sin(nt)}{nt}, \text{ 其中 } n = 1, 2, \dots \\ \text{所对应的分布函数 } F_n(x) &= \begin{cases} 0 & \text{当 } x \leq -n \\ \frac{x+n}{2n} & \text{当 } -n < x < n \\ 1 & \text{当 } x \geq n \end{cases} \end{aligned}$$

显然 $\lim_{n \rightarrow \infty} F_n(x) = 1/2$ 不是分布函数，而且 $\lim_{n \rightarrow \infty} \varphi_n(t)$ 也不是特征函数，因为

$$\lim_{n \rightarrow \infty} \varphi_n(t) = \begin{cases} 0 & \text{若 } t \neq 0 \\ 1 & \text{若 } t = 0 \end{cases}$$

定理 3.18. Lévy 连续性定理可以推广到高维，即

$$\mathbf{X}_n \xrightarrow{L} \mathbf{X} \text{ 当且仅当 } \varphi_{\mathbf{X}_n}(\mathbf{t}) \rightarrow \varphi_{\mathbf{X}}(\mathbf{t}), \forall \mathbf{t} \in \mathbb{R}^d$$

推论 3.4 (Cramér-Wold*, 1936). 若对任意 $\mathbf{t} \in \mathbb{R}^d$ 皆有 $\mathbf{t}^\top \mathbf{X}_n \xrightarrow{L} \mathbf{t}^\top \mathbf{X}$ ，则 $\mathbf{X}_n \xrightarrow{L} \mathbf{X}$ 。

证明. 利用 Lévy 连续性定理，我们有

$$\varphi_{\mathbf{X}_n}(\mathbf{t}) = E(e^{i\mathbf{t}^\top \mathbf{X}_n}) = \varphi_{\mathbf{t}^\top \mathbf{X}_n}(1) \rightarrow \varphi_{\mathbf{t}^\top \mathbf{X}}(1) = \varphi_{\mathbf{X}}(\mathbf{t}) \quad \square$$

*Herman Wold (1908-1992) 是瑞典经济学家和统计学家，他是瑞典数学家、统计学家 Harald Cramér (1893-1985) 的学生。

3.3 习题

- 3.1. 令二维随机向量 $(X, Y)^\top$ 的分布列为 $\frac{1}{6}\langle -1, -1 \rangle + \frac{1}{6}\langle -1, 1 \rangle + \frac{1}{2}\langle 1, -1 \rangle + \frac{1}{6}\langle 1, 1 \rangle$, 求该随机向量的特征函数。
- 3.2. 下列函数是否为特征函数: (1) $\sin t$, (2) $\ln(e + |t|)$, (3) $1/(1 - t^4)$ 。
- 3.3. 试证明: 连续型随机变量 X 的特征函数 $\varphi(t)$ 当 $t \rightarrow \pm\infty$ 时趋近于零。
- 3.4. 试证明: 对于随机变量 $X > 0$ 和 $t > 0$, 若其矩母函数 (3.6) 存在, 则

$$P(X \geq x) \leq M_X(t) \exp(-tx)$$

- 3.5. 已知 $X \sim N(0, 1)$, 试求随机变量 $Y = X^2$ 的特征函数 $\varphi_Y(t)$ 。
- 3.6. 如果随机变量 X 与 $-X$ 具有相同的分布, 则称 X 是对称的。证明: 随机变量 X 是对称的当且仅当其特征函数为实值函数。
- 3.7. 已知特征函数, 求相应的概率分布: (1) $\varphi(t) = \cos t$; (2) $\varphi(t) = \cos^2 t$; (3) $\varphi(t) = \sin t/t$; (4) $\varphi(t) = (\sin t/t)^2$ 。
- 3.8. 设随机变量 X 有密度函数为 $f(x) = c \exp\{-a|x|\}$, 其中 $a > 0$ 和 c 为合适的常数, 求它的特征函数。
- 3.9. 设事件 A 在第 k 次独立试验中出现的概率为 p_k , 记它在前 n 次试验中出现的次数为 X , 试求 X 的特征函数。
- 3.10. 已知 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \frac{1}{k+1}\langle 0 \rangle + \frac{1}{k+1}\langle 1 \rangle + \dots + \frac{1}{k+1}\langle k \rangle$, 求 $Y = X_1 + X_2 + \dots + X_n$ 的特征函数。根据此结果证明

$$\sum_{m=0}^{kn} \binom{n}{m}_{k+1} m = \frac{k(k+1)^n n}{2}$$

- 3.11. 设随机变量 $X \sim B(m, p)$ 和 $Y \sim B(n, p)$ 相互独立, 试证明: $Z = X + Y \sim B(m+n, p)$ 。
- 3.12. 设随机变量 X_1, X_2, \dots 独立同分布, 满足 $P(X_1 = k) = q^k p$, $k = 0, 1, \dots$, 其中 $0 < p < 1$ 且 $q = 1 - p$ 。试求 $X = \sum_{j=1}^n X_j$ 的分布。
- ☆ 3.13. 设 $\varphi(t)$ 为一个实值的特征函数, 试证明: $1 - \varphi(2t) \leq 4[1 - \varphi(t)]$ 。
- 3.14. 若 $t \rightarrow 0$ 时, 特征函数 $\varphi(t) = 1 + o(t^2)$, 试证明 $\varphi(t) \equiv 1$ 。

- 3.15. 若连续型随机变量 X 的特征函数 $\varphi(t)$ 绝对可积且二阶绝对矩存在, 试证明:
密度函数 $f_X(x)$ 满足

$$f_X(x) = \frac{1}{2\pi i x} \int_{-\infty}^{+\infty} \varphi'(t) e^{-itx} dt = -\frac{1}{2\pi x^2} \int_{-\infty}^{+\infty} \varphi''(t) e^{-itx} dt$$

$$f_X(x) \leq g(x) = \min \left\{ \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\varphi(t)| dt, \frac{1}{2\pi x^2} \int_{-\infty}^{+\infty} |\varphi''(t)| dt \right\}$$

其中 $g(x)$ 是一个平头的凹函数, 大致形态见第 719 页的图 15.8。

- ☆ 3.16. 已知 $\varphi(t)$ 是某随机变量的特征函数, 试证明 $\varphi(t)$ 是半正定函数 (见第 767 页的定义 E.5), 即对于任意的正整数 n 及任意的实数 $t_1, t_2, \dots, t_n \in \mathbb{R}$ 和任意的复数 $z_1, z_2, \dots, z_n \in \mathbb{C}$, 皆有 $\sum_{k=1}^n \sum_{j=1}^n \varphi(t_k - t_j) z_k \bar{z}_j \geq 0$ 。

第四章

一些常见的分布

胜日寻芳泗水滨，无边光景一时新。等闲识得东风面，万紫千红总是春。

朱熹《春日》

有了随机变量数字特征（期望、方差、矩等）、特征函数这些工具，人们可以研究给定分布的具体性质。本章所介绍的分布都是在实际应用中常见的，其中有些存在着数学上的联系，我们把它们放在同一小节中。

熟练掌握这些分布有助于概率建模和统计计算，也有助于对中心极限定理（第5章的主要内容）的理解。例如，若 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} U[0, 1]$ ，只要 $n \geq 4$ 就近似地有 $X_1 + X_2 + \dots + X_n \sim N(n/2, n/12)$ 。本章内容是已知工具的试验场，几乎全可付诸计算机实践，其中也牵扯到随机数的产生。

为满足统计抽样、密码分析、计算机模拟、随机设计等诸多应用的需求，如何产生高质量的随机数至今仍是物理设备和算法理论的一个重要的研究内容。给定一个具体的分布，随机数是按照该分布刻画的概率随机抽取到的实数，要求足够多的随机数能产生“群体效应”再现该分布。譬如，按照闭区间 $[0, 1]$ 上均匀分布得到的随机数，必须反映出 $[0, 1]$ 上任何实数都有相同的机会被抽取到。大体上，有两种截然不同的方法可以产生随机数：物理的方法和计算机模拟的方法。



- 通过一些物理方法（如放射性同位素的衰变、雷暴活动产生的大气噪声）可以产生无法重复的真正的随机数序列，但大都效率低且费用昂贵。近些年有所改观，利用电子器件中的热噪声作为随机源，或者利用混沌激光产生随机数取得了进展，有望生产出实用的物理设备（即硬件随机数产生器）。

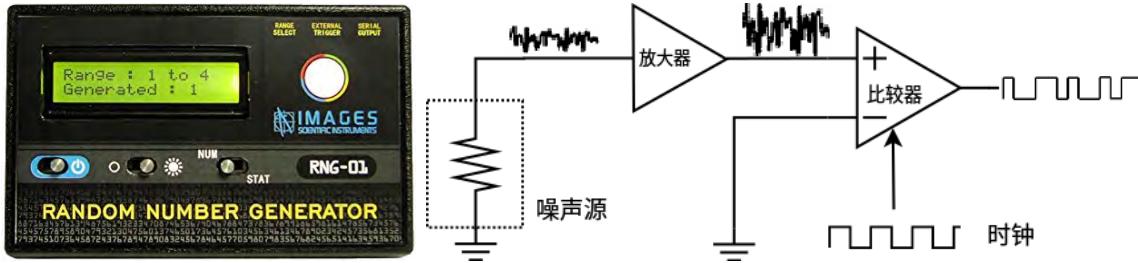


图 4.1: 硬件随机数产生器: 产生随机数的电子设备。

□ 理论上, 通过确定的算法不能产生真正意义上的随机数, 所以基于 John von Neumann (1903-1957) 思想的计算机无法产生真正的随机数而是一些“伪随机数”。人们设计精巧的算法旨在让这些伪随机数看起来更像随机数, 通常把这样的算法称作伪随机数产生器 (pseudo random number generator, PRNG)。von Neumann 说, “任何企图用确定方式产生随机数的人无疑都是逆天而行。”

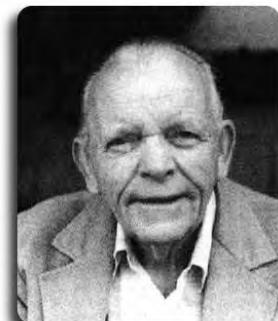
例 4.1. 在类 UNIX 操作系统中, 设备文件 `/dev/random` 允许访问采集自设备驱动或其它来源的环境噪声, 例如, 移动鼠标、敲击键盘等。FreeBSD 系统实现了 256 位的 Yarrow 算法, 为一些加密应用提供伪随机数。在终端, 可用下面的命令以 ASCII 码方式部分地读取 `/dev/random`。

```
yujiangsheng@~$ dd if=/dev/random | od -a | head -5
0000000 eot c4 esc a L f 6 nak c3 e4 % D nul f9 $ e
0000020 ^ s c syn a9 f4 | b7 1 nl w bd f3 c4 a2 c1
0000040 . del d2 ] 8f - % em d6 S e4 96 c4 9f e7 gs
0000060 b c4 cc d1 e8 fb b 8a cf 8c de Y " V c0 99
0000100 ba " N 91 90 cf rs f6 y nl aa af c8 fa 89 f0
```

目前很多伪随机数产生器的设计都源自 1949 年美国数学家 Derrick Henry Lehmer (1905-1991) 定义的下述递归关系,

$$z_{n+1} = (az_n + c) \bmod m \quad (4.1)$$

其中 $m, a, c > 0$ 都是确定的整数且 $a, c < m$, 初始整数 $0 \leq z_0 < m$ 称为种子, 由用户设定。



定义 4.1. 递归关系 (4.1) 所产生的序列 $z_0, z_1, \dots, z_n, \dots$ 称为线性同余序列 (linear congruential sequence)。

显然, 线性同余序列总是有有限周期, 周期长度不超过模数 m 且 $0 \leq z_n < m - 1$ 。

例 4.2. 递归式 (4.1) 中 $a = 3, c = 1, m = 7, z_0 = 2$ 所定义的线性同余序列是 2, 0, 1, 4, 6, 5 的循环, 周期显然为 6。

定理 4.1. 递归式 (4.1) 产生满周期 m 的线性同余序列当且仅当

- ① c 与 m 互素;
- ② $b = a - 1$ 是 p 的倍数, 其中 p 是 m 的素因子;
- ③ 若 m 是 4 的倍数, 则 b 也是 4 的倍数。

※证明. 详见 D. Knuth 的名著《计算机程序设计艺术》第二卷《半数值算法》[91] 的第 3.2 节《产生均匀的随机数》(第 17 页至第 19 页)。□

推论 4.1. 若 $m = 2^s$, 则当 c 为奇数且 $a \equiv 1 \pmod{4}$ 时式 (4.1) 产生的线性同余序列为满周期。

例 4.3. 以 $m = 2^5, a = 9, c = 1, z_0 = 2$ 为例, 递归式 (4.1) 产生的线性同余序列具有满周期, 是 2, 19, 12, 13, 22, 7, 0, 1, 10, 27, 20, 21, 30, 15, 8, 9, 18, 3, 28, 29, 6, 23, 16, 17, 26, 11, 4, 5, 14, 31, 24, 25 的不断重复。

要求线性同余序列满周期还远远不够, 我们需要定义一个衡量序列随机性程度的指标, 让在 $a = c = 1$ 的设置下产生的不带任何随机性的“坏”序列得分不高。

定义 4.2 (效能). 一个满周期的线性同余序列的效能 (potency) 定义为满足 $b^s \equiv 0 \pmod{m}$ 的最小的自然数 $s \in \mathbb{N}$, 其中 $b = a - 1$ 。

因为 0 总出现在满周期的线性同余序列中, 所以常设 $z_0 = 0$ 来考察效能。由递归式 (4.1),

$$\begin{aligned} z_n &= \frac{c(a^n - 1)}{a - 1} \pmod{m}, \text{ 令 } b = a - 1 \\ &= c(n + bC_n^2 + \cdots + b^{s-1}C_n^s) \pmod{m} \end{aligned}$$

※例 4.4. 当 $a = 1$ 时, 效能 $s = 1$, 随机性最差。当 $m = 2^{35}, a = 2^k + 1$ 时,

$$s = \begin{cases} 2 & \text{当 } k \geq 18 \\ 3 & \text{当 } k = 12, 13, \dots, 17 \\ 4 & \text{当 } k = 9, 10, 11 \end{cases}$$

当效能等于 2 时, $z_{n+1} - z_n \equiv c + cbn$, 随机性也不强; 当效能等于 3 时, 点 $(z_n, z_{n+1}, z_{n+2}) \in \mathbb{R}^3$ 一定落于平面 $x - 2y + z = e$ 之内, 其中 $e = d - 2m, d - m, d, d + m$ 而 $d = cb \pmod{m}$, 随机性也不算太强 [91]。

在实际应用中, 一般要求效能至少为 5。对随机性而言, 高效能是必要的, 但并非充分的。所以, 人们用效能来揭露差的随机性而拒绝某伪随机数产生器, 而不是用它来接受。伪随机数是否“合格”还要通过统计检验, 详见第 9 章的拟合优度检验。

定义 4.3. 我们把基于式 (4.1) 的伪随机数产生器称为 Lehmer 产生器或线性同余产生器 (linear congruential generator, LCG)。

若要让 LCG 产生 $U[0, 1]$ 的随机数, 通常 m 得选得很大, 如 $m = 2^{32}$ 或 $m = 2^{64}$, 使得伪随机数 z_n/m 看起来像是来自分布 $U[0, 1]$ 。

例 4.5. 多数编程语言都内置了产生均匀分布 $U[0, 1]$ 随机数的函数, 譬如,

- ANSI C 内置的 rand 函数采用 $m = 2^{32}, a = 1103515245, c = 12345$ 的 LCG。
- Fortran 95 及其后续版本的 LCG 采用了 Marsaglia 算法, 其中用到了

$$z_{n+1} = 69069z_n + 1327217885 \pmod{2^{32}}$$

LCG 虽然速度快但随机性不佳, 难以达到随机模拟、加密等应用的要求。1965 年, M. D. MacLaren 和 G. Marsaglia 提出了基于两个线性同余产生器 G_1, G_2 的组合产生器, 其基本思想是用一个线性同余序列来“搅乱”另一个线性同余序列, 就像是重新洗牌, 虽然没有生成新的元素, 但能使得 LCG 产生的伪随机数具有更好的随机性。

算法 4.1 (MacLaren-Marsaglia 组合产生器, 1965). 该算法需要用到一个辅助数组 $V[0], \dots, V[k-1]$, 它的初始化就是线性同余产生器 G_1 产生的前 k 个随机数。在实际应用中, 自然数 k 一般选在 100 左右。

- 置 $k \leftarrow 128, n \leftarrow 0, V[0] \leftarrow x_0, \dots, V[k-1] \leftarrow x_{k-1}$, 其中 x_0, x_1, \dots, x_{k-1} 是 G_1 产生的前 k 个随机数。
- 置 $j \leftarrow \lfloor ky_{n+1}/m \rfloor$, 其中 y_{n+1} 是 G_2 产生的第 $n+1$ 个随机数, m 是线性同余产生器 G_2 所用的模数。显然, $0 \leq j < k$ 。
- 输出 $z_n \leftarrow V[j]$; 置 $V[j] \leftarrow x_{n+1}$, 其中 x_{n+1} 是 G_1 产生的第 $n+1$ 个随机数; 置 $n \leftarrow n+1$ 。
- 重复步骤 2 至步骤 3 得到随机数序列 $z_0, z_1, \dots, z_n, \dots$

***例 4.6.** 设线性同余产生器 G_1 产生的序列 $\{x_n\}$ 是 $1, 2, \dots, 20$ 的不断重复, 随机性很差。现在用**例 4.3** 所述的线性同余产生器 G_2 来搅乱 $\{x_n\}$, 置 $k = 9$, 请读者实现**算法 4.1**。不难得到新的随机数序列如下, 其随机性有所改善。

```
[1] 6 4 3 7 2 1 7 3 8 2 11 9 5 9 15 12 8 10 13 6 5 14 3 19
[25] 4 1 7 6 20 2 5 18 17 9 15 11 8 13 19 16 12 14 3 10 4 1 7 4
[49] 20 2 5 18 17 6 15 11 16 13 19 18 12 14 17 10 9 1 7 3 20 5 11 8
[73] 4 6 15 2 16 13 19 16 12 14 17 10 9 18 7 3 8 5 11 10 4 6 9 2
[97] 1 13 19 15 12 17 3 20 16 18 7 14 8 5 11 8 4 6 9 2 1 10 19 15
[121] 20 17 3 2 16 18 1 14 13 5 11 7 4 9 15 12 8 10 19 6 20 17 3 20
[145] 16 18 1 14 13 2 11 7 12 9 15 14 8 10 13 6 5 17 3 19 16 1 7 4
[169] 20 2 11 18 12 9 15 12 8 10 13 6 5 14 3 19 4 1 7 6 20 2 5 18
[193] 17 9 15 11 8 13 19 16 12 14 3 10 4 1 7 4 20 2 5 18 17 6 15 11
```

※例 4.7 (J. A. Greenwood, 1976). 设计两个满周期的线性同余序列如下,

$$\begin{aligned}x_0 &= 5772156649, & x_{n+1} &= (3141592653x_n + 2718281829) \pmod{2^{35}} \\y_0 &= 1781072418, & y_{n+1} &= (2718281829y_n + 3141592653) \pmod{2^{35}}\end{aligned}$$

令 $k = 64$ 并舍弃序列 $y_0, y_1, \dots, y_n, \dots$ 中所有的 0, 则 MacLaren-Marsaglia 组合产生器的周期高达 $2^{35}(2^{35} - 1)$, 基本满足实用需求。

MacLaren-Marsaglia 组合产生器的周期在很多情况下为 G_1 和 G_2 周期的最小公倍数 [91]。美中不足的是如果 G_1 和 G_2 具有强相关性, 与原序列相比, 组合产生器的随机性效果可能会变差。

1976 年, C. Bays 和 S. D. Durham 借鉴 MacLaren-Marsaglia 组合产生器提出了下面的改进算法, 使得“搅乱”某序列的过程仅仅依靠该序列本身, 而且“搅乱”后的随机性不会比之前的差。

算法 4.2 (Bays-Durham 自乱产生器, 1976). 辅助数组 $V[0], \dots, V[k-1]$ 的初始化就是产生器 G 所产生的前 k 个随机数。

- 置 $k \leftarrow 128, n \leftarrow 0, V[0] \leftarrow x_0, \dots, V[k-1] \leftarrow x_{k-1}, y \leftarrow x_k$, 其中 $x_0, x_1, \dots, x_{k-1}, x_k$ 是 G 产生的前 $k+1$ 个随机数。
- 置 $j \leftarrow \lfloor ky/m \rfloor$, 其中 m 是 G 所用的模数。显然, $0 \leq j < k$ 。
- 输出 $z_n \leftarrow V[j]$; 置 $y \leftarrow V[j], V[j] \leftarrow x_{n+1}$, 其中 x_{n+1} 是 G 产生的第 $n+1$ 个随机数; 置 $n \leftarrow n+1$ 。
- 重复步骤 2 至步骤 3 得到随机数序列 $z_0, z_1, \dots, z_n, \dots$

※例 4.8. 置 $k = 9$, 用**例 4.3** 所述的线性同余产生器来搅乱自身, 算法 4.2 产生的随机数序列的随机性比起**例 4.3** 有很大的改善。

[1] 1 2 12 13 22 0 13 7 19 7 27 19 20 15 22 0 1 3 28 30 10 12 10 16
[25] 9 26 6 21 8 4 14 17 11 2 29 23 18 31 22 7 25 5 27 20 0 13 12 9
[49] 24 10 3 8 29 1 6 21 15 19 11 18 5 26 30 16 4 24 28 25 22 13 14 1
[73] 17 19 31 2 27 7 12 10 23 0 15 20 21 23 28 8 3 29 30 5 9 11 18 16
[97] 6 31 14 25 26 7 12 22 17 19 2 4 1 30 13 10 24 27 0 8 3 6 15 20
[121] 21 11 18 4 17 26 29 9 23 28 25 13 5 31 2 16 24 12 7 0 10 19 14 21
[145] 9 8 28 22 20 3 15 27 6 30 1 17 26 11 18 16 31 5 4 13 25 24 23 1
[169] 14 7 22 10 29 12 20 2 27 0 3 29 8 15 19 18 26 28 4 21 11 9 16 17
[193] 2 6 14 24 30 23 7 13 22 1 12 27 5 10 25 15 19 31 0 21 28 18 23 20
[217] 17 3 6 8 9 31 29 25 16 4 14 30 2 5 13 22 26 19 11 10 24 27 20 21
[241] 18 3 0 29 7 1 6 16 12 30 23 8 15 11 4 26 9 31 5 2 17 24 14 25
[265] 19 28 27 21 20 15 0 7 22 1 18 8 12 10 23 28 30 13 16 9 17 14 5 3
[289] 29 4 2 19 6 13 25 11 1 22 26 10 31 12 27 30 8 15 24 21 7 0 20 23

均匀分布的 PRNG 是最基本的，多数编程语言都提供了均匀分布 $U[0, 1]$ 的随机数产生函数，如 C、Fortran、LISP、R、Maxima 等，甚至绘图语言 MetaPost 中也有函数 uniformdeviate 来实现该功能（例如图 1.9 是利用 MetaPost 生成的）。对于均匀分布的 RNG 及其评估方法的研究仍在继续。

通过一些方法，如逆 CDF 法等，在均匀分布随机数的基础上，我们可以设计常见分布的 PRNG 算法。本章将具体介绍以下常见的分布及其关系。

离散型随机变量：单点分布，两点分布，二项分布，几何分布，负二项分布，Pólya 分布及其特款——超几何分布，Poisson 分布等。对它们的更深刻的理解要等到学习了随机过程理论（见第 6 章）。

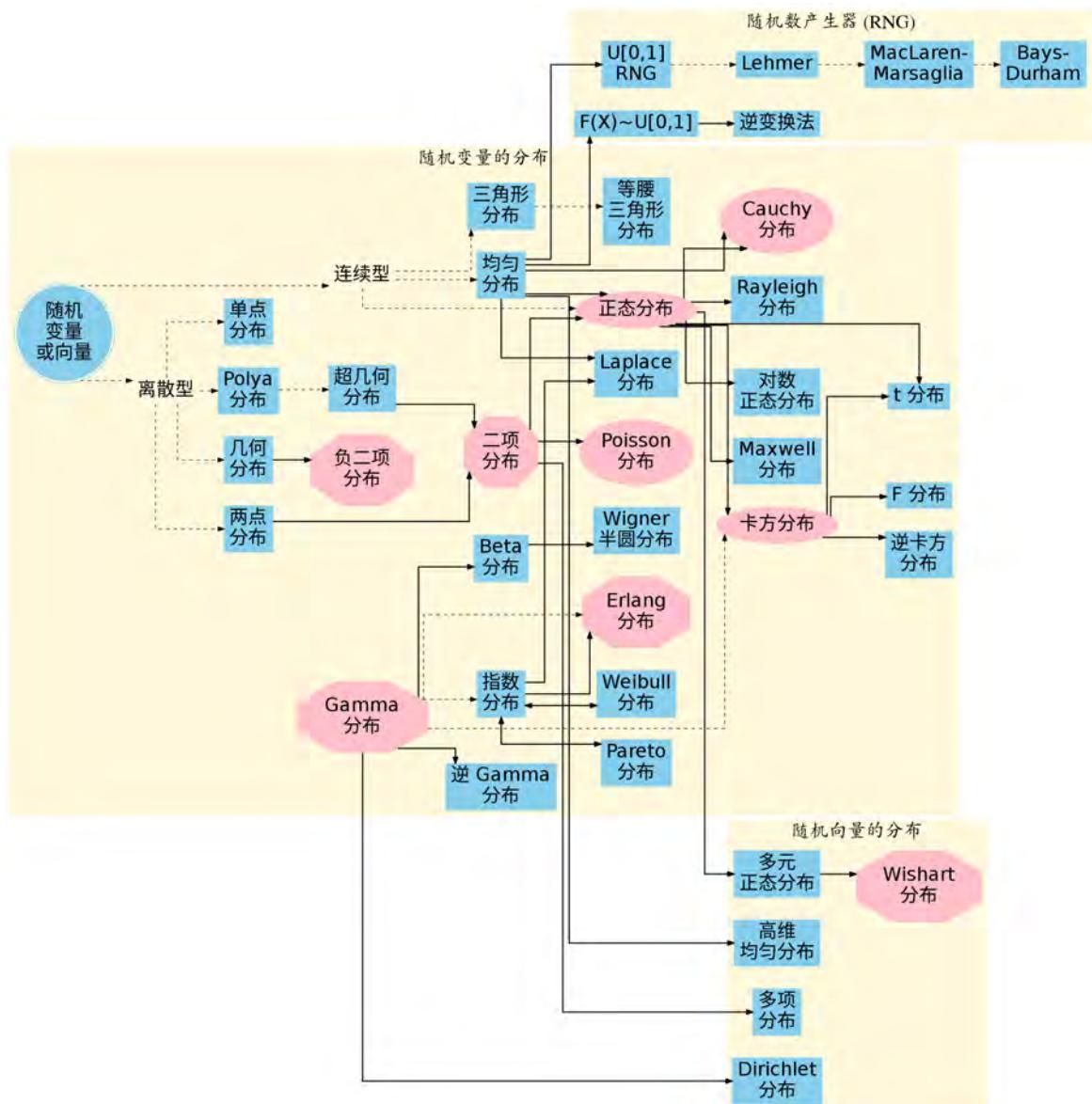
连续型随机变量：均匀分布，三角形分布，正态分布，对数正态分布，偏正态分布，Laplace 分布，Cauchy 分布，Gamma 分布及其特款（包括 χ^2 分布、指数分布、Erlang 分布），逆 Gamma 分布，Beta 分布， t 分布， F 分布，Pareto 分布，以物理学家命名的分布（包括 Boltzmann 分布，Gibbs 分布，Weibull 分布，Rayleigh 分布，Maxwell 分布，Wigner 半圆分布）等。

随机向量：高维均匀分布，多项分布，Dirichlet 分布，多元正态分布，Wishart 分布。

由于篇幅所限，本书无法对所有分布的 PRNG 展开深入的讨论。本章所提供的各种分布的随机数产生算法都是出于演示的目的，这些基本算法在算法复杂度上仍可改进。对随机数产生算法感兴趣的读者可参阅统计计算 [151]，计算统计学方面的著作 [57]，或者科学计算专著 [124] 的第七章（该书提供了很多 RNG 算法的 C++ 源码）。另外，Knuth 的《计算机程序设计艺术》第二卷对 RNG 也有所论述。



第四章的主要内容及其关系



我们把“服从同一分布类型的若干独立随机变量之和仍服从该分布类型”这一性质称为“和型不变性”，满足此性质的分布其节点形状为（粉色）椭圆，在一定条件下满足此性质的分布其节点形状为（粉色）八边形。

4.1 离散型随机变量的分布

离 散型随机变量 X 的所有信息都在它的分布列中, 为了方便起见, 我们采用 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \cdots + p_n\langle x_n \rangle + \cdots$ 或概率函数来描述它们。对于离散均匀分布 $Y \sim \frac{1}{m}\langle y_1 \rangle + \frac{1}{m}\langle y_2 \rangle + \cdots + \frac{1}{m}\langle y_m \rangle$, 其中 $y_1, y_2, \dots, y_m \in \mathbb{R}$ 两两不等, 约定用符号 $Y \sim U\{y_1, y_2, \dots, y_m\}$ 简记之。

例 4.9. 已知 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \frac{1}{k+1}\langle 0 \rangle + \frac{1}{k+1}\langle 1 \rangle + \cdots + \frac{1}{k+1}\langle k \rangle$, 由**例 1.16** 知 $Y = X_1 + X_2 + \cdots + X_n$ 的分布列是

$$P(Y=y) = \frac{1}{(k+1)^n} \binom{n}{y}_{k+1}, \text{ 其中 } y=0, 1, \dots, kn$$

多项式系数 $\binom{n}{y}_{k+1}$ 的计算可利用第 102 页的式 (1.39)。

算法 4.3. 离散分布 $X \sim p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \cdots + p_n\langle x_n \rangle + \cdots$ 的随机数 x^* 可通过下面的算法得到, 该算法的设计依赖于均匀分布的一个重要结果——**定理 4.4** 以及基于此结果的逆 CDF 法, 我们将在 §4.2.1 予以介绍。

- 抽取 $U(0, 1)$ 的随机数 u^* 。
- 令 n 是满足 $u^* \leq \sum_{j=1}^n p_j$ 的最小自然数, 输出 $x^* \leftarrow x_n$ 。

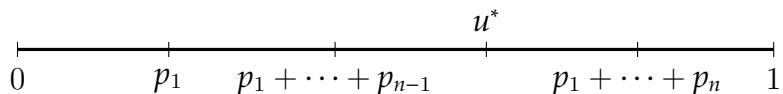


图 4.2: 若 $p_1 + \cdots + p_{n-1} < u^* \leq p_1 + \cdots + p_n$, 则输出随机数 x_n 。

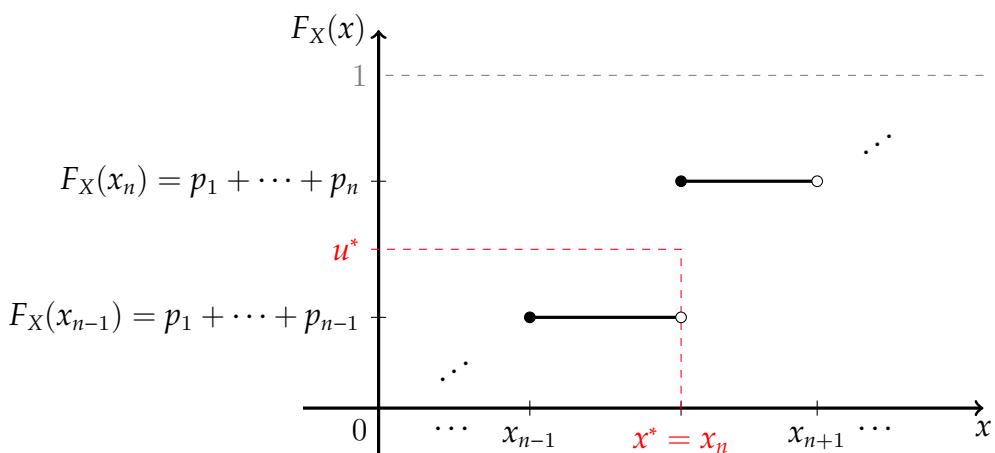


图 4.3: 算法 4.3 的直观含义: 若 $F_X(x_{n-1}) \leq u^* < F_X(x_n)$, 则 $x^* = x_n$ 。

例 4.10. 利用算法 4.3 产生 $X \sim \frac{1}{n}\langle 1 \rangle + \cdots + \frac{1}{n}\langle n \rangle$ 的随机数:

$$x^* \leftarrow \lceil u^* n \rceil, \text{ 其中 } u^* \text{ 是 } U[0, 1] \text{ 的随机数}$$

练习 4.1. 利用算法 4.3 产生 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 和 $X \sim B(n, p)$ 的随机数。

提示: 第 261 页的算法 4.4 和第 263 页的算法 4.5。

性质 4.1. 用算法 4.3 产生离散分布 $p_1\langle x_1 \rangle + p_2\langle x_2 \rangle + \cdots + p_n\langle x_n \rangle + \cdots$ 的一个随机数所用的平均执行步数为 $1 + \sum_{n=1}^{\infty} np_n$ 。

证明. 步骤 1 是必然执行的, 步骤 2 的执行次数 Y 是一个取值 $1, 2, \dots$ 的随机变量。令 $q_0 = 0$, 则

$$P(Y = n) = P(q_{n-1} < U \leq q_n) = q_n - q_{n-1} = p_n$$

即 $Y \sim p_1\langle 1 \rangle + p_2\langle 2 \rangle + \cdots + p_n\langle n \rangle + \cdots$, 因此 $E(Y) = \sum_{n=1}^{\infty} np_n$ 。 \square

练习 4.2. 基于算法 4.3, 用 R 或 Maxima 产生类别分布 $0.2\langle 0 \rangle + 0.7\langle 1 \rangle + 0.1\langle 2 \rangle$ 的 100 个随机数, 并绘出执行步数的直方图。

例 4.11. 在语料中随机选一个词 (或标点符号) w_0 , 从它出发玩“接龙游戏”得到随机文本 $w_0w_1w_2 \cdots w_n$, 要求 $w_jw_{j+1}, j = 0, 1, \dots, n-1$ 在语料中出现。譬如, $w_0 = \text{believed}$, $w_1 = \text{in}$ 或 them , 按照均匀分布产生 w_1 ; 以此类推再产生 $w_2 \dots$ 。

这个例子纯粹是一个玩具模型, 生成的句子多不合乎语法。如果对语料进行一定的句法/语义分析, 随机替换相同结构的片段所生成的文本就接近真实语言了, 可用于设计自动网聊程序。

定义 4.4 (概率母函数). 若随机变量 $X \sim p_0\langle 0 \rangle + p_1\langle 1 \rangle + \cdots + p_k\langle k \rangle + \cdots$, 则 X 的概率母函数 (probability generating function, pgf) 定义为

$$G_X(s) = E(s^X) = \sum_{k=0}^{\infty} p_k s^k \quad (4.2)$$

显然, $G_X(1) = p_0 + p_1 + \cdots + p_k + \cdots = 1$ 。

性质 4.2. 作为练习, 请读者验证概率母函数具有以下性质。

- ① 令 $\varphi_X(t), M_X(t)$ 分别是 $X \sim p_0\langle 0 \rangle + p_1\langle 1 \rangle + \cdots + p_k\langle k \rangle + \cdots$ 的特征函数和矩母函数, 定义见 (3.5) 和 (3.6), 则

$$G_X(e^{it}) = \varphi_X(t) \quad G_X(e^t) = M_X(t)$$

- ② 概率母函数 $G_X(s)$ 对分布列 $X \sim p_0\langle 0 \rangle + p_1\langle 1 \rangle + \cdots + p_k\langle k \rangle + \cdots$ 的表示是唯一的，其中

$$p_k = \frac{G_X^{(k)}(0)}{k!}, \text{ 其中 } k = 0, 1, \dots$$

即，若两个随机变量的概率母函数相同，则它们的分布列也相同。

- ③ $X \sim p_0\langle 0 \rangle + p_1\langle 1 \rangle + \cdots + p_k\langle k \rangle + \cdots$ 的期望和方差分别为

$$E(X) = G'_X(1) \quad V(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$$

- ④ 已知 X_1, \dots, X_n 是独立的，则

$$G_{\alpha_1 X_1 + \dots + \alpha_n X_n + \beta}(s) = G_{X_1}(s^{\alpha_1}) \cdots G_{X_n}(s^{\alpha_n}) \cdot s^{\beta}$$

本节内容

本节所介绍的离散分布虽用处各异，但它们之间存在着千丝万缕的联系：两点分布、二项分布、几何分布、负二项分布一脉相承，此“脉”就是 Bernoulli 试验。Poisson 分布是二项分布在一定条件下的“极限状态”，超几何分布是 Pólya 分布的特例，而 Pólya 分布在一定条件下以二项分布为“极限状态”。

关键知识

(1) 会求离散型随机变量的特征函数；(2) 掌握所介绍分布的性质；(3) 深入理解二项分布与正态分布的关系。

4.1.1 单点分布和两点分布

单点分布和两点分布的定义分别见第 120 页的例 2.8 和例 2.9。经常利用性质 2.30 和性质 2.33 来判定随机变量服从单点分布。两点分布 $X \sim p\langle a \rangle + (1-p)\langle b \rangle$ 亦可看作是两个相异单点分布的凸组合。两点分布的特例 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 称作 0-1 分布或 Bernoulli 分布。不难看出，0-1 分布 $p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 即为二项分布 $B(1, p)$ 。

性质 4.3. 单点分布 $X \sim \langle c \rangle$ 的特征函数、期望和方差分别为

$$\begin{aligned}\varphi(t) &= e^{itc} \\ E(X) &= c \\ V(X) &= 0\end{aligned}$$

性质 4.4. 0-1 分布 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 的概率函数是

$$P(X = x) = p^x(1-p)^{1-x}$$

其特征函数和概率母函数分别为

$$\varphi_X(t) = 1 + p(e^{it} - 1) \quad G_X(s) = ps + 1 - p$$

由式 (3.11) 求得各阶原点矩 $m_k = p$, 其中 $k = 1, 2, \dots$ 。进而,

$$V(X) = m_2 - m_1^2 = p(1-p)$$

练习 4.3. 试证明: $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 的偏度系数 c_s 和峰度系数 c_k 分别为

$$c_s = \frac{1-2p}{\sqrt{p(1-p)}} \quad c_k = \frac{1}{p(1-p)} - 6$$

算法 4.4. 分布 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 的随机数 x^* 可以通过 $U(0, 1)$ 的随机数产生, 即

$$x^* = J(p - u^*) = \begin{cases} 1 & \text{若 } u^* \leq p \\ 0 & \text{若 } u^* > p \end{cases}$$

其中 u^* 是取自 $U(0, 1)$ 的随机数, $J(\cdot)$ 是非负判定函数 (见第 123 页)。

例 4.12. 已知算法 A 可以产生 0-1 分布 $0.5\langle 1 \rangle + 0.5\langle 0 \rangle$ 的随机数, 如何利用 A 产生分布 $\frac{1}{6}\langle 1 \rangle + \frac{1}{6}\langle 2 \rangle + \dots + \frac{1}{6}\langle 6 \rangle$ 的随机数?

解. 连续使用算法 A 三次, 如果产生 000 或 111, 则重新再来。利用对应关系 $001 \rightarrow 1, 010 \rightarrow 2, \dots, 110 \rightarrow 6$ 得到随机数。

解法二. 输入 $I = (1, 2, \dots, 6)$, 利用 A 独立地产生 n 个随机数, 其中 n 为 I 的长度, 如果全为 0 则重新抽样。譬如抽得 $(0, 1, 1, 0, 1, 0)$, 其中 1 的位置指标是 $(2, 3, 5)$, 更新 I 为这些指标位置上的数, 即 $I = (2, 3, 5)$ 。重复上述过程直至 I 只剩一个数, 便是问题所求的随机数。

练习 4.4. 求两点分布 $X \sim p\langle a \rangle + (1 - p)\langle b \rangle$ 的期望和方差, 其中 $a \neq b$ 。

提示: $E(X) = pa + (1 - p)b, V(X) = p(1 - p)(a - b)^2$

4.1.2 二项分布

二项分布的定义见第 122 页的定义 2.11，就是 n 重 Bernoulli 试验中“成功”的次数的概率分布。如例 3.12 所描述，若 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ ，则 $X = X_1 + X_2 + \dots + X_n \sim B(n, p)$ ，其概率函数为

$$P(X = k) = C_n^k p^k (1-p)^{n-k}, \text{ 其中 } 0 < p < 1 \text{ 且 } k = 0, 1, \dots, n$$

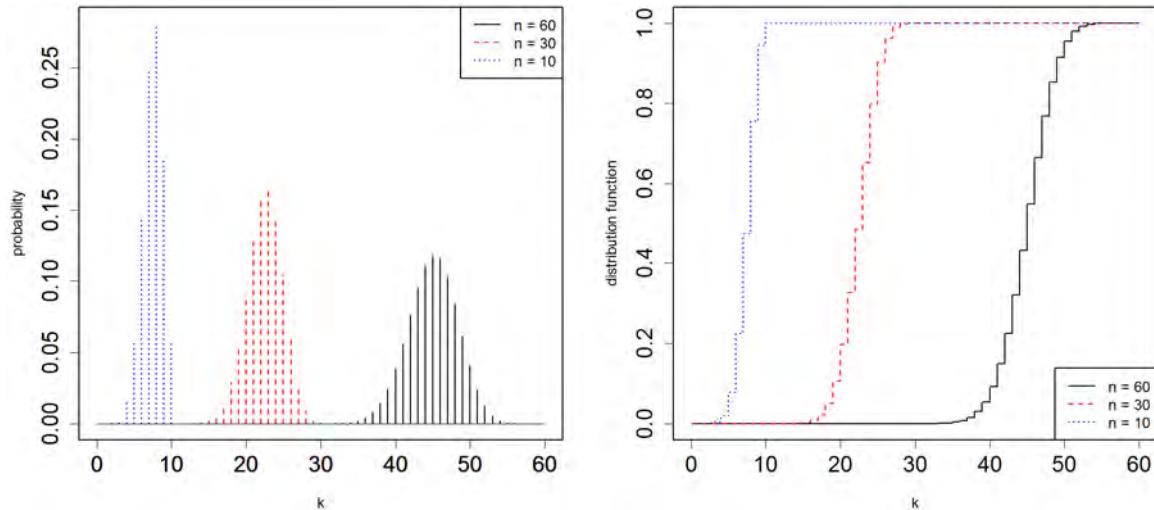


图 4.4: 二项分布 $B(n, p)$ 的概率质量函数和分布函数，其中 $p = 3/4, n = 10, 30, 60$ 。

算法 4.5. 二项分布 $X \sim B(n, p)$ 的随机数 x^* 可由 n 个独立产生的 0-1 分布 $p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 的随机数 $y_1^*, y_2^*, \dots, y_n^*$ 求和得到，即 $x^* = \sum_{j=1}^n y_j^*$ 。

性质 4.5. 二项分布 $X \sim B(n, p)$ 的特征函数、概率母函数和数字特征分别为

$$\begin{aligned} \varphi_X(t) &= [1 + p(e^{it} - 1)]^n & G_X(s) &= (ps + 1 - p)^n \\ E(X) &= np & V(X) &= np(1 - p) \\ c_s &= \frac{1 - 2p}{\sqrt{np(1 - p)}} & c_k &= \frac{1}{np(1 - p)} - \frac{6}{n} \end{aligned}$$

$P(X = k)$ 在 $\lfloor (n+1)p \rfloor$ 或 $\lfloor (n+1)p \rfloor - 1$ 取得最大值。一般情况下， $P(X = k)$ 是非对称的，见图 4.4。当 $p = 1/2$ 时， $c_s = 0$ ， $P(X = k)$ 关于 $n/2$ 对称。

性质 4.6. 如果随机变量 $X \sim B(n, p)$ 与 $Y \sim B(m, p)$ 相互独立，则

$$X + Y \sim B(n + m, p)$$

性质 4.6 在习题 3.11 中我们已经证明过了，它的直观含义是：如果考察同一事件 A 的 n 重 Bernoulli 试验与 m 重 Bernoulli 试验相互独立，则二者合起来构成考

察 A 的 $n+m$ 重 Bernoulli 试验。性质 4.6 所满足的和型不变性带有一定的条件，即事件 A 发生的概率 p 在两个试验中是一样的。

性质 4.7. 已知 $X \sim B(n, p)$, 令 $p_k = C_n^k p^k (1-p)^{n-k}, k = 0, 1, \dots, n$, 则

$$\frac{X}{n} \sim p_0 \langle 0 \rangle + \dots + p_k \langle k/n \rangle + \dots + p_n \langle 1 \rangle$$

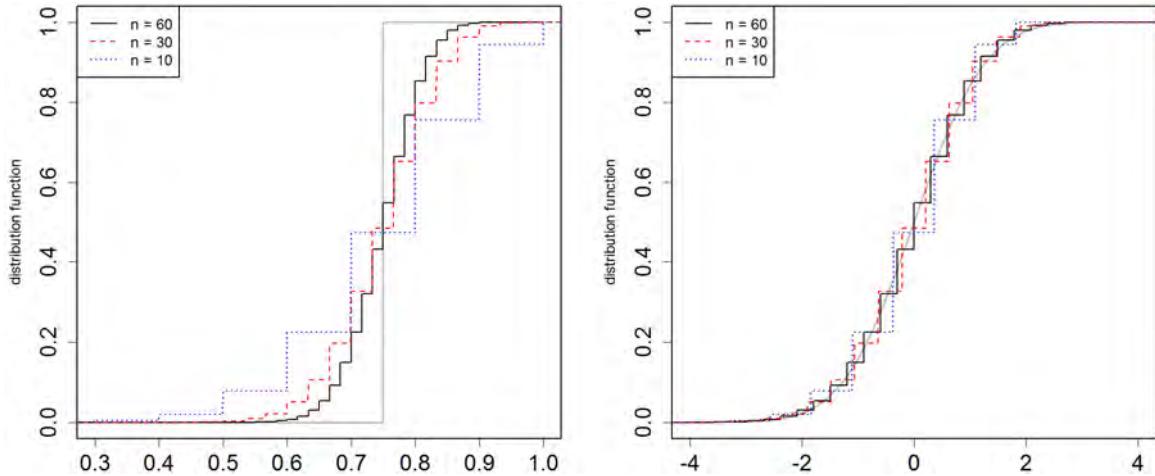


图 4.5: 若 $X \sim B(n, p)$, 随机变量 $\frac{X}{n}$ 和 $\frac{X-np}{\sqrt{np(1-p)}}$ 的分布函数如图所示, 其中 $p = 3/4, n = 10, 30, 60$ 。当 $n \rightarrow \infty$ 时, 分布函数趋向于灰色的粗线, 分别是单点分布 $\langle p \rangle$ 和标准正态分布 $N(0, 1)$ 的分布函数。

性质 4.8. 已知 $X \sim B(n, p)$, 令 $q = 1 - p$, 则

$$\begin{aligned} \frac{P(X = k+1)}{P(X = k)} &= \frac{n-k}{k+1} \cdot \frac{p}{q} \\ &= \begin{cases} > 1 & \text{当 } k+1 \leq np \\ \approx 1 & \text{当 } k < np < k+1 \\ < 1 & \text{当 } k \geq np \end{cases} \end{aligned}$$

从该性质不难看出, $P(X = k+1)/P(X = k)$ 是关于 k 的严格减函数。

练习 4.5. 若 $X \sim B(n, p)$, 则对任意 $k \leq np$ 皆有

$$P(X \leq k) \leq \exp \left\{ -\frac{2(np - k)^2}{n} \right\}$$

提示: 利用 Hoeffding 不等式 (2.80) 可证。

对于 $X \sim B(n, p)$, 当 n 很大时, 图 1.7 显示概率 $p_k = P(X = k), k = 0, 1, \dots, n$

呈现出一定的对称性。如果用一个连续函数来拟合 $\{(k, p_k) : k = 0, 1, \dots, n\}$, 它会是一个怎样的函数呢?

性质 4.9. 对于二项分布 $X \sim B(n, p)$, 若 n 很大, 则

$$P(X = np + k) \approx P(X = np) \exp \left\{ -\frac{k^2}{2npq} \right\}$$

证明. 由**性质 4.8** 和式 (3.13), 我们有

$$\begin{aligned} \ln \frac{P(X = np + j)}{P(X = np + j - 1)} &= \ln \frac{1 - \frac{j-1}{nq}}{1 + \frac{j}{np}} \\ &= \ln \left(1 - \frac{j-1}{nq} \right) - \ln \left(1 + \frac{j}{np} \right) \\ &= -\frac{j}{npq} + \frac{1}{nq} + o\left(\frac{1}{n}\right) \end{aligned}$$

两边对 j 从 1 至 k 求和, 于是

$$\begin{aligned} \ln P(X = np + k) - \ln P(X = np) &= \sum_{j=1}^k \ln \frac{P(X = np + j)}{P(X = np + j - 1)} \\ &= -\frac{k(k+1)}{2npq} + \frac{k}{nq} + o\left(\frac{1}{n}\right) \\ &\approx -\frac{k^2}{2npq} \end{aligned} \quad \square$$

上述结果促使我们联想当 n 很大时 $B(n, p)$ 可以由正态分布来近似, **附录 A** 用基于 Stirling 公式的初等方法推导出它恰是 $N(np, npq)$, 其中 $q = 1 - p$ 。法国数学家 A. de Moivre 最早研究过 $p = 1/2$ 的情形, 他发现可用正态分布 $N(n/2, n/4)$ 来近似, 而且 n 越大近似程度越高。

性质 4.10 (二项分布的正态近似). 已知 $X \sim B(n, p)$, 当 n 很大时, 满足 $np \geq 5$ 且 $n(1-p) \geq 5$, 利用图 4.6 所示的连续性修正, 概率 $P(a \leq X \leq b)$ 可由标准正态分布 $Z \sim N(0, 1)$ 通过下面的方式来看似计算。

$$\begin{aligned} P(a \leq X \leq b) &\approx P \left\{ \frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}} \right\} \\ &= \Phi \left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}} \right) - \Phi \left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \right) \end{aligned}$$

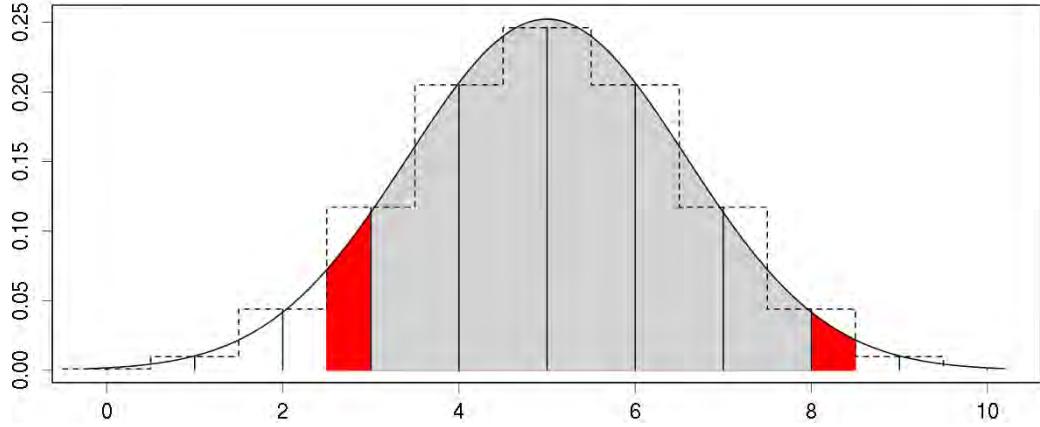


图 4.6: 已知二项分布 $X \sim B(10, 0.5)$, 概率 $P(3 \leq X \leq 8)$ 可用正态分布 $Y \sim N(5, 2.5)$ 来近似计算。从图中不难发现, $P(3 - \frac{1}{2} \leq Y \leq 8 + \frac{1}{2})$ 比 $P(3 \leq Y \leq 8)$ 要更精确一些, 这种上下限的修正称为连续性修正 (continuity correction)。

例 4.13. 一个大规模语料库有 n 个语句, 考察包含单词 w 的语句的个数 C_w 。自然语言处理常假设 $C_w \sim B(n, p_w)$, 其中 p_w 是单词 w 在语句中出现的概率。类似地, 两个单词 w, w' 共同出现在一个句子里的次数 $C_{w,w'}$ 也可假设服从某个二项分布。

例 4.14. 连续抛一枚均匀的硬币 100 次, 出现正面的次数介于 40 和 60 之间的概率?

解. 出现正面的次数 $X \sim B(100, 0.5)$, 则 $P(40 \leq X \leq 60) = P(X \leq 60) - P(X \leq 39) \approx 0.9647998$ 。利用性质 4.10 求得近似值 0.9642712, 比未经过修正的结果要精确些。

练习 4.6. 利用性质 4.10, 证明下面的经过连续性修正的结果。

$$\begin{aligned} P(a < X \leq b) &\approx P\left\{\frac{a + \frac{1}{2} - np}{\sqrt{np(1-p)}} < Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right\} \\ P(a \leq X < b) &\approx P\left\{\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq Z < \frac{b - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right\} \\ P(a < X < b) &\approx P\left\{\frac{a + \frac{1}{2} - np}{\sqrt{np(1-p)}} < Z < \frac{b - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right\} \end{aligned}$$

定义 4.5 (广义二项分布). 已知随机变量 $X_k \sim p_k \langle 1 \rangle + (1 - p_k) \langle 0 \rangle, k = 1, 2, \dots, n$ 相互独立, 则称 $X = X_1 + X_2 + \dots + X_n$ 服从广义二项分布。

练习 4.7. 请读者验证如上定义的广义二项分布的特征函数、期望和方差分别为

$$\varphi(t) = \prod_{k=1}^n [1 + p_k(e^{it} - 1)]$$

$$E(X) = \sum_{k=1}^n p_k$$

$$V(X) = \sum_{k=1}^n p_k(1 - p_k)$$

4.1.3 Pólya 分布及其特例（超几何分布）

1923 年，美籍匈牙利裔数学家 G. Pólya (1887-1985) 考虑了下面的球-盒子模型。

例 4.15 (Pólya 的球-盒子模型). 在一个盒子里有 N 个球，其中 w 个白球和 b 个黑球。从盒子里随机取出一个球，往盒子里返回 $s+1$ 个同颜色的球，将此过程重复 n 次，令随机变量 X_n 表示取出黑球的数量，试求 $P(X_n = k) = ?$

解. 令 $p = b/N, q = w/N, a = s/N$ ，显然 p, q 满足 $p + q = 1$ ，其中 $0 < p < 1$ ，对于 $k = 1, 2, \dots, n-1$ 有

$$\begin{aligned} P(X_n = k) &= C_n^k \frac{b(b+s) \cdots [b+(k-1)s]w(w+s) \cdots [w+(n-k-1)s]}{N(N+s) \cdots [N+(n-1)s]} \\ &= C_n^k \frac{p(p+a) \cdots [p+(k-1)a]q(q+a) \cdots [q+(n-k-1)a]}{(1+a) \cdots [1+(n-1)a]} \\ P(X_n = 0) &= \frac{w(w+s) \cdots [w+(n-1)s]}{N(N+s) \cdots [N+(n-1)s]} = \frac{q(q+a) \cdots [q+(n-1)a]}{(1+a) \cdots [1+(n-1)a]} \\ P(X_n = n) &= \frac{b(b+s) \cdots [b+(n-1)s]}{N(N+s) \cdots [N+(n-1)s]} = \frac{p(p+a) \cdots [p+(n-1)a]}{(1+a) \cdots [1+(n-1)a]} \end{aligned}$$

定义 4.6 (Pólya 分布). 称例 4.15 中 X_n 所服从的分布为 Pólya 分布，记作 $X_n \sim \text{Pólya}(n, p, a)$ ，它常用作传染病模型。特别地，当 $a = 0$ ，即 $s = 0$ 时， $X_n \sim \text{B}(n, p)$ 。另外，已知前 n 次试验中共取到 k 次黑球，则第 $n+1$ 次试验取到黑球的概率是

$$\begin{aligned} P(X_{n+1} = k+1 | X_n = k) &= \frac{b+ks}{N+ns} \\ &= \frac{p+ka}{1+na} \end{aligned}$$



练习 4.8. 请读者验证：在例 4.15 中，

$$E(X_n) = np \quad V(X_n) = \frac{npq(1+na)}{1+a}$$

性质 4.11. 令 $N \rightarrow \infty$ ，若 $p = b/N, q = 1-p$ 为常数且 $\lim_{N \rightarrow \infty} a = 0$ ，则分布 $\text{Pólya}(n, p, a)$ 的极限是二项分布 $\text{B}(n, p)$ ，即

$$\lim_{a \rightarrow 0} P(X_n = k) = C_n^k p^k q^{n-k}$$

定义 4.7 (超几何分布). 令 $s = -1$ ，Pólya 分布的特例称为超几何分布 (hypergeometric distribution)，记作 $X_n \sim \text{Hyper}(b, w, n)$ ， X_n 表示在例 4.15 的球-盒子模型中，不放

回地取 n 个球所含黑球的数量，其概率函数是

$$\begin{aligned} P(X_n = k) &= \frac{C_b^k C_w^{n-k}}{C_{b+w}^n} \\ &= \frac{C_b^k C_{N-b}^{n-k}}{C_N^n}, \text{ 其中 } N = b + w, \max(0, n - w) \leq k \leq \min(b, n) \end{aligned}$$

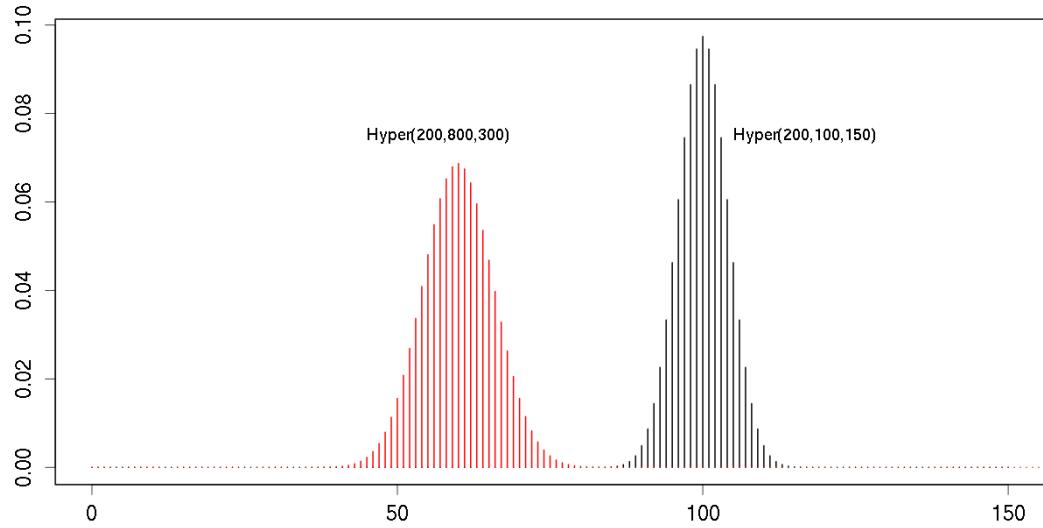


图 4.7: 超几何分布的竖线图：当 $N = b + w$ 很大时，超几何分布 $\text{Hyper}(b, w, n)$ 与二项分布 $B(n, p)$ 近似，其中 $p = b/N$ 。

练习 4.9. 求超几何分布 $\text{Hyper}(b, w, n)$ 的期望和方差，通过 Maxima 查看它的偏度系数和峰度系数。答案： $E(X) = np$, $V(X) = npq\frac{N-n}{N-1}$, 其中 $N = b + w$, $p = b/N$, $q = 1 - p$ 。

4.1.4 几何分布和负二项分布

设 Bernoulli 试验中某事件 A 出现的概率为 $p \in (0, 1)$ 。无限地重复该试验，事件 A 在第 $n+1$ 次试验中头次出现（即前 n 次试验中 A^c 连续出现，第 $n+1$ 次试验中 A 才第一次出现）的概率为 pq^n ，其中 $q = 1 - p, n = 0, 1, 2, \dots$ 。

定义 4.8 (几何分布). 令随机变量 X 表示事件 A 头次出现之前 A^c 出现的次数，则

$$X \sim p\langle 0 \rangle + pq\langle 1 \rangle + \dots + pq^n\langle n \rangle + \dots$$

我们称 X 服从几何分布 (geometric distribution)^{*}，记作 $X \sim \text{Geom}(p)$ 。

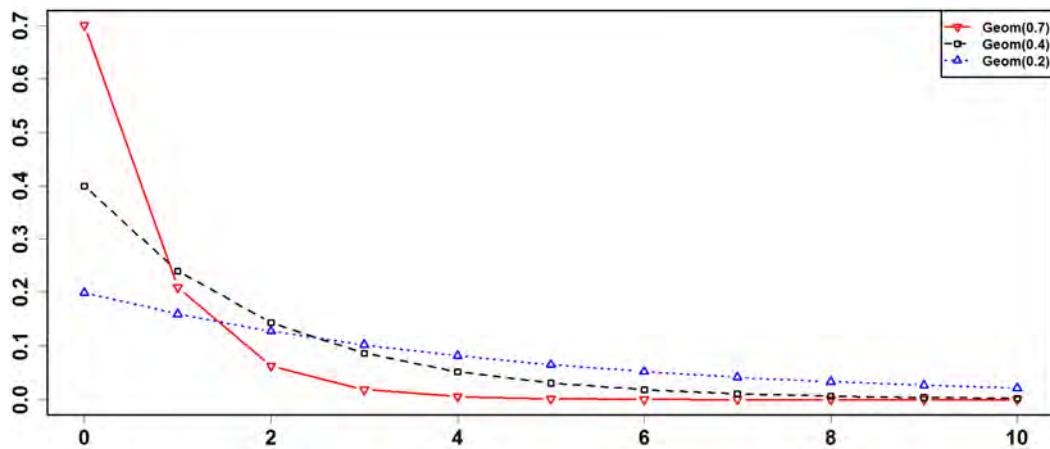


图 4.8: 几何分布 $\text{Geom}(p)$ 的概率函数之间的比较，其中 $p = 0.7, 0.4, 0.2$ 。

练习 4.10. 若 $X \sim \text{Geom}(p)$ ，则 $P(X \geq n) = (1 - p)^n$ ，其中 $n = 0, 1, 2, \dots$ 。

例 4.16. 某抽奖活动的中奖概率为 $p = 0.01$ ，每次只允许抽一张彩券。以不小于 95% 的概率至少中奖一次，需要抽多少次彩券？

解. 第一次中奖之前，未中奖的抽彩次数 $X \sim \text{Geom}(0.01)$ 。连抽 n 次都没中的概率是 $P(X \geq n) = (1 - p)^n$ ，由题意从 $P(X \geq n) \leq 1 - 0.95$ 解出需要抽 $n \geq 299$ 次。

练习 4.11. 几何分布 $X \sim \text{Geom}(p)$ 的特征函数、概率母函数和常见数字特征分别为

$$\begin{aligned} \varphi_X(t) &= p(1 - qe^{it})^{-1} & G_X(s) &= \frac{p}{1 - qs}, \text{ 其中 } |s| < \frac{1}{q} \\ E(X) &= \frac{q}{p} & V(X) &= \frac{q}{p^2} \\ c_s &= \frac{1+q}{\sqrt{q}} & c_k &= 6 + \frac{p^2}{q} \end{aligned}$$

^{*} “几何分布”一词缘自几何数列 $p, pq, \dots, pq^n, \dots$ 。有些教科书中把几何分布的随机变量 X 定义为事件 A 第一次发生时所需的试验次数，即 $X \sim p\langle 1 \rangle + pq\langle 2 \rangle + \dots + pq^{n-1}\langle n \rangle + \dots$ 。

在多重 Bernoulli 试验中，事件 A^c 的出现称为“失败”。连续失败 m 次后，对后续的试验而言这些失败的记忆都可以被遗忘。也就是说，在已有至少 m 次 A^c 连续出现的“历史条件”下，再有至少 n 次 A^c 连续出现的概率 $P(X \geq m+n | X \geq m)$ 等于不计“历史条件”的 $P(X \geq n)$ 。

定理 4.2 (无记忆性). 几何分布 $X \sim \text{Geom}(p)$ 满足所谓的“无记忆性”，即

$$P(X \geq m+n | X \geq m) = P(X \geq n)$$

其中 m, n 皆为非负整数。反之，若取值 $0, 1, 2, \dots$ 的离散型随机变量 X 具有无记忆性，则 X 服从几何分布。

证明. 首先往证“ \Rightarrow ”：由练习 4.10 知，

$$P(X \geq m+n | X \geq m) = \frac{P(X \geq m+n)}{P(X \geq m)} = (1-p)^n = P(X \geq n)$$

下面往证“ \Leftarrow ”：不妨设 $X \sim p_0\langle 0 \rangle + p_1\langle 1 \rangle + \dots + p_n\langle n \rangle + \dots$ ，其中 $0 < p_0 < 1$ ，则 $p_0 + p_1 + \dots + p_n + \dots = 1$ 。从 $P(X \geq 2 | X \geq 1) = P(X \geq 1)$ 可以推出

$$\frac{1 - p_0 - p_1}{1 - p_0} = 1 - p_0, \text{ 进而 } p_1 = p_0(1 - p_0)$$

用归纳法可证 $p_n = p_0(1 - p_0)^n$ ，即 $X \sim \text{Geom}(p_0)$ ，得证。 \square

定理 4.2 的直观含义是：历史上连续观察到了 m 个 A^c ，再观察到 n 个 A^c 的概率就等于没有历史观察到 n 个 A^c 的概率。在这无限重复的 Bernoulli 试验中，大自然对历史是没有记忆的。

性质 4.12. 若 $X_j \sim \text{Geom}(p_j), j = 1, 2, \dots, k$ 相互独立，则

$$X = \min(X_1, \dots, X_k) \sim \text{Geom}\left(1 - \prod_{j=1}^k (1 - p_j)\right)$$

证明. 在抛硬币试验 \mathcal{E}_j 中出现正面的概率是 p_j ，在试验 $\mathcal{E} = \mathcal{E}_1 \times \dots \times \mathcal{E}_k$ 中，有正面的概率是 $p = 1 - \prod_{j=1}^k (1 - p_j)$ 。无限地重复 \mathcal{E} ， $X = n$ 的概率是 $p(1 - p)^n$ 。 \square

几何分布的“连续版本”是指数分布（见第 296 页的定义 4.17） $\text{Expon}(\beta)$ ，之所以提前介绍它是因为几何分布的 RNG 要用到下面的结果。

\rightsquigarrow **性质 4.13.** 如果 $X \sim \text{Expon}(\beta)$ ，则

$$Y = \lfloor X \rfloor \sim \text{Geom}(1 - e^{-\beta})$$

证明. 根据式 (4.17), $Y = n$ 的概率是

$$P(Y = n) = \int_n^{n+1} \beta e^{-\beta x} dx = (1 - e^{-\beta})e^{-\beta n}, \text{ 其中 } n = 0, 1, 2, \dots$$

令 $p = 1 - e^{-\beta}$, 不难看出 $Y \sim \text{Geom}(p)$ 。 \square

算法 4.6. 利用 $U(0, 1)$ 的随机数 u^* 产生几何分布 $X \sim \text{Geom}(p)$ 的随机数 x^* :

$$x^* = \left\lfloor \frac{\ln u^*}{\ln(1-p)} \right\rfloor$$

证明. 根据性质 4.13, 由 $p = 1 - e^{-\beta}$ 得到 $\beta = -\ln(1-p)$, 参考算法 4.17 即可。 \square

几何分布和许多分布有关系, 除了刚刚介绍的指数分布, 还有下面重点讨论的 Pascal 分布, 也称作负二项分布。为什么要研究几何分布、负二项分布? 它的背景见 Bernoulli 过程 (见第 383 页的例 6.2)。

定义 4.9 (负二项分布, 亦称 Pascal 分布). 如果随机变量 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Geom}(p)$, 则随机变量 $X = X_1 + X_2 + \dots + X_n$ 服从负二项分布 (negative binomial distribution), 记作 $X \sim \text{NegB}(n, p)$ 。显然, 几何分布 $\text{Geom}(p)$ 就是 $\text{NegB}(1, p)$ 。



该分布刻画的是多重 Bernoulli 试验中, 某事件 A 第 n 次出现之前 A^c 出现 $X = k$ 次的概率, 其中 $k = 0, 1, 2, \dots$ 。

练习 4.12. 设 $0 < p < 1$, 负二项分布 $X \sim \text{NegB}(n, p)$ 的概率函数为

$$P(X = k) = C_{n+k-1}^k p^n q^k, \text{ 其中 } q = 1 - p, k = 0, 1, 2, \dots \quad (4.3)$$

它的特征函数、概率母函数和数字特征分别为

$$\begin{aligned} \varphi_X(t) &= p^n (1 - q e^{it})^{-n} & G_X(s) &= \left(\frac{ps}{1 - qs} \right)^n \\ E(X) &= \frac{nq}{p} & V(X) &= \frac{nq}{p^2} \\ c_s &= \frac{1+q}{\sqrt{nq}} & c_k &= \frac{6}{n} + \frac{p^2}{nq} \end{aligned}$$

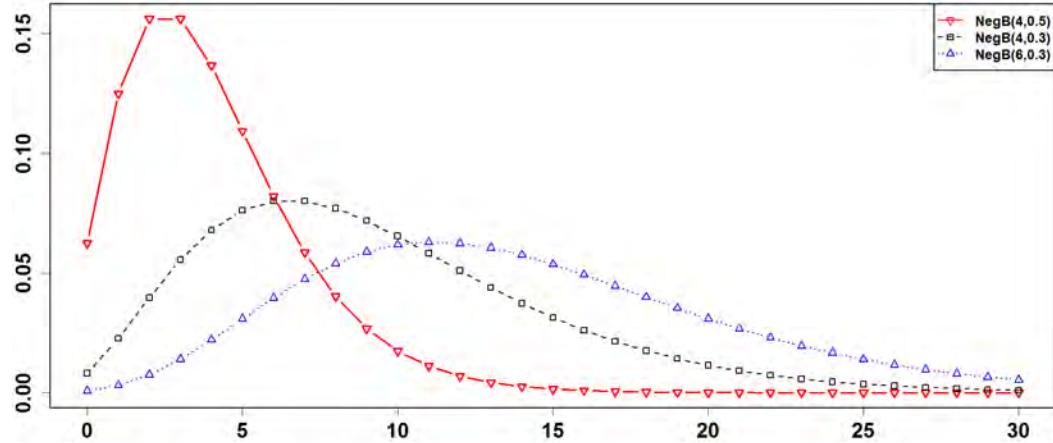


图 4.9: 负二项分布 $\text{NegB}(4, 0.5)$, $\text{NegB}(4, 0.3)$ 和 $\text{NegB}(6, 0.3)$ 的概率函数。

练习 4.13. 与性质 4.6 类似, 负二项分布在一定条件下也满足和型不变性。即, 如果随机变量 $X \sim \text{NegB}(n, p)$ 与 $Y \sim \text{NegB}(m, p)$ 相互独立, 则 $X + Y \sim \text{NegB}(m + n, p)$ 。

例 4.17. 已知 $X_1, X_2 \stackrel{\text{iid}}{\sim} \text{Geom}(p)$, 试证明:

$$P(X_1 = k | X_1 + X_2 = n) = \frac{1}{n+1}, \quad \text{其中 } k = 0, 1, \dots, n$$

证明. 由定义 4.9 知, $X_1 + X_2 \sim \text{NegB}(2, p)$ 。因为 X_1, X_2 独立, 所以

$$\begin{aligned} P(X_1 = k, X_1 + X_2 = n) &= P(X_1 = k, X_2 = n - k) \\ &= P(X_1 = k)P(X_2 = n - k) \\ \text{进而, } P(X_1 = k | X_1 + X_2 = n) &= \frac{P(X_1 = k)P(X_2 = n - k)}{P(X_1 + X_2 = n)} \\ &= \frac{pq^k \cdot pq^{n-k}}{C_{n+1}^n p^2 q^n} = \frac{1}{n+1} \end{aligned} \quad \square$$

算法 4.7. 由负二项分布的定义, $X \sim \text{NegB}(n, p)$ 的随机数 x^* 可利用几何分布 $\text{Geom}(p)$ 的随机数 $x_1^*, x_2^*, \dots, x_n^*$ 通过 $x^* = x_1^* + x_2^* + \dots + x_n^*$ 得到。

4.1.5 Poisson 分布

1837 年, 法国数学家 S. D. Poisson (1781-1840) 发表著作《关于刑事和民事案件审判的概率研究》, 其中 Poisson 研究了这样一个随机变量, 它刻画的是在一个给定长度的时间段内某随机事件发生的次数。譬如, 在长度为 t 的时间段内电话被呼叫的次数 (见第 90 页的例 1.70)、服务器被攻击的次数等。有人甚至用这样的随机变量为 1875 至 1894 年间普鲁士军队中每年被马踢死的士兵数建模, 理论结果和实际观察惊人地吻合。

换一种说法, Poisson 研究的是随机变量序列 $X_n \sim B(n, p_n), n = 1, 2, \dots$ 在附加约束条件 $\lim_{n \rightarrow \infty} np_n = \lambda > 0$ (其中 λ 为常数) 之下的“极限状态”。Poisson 发现, 随着 $n \rightarrow \infty$, 概率 $P(X_n = k) = C_n^k p_n^k (1 - p_n)^{n-k}$ 具有如下变化趋势。

$$\frac{(np_n)^k}{k!} \left(1 - \frac{np_n}{n}\right)^n \frac{n(n-1)\cdots(n-k+1)}{(n-np_n)^k} \rightarrow \frac{\lambda^k}{k!} e^{-\lambda} \quad (4.4)$$



定义 4.10 (Poisson 分布). 如果取值为 $0, 1, 2, \dots$ 的离散型随机变量 X 具有如下的概率函数,

$$P_\lambda(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ 其中 } k = 0, 1, 2, \dots \quad (4.5)$$

则称 X 服从参数为 λ 的 Poisson 分布, 记作 $X \sim \text{Poisson}(\lambda)$ 。当 $\lambda = 2, 4, 6$ 时, 概率函数 (4.5) 的折线图见第 91 页的图 1.33。Poisson 分布在随机过程、排队论、可靠性理论中非常有用。第 419 页讲了 Poisson 分布的由来。

练习 4.14. 已知 $X \sim \text{Poisson}(\lambda)$, 试证明: 对于任一固定的 $k = 0, 1, 2, \dots$, 函数 $P_\lambda(X \leq k)$ 关于 λ 是非增的。提示:

$$\frac{d}{d\lambda} P_\lambda(X \leq k) = \frac{d}{d\lambda} \sum_{j=0}^k \frac{\lambda^j e^{-\lambda}}{j!} = -\frac{\lambda^k e^{-\lambda}}{k!} < 0$$

练习 4.15. 验证 $X \sim \text{Poisson}(\lambda)$ 满足递归关系

$$P(X = k + 1) = \frac{\lambda}{k + 1} P(X = k)$$

例 4.18. Poisson 分布常用于描述一些物理现象, 详见 W. Feller 的《概率论及其应用》上卷 [45] 第六章。例如, 考虑放射性元素钋在 7.5 秒的时间段内释放的 α 粒子数, 可能是 $0, 1, 2, \dots$ 。1910 年, Rutherford-Geiger 实验独立地观察了 $N = 2608$ 次,

结果是 57 次试验发现粒子个数为 0, 203 次试验发现粒子个数为 1, ……, 1 次试验发现粒子个数为 14。平均每次试验发现的粒子个数为

$$\hat{\lambda} = \frac{1}{2608}(0 \times 57 + 1 \times 203 + \cdots + 14 \times 1) = 3.871549$$

令 $X \sim \text{Poisson}(\hat{\lambda})$ 且 $P(X = k) = p_k$, 将 Np_k 作为观察到 k 个粒子的试验次数的拟合值, 其中 $k = 0, 1, 2, \dots, 14$ 。通过对比, 读者不难发现拟合值与真实值很接近。

表 4.1: Rutherford-Geiger 实验 (1910) 及其拟合结果。

粒子个数 k	发现次数	拟合结果 Np_k
0	57	54.3144249
1	203	210.2809617
2	383	407.0565318
3	525	525.3131137
4	532	508.4438755
5	408	393.6930837
6	273	254.0336826
7	139	140.5005529
8	45	67.9943483
9	27	29.2492729
10	10	11.3239996
11	4	3.9855836
12	0	1.2858652
13	1	0.3829454
14	1	0.1058994

例 4.19. 假设一年 365 天的出生率都相同。随机地选择 400 个人, 生日是 6 月 3 日的人数 $Y \sim B(400, 1/365)$, 与 $X \sim \text{Poisson}(1)$ 的概率分布进行比较, 发现二者很接近。

表 4.2: 二项分布与 Poisson 分布的比较。

k	0	1	2	3	4	5
$B(400, 1/365)$	0.3337	0.3667	0.2010	0.0733	0.0200	0.0043
$\text{Poisson}(1)$	0.3679	0.3679	0.1839	0.0613	0.0153	0.0031

性质 4.14. 设随机变量序列 $X_n \sim B(n, p_n), n = 1, 2, \dots$ 满足 $\lim_{n \rightarrow \infty} np_n = \lambda > 0$, 则 $X_n \xrightarrow{L} X$, 其中 $X \sim \text{Poisson}(\lambda)$ 。该性质的直观解释见图 4.10 的左图。

练习 4.16. 仿照式 (4.4), 可得到与性质 4.14 类似的结果如下 (其直观解释见图 4.10 的右图): 设随机变量序列 $X_n \sim \text{NegB}(n, p_n), n = 1, 2, \dots$ 满足 $\lim_{n \rightarrow \infty} n(1 - p_n) = \lambda > 0$, 则 $X_n \xrightarrow{L} X$, 其中 $X \sim \text{Poisson}(\lambda)$ 。

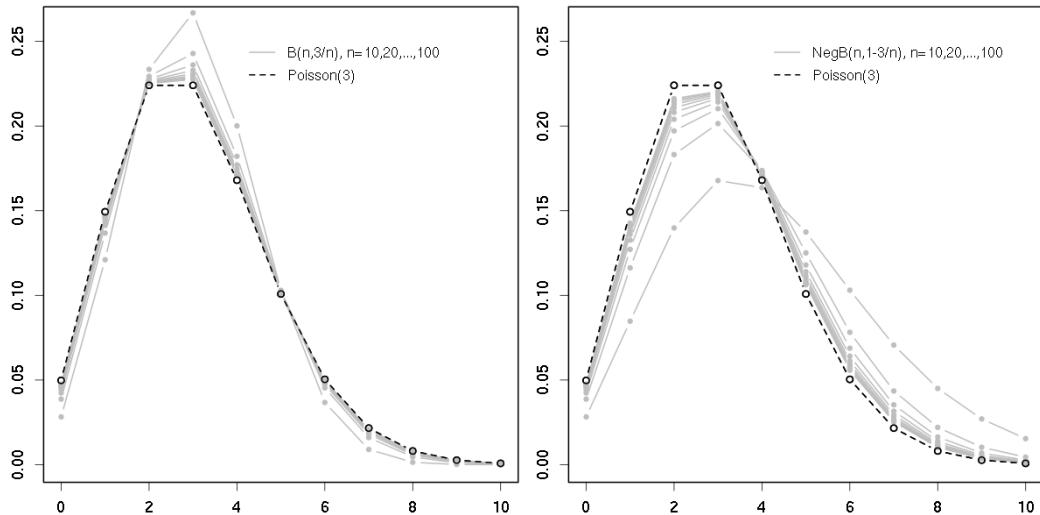


图 4.10: 当 n 很大而 p 很小时, 二项分布 $B(n, p)$ 可用分布 $Poisson(np)$ 来近似。左图中, 虚线是 $Poisson(3)$ 的折线图, 实线是 $B(n, 3/n)$ 的折线图, 性质 4.14 保证 n 越大二者越接近。右图是对练习 4.16 结果的直观解释。

性质 4.15. 由例 3.14 知, $X \sim Poisson(\lambda)$ 的特征函数、概率母函数和数字特征分别为

$$\begin{aligned}\varphi_X(t) &= \exp\{\lambda(e^{it} - 1)\} & G_X(s) &= \exp\{\lambda(s - 1)\} \\ E(X) &= \lambda & V(X) &= \lambda \\ c_s &= \frac{1}{\sqrt{\lambda}} & c_k &= \frac{1}{\lambda}\end{aligned}$$

例 4.20. 已知随机变量 $X \sim Poisson(\lambda)$, 试证明: 当 $\lambda \rightarrow \infty$ 时, 有

$$Y = \frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow{L} N(0, 1)$$

证明. 由式 (3.9) 知, 随机变量 $Y = \frac{X - \lambda}{\sqrt{\lambda}}$ 的特征函数为

$$\varphi_Y(t) = \exp\{-it\sqrt{\lambda}\} \exp\{\lambda(e^{it/\sqrt{\lambda}} - 1)\} = \exp\left\{-\frac{t^2}{2} + O\left(\frac{1}{\sqrt{\lambda}}\right)\right\}$$

当 $\lambda \rightarrow \infty$ 时, $\varphi_Y(t)$ 趋于 $\exp\{-t^2/2\}$, 即标准正态分布的特征函数。 \square

※例 4.21. 对于随机变量 $X \sim Poisson(\lambda)$, 有

$$\psi(t) = \ln \varphi(t) = \lambda \sum_{k=1}^{\infty} \frac{(it)^k}{k!}$$

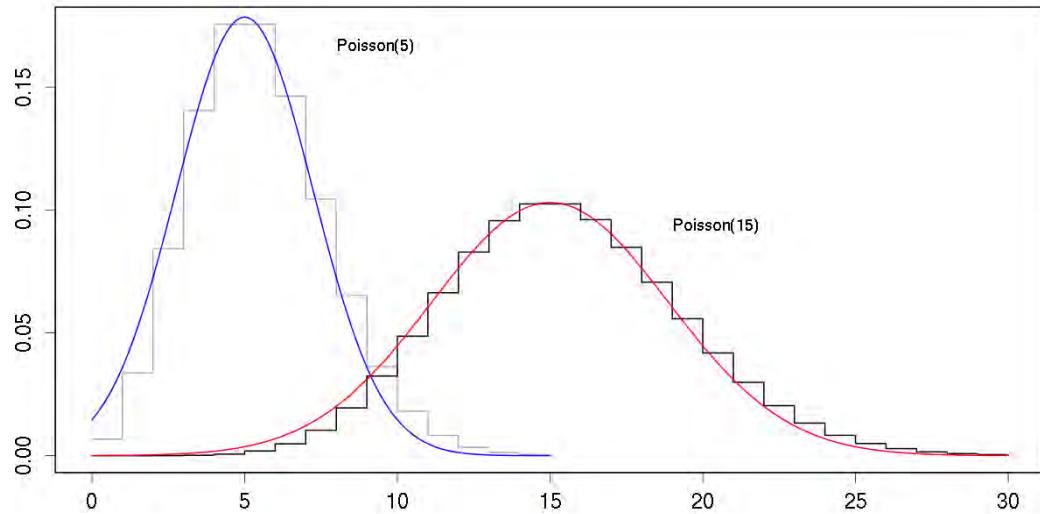


图 4.11: Poisson 分布的正态近似: 参数 $\lambda = 5$ 和 15 时, 概率分布 $\text{Poisson}(\lambda)$ 的阶梯图与 $N(\lambda, \lambda)$ 的密度函数曲线。不难发现 λ 越大, 二者越接近。

由式 (3.14) 可得 k 阶半不变量 $\varkappa_k = \lambda$, 其中 $k = 1, 2, 3, \dots$ 。

练习 4.17 (和型不变性). 已知随机变量 X_j 相互独立并且 $X_j \sim \text{Poisson}(\lambda_j)$, 其中 $j = 1, 2, \dots, n$, 试证明:

$$X_1 + X_2 + \dots + X_n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \dots + \lambda_n)$$

提示: 由定理 3.4 和 Poisson 分布的特征函数易证。

练习 4.18. 如果 $X_1 \sim \text{Poisson}(\lambda_1)$ 与 $X_2 \sim \text{Poisson}(\lambda_2)$ 相互独立, 则

$$X_1 | (X_1 + X_2 = n) \sim B\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$$

提示: 利用练习 4.17 的结果, 仿照例 4.17 可证。

作为练习 4.17 的逆命题, 下面我们不加证明地介绍苏联数学家 Dmitry Abramovich Raikov (1905-1981) 证得的有关 Poisson 分布的一个非常有趣的性质。

定理 4.3 (Raikov, 1937). 两个独立随机变量之和服从 Poisson 分布, 则这两个随机变量也是服从 Poisson 分布的。

例 4.22 (复合分布). 已知 $X \sim B(N, p)$ 且 $N \sim \text{Poisson}(\lambda)$, 试证明: $X \sim \text{Poisson}(\lambda p)$ 。

证明. 对于任意的 $s \in \{0, 1, 2, \dots\}$, 皆有

$$\begin{aligned}
 P(X = s) &= \sum_{n=0}^{\infty} P(X = s | N = n) P(N = n) \\
 &= \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} C_n^s p^s (1-p)^{n-s} \\
 &= \frac{e^{-\lambda} p^s \lambda^s}{s!} \sum_{n=s}^{\infty} \frac{(\lambda - \lambda p)^{n-s}}{(n-s)!} \\
 &= \frac{e^{-\lambda} p^s \lambda^s}{s!} e^{\lambda(1-p)} \\
 &= \frac{(\lambda p)^s}{s!} e^{-\lambda p}
 \end{aligned}
 \quad \square$$

上例中的 X 称为服从复合的 Poisson 分布。关于复合的 Poisson 分布有下面更一般的结果, 请读者验证上例是下例的特款。

例 4.23. 已知 $N \sim \text{Poisson}(\lambda)$ 。假设随机变量 X_1, \dots, X_N 独立同分布, 特征函数都是 $\varphi(t)$ 。若 N, X_1, \dots, X_N 相互独立, 则随机变量 $Y = X_1 + \dots + X_N$ 的特征函数为

$$\varphi_Y(t) = \exp\{\lambda[\varphi(t) - 1]\}$$

证明. 由 $E[e^{it(X_1 + \dots + X_N)} | N = n] = [\varphi(t)]^n$ 和双期望定理 2.18 得到

$$\varphi_Y(t) = E\{E[e^{it(X_1 + \dots + X_N)} | N]\} = \sum_{n=1}^{\infty} [\varphi(t)]^n \frac{\lambda^n}{n!} e^{-\lambda} = \exp\{\lambda[\varphi(t) - 1]\} \quad \square$$

4.2 连续型随机变量的分布

连续型随机变量 X 的所有“信息”都在它的密度函数 $f_X(x)$ 中，因为 X 落在任意 Borel 集 $B \in \mathfrak{B}_1$ 上的概率 $P(X \in B)$ 可由 $f_X(x)$ 表示为

$$P(X \in B) = \int_B f_X(x) dx$$

我们在第 2 章已经介绍过连续型随机变量的分布，如均匀分布、正态分布等，本章还将详细介绍 Laplace 分布、Cauchy 分布、Gamma 分布、Beta 分布、 t 分布、 F 分布、Pareto 分布，以及物理学中常见连续型随机变量的分布。

标准正态分布的密度函数 $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}$ 是一个很特殊的函数，中间高两边低的密度函数中为何单单它成为万众瞩目的焦点？原因是第 5 章即将介绍的中心极限定理，附录 A 详述了正态分布的由来，它的密度函数是由二项分布 $B(n, p)$ 推导出来的。

本章的开篇介绍了利用 LCG 来产生均匀分布 $U(0, 1)$ 的伪随机数，很多分布的随机数都可以通过它来构造。如果有某种物理的方法（如光电效应、大气噪声等随机现象）能够生成 0-1 分布 $\frac{1}{2}\langle 1 \rangle + \frac{1}{2}\langle 0 \rangle$ 的随机数，则可以通过下面的方法产生 $U(0, 1)$ 分布的真随机数。

算法 4.8. 令随机变量 $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} \frac{1}{2}\langle 1 \rangle + \frac{1}{2}\langle 0 \rangle$ ，则 $X \sim U[0, 1]$ 的随机数可按下列方式构造： $X = \sum_{k=1}^{\infty} 2^{-k}X_k$ ，即实数的二进制表示。

练习 4.19. 请问：第 124 页的练习 2.6 中分布 $g(x)$ 的随机数如何产生？

提示：先产生 Y 的随机数 j ，再产生 $f_j(x)$ 的随机数即为所求。

算法 4.9. 已知概率密度函数 $f_1(x), \dots, f_n(x)$ 以及非负实数 p_1, \dots, p_n 满足 $p_1 + \dots + p_n = 1$ ，下面的方法可产生 $X \sim p_1f_1(x) + \dots + p_nf_n(x)$ 的随机数。

- 产生 $K \sim p_1\langle 1 \rangle + \dots + p_n\langle n \rangle$ 的随机数 k^* 。
- 产生 $X \sim f_{k^*}(x)$ 的随机数 x^* 即是所求。

本节内容

均匀分布之所以重要，是因为很多分布的随机数都可由 $U[0, 1]$ 的随机数构造，例如，若 $X \sim U(0, 1)$ ，则 $Y = -\ln X \sim \text{Expon}(1)$ ；正态分布因中心极限定理而变成重中之重；Beta 分布和 Gamma 分布来自 Euler 的两类积分， χ^2 分布和指数分布是 Gamma 分布的特例； χ^2 分布（见第 162 页的例 2.45）、 t 分布和 F 分布（见习题 2.32 和 2.33）来自统计学；Pareto 分布来自经济学；Boltzmann 分布、Gibbs 分布、Weibull 分布、Rayleigh 分布、Maxwell 分布（见例 2.45）和 Wigner 半圆分布都来自物理学。

关键知识

(1) 了解这些常见连续型随机变量的由来, 熟悉它们的密度函数、特征函数以及常见的数字特征; (2) 掌握这些常见分布的性质, 如独立的正态分布(或 Cauchy 分布、 χ^2 分布)的随机变量之和依然服从正态分布(或 Cauchy 分布、 χ^2 分布); (3) 掌握逆 CDF 法, 了解常见分布的随机数产生算法。

4.2.1 均匀分布

性质 4.16. 均匀分布（见第 124 页的例 2.12） $X \sim U[a, b]$ 是“等概率”的连续情形，其特征函数和一些常见的数字特征分别为

$$\begin{aligned}\varphi(t) &= \frac{\exp(itb) - \exp(ita)}{it(b-a)} \\ E(X) = M(X) &= \frac{a+b}{2}, & V(X) &= \frac{(b-a)^2}{12} \\ c_s &= 0, & c_k &= -\frac{6}{5}\end{aligned}$$

练习 4.20. 求 $X \sim U[0, 1]$ 的特征函数。答案： $i(1 - e^{it})/t$ 。

性质 4.17. 如果 $X \sim U[0, 1]$ 且 $b > a$ ，则

$$a + (b-a)X \sim U[a, b]$$

证明. 随机变量 $Y = a + (b-a)X$ 的密度函数为

$$\begin{aligned}f_Y(y) &= \frac{1}{b-a} f_X\left(\frac{y-a}{b-a}\right) \\ &= \begin{cases} \frac{1}{b-a} & \text{当 } a \leq y \leq b \\ 0 & \text{其他} \end{cases} \quad \square\end{aligned}$$

若 $X \sim U[a, b]$ ，当 $b \rightarrow a$ 时， X 退化为单点分布 $X \sim \langle a \rangle$ 。为了保证连续型随机变量 X 的密度函数 $f(x)$ 在退化的状态下依然存在，我们引入 delta 函数将之表示为 $f(x) = \delta(x - a)$ 。由式 (3.4)，读者不难验证 $E(X) = a, V(X) = 0$ 与单点分布吻合。

定理 4.4. 已知随机变量 X 的分布函数 $F_X(x)$ 连续，则

$$Y = F_X(X) \sim U[0, 1]$$

证明. 对任意 $y \in [0, 1]$ ，至少存在一个 x 使得 $y = F_X(x) = P\{X \leq x\}$ ，记 $F_X^{-1}(y)$ 为这些 x 中的最小者。随机变量 Y 的分布函数为

$$F_Y(y) = P\{F_X(X) \leq y\} = \begin{cases} 0 & \text{当 } y \leq 0 \\ P\{X \leq F_X^{-1}(y)\} = y & \text{当 } 0 < y \leq 1 \\ 1 & \text{当 } y > 1 \end{cases} \quad \square$$

算法 4.10 (逆 CDF 法). 定理 4.4 为 $X \sim F_X(x)$ 的随机数的提供了一个可行的产生方法，称为逆 CDF 法，它由以下两个步骤组成。

- 产生 $Y \sim [0, 1]$ 的随机数 y^* ;
- 通过 $x^* = F_X^{-1}(y^*)$ 求得 X 的随机数 x^* 。

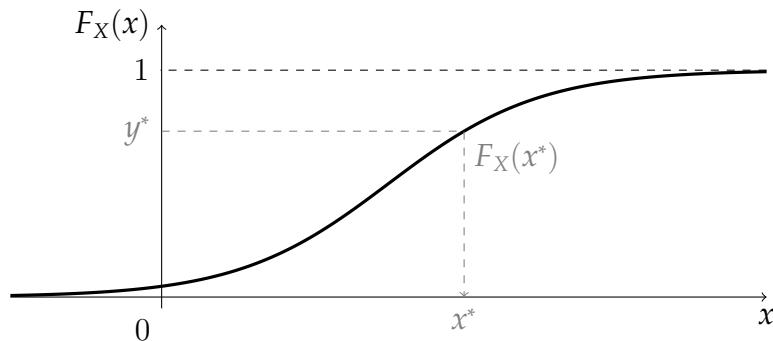


图 4.12: 算法 4.10 (逆 CDF 法) 的直观含义: 基于定理 4.4, $Y = F_X(X) \sim U[0, 1]$ 。一般地, $F_X^{-1}(\cdot)$ 不是显式的, 要借助数值分析的方法才能由 y^* 得到 x^* 。

练习 4.21. 利用逆 CDF 法给出算法 4.3 和算法 4.6。

练习 4.22. 利用逆 CDF 法给出 Logistic(0,1) 分布的随机数产生算法。随机变量 X 服从 Logistic 分布, 记作 $X \sim \text{Logistic}(m, s)$, 其中 m 为位置参数, s 为尺度参数, 当且仅当其密度函数为

$$f_X(x) = \frac{\exp\{-(x-m)/s\}}{s[1 + \exp\{-(x-m)/s\}]^2}$$

提示: $X \sim \text{Logistic}(m, s)$ 的分布函数为

$$F_X(x) = \frac{1}{1 + \exp\{-(x-m)/s\}}$$

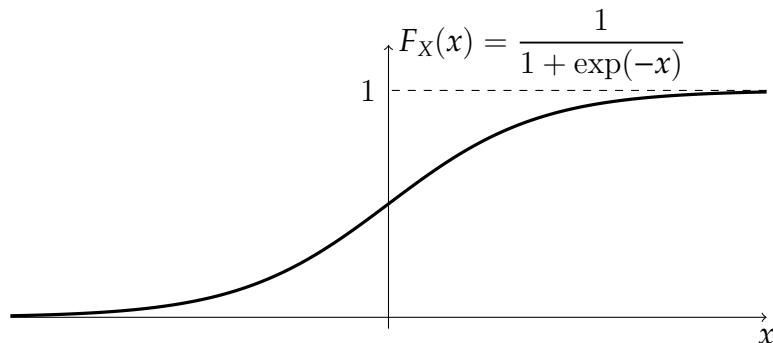


图 4.13: $X \sim \text{Logistic}(0, 1)$ 的分布函数在人工神经网络、机器学习中俗称 sigmoid 函数或 logistic 函数, 它是定义在 \mathbb{R} 上、取值在 $(0, 1)$ 内的、严格增的光滑函数。

 **定理 4.4** 虽然提供了一种产生随机数的通用方法，但如果分布函数 $F_X(x)$ 没有解析表达式或其逆映射很难求得，譬如正态分布，还需要借助其他数值计算的方法才能算得随机数。

定理 4.5. 已知随机变量 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} U[0, 1]$ ，则随机变量 $X = X_1 + X_2 + \dots + X_n$ 的特征函数为 $\varphi_X(t) = [i(1 - e^{it})/t]^n$ ，期望和方差分别为 $E(X) = n/2$ 和 $V(X) = n/12$ ，且 X 具有密度函数

$$f_n(x) = \begin{cases} \sum_{k=0}^n (-1)^k \frac{C_n^k (x-k)_+^{n-1}}{(n-1)!} & \text{当 } x \in [0, n] \\ 0 & \text{当 } x \notin [0, n] \end{cases} \quad (4.6)$$

其中正截尾函数 x_+ 定义如下：

$$x_+ = xJ(x) = \begin{cases} x, & \text{若 } x > 0 \\ 0, & \text{若 } x \leq 0 \end{cases}$$

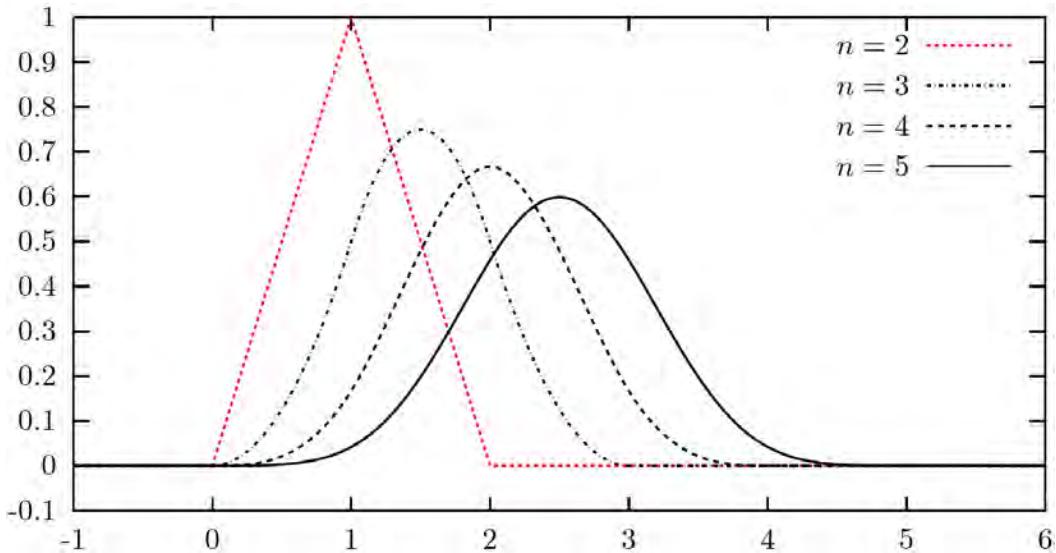


图 4.14：随着 n 的增大，密度函数 (4.6) 越来越逼近正态分布 $\phi(x|n/2, n/12)$ 。事实上，当 $n = 4$ 时，逼近效果就已经非常好了。

例 4.24. 定理 4.5 中，当 $n = 2$ 时 X 的分布是区间 $[0, 2]$ 上的三角形分布，由第 157 页的例 2.39 知其密度函数为

$$f_2(x) = \begin{cases} 1 - |x - 1| & \text{当 } x \in [0, 2] \\ 0 & \text{当 } x \notin [0, 2] \end{cases} = \begin{cases} x & \text{当 } x \in [0, 1] \\ 2 - x & \text{当 } x \in [1, 2] \\ 0 & \text{当 } x \notin [0, 2] \end{cases}$$

练习 4.23. 求例 4.24 中 $X \sim \Delta[0, 2]$ 的特征函数。答案： $-(e^{it} - 1)^2/t^2$ 。

练习 4.24. 定理 4.5 中, 求 $n = 3$ 时 X 的密度函数 $f(x)$ 。答案:

$$f(x) = \begin{cases} \frac{1}{2}x^2 & \text{当 } x \in [0, 1) \\ -x^2 + 3x - \frac{3}{2} & \text{当 } x \in [1, 2] \\ \frac{1}{2}x^2 - 3x + \frac{9}{2} & \text{当 } x \in [2, 3] \\ 0 & \text{其他} \end{cases}$$

练习 4.25. 令 X 是定理 4.5 中所定义的随机变量, 试证明: 当 $n \rightarrow \infty$ 时, X 的分布趋向 $N(n/2, n/12)$ 。提示: 利用式 (3.9) 验证标准化的随机变量 $(X - n/2)/\sqrt{n/12}$ 的特征函数为 $\exp\{-t^2/2 + O(1/n)\}$ 。

4.2.2 三角形分布

定义 4.11. 连续型随机变量 X 服从区间 $[a, b]$ 上众数为 c 的三角形分布 (triangular distribution), 记作 $X \sim \Delta[a, b; c]$, 当且仅当其密度函数为

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{当 } a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{当 } c < x \leq b \\ 0 & \text{其他} \end{cases} \quad (4.7)$$

特别地, 当 $c = (a+b)/2$, 即区间 $[a, b]$ 的中点时, $\Delta[a, b; c]$ 也称为等腰三角形分布, 简记作 $\Delta[a, b]$ 。

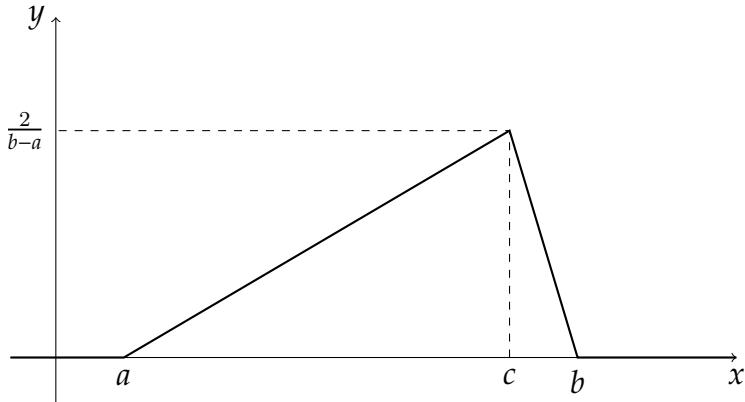


图 4.15: 三角形分布 $\Delta[a, b; c]$ 的众数是 c , 即密度函数在 c 取得最大值。

练习 4.26. 请验证三角形分布 $X \sim \Delta[a, b; c]$ 的分布函数、期望和方差分别为

$$F(x) = \begin{cases} 0 & \text{当 } x < a \\ \frac{(x-a)^2}{(b-a)(c-a)} & \text{当 } a \leq x \leq c \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)} & \text{当 } c < x \leq b \\ 1 & \text{当 } x > b \end{cases}$$

$$\mathbb{E}(X) = \frac{a+b+c}{3}, \quad \mathbb{V}(X) = \frac{a^2+b^2+c^2-ab-ac-bc}{18}$$

算法 4.11. 基于 $U[0, 1]$ 的随机数 u^* , 利用逆 CDF 法构造三角形分布 $X \sim \Delta[a, b; c]$ 的随机数 x^* 如下。

$$x^* = \begin{cases} a + \sqrt{u^*(b-a)(c-a)} & \text{如果 } 0 \leq u^* \leq \frac{c-a}{b-a} \\ b + \sqrt{(1-u^*)(b-a)(b-c)} & \text{如果 } \frac{c-a}{b-a} < u^* \leq 1 \end{cases}$$

4.2.3 正态分布、对数正态分布和偏正态分布

由于中心极限定理，正态分布（见第 125 页的 [定义 2.13](#)）是最重要的分布，历史上首位发现函数 $\phi(x|\mu, \sigma^2)$ 重要价值的是法国数学家 A. de Moivre。



De Moivre 于 1718 年出版了史上第一部概率论教材《机遇论》，首次给出了二项分布的概率函数，并描绘了正态分布的密度函数 $\phi(x|\mu, \sigma^2)$ 。1733 年，de Moivre 用 $\phi(x|\mu, \sigma^2)$ 来逼近二项分布（见[附录 A](#)）。该函数在物理学也有重要的应用，譬如

$$u(x, t) = \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{x^2}{4t}\right\}, \text{ 其中 } t > 0$$

就是热传导方程 $u_t = u_{xx}$ 的基本解，它的物理含义是开始集中于原点的单位热源所造成的 t 时刻 x 轴上的热量分布。

性质 4.18. 正态分布 $X \sim N(\mu, \sigma^2)$ 的特征函数为 $\varphi(t) = \exp(it\mu - \sigma^2 t^2 / 2)$ ，其数字特征分别为 $E(X) = \mu, V(X) = \sigma^2$ ，偏度系数和峰度系数都是 0。

~定理 4.6 (和型不变性). 如果随机变量 $X_j \sim N(\mu_j, \sigma_j^2), j = 1, 2, \dots, n$ 相互独立，则

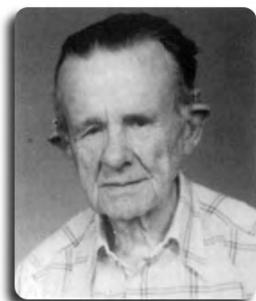
$$\sum_{j=1}^n X_j \sim N\left(\sum_{j=1}^n \mu_j, \sum_{j=1}^n \sigma_j^2\right)$$

证明. 根据[定理 3.4](#)，随机变量 $X = \sum_{j=1}^n X_j$ 的特征函数为

$$\varphi_X(t) = \exp\left\{it \sum_{j=1}^n \mu_j - \frac{t^2}{2} \sum_{j=1}^n \sigma_j^2\right\}$$

恰为正态分布 $N(\sum_{j=1}^n \mu_j, \sum_{j=1}^n \sigma_j^2)$ 的特征函数。 □

与 Raikov 定理 4.3 类似，正态分布也有类似的结果。该结果先被 P. Lévy 猜测到，后于 1936 年被瑞典数学家、统计学家 Harald Cramér (1893-1985, 照片见右) 证得 Cramér-Lévy 定理。1946 年 Cramér 出版的《统计学数学方法》[29] 是数理统计学的第一部经典名著，标志着统计学进入了成熟的阶段。



~定理 4.7 (Cramér-Lévy, 1936). 如果随机变量 X_1, X_2 相互独立且 $X = X_1 + X_2$ 服从正态分布，则 X_1, X_2 都服从正态分布。

※证明. 详见 Feller 的《概率论及其应用》下卷第十五章第八节。 □

 与定理 4.7 等价的结果是：若两个密度函数的卷积是正态分布，则这两个密度函数都是正态的。 $X = X_1 + X_2$ 即便是接近正态分布的，也能推出 X_1, X_2 是接近正态分布的。这种稳定性源于正态分布往往是一些无关紧要的独立因素集体作用的结果。对正态分布更深刻的理解要等到学习中心极限定理（详见本书的第 5 章）。

下面介绍两个从均匀分布的随机数“构造”出正态分布的随机数的常用方法，它们的理论依据分别是如下的定理 4.8 和定理 4.5。

定理 4.8 (Box-Muller 变换法^{*}, 1958). 如果随机变量 $U_1, U_2 \stackrel{\text{iid}}{\sim} U(0, 1)$ ，定义新的随机变量 X_1, X_2 如下，则 $X_1, X_2 \stackrel{\text{iid}}{\sim} N(0, 1)$ 。

$$\begin{aligned} X_1 &= \sqrt{-2 \ln U_1} \cos(2\pi U_2) \\ X_2 &= \sqrt{-2 \ln U_1} \sin(2\pi U_2) \end{aligned}$$

证明. 利用第 154 页的定理 2.15 求解 $(X_1, X_2)^\top$ 的密度函数 $f(x_1, x_2)$ 如下：Box-Muller 变换的逆变换为

$$U_1 = \exp \left\{ -\frac{1}{2}(X_1^2 + X_2^2) \right\}, \quad U_2 = \frac{1}{2\pi} \arctan \frac{X_2}{X_1}$$

请读者验证其雅可比行列式的绝对值为

$$|J| = \frac{1}{2\pi} \exp \left\{ -\frac{1}{2}(x_1^2 + x_2^2) \right\}$$

于是 $f(x_1, x_2) = \phi(x_1)\phi(x_2)$ ，进而 $X_1, X_2 \stackrel{\text{iid}}{\sim} N(0, 1)$ 得证。 \square

练习 4.27. 根据定理 4.8 设计分布 $N(0, 1)$ 的伪随机数产生器。

算法 4.12 (正态分布的伪随机数产生器). 令 $(X, Y)^\top$ 是上半单位圆内均匀分布的随机点， α 是该点的弧度，显然 $\alpha \sim U(0, \pi)$ ，并有

$$\begin{aligned} \sin(2\alpha) &= \frac{2XY}{X^2 + Y^2} \\ \cos(2\alpha) &= \frac{X^2 - Y^2}{X^2 + Y^2} \end{aligned}$$

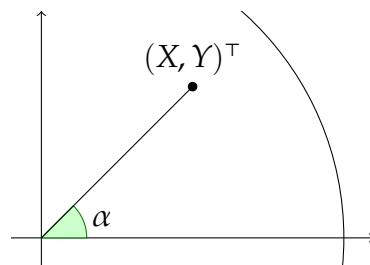


图 4.16: 若 $(X, Y)^\top$ 服从上半单位圆内均匀分布，则 $\alpha \sim U(0, \pi)$ 。

*Box-Muller 变换法由英国统计学家 George Edward Pelham Box (1919-) 和 Mervin Edgar Muller 于 1958 年提出 [19]。

在定理 4.8 中, 令 $U_2 = \alpha/\pi$ 。下面的算法是 Box-Muller 变换法的变体, 避免了调用三角函数 \sin 和 \cos , 付出的代价是花费一些额外的时间用于产生上半单位圆内均匀分布的随机点。

- 产生均匀分布 $U(0, 1)$ 的随机数 u^*, x_*, y_* ; 置 $x_* \leftarrow 2x_* - 1$ 。
- 若 $x_*^2 + y_*^2 > 1$, 则回到步骤 1; 否则, 置

$$\begin{aligned} x_1 &\leftarrow \frac{x_*^2 - y_*^2}{x_*^2 + y_*^2} \sqrt{-2 \ln u^*} \\ x_2 &\leftarrow \frac{2x_*y_*}{x_*^2 + y_*^2} \sqrt{-2 \ln u^*} \end{aligned}$$

定义 4.12 (对数正态分布). 如果随机变量 $Y \sim N(\mu, \sigma^2)$, 则称 $X = e^Y$ 服从对数正态分布, 记作 $X \sim \log N(\mu, \sigma^2)$, 其中参数 $\sigma > 0$ 。分布 $\log N(\mu, \sigma^2)$ 的密度函数为

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\} & \text{当 } x > 0 \\ 0 & \text{当 } x \leq 0 \end{cases} \quad (4.8)$$

对数正态分布用于定义一个重要的随机过程——几何布朗运动 (见定义 6.33), 常用于金融数学。另外, 该分布不能由各阶矩的信息唯一确定。

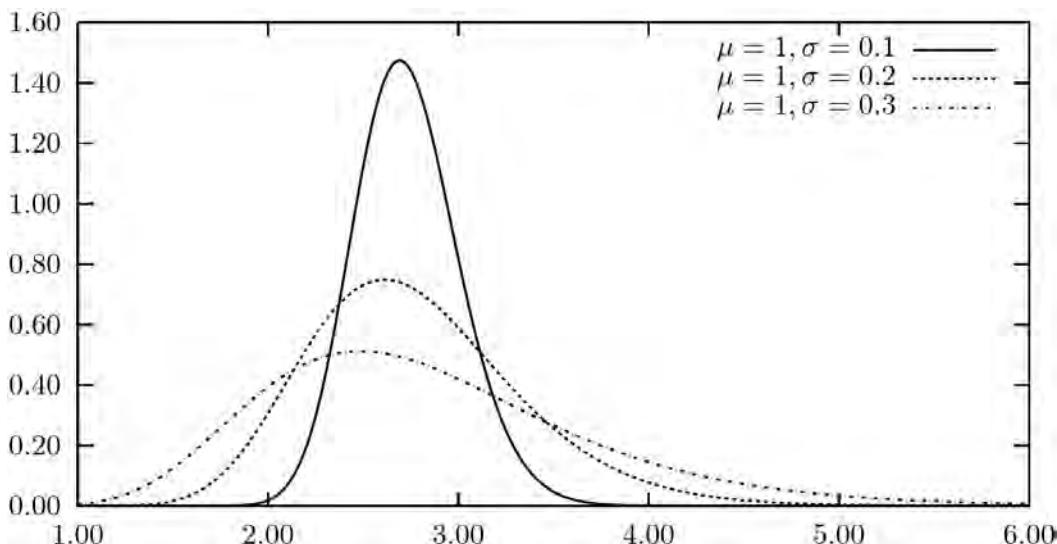


图 4.17: 参数 σ 越小, 对数正态分布的密度函数曲线显得越“高瘦”。对数正态分布不能由各阶矩唯一确定。

练习 4.28. 试证明: 若随机变量 $X_1 \sim \log N(\mu_1, \sigma_1^2)$ 和 $X_2 \sim \log N(\mu_2, \sigma_2^2)$ 相互独立, 则 $X_1 X_2 \sim \log N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ 。提示: 由定理 4.6 立得。

练习 4.29. 请验证随机变量 $X \sim \log N(\mu, \sigma^2)$ 的众数为 $e^{\mu-\sigma^2}$, 其他数字特征分别为

$$M(X) = e^\mu \quad E(X) = e^{\mu+\sigma^2/2} \quad V(X) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

定义 4.13 (偏正态分布). 由第 128 页的**练习 2.9** 的结果不难得到

$$\int_{-\infty}^{+\infty} \phi\left(\frac{x-\mu}{\sigma}\right) \Phi\left(\beta \frac{x-\mu}{\sigma}\right) dx = \frac{\sigma}{2}, \text{ 其中 } \sigma > 0$$

如果随机变量 X 的密度函数是

$$f_X(x) = \frac{2}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \Phi\left(\beta \frac{x-\mu}{\sigma}\right), \text{ 其中 } \sigma > 0$$

则称 X 服从偏正态分布 (skewed normal distribution), 记作 $X \sim SN(\mu, \sigma^2; \beta)$ 。例如, 第 189 页的**例 2.74** 所描述的分布就是 $SN(0, 1/\alpha^2; \eta/\alpha)$ 。

特别地, $SN(0, 1; \beta)$ 称为标准偏正态分布, 其密度函数为 $f(x) = 2\phi(x)\Phi(\beta x)$ 。显然, $\beta = 0$ 时, 偏正态分布 $SN(\mu, \sigma^2; \beta)$ 就是 $N(\mu, \sigma^2)$ 。当 $\beta > 0$ 时, $SN(\mu, \sigma^2; \beta)$ 称为左偏; 当 $\beta < 0$ 时称为右偏 (见下图)。

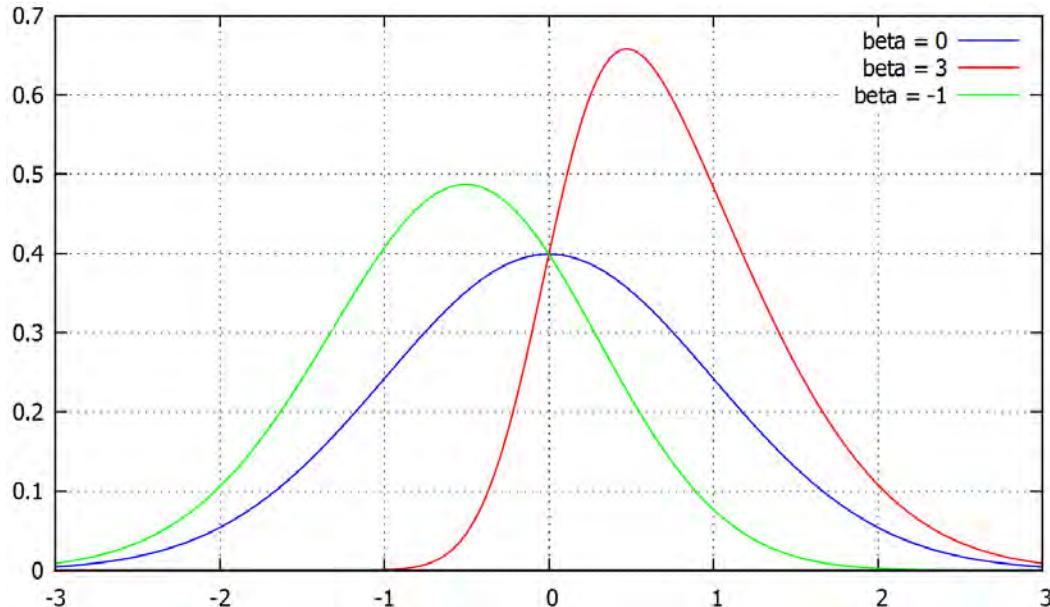


图 4.18: 标准偏正态分布 $SN(0, 1; \beta)$ 的密度函数曲线: 参数 β 的正负决定了密度函数往左偏还是往右偏。另外, $|\beta|$ 越大, 偏的程度就越大。

练习 4.30. 参考**例 2.74**, 求随机变量 $X \sim SN(\mu, \sigma^2; \beta)$ 的期望和方差。答案:

$$E(X) = \mu + \beta\sigma \sqrt{\frac{2}{\pi(1+\beta^2)}} \quad V(X) = \sigma^2 \left[1 - \frac{2\beta^2}{\pi(1+\beta^2)} \right]$$

性质 4.19. 已知 $X_1, X_2 \stackrel{\text{iid}}{\sim} N(0, 1)$ 和实数 β , 则随机变量 $Y = \begin{cases} -X_1 & \text{如果 } X_2 > \beta X_1 \\ X_1 & \text{否则} \end{cases}$ 服从标准偏正态分布 $SN(0, 1; \beta)$ 。

证明. 按照定义, Y 的分布函数是

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\{-X_1 \leq y, X_2 > \beta X_1\} + P\{X_1 < y, X_2 \leq \beta X_1\} \\ &= \int_{-y}^{+\infty} \int_{\beta x_1}^{+\infty} \phi(x_1)\phi(x_2)dx_2dx_1 + \int_{-\infty}^y \int_{-\infty}^{\beta x_1} \phi(x_1)\phi(x_2)dx_2dx_1 \\ &= \int_{-y}^{+\infty} \phi(x_1)[1 - \Phi(\beta x_1)]dx_1 + \int_{-\infty}^y \phi(x_1)\Phi(\beta x_1)dx_1 \end{aligned}$$

只需验证 $F'_Y(y)$ 是标准偏正态分布的密度函数即可。请读者补全证明。 □

算法 4.13. 标准偏正态分布 $Y \sim SN(0, 1; \beta)$ 的伪随机数 y^* 可以通过下面的方法产生:

- 独立产生标准正态分布 $N(0, 1)$ 的两个随机数, x_1^* 和 x_2^* 。
- 如果 $x_2^* > \beta x_1^*$, 则令 $y^* = -x_1^*$, 否则令 $y^* = x_1^*$ 。

4.2.4 Laplace 分布

1812 年, 法国数学家、天文学家、概率论的先驱 P. S. Laplace (1749-1827) 在《概率的分析理论》中提出了一个与正态分布类似的分布, 称作 Laplace 分布。

Laplace 分布和均匀分布、指数分布、 F 分布等都有联系, 适合描述材料的拉伸强度、断裂强度等, 在经济学、通信工程等领域也有应用。Laplace 分布有两个参数, 一般记作 $X \sim \text{Laplace}(\mu, \sigma)$, 它的密度函数是

$$f(x) = \frac{1}{2\sigma} \exp\left\{-\frac{|x - \mu|}{\sigma}\right\}, \text{ 其中 } \sigma > 0 \quad (4.9)$$

练习 4.31. 已知 $X \sim \text{Laplace}(\mu, \sigma)$ 且 $a \neq 0$, 则 $aX + b \sim \text{Laplace}(a\mu + b, |a|\sigma)$ 。提示: 利用例 2.17 的结果。

练习 4.32. 请读者验证 $X \sim \text{Laplace}(\mu, \sigma)$ 的分布函数是

$$F(x) = \begin{cases} \frac{1}{2} \exp\left(\frac{x - \mu}{\sigma}\right) & \text{如果 } x < \mu \\ 1 - \frac{1}{2} \exp\left(-\frac{x - \mu}{\sigma}\right) & \text{如果 } x \geq \mu \end{cases}$$

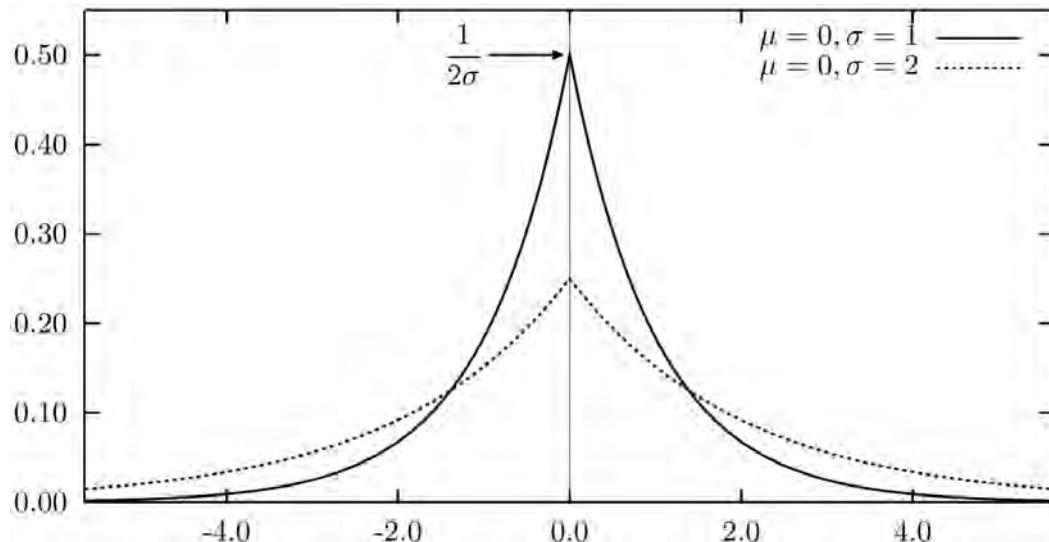


图 4.19: 参数 σ 越小, Laplace 分布的密度函数曲线显得越“高瘦”。Laplace 分布又称为双侧指数分布, 它与指数分布的关系见稍后的性质 4.29。

性质 4.20. 已知 $X, Y \stackrel{\text{iid}}{\sim} U(0, 1)$, 则 $Z = \ln(X/Y) \sim \text{Laplace}(0, 1)$ 。

证明. 仿照例 2.41, $F(z) = P\{\ln(X/Y) \leq z\} = P(Y \geq e^{-z}X)$ 。分别考虑 $z < 0$ 和 $z > 0$

两种情形，见下图。经过简单计算，得到

$$F(z) = \begin{cases} \frac{1}{2} \exp(z) & \text{如果 } z < 0 \\ 1 - \frac{1}{2} \exp(-z) & \text{如果 } z \geq 0 \end{cases}$$
□

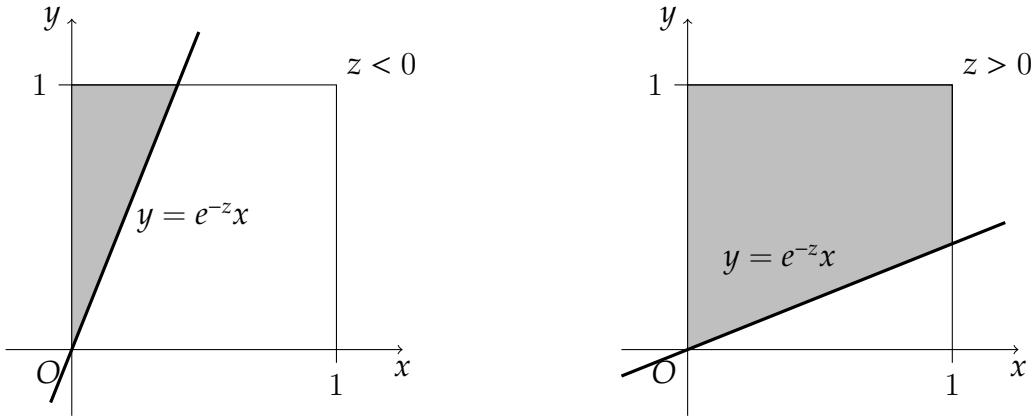


图 4.20: 性质 4.20 的直观证明：图中阴影部分的面积即是概率 $P(Y \geq e^{-z}X)$ 。

练习 4.33. 请读者验证 Laplace(μ, σ) 的特征函数、常见数字特征如下：

$$\varphi(t) = \frac{\exp(it\mu)}{1 + \sigma^2 t^2}, \text{ 并且 } E(X) = M(X) = \mu, V(X) = 2\sigma^2, c_s = 0, c_k = 3$$

算法 4.14. 利用逆 CDF 法，Laplace 分布 $X \sim \text{Laplace}(\mu, \sigma)$ 的随机数 x^* 可按如下方法产生：产生 $U(0, 1)$ 的随机数 u^* ；若 $u^* \geq 0.5$ ，则置 $x^* \leftarrow \mu - \sigma \ln(2 - 2u^*)$ ，否则置 $x^* \leftarrow \mu + \sigma \ln(2u^*)$ 。

练习 4.34. 请读者验证 Laplace 分布的任意阶矩存在且有限，Cauchy 分布的任意阶矩都不存在。然而 Laplace 分布 $\exp(-|x|)/2$ 与 Cauchy 分布 $1/[\pi(1+x^2)]$ 却由 §3.2.1 的反演公式牵线搭桥建立起“美妙的”关系。

$$\int_{-\infty}^{+\infty} e^{itx} \frac{\exp(-|x|)}{2} dx = \frac{1}{1+t^2}$$

并且其中 $\int_{-\infty}^{+\infty} \frac{e^{-itx}}{\pi(1+t^2)} dt = \exp(-|x|)$

提示：分布 $\text{Cauchy}(0, 1)$ 的特征函数为 $\exp(-|t|)$ 。

4.2.5 Cauchy 分布

以法国数学家 Baron Augustin-Louis Cauchy (1789-1857) 命名的 Cauchy 分布 $\text{Cauchy}(\mu, \lambda)$ 的密度函数及其图像见第 171 页的例 2.59。

Cauchy 分布没有期望、方差和任何阶的矩，是一类非常特殊的分布。我们把 $X \sim \text{Cauchy}(0, 1)$ 称为标准 Cauchy 分布，总有 $\lambda X + \mu \sim \text{Cauchy}(\mu, \lambda)$ 。

练习 4.35. 随机变量 $X \sim \text{Cauchy}(\mu, \lambda)$ 的特征函数为 $\varphi(t) = \exp\{it\mu - \lambda|t|\}$ ，其数字特征为 $M(X) = \mu$ ，分布函数为

$$F(x) = \frac{1}{\pi} \arctan \frac{x - \mu}{\lambda} + \frac{1}{2}$$

性质 4.21. 如果随机变量 $X, Y \stackrel{\text{iid}}{\sim} N(0, 1)$ ，则 $Z = X/Y \sim \text{Cauchy}(0, 1)$ 。

证明. 利用第 154 页的例 2.35 的结果，可得 $Z = X/Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} \phi(yz)\phi(y)|y|dy = \frac{1}{\pi(z^2 + 1)} \quad \square$$

练习 4.36. 如果 $X \sim N(0, \lambda^2)$ 和 $Y \sim N(0, 1)$ 相互独立，则 $\mu + X/Y \sim \text{Cauchy}(\mu, \lambda)$ 。

练习 4.37. 已知随机变量 $\theta \sim U(-\pi/2, \pi/2)$ ，试证明 $X = \tan \theta \sim \text{Cauchy}(0, 1)$ ，并且 $Y = \mu + \lambda \tan \theta \sim \text{Cauchy}(\mu, \lambda)$ 。

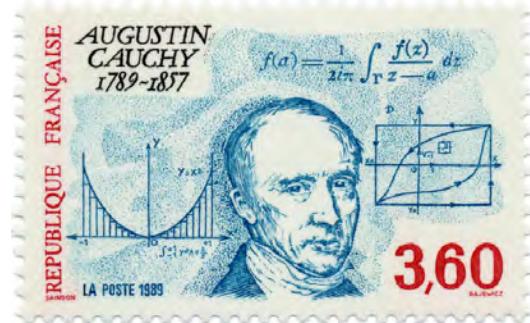
练习 4.38. 利用练习 4.36 或练习 4.37 给出分布 $\text{Cauchy}(\mu, \lambda)$ 的随机数产生算法。

性质 4.22 (和型不变性). 若随机变量 $X_j \sim \text{Cauchy}(\mu_j, \lambda_j), j = 1, 2, \dots, n$ 相互独立，则

$$\sum_{j=1}^n c_j X_j \sim \text{Cauchy}\left(\sum_{j=1}^n c_j \mu_j, \sum_{j=1}^n |c_j| \lambda_j\right)$$

其中 c_1, c_2, \dots, c_n 为实数。特别地，若 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Cauchy}(\mu, \lambda)$ ，则

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) \sim \text{Cauchy}(\mu, \lambda)$$



4.2.6 Gamma 分布及其特例 (χ^2 分布和指数分布)

1729-1731 年, 数学分析大师 L. Euler (1707-1783) 研究了两个由积分定义的高等超越函数——Beta 函数和 Gamma 函数*, 这两类特殊函数在数学物理里有广泛的应用。此外, 由它们可以定义两类非常重要的分布——Beta 分布和 Gamma 分布。

考虑到 Beta 函数可由 Gamma 函数定义, 见式 (4.20), 我们首先介绍 Gamma 函数和 Gamma 分布, 而把 Beta 函数和 Beta 分布的内容放到下一小节。另外, Gamma 分布的重要性还体现在由它可以诱导出逆 Gamma 分布、 χ^2 分布和指数分布, 它们在应用中也都是很常见的。

定义 4.14 (第二类 Euler 积分). 1730 年 1 月 8 日, Euler 在给德国数学家 Christian Goldbach (1690-1764) 的信中提到了他近期一直关注的特殊函数 $\Gamma(x)$, 其中 $\forall x \in \mathbb{R}$ 且 $x \neq 0, -1, -2, \dots$ 。该函数有着重要的性质, 具体定义如下。



$$\Gamma(x) = \int_0^1 (-\ln t)^{x-1} dt \stackrel{u=-\ln t}{=} \int_0^{+\infty} u^{x-1} e^{-u} du \quad (4.10)$$

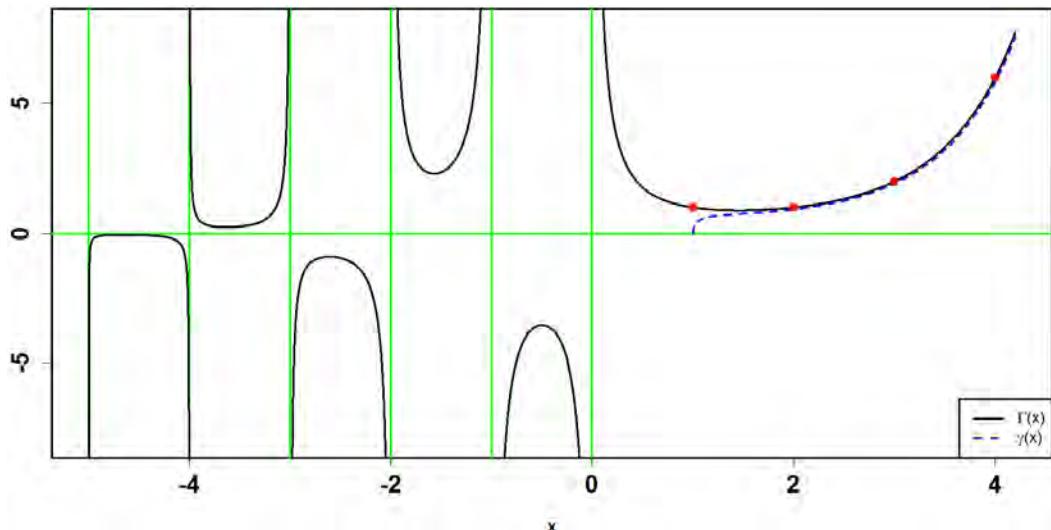


图 4.21: Gamma 函数 $\Gamma(x)$ 在点 $x = 0, -1, -2, \dots$ 处不连续, 曲线穿过点 $(n+1, n!), n = 0, 1, 2, \dots$ 。初等函数 $\gamma(x)$ 是对 $\Gamma(x)$ 的近似, 具体定义见 (4.12)。

*1809 年, A. M. Legendre 将第二类 Euler 积分所定义的函数定名为 Gamma 函数并采用记号 $\Gamma(x)$ 。Gamma 函数与神秘的 Riemann ζ 函数 (1.37) 有着美丽的联系 [71], 例如,

$$\zeta(x)\Gamma(x) = \int_0^{+\infty} \frac{u^{x-1}}{e^u - 1} du, \text{ 其中 } x \notin \{1, 0, -1, -2, \dots\}$$

为何要考虑式 (4.10) 之类的积分呢？不难证明下面的递归关系，

$$\Gamma(x+1) = x\Gamma(x), \text{ 且 } \Gamma(1) = 1 \quad (4.11)$$

Euler 揭示了超越函数 $\Gamma(x)$ 是对阶乘的推广。事实上，当 $n \in \mathbb{N}$ 时，

$$n! = \int_0^1 (-\ln t)^n dt = \Gamma(n+1)$$



图 4.22: 瑞士法郎上的 Euler 头像和 Γ 函数。Euler (1707-1783) 是伟大的数学家、物理学家和天文学家，一生中多数工作时间在俄国皇家科学院（圣彼得堡）和柏林科学院度过。他是史上最多产的数学家之一，平均每年发表的论文达八百页之多。1783 年 9 月 18 日，Euler 停止了计算和生命。他留下的著作已整理出版了七十多卷（每卷五百多页），整理工作仍在继续。Laplace 曾说，Lisez Euler, lisez Euler, c'est notre maître à tous (读 Euler 吧，读 Euler 吧，他是我们所有人的导师)。

另外，由 Stirling 公式 (1.5)，不难得到

$$\Gamma(x) \sim \gamma(x) = \sqrt{2\pi(x-1)} \left(\frac{x-1}{e} \right)^{x-1}, \text{ 其中 } x \geq 2 \quad (4.12)$$

$$\lim_{x \rightarrow +\infty} \frac{\Gamma(x)}{\Gamma(x+\alpha)} x^\alpha = 1, \forall \alpha \in \mathbb{R} \quad (4.13)$$

性质 4.23 (反射公式). Euler 发现了下面的反射公式 (reflection formula):

$$\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin(\pi x)}, \forall x \notin \mathbb{Z} \quad (4.14)$$

※证明. 详情见 [71] 的第 59 页，用到了下面两个事实。

$$\begin{aligned} \frac{1}{\Gamma(x)\Gamma(1-x)} &= x \prod_{r=1}^{\infty} \left(1 - \frac{x^2}{r^2} \right) \\ \sin(\pi x) &= \pi x \prod_{r=1}^{\infty} \left(1 - \frac{x^2}{r^2} \right) \end{aligned}$$

练习 4.39. 请读者验证 $\Gamma(1/2) = \sqrt{\pi}$ 。更一般地, $\forall n \in \mathbb{N}$,

$$\begin{aligned}\Gamma\left(\frac{1}{2} + n\right) &= \frac{(2n)!}{4^n n!} \sqrt{\pi} \\ \Gamma\left(\frac{1}{2} - n\right) &= \frac{(-4)^n n!}{(2n)!} \sqrt{\pi}\end{aligned}$$

提示: 利用反射公式 (4.14) 和递归公式 (4.11)。

练习 4.40. 请读者验证 Gamma 函数的如下性质,

$$\frac{\Gamma(\alpha)}{\beta^\alpha} = \int_0^{+\infty} x^{\alpha-1} e^{-\beta x} dx, \text{ 其中 } \alpha > 0, \beta > 0$$

定义 4.15 (Gamma 分布). 源于 Gamma 函数的定义式 (4.10), 如果连续型随机变量 X 的密度函数如下所示, 则称 X 服从参数为 (α, β) 的 Gamma 分布, 记作 $X \sim \text{Gamma}(\alpha, \beta)$ 。

$$g_{\alpha,\beta}(x) = \begin{cases} 0 & \text{当 } x \leq 0 \\ \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{当 } x > 0, \alpha > 0, \beta > 0 \end{cases} \quad (4.15)$$

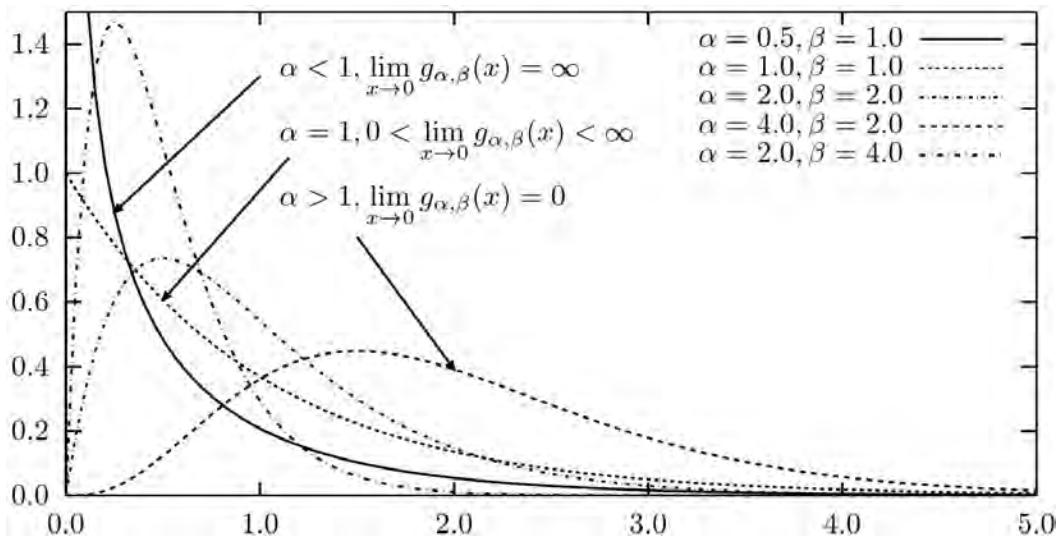


图 4.23: 分布 $\text{Gamma}(\alpha, \beta)$ 的参数 α 称为形状参数, 它决定了密度曲线 $g_{\alpha,\beta}(x)$ 的形状; β 称为尺度参数, 当 α 固定时, β 越大曲线在 0 附近越“高瘦”。

定义 4.16 (逆 Gamma 分布). 已知 $X \sim \text{Gamma}(\alpha, \beta)$, 随机变量 $Y = 1/X$ 的分布称

为参数为 (α, β) 的逆 Gamma 分布, 记作 $Y \sim \text{Inv-Gamma}(\alpha, \beta)$, 其密度函数为

$$f_{\alpha, \beta}(y) = \begin{cases} 0 & \text{当 } y \leq 0 \\ \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\beta/y} & \text{当 } y > 0, \alpha > 0, \beta > 0 \end{cases}$$

性质 4.24. $X \sim \text{Gamma}(\alpha, \beta)$ 的特征函数为 $\varphi(t) = (1 - it/\beta)^{-\alpha}$, 其 k 阶矩为

$$m_k = \frac{\Gamma(\alpha + k)}{\beta^k \Gamma(\alpha)} = \frac{1}{\beta^k} \alpha(\alpha + 1)(\alpha + 2) \cdots (\alpha + k - 1)$$

请读者验证 $X \sim \text{Gamma}(\alpha, \beta)$ 的其他数字特征为

$$\begin{aligned} E(X) &= \frac{\alpha}{\beta} & V(X) &= \frac{\alpha}{\beta^2} \\ c_s &= \frac{2}{\sqrt{\alpha}} & c_k &= \frac{6}{\alpha} \end{aligned}$$

逆 Gamma 分布 $Y \sim \text{Inv-Gamma}(\alpha, \beta)$ 的期望和方差分别为

$$E(Y) = \frac{\beta}{\alpha - 1} \quad V(Y) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)}$$

练习 4.41. 若随机变量 $Y \sim \text{Gamma}(\alpha, 1)$, 则 $X = Y/\beta \sim \text{Gamma}(\alpha, \beta)$ 。

性质 4.25. 若随机变量 $X_j \sim \text{Gamma}(\alpha_j, \beta), j = 1, 2, \dots, n$ 相互独立, 则

$$\sum_{j=1}^n X_j \sim \text{Gamma}\left(\sum_{j=1}^n \alpha_j, \beta\right)$$

算法 4.15. 因为**练习 4.41** 的结果, 要产生 $X \sim \text{Gamma}(\alpha, \beta)$ 的随机数 x^* 只需要考虑产生 $Y \sim \text{Gamma}(\alpha, 1)$ 的随机数 y^* , 然后令 $x^* = y^*/\beta$ 即可。

□ 若 α 是自然数, 由**性质 4.25** 知 $\text{Gamma}(\alpha, 1)$ 分布的随机数 y^* 就是 α 个 $\text{Gamma}(1, 1)$ 分布的随机数 x_1, \dots, x_α 之和, 即 $y^* = x_1 + \dots + x_\alpha$ 。分布 $\text{Gamma}(1, 1)$ 的随机数产生算法见第 298 页的**算法 4.17**。

□ 若 α 不是自然数, 令 $m = \lfloor \alpha \rfloor$ 是不超过 α 的最大整数, 令 $p = \alpha - \lfloor \alpha \rfloor$, 则 $\text{Gamma}(\alpha, 1)$ 的随机数 y^* 就是 m 个 $\text{Gamma}(1, 1)$ 分布的随机数与一个 $\text{Gamma}(p, 1)$ 分布的随机数之和。第 717 页的例 15.5 给出了产生 $\text{Gamma}(\alpha, 1)$ 随机数的另外一种方法。

定义 4.17 (χ^2 分布、指数分布与 Erlang 分布). 把 Gamma 分布稍作限制, 便可得到下面三个常见的分布, 请读者写出它们的特征函数。

□ 分布 $X \sim \text{Gamma}(\eta/2, 1/2)$ 特称为自由度为 η 的 χ^2 分布, 它是德国的大地测量学家 Friedrich Robert Helmert (1843-1917) 在研究正态总体的样本方差时发现的, 记作 $X \sim \chi_{\eta}^2$, 其密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } x \leq 0 \\ \frac{x^{\eta/2-1} e^{-x/2}}{2^{\eta/2} \Gamma(\eta/2)} & \text{当 } x > 0, \eta > 0 \end{cases} \quad (4.16)$$

由性质 4.24 知, $E(X) = \eta, V(X) = 2\eta$ 。

□ 分布 $X \sim \text{Gamma}(1, \beta)$ 特称为参数为 β 的指数分布 (exponential distribution), 记作 $X \sim \text{Expon}(\beta)$, 其分布函数为 $\max(1 - \exp\{-\beta x\}, 0)$, 密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } x \leq 0 \\ \beta e^{-\beta x} & \text{当 } x > 0, \beta > 0 \end{cases} \quad (4.17)$$

由性质 4.24 知, $E(X) = 1/\beta, V(X) = 1/\beta^2$ 。

□ 令 n 为自然数, β 为正实数, 在排队论中, Gamma 分布的特款 $\text{Gamma}(n, \beta)$ 被称为 Erlang 分布, 以丹麦数学家 Agner Krarup Erlang (1878-1929) 命名, 记作 $\text{Erlang}(n, \beta)$ 。显然, 指数分布也是 Erlang 分布的特款。后续学习中我们将知道 Erlang 分布的研究背景是 Poisson 过程 (详见 §6.2.1)。

练习 4.42. 若 $Y \sim \text{Erlang}(n, 1)$, 则 $2Y \sim \chi_{2n}^2$ 。提示: 利用练习 4.41 的结果。

练习 4.43. 如果 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(\beta)$, 则 $X_1 + \dots + X_n \sim \text{Erlang}(n, \beta)$ 。

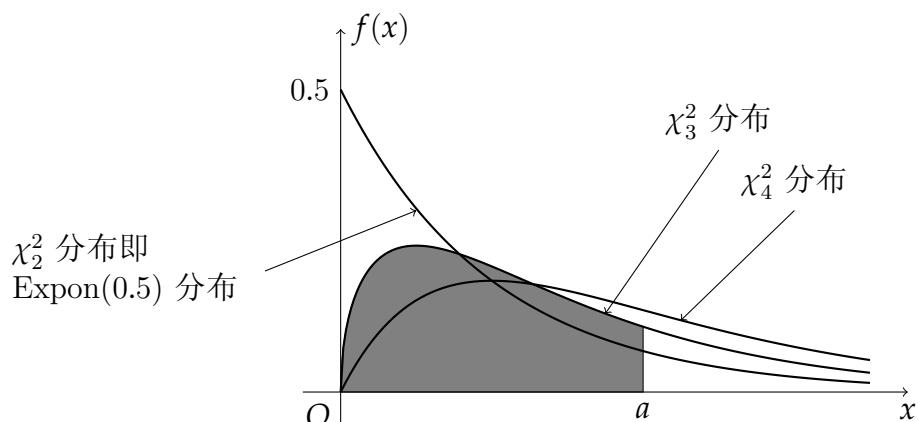


图 4.24: 通过绘制 χ_{η}^2 分布的密度函数曲线不难发现, $\eta > 0$ 越小, $X \sim \chi_{\eta}^2$ 分布的概率就越集中在 $(0, a]$ 内, 其中 $a > 0$ 。

定义 4.18 (逆 χ^2 分布). 分布 $Y \sim \text{Inv-Gamma}(\eta/2, 1/2)$ 特称为自由度为 η 的 逆 χ^2 分布, 记作 $Y \sim \chi_{\eta}^{-2}$, 其密度函数为

$$f(y) = \begin{cases} 0 & \text{当 } y \leq 0 \\ \frac{y^{-(\eta/2+1)} e^{-1/(2y)}}{2^{\eta/2} \Gamma(\eta/2)} & \text{当 } y > 0, \eta > 0 \end{cases} \quad (4.18)$$

性质 4.26. 若 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1)$, 则 $X = X_1^2 + X_2^2 + \dots + X_n^2 \sim \chi_n^2$ 。

证明. 随机变量 $Y = X_1^2$ 的密度函数见第 131 页的例 2.18 的结果。由 χ^2 分布的定义 4.17 知 $X_1^2, X_2^2, \dots, X_n^2 \stackrel{\text{iid}}{\sim} \chi_1^2$, 进而利用性质 4.25 可证得。或者, 直接利用第 162 页的例 2.45 的结果亦可证得。□

练习 4.44. 计算随机变量 $X \sim \chi_n^2$ 的特征函数, 其中 $n \in \mathbb{N}$ 。提示: 由习题 3.5 知 χ_1^2 分布的特征函数是 $(1 - 2it)^{-1/2}$, 进而由性质 4.26 得到 χ_n^2 分布的特征函数是 $(1 - 2it)^{-n/2}$ 。

算法 4.16. 若 n 为自然数, 根据性质 4.26, χ_n^2 分布的随机数 x^* 可由 $N(0, 1)$ 分布的 n 个随机数 x_1, x_2, \dots, x_n 通过 $x^* = \sum_{j=1}^n x_j^2$ 得到, 也可以由 $\text{Gamma}(n/2, 1)$ 的随机数 y^* 通过 $x^* = 2y^*$ 得到。

当 $n \rightarrow \infty$ 时, 随机变量 χ_n^2 , $\sqrt{2\chi_n^2}$ 和 $\sqrt[3]{\chi_n^2/n}$ 都趋向正态分布。 χ_n^2 分布因 χ^2 统计量 (见 §7.2.1) 和拟合优度的 Pearson χ^2 检验 (详见第 575 页的引理 9.2) 而成为常见分布。有关 χ^2 分布的近似计算见定理 5.24。

算法 4.17. 根据练习 4.41, 只需要考虑标准指数分布 $X \sim \text{Expon}(1)$ 或 $\text{Gamma}(1, 1)$ 的伪随机数产生器便可得到 $Y = X/\beta \sim \text{Expon}(\beta)$ 的随机数。利用逆 CDF 法产生 $X \sim \text{Expon}(1)$ 的随机数 $x^* = -\ln u^*$, 其中 u^* 是 $U(0, 1)$ 的随机数。

指数分布常用于描述机械或电子元器件的使用寿命、系统的稳定时间、随机事件发生的时间间隔等, 它与几何分布类似, 同样具有“无记忆性”(见定理 4.9), 形容为“一个能用如初的旧灯泡, 其寿命分布和新灯泡一样”。性质 4.27 是指数分布无记忆性的根本原因, 这个特殊性质致使指数分布被用于定义 Poisson 过程 (见 §6.2.1)。另外, 指数分布还被用来判断一个分布是否是重尾的 (见定义 4.24)。由此可见, 指数分布是一个非常重要的分布。

性质 4.27. 对于指数分布 $X \sim \text{Expon}(\beta)$, 有 $P(X > t) = \exp\{-\beta t\}$ 。

性质 4.28 (无记忆性). 只取正值的连续型随机变量 X 服从指数分布当且仅当

$$P(X > s + t | X > s) = P(X > t), \text{ 其中 } s \geq 0, t \geq 0$$

※证明. 首先往证 “ \Rightarrow ”: 已知 $X \sim \text{Expon}(\beta)$, 于是

$$\begin{aligned} P(X > s + t | X > s) &= \frac{P(X > s + t, X > s)}{P(X > s)} \\ &= \frac{P(X > s + t)}{P(X > s)}, \text{ 利用性质 4.27} \\ &= \frac{\exp\{-\beta(s + t)\}}{\exp\{-\beta s\}} \\ &= P(X > t) \end{aligned}$$

下面往证 “ \Leftarrow ” (请读者将证明补全): 令 $g(x) = P(X > x)$, 其中 $x \geq 0$, 则 $g(x)$ 是减函数。由已知条件得到 $g(s + t) = g(s)g(t)$, 对于任意的有理数 m/n , 往证

$$g\left(\frac{m}{n}\right) = \left[g\left(\frac{1}{n}\right)\right]^m = [g(1)]^{m/n}$$

再说明对于实数 $x \geq 0$ 有 $g(x) = [g(1)]^x$ 。令 $\beta = -\ln g(1)$, 则有 $P(X \leq x) = 1 - \exp\{-\beta x\}$ 。 \square

性质 4.28. 若随机变量 $X \sim \text{Expon}(\alpha)$ 与 $Y \sim \text{Expon}(\beta)$ 相互独立, 则

$$\begin{aligned} \min(X, Y) &\sim \text{Expon}(\alpha + \beta) \\ P(X < Y) &= \frac{\alpha}{\alpha + \beta} \end{aligned}$$

即, 分别对 X, Y 进行抽样, 所得样本满足 $X < Y$ 的概率是 $\alpha/(\alpha + \beta)$ 。

证明. 根据独立假设和指数分布的性质 4.27, 我们有

$$\begin{aligned} P(\min(X, Y) > t) &= P(X > t, Y > t) = P(X > t)P(Y > t) = \exp\{-(\alpha + \beta)t\} \\ P(X < Y) &= \int_0^{+\infty} f_X(x)P(Y > x)dx = \int_0^{+\infty} \alpha e^{-\alpha x} e^{-\beta x} dx = \frac{\alpha}{\alpha + \beta} \quad \square \end{aligned}$$

性质 4.29. 已知随机变量 $X, Y \stackrel{\text{iid}}{\sim} \text{Expon}(\beta)$, 则对于任意实数 μ 皆有

$$\mu + X - Y \sim \text{Laplace}(\mu, 1/\beta)$$

证明. 先往证 $Z = X - Y \sim \text{Laplace}(0, 1/\beta)$: 由第 154 页的例 2.35 的结果, 不难得到

Z 的密度函数为

$$f_Z(z) = \begin{cases} \int_0^{+\infty} \beta e^{-\beta x} \beta e^{-\beta(x-z)} dx = \frac{\beta}{2} \exp(\beta z) & \text{当 } z < 0 \\ \int_z^{+\infty} \beta e^{-\beta x} \beta e^{-\beta(x-z)} dx = \frac{\beta}{2} \exp(-\beta z) & \text{当 } z \geq 0 \end{cases}$$

$$= \frac{\beta}{2} \exp(-\beta|z|)$$

再由第 131 页的例 2.17 知 $\mu + Z$ 的密度函数为 $f_Z(z - \mu)$, 得证。 \square

4.2.7 Beta 分布

Beta 分布在第 12 章所介绍的贝叶斯统计学中常作为二项分布和几何分布的共轭先验分布（见第 655 页的例 12.17），该分布由 Beta 函数定义。

定义 4.19 (第一类 Euler 积分). $\forall \alpha > 0, \beta > 0$, 定义特殊函数 $B(\alpha, \beta)$ 如下。

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx \quad (4.19)$$

练习 4.45. 由定义 (4.19), 试证明:



性质 4.30. 第一类 Euler 积分 (4.19) 可由 Gamma 函数表示为

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (4.20)$$

证明. 模仿结果 (4.14) 的证明, 下面往证 $\Gamma(\alpha)\Gamma(\beta) = \Gamma(\alpha + \beta)B(\alpha, \beta)$ 。

$$\begin{aligned} \Gamma(\alpha)\Gamma(\beta) &= \int_0^{+\infty} x^{\alpha-1} e^{-x} dx \int_0^{+\infty} y^{\beta-1} e^{-y} dy \\ &= \int_0^{+\infty} \int_0^{+\infty} x^{\alpha-1} y^{\beta-1} e^{-x-y} dx dy \\ &\stackrel{x=uv}{=} \int_0^{+\infty} \left\{ \int_0^1 (uv)^{\alpha-1} [u(1-v)]^{\beta-1} e^{-u} u dv \right\} du \\ &= \int_0^{+\infty} u^{\alpha+\beta-1} e^{-u} du \int_0^1 v^{\alpha-1} (1-v)^{\beta-1} dv \\ &= \Gamma(\alpha + \beta)B(\alpha, \beta) \end{aligned}$$

□

练习 4.46. 验证 $B(\alpha, \beta) = B(\beta, \alpha)$ 。根据结果 (4.20) 证明 $B(\alpha, \beta)$ 满足下面的递归关系。

$$\begin{aligned} B(\alpha, \beta) &= B(\alpha, \beta + 1) + B(\alpha + 1, \beta) \\ B(\alpha + 1, \beta) &= \frac{\alpha}{\alpha + \beta} B(\alpha, \beta), \quad B(1, 1) = 1 \end{aligned}$$

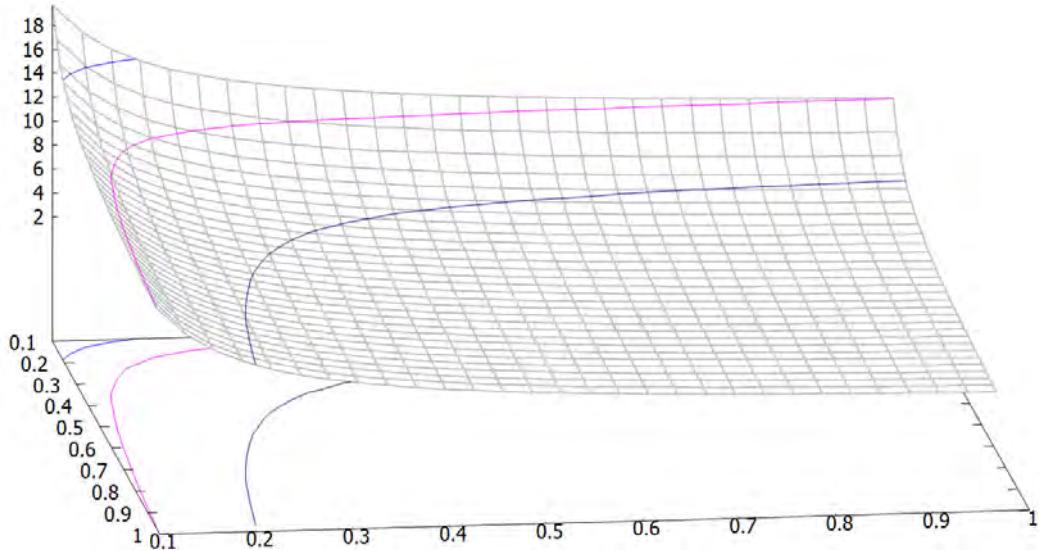


图 4.25: 函数 $B(\alpha, \beta)$ 的曲面: 对于固定的 β , 总有 $\lim_{\alpha \rightarrow 0} B(\alpha, \beta) = +\infty$, $\lim_{\alpha \rightarrow \infty} B(\alpha, \beta) = 0$.

定义 4.20 (Beta 分布). 如下定义的函数 $b_{\alpha, \beta}(x)$ 为某一连续型随机变量 X 的密度函数, 称 X 服从参数为 (α, β) 的 Beta 分布, 记作 $X \sim \text{Beta}(\alpha, \beta)$ 。

$$b_{\alpha, \beta}(x) = \begin{cases} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} & \text{当 } 0 < x < 1, \alpha > 0, \beta > 0 \\ 0 & \text{当 } x \leq 0 \text{ 或 } x \geq 1 \end{cases} \quad (4.21)$$

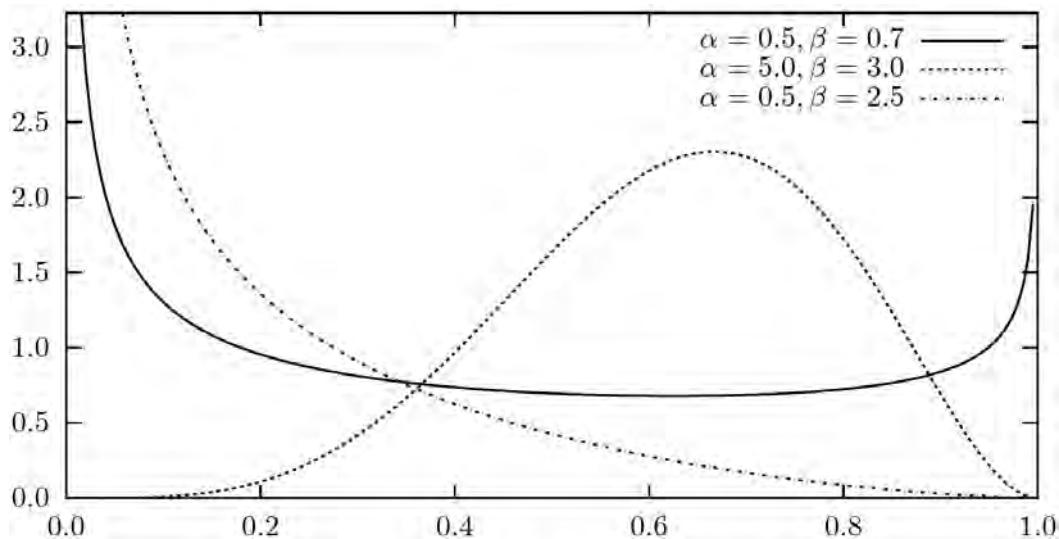


图 4.26: 当 $\alpha < 1, \beta < 1$ 时, 密度曲线呈现 U 型。当 α, β 中只有一个小于 1 时, 则密度曲线只有一头翘向无穷。极端的情形是: $\text{Beta}(1, 1) = \text{U}(0, 1)$ 。

性质 4.31. 如果随机变量 $X \sim \text{Beta}(\alpha, \beta)$, 则 $1 - X \sim \text{Beta}(\beta, \alpha)$ 。

证明. 验证 $1 - X$ 的密度函数仍为式 (4.21) 即可。 \square

练习 4.47. 请读者验证 Beta 分布 $X \sim \text{Beta}(\alpha, \beta)$ 的期望和方差分别为

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta} \quad \mathbb{V}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

提示: X 的 k 阶矩 m_k 为

$$\begin{aligned} m_k &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1} (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} B(\alpha + k, \beta) \\ &= \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(\alpha + \beta + k)} \end{aligned}$$

性质 4.32. 若随机变量 $X \sim \text{Gamma}(\alpha, 1)$ 与 $Y \sim \text{Gamma}(\beta, 1)$ 相互独立, 则随机变量

$$Z = \frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)$$

证明. 随机向量 $(X, Y)^\top$ 的密度函数 $f(x, y)$ 为

$$f(x, y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} y^{\beta-1} \exp\{-x-y\}$$

下面求随机向量 $(X, Z)^\top$ 的密度函数: 由第 154 页的定理 2.15 知随机向量 $(X, Z)^\top$ 的密度函数 $g(x, z)$ 为

$$g(x, z) = xz^{-2} f(x, x/z - x) = \frac{x^{\alpha+\beta-1} \exp\{-x/z\}}{\Gamma(\alpha)\Gamma(\beta)} z^{-1-\beta} (1-z)^{\beta-1}$$

于是, 随机变量 Z 的密度函数为

$$\begin{aligned} f_Z(z) &= \int_0^{+\infty} g(x, z) dx \\ &= z^{-1-\beta} (1-z)^{\beta-1} \int_0^{+\infty} \frac{x^{\alpha+\beta-1} \exp\{-x/z\}}{\Gamma(\alpha)\Gamma(\beta)} dx, \text{ 利用练习 4.40} \\ &= z^{-1-\beta} (1-z)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha+\beta} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1} (1-z)^{\beta-1} \end{aligned} \quad \square$$

练习 4.48. 若随机变量 $X \sim \chi_m^2$ 与 $Y \sim \chi_n^2$ 相互独立, 则随机变量

$$Z = \frac{X}{X+Y} \sim \text{Beta}(m/2, n/2)$$

提示: 仿照性质 4.32 的证明。

练习 4.49. 请读者利用性质 4.32 和练习 4.48 设计产生 $\text{Beta}(\alpha, \beta)$ 随机数的算法。

例 4.25. 试证明: 若 $X \sim U(0, 1]$, 则 $X^2 \sim \text{Beta}(1/2, 1)$ 。

证明. 利用例 2.17 的结果, 得到 $Y = X^2$ 上的密度函数是 $\frac{1}{2}y^{-1/2}$, 恰是 $\text{Beta}(1/2, 1)$ 的密度函数。 \square

4.2.8 t 分布和 F 分布

1908 年, 英国统计学家兼化学家 William Sealy Gosset (1876-1937) 以笔名 Student 在《生物统计》(Biometrika) 学报上发表重要论文《均值的或然误差》[61], 提出了 t 分布 (亦称学生 t 分布)。

Gosset 将 t 分布用于估计小样本时正态总体的均值, 开创了小样本分析的先河 (见第 7 章、第 8 章)。 t 分布是统计学中最常用的分布之一。

定义 4.21 (t 分布). 如果随机变量 $X \sim N(0, 1)$ 与 $Y \sim \chi_n^2$ 相互独立, 则随机变量 $T = X / \sqrt{Y/n}$ 的分布称为自由度为 n 的 t 分布, 记作 $T \sim t_n$ 。在不引起歧义的情况下, t_n 分布的定义也简记作

$$t_n = \frac{N(0, 1)}{\sqrt{\chi_n^2/n}}$$



由习题 2.32, 随机变量 $T \sim t_n$ 的密度函数为

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \sqrt{n\pi}} \left(\frac{t^2}{n} + 1\right)^{-\frac{n+1}{2}}$$

显然, t_1 就是 Cauchy($0, 1$), 见性质 4.21 和式 (2.46)。

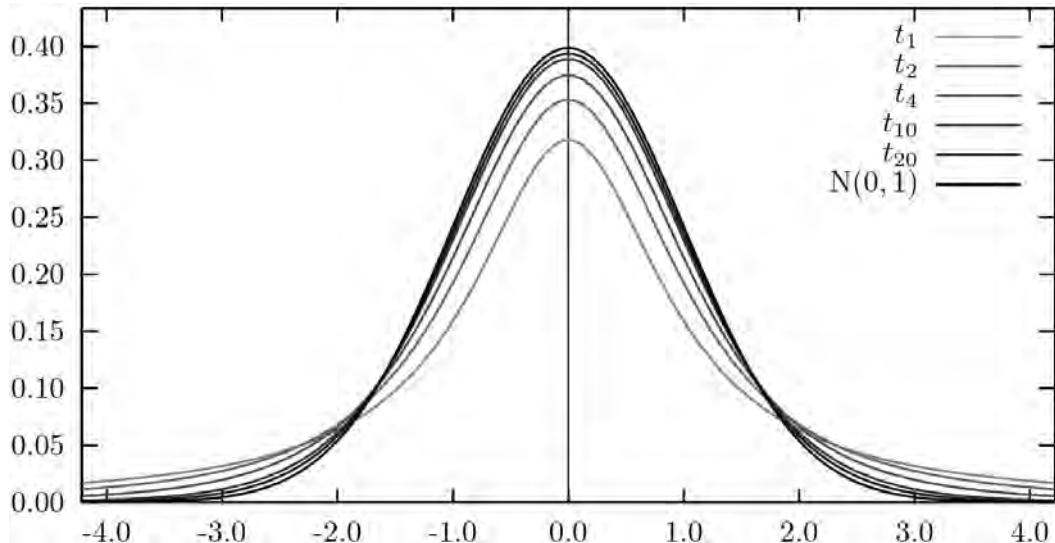


图 4.27: 随着 n 的增大, t_n 分布的密度函数越来越显得“高瘦”。当 n 趋向无穷时, t_n 分布的极限就是标准正态分布, 见例 5.10。

练习 4.50. t 分布 $X \sim t_n$ 的数字特征分别为

$$\begin{aligned} E(X) &= M(X) = 0 \\ V(X) &= \frac{n}{n-2}, \text{ 其中 } n > 2 \end{aligned}$$

另外, t_n 分布的 k 阶矩仅当 $k < n$ 时存在, 且奇数阶矩都为 0。

在统计学中, F 分布也是最常用的分布之一。Fisher 在方差分析方面的研究工作与 F 分布有关, 在此基础上美国统计学家 G. W. Snedecor (1881-1974) 于 1934 年定义了 F 分布。为纪念 Fisher, 人们以 Fisher 的首字母命名了该分布。

定义 4.22 (F 分布). 如果随机变量 $X \sim \chi_m^2$ 与 $Y \sim \chi_n^2$ 相互独立, 则随机变量 $Z = \frac{X/m}{Y/n}$ 的分布称为自由度为 (m, n) 的 F 分布, 亦称 Fisher-Snedecor 分布, 记作 $Z \sim F_{m,n}$ 。为方便记忆, $F_{m,n}$ 分布的定义也简记作

$$F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n}$$

令参数 $m, n > 0$, 随机变量 $Z \sim F_{m,n}$ 的密度函数为

$$f(z) = \begin{cases} C_{m,n} \frac{z^{\frac{m}{2}-1}}{(mz+n)^{\frac{m+n}{2}}} & \text{当 } z > 0, \text{ 其中 } C_{m,n} = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \\ 0 & \text{当 } z \leq 0 \end{cases}$$

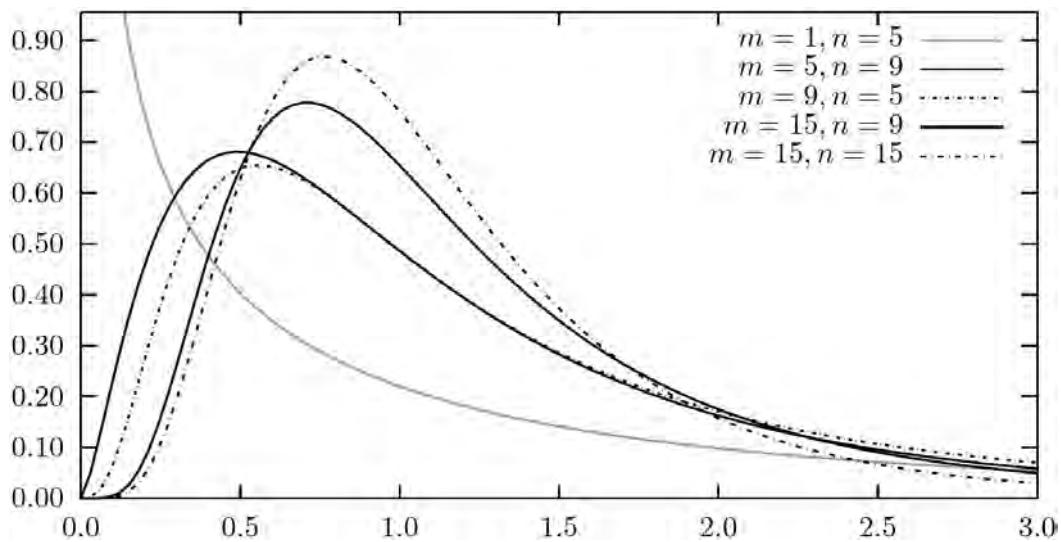


图 4.28: 不同自由度的 $F_{m,n}$ 分布的密度函数: $m = 1$ 与 $m \geq 2$ 时的形状不同。

性质 4.33. 若随机变量 $T \sim t_n$, 则 $T^2 \sim F_{1,n}$ 。

证明. 不妨设 $T = X/\sqrt{Y/n}$, 其中 $X \sim N(0, 1)$ 与 $Y \sim \chi_n^2$ 相互独立, 进而 $X^2 \sim \chi_1^2$ 与 Y 相互独立, 由**定义 4.22** 得到 $T^2 = X^2/(Y/n) \sim F_{1,n}$ 。 \square

练习 4.51. 已知随机变量 $F \sim F_{m,n}$, 则 $1/F \sim F_{n,m}$ 并且

$$\lim_{n \rightarrow \infty} mF \sim \chi_m^2, \text{ 还有 } \lim_{m \rightarrow \infty} \frac{n}{F} \sim \chi_n^2$$

练习 4.52. 随机变量 $X \sim F_{m,n}$ 的期望和方差分别为

$$\begin{aligned} E(X) &= \frac{n}{n-2}, \text{ 其中 } n > 2 \\ V(X) &= \frac{2n^2(n+m-2)}{m(n-4)(n-2)^2}, \text{ 其中 } n > 4 \end{aligned}$$

练习 4.53. 试证明: 如果随机变量 $F \sim F_{m,n}$, 则 $\frac{mF}{mF+n} \sim \text{Beta}(m/2, n/2)$ 。反之, 如果 $X \sim \text{Beta}(m/2, n/2)$, 则 $\frac{n}{m(1-X)} \sim F_{m,n}$ 。提示: 利用**练习 4.48** 的结果。

4.2.9 Pareto 分布

意大利经济学家 Vilfredo Pareto (1848-1923) 发现了这样一个残酷的事实：20% 的人口控制着 80% 的财富，而在这 20% 的富人及其财富当中，最富裕的 20% 又拥有着 80% 的财富，以此类推……。这一事实被称为 Pareto 法则或“80-20”法则。

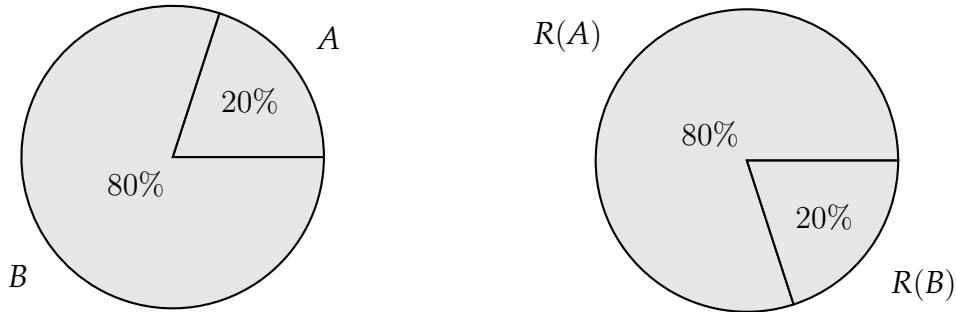


图 4.29: 占总体 20% 的人（时间、资源、努力等）获得 80% 的财富（效果、回馈等）。

为描述财富的分配，Pareto 提出了 Pareto 分布，常用于刻画经济学和社会学领域中大部分资源、财富掌握在小部分人手中的现象。

定义 4.23 (Pareto 分布). 令 $\alpha > 0, \mu > 0$, Pareto 分布 $X \sim \text{Pareto}(\alpha, \mu)$ 的分布函数定义为

$$F(x) = \begin{cases} 1 - (\mu/x)^\alpha & \text{当 } x \geq \mu \\ 0 & \text{当 } x < \mu \end{cases}$$



定义 4.24. 如果分布函数 $F(x)$ 的尾函数 $\bar{F}(x) = 1 - F(x)$ 比任意指数分布 $X \sim \text{Expon}(\lambda)$ 的尾函数 $\bar{F}_X(x) = 1 - F_X(x)$ 要厚重，即 $x \rightarrow \infty$ 时， $\bar{F}(x)$ 趋于零的速度比 $\bar{F}_X(x)$ 要慢许多，则称之为重尾分布 (heavy-tailed distribution)。即

$$\lim_{x \rightarrow \infty} \frac{\bar{F}(x)}{\bar{F}_X(x)} = \infty, \text{ 其中 } \bar{F}_X(x) \text{ 是 } X \sim \text{Expon}(\lambda) \text{ 的尾函数}$$

$$\text{等价地, } \lim_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = \infty, \text{ 其中 } \forall \lambda > 0$$

练习 4.54. 试证明：Pareto 分布是重尾分布。

在金融、保险、电信、网络、环境等领域，重尾分布常用来刻画真实数据，而不是用正态分布。譬如，手机用户每月打电话的次数、时长呈现重尾分布。

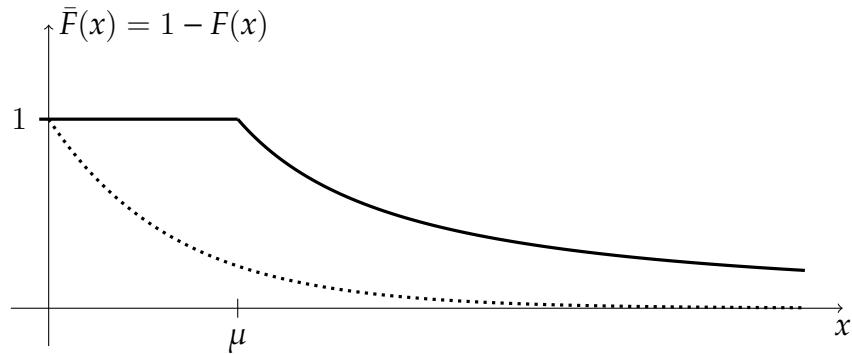


图 4.30: 实线是重尾分布 $\text{Pareto}(\alpha, \mu)$ 的尾函数 $\bar{F}(x) = 1 - F(x)$, 其中 $\alpha > 1$ 。若横轴表示财富, 尾函数 $\bar{F}(x)$ 表示财富大于 x 的人群比例。虚线是指数分布 $X \sim \text{Expon}(\lambda)$ 的尾函数 $\bar{F}_X(x)$, 它趋向于零的速度比 $\bar{F}(x)$ 更快一些。

练习 4.55. 试证明: Pareto 分布 $X \sim \text{Pareto}(\alpha, \mu)$ 的密度函数是

$$f(x) = \begin{cases} \alpha\mu^\alpha/x^{\alpha+1} & \text{当 } x \geq \mu \\ 0 & \text{当 } x < \mu \end{cases}$$

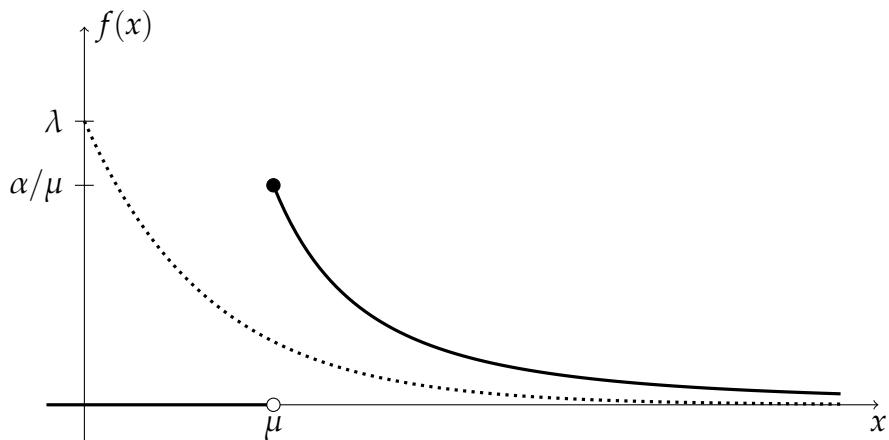


图 4.31: 实线是重尾分布 $\text{Pareto}(\alpha, \mu)$ 的密度函数曲线 $f(x)$, 其中 $\alpha > 1$ 。虚线是指数分布 $\text{Expon}(\lambda)$ 的密度函数。

性质 4.34. 分布 $\text{Pareto}(\alpha, \mu)$ 的密度函数 $f(x|\alpha, \mu)$ 满足性质

$$\lim_{x \rightarrow \infty} f(x|\alpha, \mu) = \delta(x - \mu), \text{ 其中 } \delta(x) \text{ 是 delta 函数}$$

练习 4.56. 随机变量 $X \sim \text{Pareto}(\alpha, \mu)$ 的 k 阶矩

$$m_k = \begin{cases} \frac{\alpha\mu^k}{\alpha - k} & \text{如果 } k < \alpha \\ \infty & \text{否则} \end{cases}$$

特别地，期望为 $E(X) = \alpha\mu/(\alpha - 1)$ ，其中 $\alpha > 1$ ；方差为 $\alpha\mu^2/[(\alpha - 1)^2(\alpha - 2)]$ ，其中 $\alpha > 2$ 。

性质 4.35. 若随机变量 $X \sim \text{Pareto}(\alpha, \mu)$ ，则 $Y = \ln(X/\mu) \sim \text{Expon}(\alpha)$ 。反之，若随机变量 $Y \sim \text{Expon}(\alpha)$ ，则 $X = \mu e^Y \sim \text{Pareto}(\alpha, \mu)$ 。

证明. 由定理 2.9 不难证得，留给读者补全。 \square

性质 4.36. 分布 $X \sim F(x)$ 是重尾的当且仅当

$$Ee^{\lambda X} = \int_{-\infty}^{+\infty} e^{\lambda x} dF(x) = \infty$$

定义 4.25. 如果分布函数 $F(x)$ 的尾函数 $\bar{F}(x) = 1 - F(x)$ 在右边足够远的地方非常平坦变化不大，则称之为长尾分布 (long-tailed distribution)。即

$$\lim_{x \rightarrow \infty} \frac{\bar{F}(x + t)}{\bar{F}(x)} = 1, \text{ 其中 } \forall t > 0$$

练习 4.57. 试证明：Pareto 分布是长尾分布。

性质 4.37. 长尾分布都是重尾分布，但反之不成立。

4.2.10 以物理学家命名的分布

本节所介绍的概率分布在物理学中有着重要的应用，它们以一些知名的物理学家命名。考虑到这些分布在后续的章节中很少用到，本节的内容可作为选读。



热力学和统计力学的奠基者之一、奥地利物理学家 Ludwig Eduard Boltzmann (1844-1906) 揭示了热平衡条件下系统状态的分布，称为 Boltzmann 分布。

定义 4.26 (Boltzmann 分布). 在热平衡条件之下，若所有可能状态的集合 S 有限，每个状态 $s \in S$ 具有能量 E_s ，则系统处于状态 s 的概率为

$$p_s = \frac{1}{Z} \exp\left\{-\frac{E_s}{k_B T}\right\}$$

其中 T 为绝对温度， $k_B = 1.380662 \times 10^{-23} \text{ J/K}$ 为 Boltzmann 常数， Z 为归一因子，即

$$Z = \sum_{s \in S} \exp\left\{-\frac{E_s}{k_B T}\right\}$$

通常状态的数目巨大，某一温度下大多数状态的 p_s 都接近零，第 15 章将介绍 Boltzmann 分布的抽样。

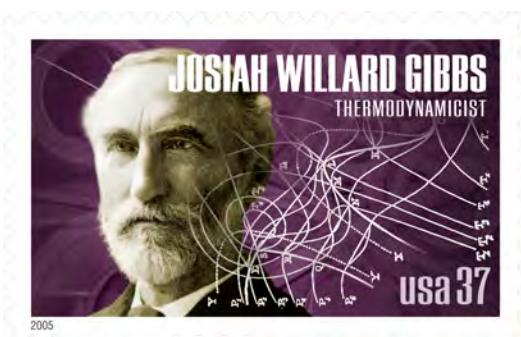
美国理论物理学家 Josiah Willard Gibbs (1839-1903) 发展了 Boltzmann 的理论，1902 年，Gibbs 发表了著作《统计力学的基本原理》，提出了 Gibbs 分布。

定义 4.27. Gibbs 分布的密度函数为

$$f(x) = \frac{1}{Z} \exp\left\{-\frac{E(x)}{k_B T}\right\}$$

其中， $E(x)$ 是状态 x (可以是向量) 的能量， Z 是归一因子，即

$$Z = \int_{\mathbb{R}} \exp\left\{-\frac{E(x)}{k_B T}\right\} dx$$



更一般地，Gibbs 分布具有如下形式的密度函数。

$$f(x) = \frac{1}{Z(\beta)} \exp\{-\beta E(x)\}$$

其中, $E(x)$ 是实值函数, β 是参数, $Z(\beta)$ 是归一因子。

1939 年, 瑞典物理学家 Wallodi Weibull (1887-1979) 在其著作《材料强度的统计理论》中研究材料的断裂强度时用到了法国数学家 Maurice Fréchet (1878-1973) 于 1927 年提出的一个分布, 后人习惯地将此分布称作 Weibull 分布。该分布常用于可靠性分析和寿命数据分析, 如金属的疲劳寿命等。

定义 4.28 (Weibull 分布, 1939). 标准指数分布 $Y \sim \text{Expon}(1)$ 经过变换 $X = \lambda^{-1}Y^{1/\alpha}$ 所得的随机变量 X 称为服从 Weibull 分布, 记作 $X \sim \text{Weibull}(\lambda, \alpha)$, 其中 $\alpha > 0$ 称为形状参数, $\lambda > 0$ 称为尺度参数。 X 的密度函数是

$$f(x) = \begin{cases} 0 & \text{当 } x \leq 0 \\ \alpha\lambda^\alpha x^{\alpha-1} \exp\{-(\lambda x)^\alpha\} & \text{当 } x > 0, \alpha > 0, \lambda > 0 \end{cases}$$

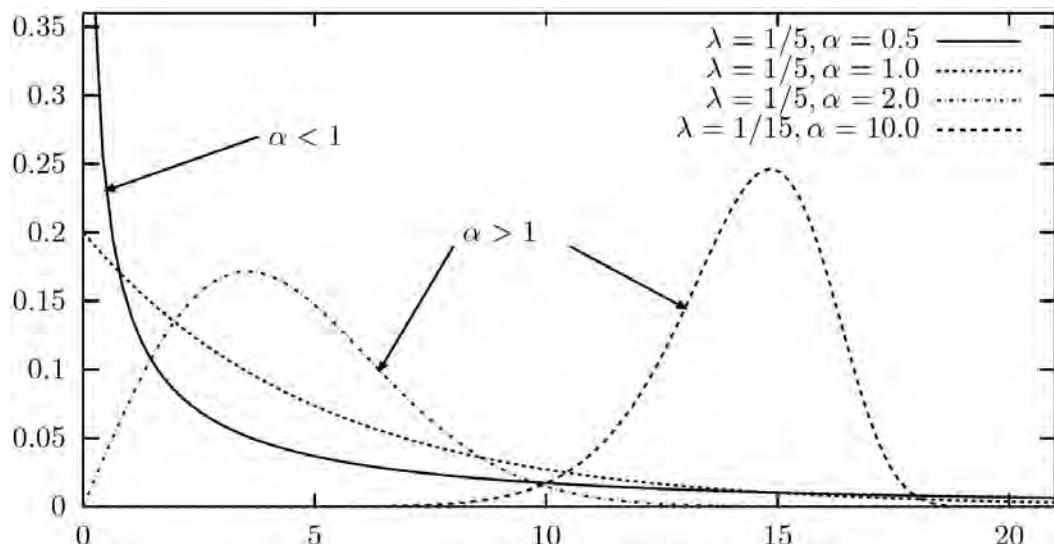


图 4.32: 分布 $\text{Weibull}(\lambda, \alpha)$ 的参数 α 称为形状参数, 当 $\alpha < 1, \alpha = 1, \alpha > 1$ 时, 密度函数曲线有三种不同的形状; λ 称为尺度参数。

练习 4.58. 请验证: 随机变量 $X \sim \text{Weibull}(\lambda, \alpha)$ 的分布函数是

$$F(x) = 1 - \exp\{-(\lambda x)^\alpha\}$$

并且, $X \sim \text{Weibull}(\lambda, \alpha)$ 的期望和方差分别是

$$E(X) = \frac{1}{\lambda} \Gamma\left(1 + \frac{1}{\alpha}\right) \quad V(X) = \frac{1}{\lambda^2} \left[\Gamma\left(1 + \frac{1}{\alpha^2}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right) \right]$$

练习 4.59. 指数分布 $\text{Expon}(\beta)$ 就是 $\text{Weibull}(\beta, 1)$ 分布。

练习 4.60. 请按照逆 CDF 法给出 $\text{Weibull}(\lambda, \alpha)$ 分布的随机数产生算法。

1880 年, 英国物理学家 Lord Rayleigh (1842-1919, 又名 John William Strutt) 在研究谐振分量叠加而形成的振幅分布时发现了 Rayleigh 分布, 其密度函数的推导详见第 162 页的例 2.45。Rayleigh 分布还常用于统计通信理论, 两个正交高斯噪声信号之和的包络服从 Rayleigh 分布。

定义 4.29 (Rayleigh 分布, 1880). 如果随机变量 $X_1, X_2 \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, 则二维随机向量 $(X_1, X_2)^\top$ 的模长 $X = \sqrt{X_1^2 + X_2^2}$ 所服从的分布被称为 Rayleigh 分布, 记作 $X \sim \text{Rayleigh}(\sigma)$, 其中 $\sigma > 0$ 是尺度参数。 X 的密度函数是

$$f(x) = \begin{cases} 0 & \text{当 } x \leq 0 \\ \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) & \text{当 } x > 0 \end{cases}$$

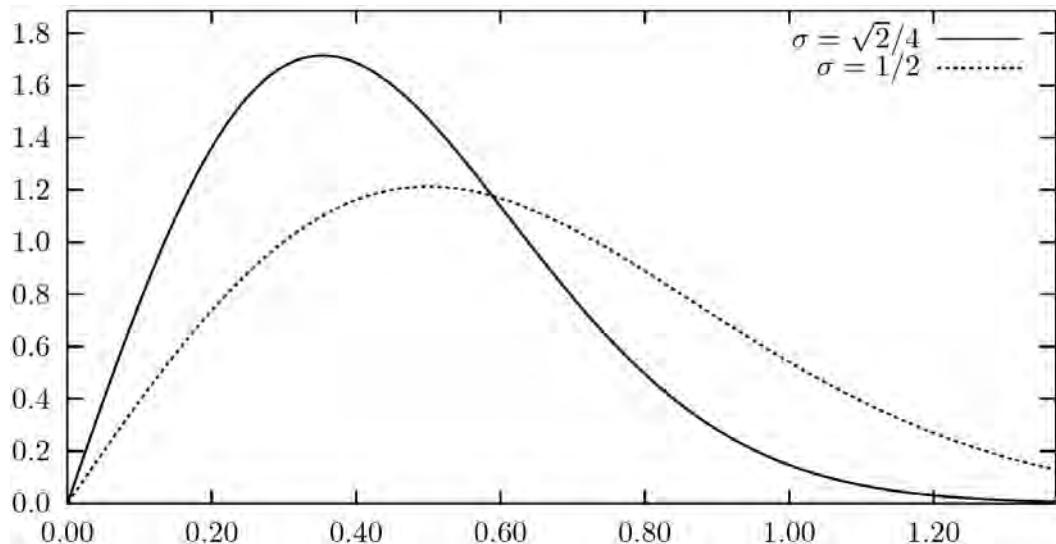
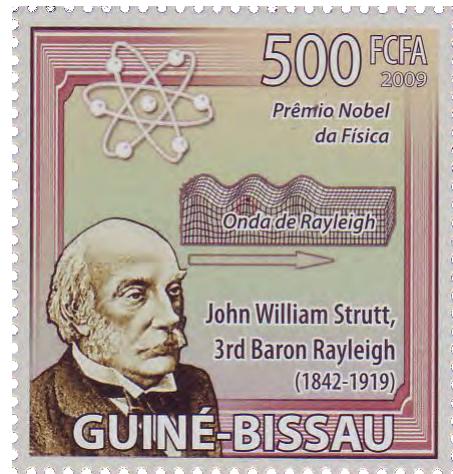


图 4.33: 尺度参数 σ 越小, Rayleigh(σ) 分布的密度函数曲线显得越“高瘦”。

练习 4.61. 请读者验证 Rayleigh 分布的下述性质。

分布 $\text{Rayleigh}(\sigma)$ 就是分布 $\text{Weibull}(1/(\sqrt{2}\sigma), 2)$ 。

随机变量 $X \sim \text{Rayleigh}(\sigma)$ 的数字特征分别为

$$\mathbb{E}(X) = \sigma \sqrt{\frac{\pi}{2}} \quad \mathbb{V}(X) = \frac{(4 - \pi)\sigma^2}{2} \quad \mathbb{M}(X) = \sigma \sqrt{\ln 4}$$

- 如果 $X \sim \text{Rayleigh}(1)$, 则 $X^2 \sim \chi_2^2$ 。
- 若复数 $Z = X + Yi$ 的实部和虚部独立同分布于 $N(0, \sigma^2)$, 则复数 Z 的模长 $|Z| = \sqrt{X^2 + Y^2}$ 服从分布 $\text{Rayleigh}(\sigma)$ 。
- 如果 $X \sim \text{Expon}(\lambda)$, 则 $Y = \sigma \sqrt{2\lambda X} \sim \text{Rayleigh}(\sigma)$ 。

算法 4.18. 分布 $X \sim \text{Rayleigh}(\sigma)$ 的随机数 $x^* = \sigma \sqrt{-2 \ln u^*}$, 其中 u^* 是均匀分布 $U(0, 1)$ 的随机数。

1859 年, 电磁理论与统计热力学的奠基者、英国著名的理论物理学家兼数学家 James Clerk Maxwell (1831-1879) 揭示了平衡态下理想气体分子的速率所服从的分布, 因此得名。Maxwell 分布的定义与 Rayleigh 分布类似, Maxwell 分布的密度函数的推导详见例 2.45。

定义 4.30 (Maxwell 分布, 1859). 已知随机变量 $X_1, X_2, X_3 \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, 三维随机向量 $(X_1, X_2, X_3)^\top$ 的长度 $X = \sqrt{X_1^2 + X_2^2 + X_3^2}$ 所服从的分布被称为 Maxwell 分布, 记作 $X \sim \text{Maxwell}(\sigma)$, 其中 $\sigma > 0$ 是尺度参数。 X 的密度函数是

$$f(x) = \begin{cases} 0 & \text{当 } x \leq 0 \\ \sqrt{\frac{2}{\pi}} \frac{x^2}{\sigma^3} \exp\left(-\frac{x^2}{2\sigma^2}\right) & \text{当 } x > 0 \end{cases}$$

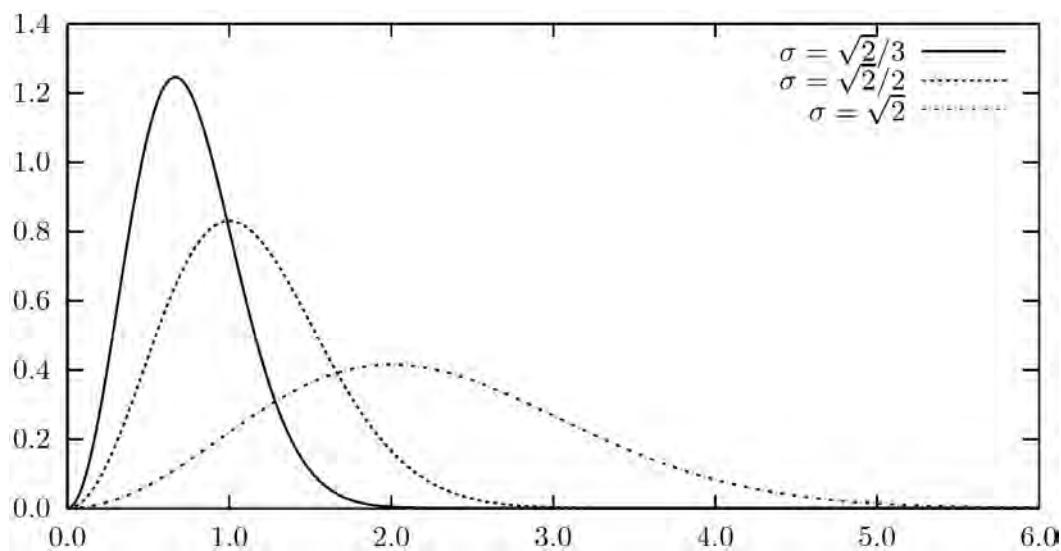


图 4.34: 尺度参数 σ 越小, $\text{Maxwell}(\sigma)$ 分布的密度函数曲线显得越“高瘦”。

练习 4.62. 请读者验证：随机变量 $X \sim \text{Maxwell}(\sigma)$ 的期望和方差分别是

$$\mathbb{E}(X) = \frac{2\sqrt{2}\sigma}{\sqrt{\pi}} \quad \mathbb{V}(X) = \sigma^2 \left(3 - \frac{8}{\pi}\right)$$

很多物理系统的性质可通过矩阵这一工具加以研究，例如，核物理中利用随机矩阵（即取值为矩阵的随机变量）对能谱进行分析。1932 年，美籍匈牙利裔物理学家、数学家 Eugene Paul Wigner (1902-1995) 在量子力学的研究中发现了一类重要的分布，后来以他的名字命名为 Wigner 分布，也称作 Wigner 半圆分布。该分布因著名的 Wigner 半圆律 (semicircular law, 见例 4.26) 而在应用日益广泛的随机矩阵理论中占有重要的地位。



定义 4.31 (Wigner 半圆分布, 1932). 如果连续型随机变量 X 具有以下的密度函数，则称 X 服从 Wigner 半圆分布，并记作 $X \sim \text{Wigner}(r)$ 。

$$f(x) = \begin{cases} \frac{2}{\pi r^2} \sqrt{r^2 - x^2} & \text{若 } |x| < r, \text{ 其中 } r > 0 \\ 0 & \text{若 } |x| \geq r \end{cases}$$

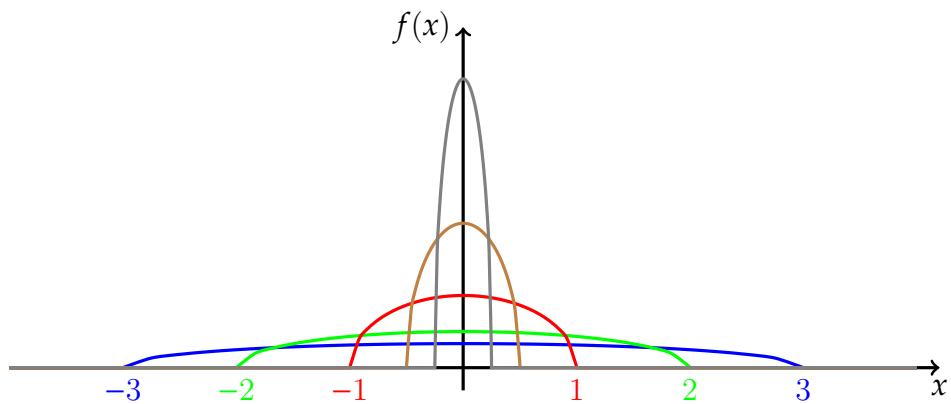


图 4.35: 当 $r = 3, 2, 1, 0.5, 0.25$ 时 Wigner 半圆分布的密度函数曲线。

练习 4.63. 请读者验证 $X \sim \text{Wigner}(r)$ 的期望和方差分别为

$$\mathbb{E}(X) = 0 \quad \mathbb{V}(X) = \frac{r^2}{4}$$

练习 4.64. 试证明：若 $r > 0$ 且 $X \sim \text{Beta}(3/2, 3/2)$ ，则

$$r(2X - 1) \sim \text{Wigner}(r)$$

提示：利用例 2.17 的结果。

例 4.26 (Wigner 半圆律). 已经证明，当阶数趋近无穷时，许多随机对称矩阵的特征值的极限分布就是 Wigner 半圆分布，这就是著名的 Wigner 半圆律。

为直观地了解 Wigner 半圆律，先随机地产生一个 $n \times n$ 阶的对称矩阵，不妨设 $n = 5000$ ，再考察它的特征值的分布情况。

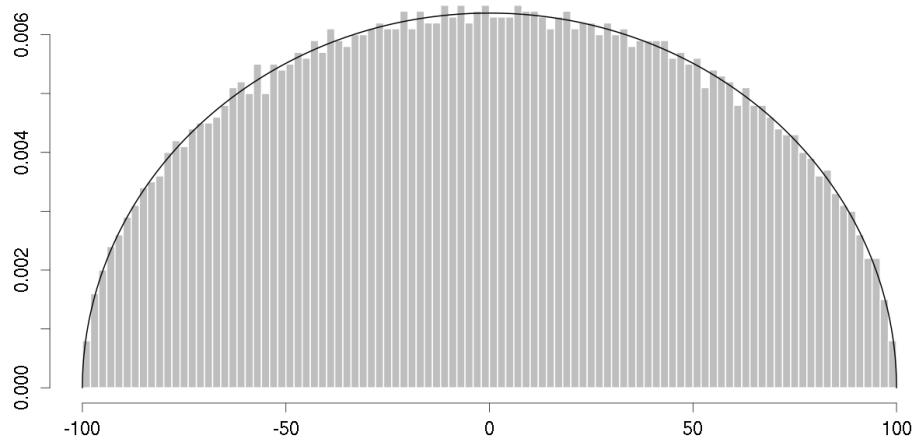


图 4.36: 一个 5000×5000 的随机对称矩阵的特征值的直方图，实线是 Wigner 半圆分布的密度函数曲线。

4.3 随机向量的分布

随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 即是取值为向量的随机变量，具有更多的实用性，例如回归分析、投资组合理论等。多元统计学 (multivariate statistics) 对随机向量进行了专门的研究，读者可参阅 R. A. Johnson 和 D. W. Wichern 合著的《应用多元统计分析》[83]，或者 C. R. Rao 的《线性统计推断及其应用》[129]。

本书只粗略地介绍几个常见的随机向量的分布，如高维均匀分布、多项分布、Dirichlet 分布和多元正态分布，它们分别是一维均匀分布、二项分布、Beta 分布和一元正态分布向高维的推广。最后，简要地介绍一下在多元统计分析中最常用的 Wishart 分布。

在具体讨论这些分布之前，先介绍随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \pi(\mathbf{x})$ 的一个基于逆 CDF 法的通用 PRNG 算法如下，该方法依然用到了 $U(0, 1)$ 的随机数，还需要计算条件分布函数。

算法 4.19. 令 $F_j(x_j|x_1, \dots, x_{j-1})$ 是在给定 $X_1 = x_1, \dots, X_{j-1} = x_{j-1}$ 的条件下 X_j 的条件分布函数，设随机变量 $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} U(0, 1)$ 抽样的结果是 u_1, \dots, u_n 。

利用逆 CDF 法（见第 280 页的 [算法 4.10](#)），随机向量 $\mathbf{X} \sim \pi(\mathbf{x})$ 的随机数 $\mathbf{x}^* = (x_1, \dots, x_n)^\top$ 可通过解下述方程组求得。

$$\begin{cases} F_1(x_1) = u_1 \\ F_j(x_j|x_1, \dots, x_{j-1}) = u_j, \text{ 其中 } j = 2, \dots, n \end{cases}$$

例 4.27. 设二维随机向量 $\mathbf{X} = (X_1, X_2)^\top$ 的联合密度函数是

$$\pi(x_1, x_2) = \begin{cases} 6x_1 & \text{如果 } x_1 + x_2 \leq 1, x_1 \geq 0, x_2 \geq 0 \\ 0 & \text{其他} \end{cases}$$

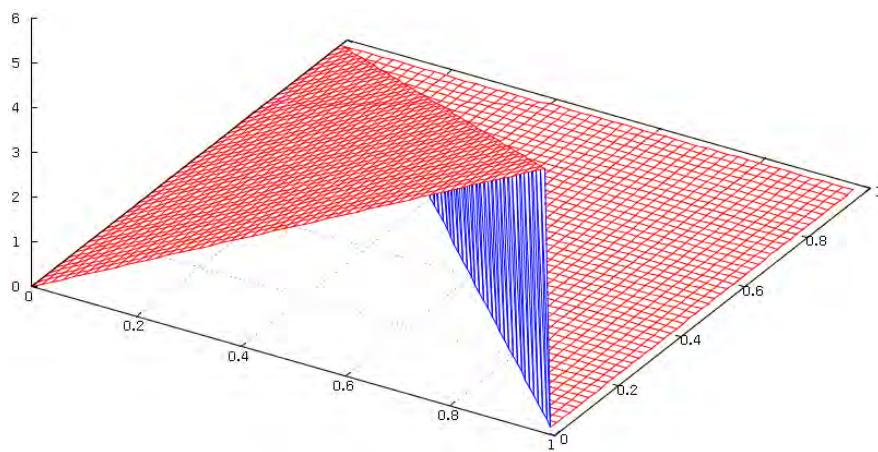


图 4.37: 例 4.27 中联合分布 $\pi(x_1, x_2)$ 的密度函数曲面。

试利用[算法 4.19](#) 给出 \mathbf{X} 的随机数 $\mathbf{x}^* = (x_1, x_2)^\top$ 的产生算法。

解. 将 $\pi(\mathbf{x})$ 分解为 $\pi(\mathbf{x}) = \pi_2(x_2)\pi_1(x_1|x_2)$, 其中

$$\begin{aligned}\pi_2(x_2) &= \int_0^{1-x_2} \pi(x_1, x_2) dx_1 = 3(1-x_2)^2, \text{ 其中 } 0 \leq x_2 \leq 1 \\ \pi_1(x_1|x_2) &= \frac{\pi(x_1, x_2)}{\pi_2(x_2)} = \frac{2x_1}{(1-x_1)^2}, \text{ 其中 } 0 \leq x_1 \leq x_2\end{aligned}$$

与上述密度函数相对应的分布函数分别为

$$\begin{aligned}F_2(x_2) &= \int_0^{x_2} \pi_2(t) dt = 1 - (1-x_2)^3, \text{ 其中 } 0 \leq x_2 \leq 1 \\ F_1(x_1|x_2) &= \int_0^{x_1} \pi_1(t|x_2) dt = \frac{x_1^2}{(1-x_2)^2}, \text{ 其中 } 0 \leq x_1 \leq x_2\end{aligned}$$

根据[算法 4.19](#), 解下面的方程组, 其中 u_1, u_2 分别是 $U_1, U_2 \stackrel{\text{iid}}{\sim} \text{U}(0, 1)$ 的随机数。所求得的解就是 \mathbf{X} 的随机数 $\mathbf{x}^* = (x_1, x_2)^\top$ 。

$$\left\{ \begin{array}{l} 1 - (1-x_2)^3 = u_1 \\ x_1^2/(1-x_2)^2 = u_2 \end{array} \right. \xrightarrow{\text{求解}} \left\{ \begin{array}{l} x_1 = \sqrt{u_2} \sqrt[3]{1-u_1} \\ x_2 = 1 - \sqrt[3]{1-u_1} \end{array} \right.$$

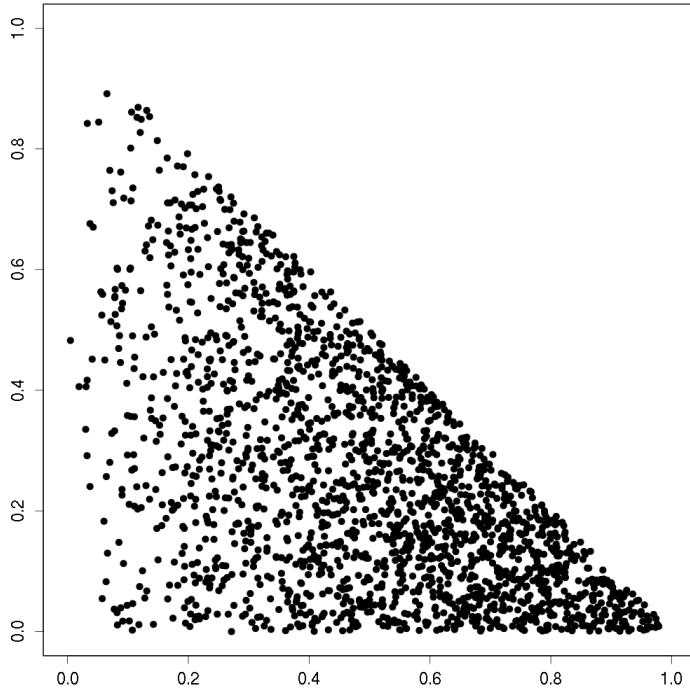


图 4.38: 利用[算法 4.19](#) 给出[例 4.27](#) 中联合分布 $\pi(x_1, x_2)$ 的随机数的散点图。

练习 4.65. 接着[例 4.27](#), 如果采用分解 $\pi(\mathbf{x}) = \pi_1(x_1)\pi_2(x_2|x_1)$, 利用[算法 4.19](#) 给出 \mathbf{X} 的随机数的产生算法。通过该练习, 读者体会 $\pi(\mathbf{x})$ 不同的分解方式, 导致

实现[算法 4.19](#) 的难易程度也不同。提示: $F_1(x_1) = 3x_1^2 - 2x_1^3$, 其中 $0 \leq x_1 \leq 1$;
 $F_2(x_1|x_2) = x_2/(1-x_1)$, 其中 $0 \leq x_2 \leq 1-x_1$ 。

本节内容

第一、二小节介绍了多项分布和 Dirichlet 分布, 它们分别是二项分布和 Beta 分布向高维的推广。第一小节证明了多项分布的随机向量的若干分量之和服从二项分布。第三小节重点讨论有着颇多应用的多元正态分布及其性质, 证明了线性变换不改变正态性, 并给出了正态随机向量的任一子向量的条件分布。第四小节初步介绍了 Wishart 分布, 它是对 χ_n^2 分布的推广。

关键知识

(1) 大致了解 Dirichlet 分布和 Wishart 分布。(2) 掌握多项分布、多元正态分布的性质, 特别是那些有关特征函数、线性变换和条件分布的结果。

4.3.1 高维均匀分布

定义 4.32 (高维均匀分布). 若 $\Omega \subset \mathbb{R}^n$ 是有界区域, 其体积或测度存在, 设为 $m(\Omega)$ 。区域 Ω 上的 (高维) 均匀分布 $\mathbf{X} \sim U(\Omega)$ 的密度函数为

$$f(\mathbf{x}) = \begin{cases} \frac{1}{m(\Omega)} & \text{当 } \mathbf{x} = (x_1, \dots, x_n)^\top \in \Omega \\ 0 & \text{其他} \end{cases}$$

例 4.28. 由多元微积分的知识, n 维超立方体 $C_n = [-1, 1]^n$ 的体积为 $m(C_n) = 2^n$; n 维单位超球体 $B_n = \{(x_1, \dots, x_n)^\top \in \mathbb{R}^n : \sum_{j=1}^n x_j^2 \leq 1\}$ 的体积为 $m(B_n) = \pi^{k/(k\Gamma(k))}$, 其中 $k = n/2$ 。进而, 不难得出分布 $U(C_n), U(B_n)$ 的密度函数。

下面, 我们将介绍如何产生高维均匀分布的随机数, 万变不离其宗, 依然要用到 $U[0, 1]$ 的随机数产生算法。

算法 4.20. n 维超长方体 $D_n = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$ 上均匀分布的随机数是这样产生的: 独立地产生 $U[a_j, b_j]$ 的随机数 $r_j, j = 1, 2, \dots, n$, 得到的 n 维向量 $\mathbf{r} = (r_1, r_2, \dots, r_n)^\top$ 便是 $U(D_n)$ 的一个随机数。

算法 4.21. 有界区域 $\Omega \subset \mathbb{R}^n$ 上均匀分布 $U(\Omega)$ 的一个随机数是这样产生的:

- 找一个合适的 n 维超长方体 D_n , 使得 $\Omega \subseteq D_n$ 。
- 利用算法 4.20 产生 $U(D_n)$ 的一个随机数 \mathbf{r} 。如果 $\mathbf{r} \in \Omega$, 则它是 $U(\Omega)$ 的一个随机数。否则, 重复该步骤。

例 4.29. 图 4.39 的左图是立方体 $C_3 = [-1, 1]^3$ 上均匀分布的一些随机数。加上限制 $x_1^2 + x_2^2 + x_3^2 \leq 1$ 后, 便产生出 $U(B_3)$ 的随机数, 其结果见图 4.39 的右图。

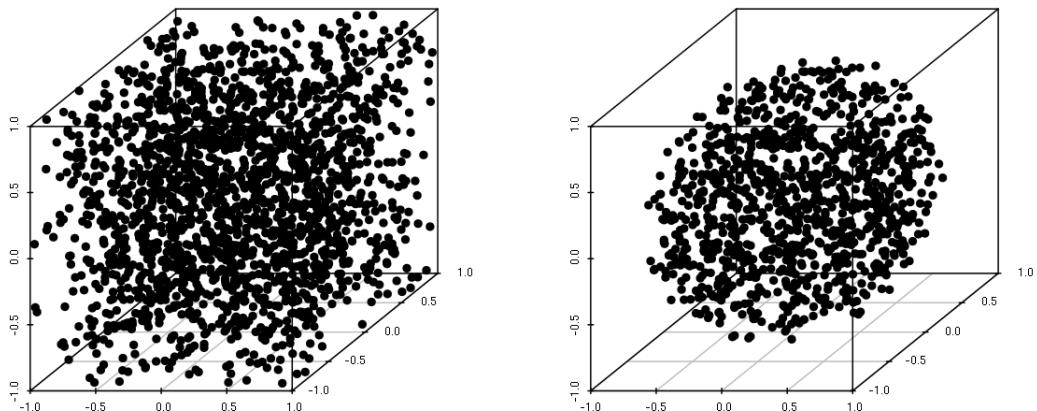


图 4.39: 立方体 $C_3 = [-1, 1]^3$ 和单位球体 B_3 上均匀分布的随机数的散点图。

例 4.30. 利用**算法 4.21** 分别产生圆环体 $O_3(R, r) = \{(x_1, x_2, x_3)^\top \in \mathbb{R}^3 : (R - \sqrt{x_1^2 + x_3^2})^2 + x_2^2 \leq r^2\}$, 并且 $R > r$ 和圆锥体 $\Delta_3 = \{(x_1, x_2, x_3)^\top \in \mathbb{R}^3 : x_1^2 + x_2^2 \leq (x_3 - 1)^2/4\}$, 并且 $x_3 \in [-1, 1]\}$ 上均匀分布的随机数, 结果如下图所示。

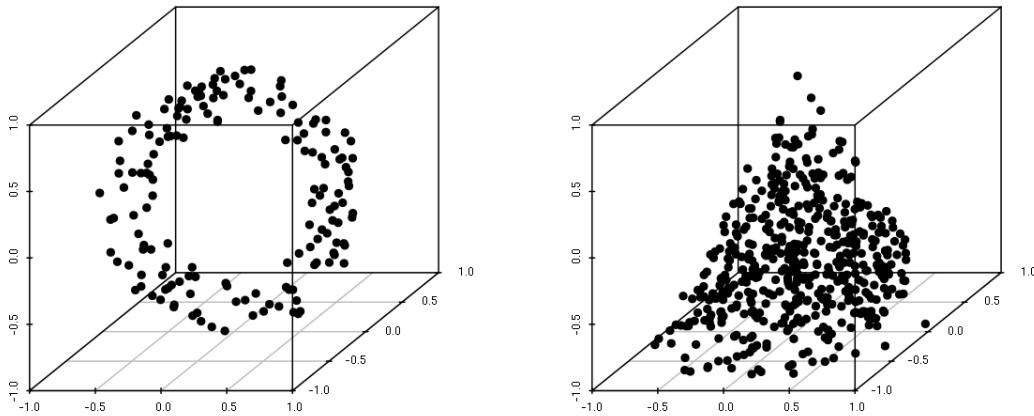


图 4.40: 圆环体 $O_3(0.8, 0.2)$ 和圆锥体 Δ_3 上均匀分布的随机数的散点图。

练习 4.66. 试说明: 超立方体上的均匀分布的边缘分布仍然是某低维超立方体上的均匀分布, 但单位超球体上的均匀分布的边缘分布不是低维单位超球体上的均匀分布。提示: 考虑投射 $(x_1, \dots, x_{n-1}, x_n)^\top \mapsto (x_1, \dots, x_{n-1})^\top$ 。

 **算法 4.21** 虽然是一个普适的方法, 但很多时候效率不高, 这是因为落入区域 $C_n - \Omega$ 里的随机数是要被丢弃的, 数量过多显然会影响算法的效率。为此, 我们建议具体问题具体分析, 譬如下面的例子。

例 4.31. 如何产生二维单位球面 $S_2 = \{(x_1, x_2, x_3)^\top \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 = 1\}$ 和环面 $T_2(R, r)$ (即, 环面体 $O_3(R, r)$ 的面) 上均匀分布的随机数?

解. (1) 将图 4.39 的右图中 $U(B_3)$ 的随机数 $\mathbf{x} = (x_1, x_2, x_3)^\top \neq \mathbf{0}$ 投射到单位球面上便得到 $U(S_2)$ 的一个随机数 $\mathbf{x}' = (x'_1, x'_2, x'_3)^\top$ 。事实上, \mathbf{x}' 就是 \mathbf{x} 的标准化, 即

$$\mathbf{x}' = \frac{\mathbf{x}}{\|\mathbf{x}\|}, \text{ 其中 } \|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + x_3^2} \text{ 是向量 } \mathbf{x} \text{ 的长度}$$

该方法用到了**算法 4.21**, 其结果见图 4.41 的左图。下面, 我们推荐另外一种算法, 在效率上要胜出一筹。

独立地产生 $U[0, 2\pi)$ 的随机数 θ 和 $U[0, \pi)$ 的随机数 φ , 则 $U(S_2)$ 的一个随机数 $\mathbf{x} = (x_1, x_2, x_3)^\top$ 可按下面球面坐标的方法产生。

$$\begin{cases} x_1 = \cos \theta \sin \varphi \\ x_2 = \sin \theta \sin \varphi \\ x_3 = \cos \varphi \end{cases}$$

(2) 令 θ, φ 是独立取自 $U[0, 2\pi]$ 的随机数, 则环面 $T_2(R, r)$ 上均匀分布的一个随机数 $x = (x_1, x_2, x_3)^\top$ 可按下面的方法产生, 结果见图 4.41 的右图。

$$\begin{cases} x_1 = (R + r \cos \theta) \cos \varphi \\ x_2 = r \sin \theta \\ x_3 = (R + r \cos \theta) \sin \varphi \end{cases}$$

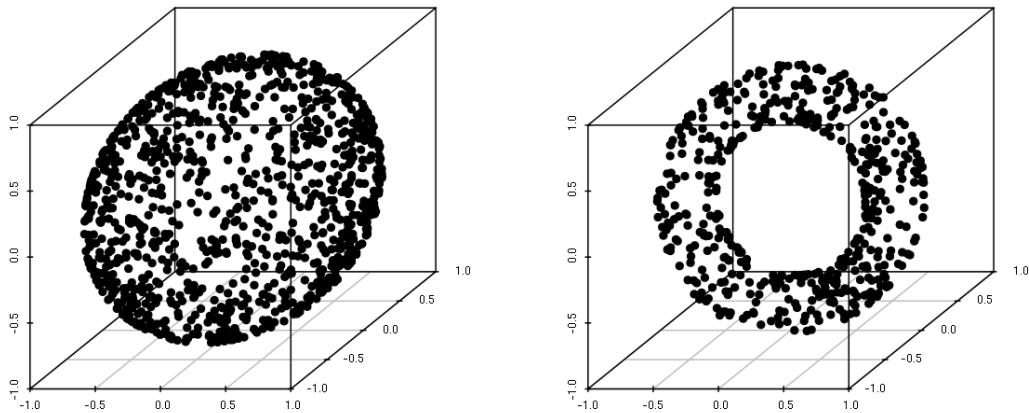


图 4.41: 二维单位球面 S_2 和二维环面 $T_2(0.8, 0.2)$ 上均匀分布的随机数的散点图。

练习 4.67. 如何产生圆锥面 $\Omega = \{(x_1, x_2, x_3)^\top \in \mathbb{R}^3 : x_1^2 + x_2^2 = (x_3 - 1)^2/4\}$, 并且 $x_3 \in [-1, 1]\}$ 上均匀分布的随机数? 提示: 圆锥面展开后是一个扇形。

※例 4.32. 利用第 10 页的例 1.2 中所描述的混沌系统 $f(x) = 2x^2 - 1$, 选初始值 $x_0 = 0.4$, 定义序列 $x_{n+1} = f(x_n)$, 做如下的变换。

$$y_n = \frac{1}{2} + \frac{\arcsin(x_n)}{\pi}, \text{ 其中 } n = 0, 1, 2, \dots$$

不妨设 $y_0, y_1, y_2, \dots, y_{n-1}, y_n, \dots$ 是变换后得到的序列。为了得到更好的“随机性”而预烧^{*}掉 y_0, y_1, \dots, y_{n-1} 。

图 4.42 显示了区域 $[0, 1] \times [0, 1]$ 上的均匀分布的“随机数” $\{(y_n, y_{n+k}) : n = 201, \dots, 1900\}$ 的散点图, 其中 $k = 2$ 时见图 (a), $k = 100$ 时见图 (b)。不难看出, 后者更像区域 $[0, 1] \times [0, 1]$ 上的均匀分布的随机数, 这是因为 k 越大, y_n 和 y_{n+k} 之间越显得没有什么关系。

另外, 图 4.42 中在水平和垂直方向还分别绘出了 $\{y_n : n = 201, \dots, 1900\}$ 和 $\{y_{n+k} : n = 201, \dots, 1900\}$ 的直方图。很显然, 通过它们是无法判断所产生的随机数是否足够地随机, 也再一次说明了“从边缘分布不能重构联合分布”(见第 141 页)。

^{*}预烧 (burn-in) 原义是让电子设备快速老化稳定的一种措施, 即让电子设备连续工作一段时间后使之进入最佳状态。在概率统计里, 为了使算法产生的伪随机数显得“更随机”一些, 往往需要扔掉最初产生的一些随机数, 这种作法被形象地类比为“预烧”。

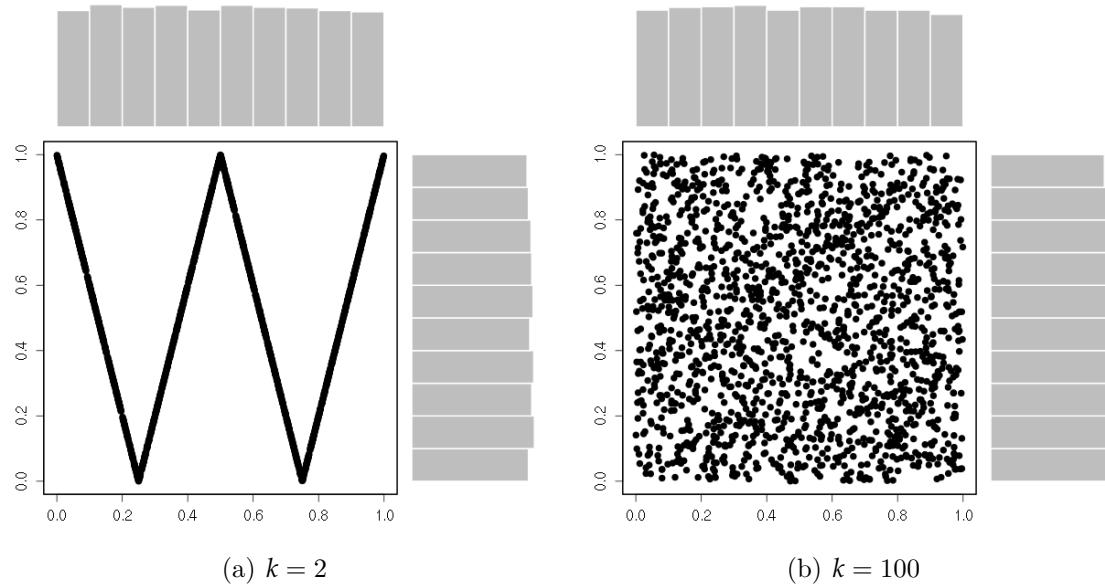


图 4.42: 利用混沌系统 (见例 4.32) 产生区域 $[0, 1] \times [0, 1]$ 上的均匀分布的“随机数”的散点图, 该方法虽不是普适的, 但很有趣。

4.3.2 多项分布

二项分布 $B(n, p)$ 源自 n 重 Bernoulli 试验（见例 1.47），多项分布是二项分布向高维的推广，它源自如下定义的 k 分类随机试验。

定义 4.33 (k -分类试验). 若随机试验共有 k 个可能的结果 A_1, \dots, A_k 且 $P(A_j) = p_j \neq 0, j = 1, 2, \dots, k$, 其中 A_1, \dots, A_k 两两互斥且 $\sum_{j=1}^k p_j = 1$, 我们把如此定义的随机试验称为 k -分类试验。

例 4.33. 事先定义好 k 种面部表情^{*}，对每种表情都收集一些图片，从中随机抽取一张图片观察其表情类别的试验就是 k -分类试验。



图 4.43: 某日本女性的七种面部表情，依次为生气、失望、恐惧、高兴、常态、悲伤、惊讶。该图片取自 JAFFE 数据库 (<http://www.kasrl.org/jaffe.html>)。

定义 4.34 (多项分布). 独立地重复 n 次 k -分类试验，令 X_j 表示事件 A_j 发生的次数。如果离散型随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_k)^\top$ 的概率函数如下所示，则称 \mathbf{X} 服从多项分布 (multinomial distribution)，记作 $\mathbf{X} \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$ 。

$$P(X_1 = n_1, X_2 = n_2, \dots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \quad (4.22)$$

其中， $n_1 \geq 0, n_2 \geq 0, \dots, n_k \geq 0$ 且 $\sum_{j=1}^k n_j = n, \sum_{j=1}^k p_j = 1$ 。有时，式 (4.22) 也被具体展开为

$$\begin{aligned} & P(X_1 = n_1, \dots, X_{k-1} = n_{k-1}, X_k = n - n_1 - \cdots - n_{k-1}) \\ &= \frac{n!}{n_1! \cdots n_{k-1}! (n - n_1 - \cdots - n_{k-1})!} p_1^{n_1} \cdots p_{k-1}^{n_{k-1}} (1 - p_1 - \cdots - p_{k-1})^{n - n_1 - \cdots - n_{k-1}} \end{aligned}$$

显然，式 (4.22) 即为 $(p_1 + p_2 + \cdots + p_k)^n$ 多项式展开中的一般项，这是术语“多项分布”的由来。当 $n = 2$ 时，即是二项分布。

例 4.34. 某大规模语料库中搜集了四类文本，其中 10% 为财经类 W ，20% 为科技类 T ，15% 为体育类 S ，55% 为娱乐类 E 。随机抽取 20 篇文本，试求概率 $P(W = 2, T = 5, S = 4, E = 9)$ 。

^{*}人类面部表情的识别 (facial expression recognition, FER) 是模式识别和机器学习领域的研究课题，其目标是让机器“看懂”人类的表情，从而更好地与人类“交流”。FER 可应用于病患护理监控、人机交互、数据驱动的动画设计、图像检索、人类情感分析等领域。

解. 已知 $(W, T, S, E)^\top \sim \text{Multin}(20; 0.1, 0.2, 0.15, 0.55)$, 则

$$P(W = 2, T = 5, S = 4, E = 9) = \frac{20!}{2!5!4!9!} (0.1)^2 (0.2)^5 (0.15)^4 (0.55)^9 = 0.00868397$$

性质 4.38. 随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_k)^\top \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$ 的特征函数为 $\varphi(t_1, t_2, \dots, t_k) = (p_1 e^{it_1} + p_2 e^{it_2} + \dots + p_k e^{it_k})^n$, 期望为 $E(\mathbf{X}) = (np_1, np_2, \dots, np_k)^\top$, 协方差矩阵为 $\Sigma = (\sigma_{ij})_{k \times k}$, 其中

$$\sigma_{ij} = \begin{cases} np_i(1-p_i) & \text{若 } i = j \\ -np_i p_j & \text{若 } i \neq j \end{cases}$$

随机向量 $(X_1, X_2, \dots, X_k)^\top \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$ 中, 因为分量 $X_k = n - X_1 - \dots - X_{k-1}$, 所以在不引起歧义的时候该多项分布也简记作 $(X_1, X_2, \dots, X_{k-1})^\top \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$ 。特别地, 当 $k = 2$ 时, $\text{Multin}(n; p_1, p_2)$ 就是二项分布 $B(n, p_1)$ 。

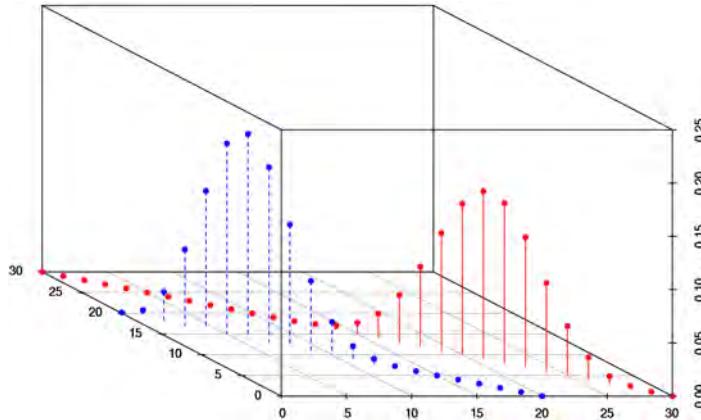


图 4.44: 多项分布 $\text{Multin}(n; \mathbf{p})$ 的概率函数的竖线图, 其中参数分别是 $n = 20, \mathbf{p} = (0.3, 0.7)^\top$ (虚线) 和 $n = 30, \mathbf{p} = (0.7, 0.3)^\top$ (实线)。

性质 4.39. 多项分布 $(X_1, \dots, X_k)^\top \sim \text{Multin}(n; p_1, \dots, p_k)$ 的任何边缘分布仍是多项分布, 具体来说,

$$(X_1, \dots, X_m, X'_{m+1})^\top \sim \text{Multin}(n; p_1, \dots, p_m, p'_{m+1}), \text{ 其中 } 1 \leq m < k \text{ 且}$$

$$X'_{m+1} = X_{m+1} + \dots + X_k = n - X_1 - \dots - X_m$$

$$p'_{m+1} = p_{m+1} + \dots + p_k = 1 - p_1 - \dots - p_m$$

特别地, $X_j \sim B(n, p_j) = \text{Multin}(n; p_j, 1 - p_j)$, 其中 $j = 1, 2, \dots, k$ 。

证明. 将 k -分类试验中的类 $m+1, \dots, k$ 合并为一个新类, 按照定义 4.34, 对于这个 $(m+1)$ -分类试验, 随机向量 $(X_1, \dots, X_m, X'_{m+1})^\top$ 服从多项分布。 \square

推论 4.2. 已知随机向量 $(X_1, X_2, \dots, X_k)^\top \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$ 且 $1 \leq m \leq k$, 则

$$(X_1 + \dots + X_m, X_{m+1} + \dots + X_k)^\top \sim \text{Multin}(n; p_1 + \dots + p_m, p_{m+1} + \dots + p_k)$$

或者, 等价地 $X_1 + X_2 + \dots + X_m \sim \text{B}(n, p_1 + p_2 + \dots + p_m)$

性质 4.40. 已知随机向量 $\mathbf{X} = (X_1, \dots, X_k)^\top \sim \text{Multin}(n; p_1, \dots, p_k)$ 且 $m < k$, 令 $\mathbf{X}_{(1)} = (X_1, \dots, X_m)^\top, \mathbf{X}_{(2)} = (X_{m+1}, \dots, X_k)^\top$, 则

$$\mathbf{X}_{(1)} | \mathbf{X}_{(2)} = (x_{m+1}, \dots, x_k)^\top \sim \text{Multin}\left(n - x_{m+1} - \dots - x_k; \frac{p_1}{p_1 + \dots + p_m}, \dots, \frac{p_m}{p_1 + \dots + p_m}\right)$$

证明. 由**性质 4.39**, $\mathbf{X}_{(2)} = (X_{m+1}, \dots, X_k)^\top \sim \text{Multin}(n; p_{m+1}, \dots, p_k, 1-p_{m+1}-\dots-p_k)$ 。基于该结果, 下面计算条件分布 $\mathbf{X}_{(1)} | \mathbf{X}_{(2)} = (x_{m+1}, \dots, x_k)^\top$ 的密度函数,

$$\begin{aligned} f(\mathbf{x}_{(1)} | \mathbf{x}_{(2)}) &= \frac{\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}}{\frac{n!}{x_{m+1}! \cdots x_k! (n - x_{m+1} - \cdots - x_k)!} p_{m+1}^{x_{m+1}} \cdots p_k^{x_k} (1 - p_{m+1} - \cdots - p_k)^{n - x_{m+1} - \cdots - x_k}} \\ &= \frac{(n - x_{m+1} - \cdots - x_k)!}{x_1! \cdots x_m!} \left(\frac{p_1}{p_1 + \cdots + p_m}\right)^{x_1} \cdots \left(\frac{p_m}{p_1 + \cdots + p_m}\right)^{x_m} \quad \square \end{aligned}$$



性质 4.40 是直观的: 独立地进行 $n - x_{m+1} - \cdots - x_k$ 次 m -分类试验, 观察到类 $1, 2, \dots, m$ 的机会之比依然是 $p_1 : p_2 : \cdots : p_m$, 要得到相应的概率, 必须乘上归一因子 $(p_1 + p_2 + \cdots + p_m)^{-1}$ 。另外, $\mathbf{X}_{(2)} = (x_{m+1}, \dots, x_k)^\top$ 所含的“信息量”对于计算条件分布是绰绰有余的了, 事实上, 我们只需知道 $x_{m+1} + \cdots + x_k$ 就行。

推论 4.3. 已知随机向量 $\mathbf{X} = (X_1, \dots, X_k)^\top \sim \text{Multin}(n; p_1, \dots, p_k)$ 且 $1 < m < k$, 令 $Y_m = X_1 + \cdots + X_{m-1}$, 则

$$X_m | Y_m = y_m \sim \text{B}\left(n - y_m, \frac{p_m}{1 - p_1 - \cdots - p_{m-1}}\right)$$

证明. 利用**性质 4.39** 和**性质 4.40** 可证, 留作习题请读者补全证明。 □

性质 4.41. 已知随机变量 $X_j \sim \text{Poisson}(\lambda_j), j = 1, \dots, k$ 相互独立, 则随机向量 $\mathbf{X} = (X_1, \dots, X_k)^\top$ 在条件 $X_1 + \cdots + X_k = n$ 之下服从多项分布, 即

$$\mathbf{X} | X_1 + \cdots + X_k = n \sim \text{Multin}\left(n; \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_k}, \dots, \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_k}\right)$$

证明. 由第 276 页的**练习 4.17** 的结果, $X_1 + \cdots + X_k \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_k)$ 。下面,

计算 $\mathbf{X}|X_1 + \dots + X_k = n$ 的条件概率函数。

$$\begin{aligned} & P(X_1 = x_1, \dots, X_k = x_k | X_1 + \dots + X_k = n), \text{ 其中 } x_k = n - x_1 - \dots - x_{k-1} \\ &= \frac{\frac{\lambda_1^{x_1}}{x_1!} \exp(-\lambda_1) \dots \frac{\lambda_k^{x_k}}{x_k!} \exp(-\lambda_k)}{\frac{(\lambda_1 + \dots + \lambda_k)^n}{n!} \exp(-\lambda_1 - \dots - \lambda_k)} \\ &= \frac{n!}{x_1! \dots x_k!} \left(\frac{\lambda_1}{\lambda_1 + \dots + \lambda_k} \right)^{x_1} \dots \left(\frac{\lambda_k}{\lambda_1 + \dots + \lambda_k} \right)^{x_k} \end{aligned}$$

□

理论上，可以利用性质 4.41 产生多项分布的随机数，但性质 4.41 中的条件 $X_1 + \dots + X_k = n$ 实现起来并不经济。所以，我们还是从多项分布的原始定义出发，给出随机数的产生算法。

算法 4.22. 多项分布 $(X_1, X_2, \dots, X_k)^\top \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$ 的随机数 $\mathbf{x}^* = (x_1, x_2, \dots, x_k)^\top$ 通过如下算法产生。

- 初始化 $(x_1, x_2, \dots, x_k)^\top = \mathbf{0}$ 。
- 将下述过程独立重复 n 次：
 - 按照分布 $p_1\langle 1 \rangle + p_2\langle 2 \rangle + \dots + p_k\langle k \rangle$ 产生随机数 j^* ;
 - 置 $x_{j^*} \leftarrow x_{j^*} + 1$ 。

4.3.3 Dirichlet 分布

德国数学家 Gustav Lejeune Dirichlet (1805-1859) 是解析数论的奠基者之一，对分析学和数学物理也有很多重大的贡献。

Dirichlet 分布是 Beta 分布向高维的推广，对非参数统计学中次序统计量理论非常重要（见第 460 页的性质 7.4），在贝叶斯统计学（第 12 章）中也常用作多项分布的共轭先验分布（见第 655 页的例 12.17）。

定义 4.35 (Dirichlet 分布). 如果连续型随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的密度函数如下所示，则称 \mathbf{X} 服从参数为 $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)^\top$ 的 Dirichlet 分布，并记作 $\mathbf{X} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ 或 $\mathbf{X} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n)$ 。

$$f(x_1, x_2, \dots, x_n) = \begin{cases} \frac{\Gamma(\alpha_1 + \dots + \alpha_n)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_n)} \prod_{j=1}^n x_j^{\alpha_j-1} & \text{其中 } x_j \geq 0 \text{ 满足 } \sum_{j=1}^n x_j = 1, \alpha_j > 0 \\ 0 & \text{其他} \end{cases} \quad (4.23)$$

 显然， $n = 2$ 时 $\text{Dirichlet}(\alpha_1, \alpha_2)$ 就是 $\text{Beta}(\alpha_1, \alpha_2)$ 。我们把 $\frac{\Gamma(\alpha_1) \dots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \dots + \alpha_n)}$ 简记为 $B(\alpha_1, \dots, \alpha_n)$ 或 $B(\boldsymbol{\alpha})$ ，它是对第 301 页 Beta 函数 (4.20) 的推广。

定义 4.36. 区域 $\Delta_{n-1} = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n : \sum_{j=1}^n x_j = 1, x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0\}$ 称为 $(n-1)$ 维单纯形 (simplex)。例如，0, 1, 2, 3 维单纯形分别是点、线段、正三角形和正四面体。单纯形具有一些简单而有趣的几何性质，如 Δ_n 的面就是一些 Δ_{n-1} 。

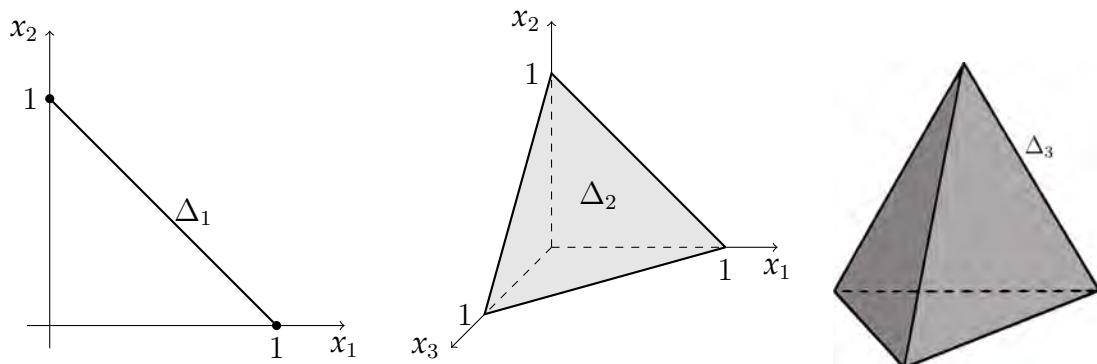
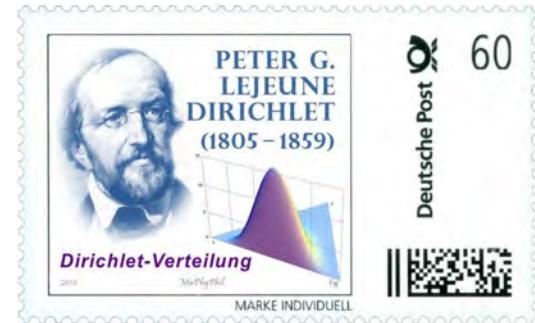


图 4.45: 一维、二维、三维单纯形 $\Delta_1, \Delta_2, \Delta_3$ 。分布 $\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n)$ 的密度函数 (4.23) 在 Δ_{n-1} 上非零。



式 (4.23) 之所以是密度函数, 依赖于多元微积分中的下述事实:

$$\int_{\Delta_{n-1}} \prod_{j=1}^n x_j^{\alpha_j-1} dx_1 \cdots dx_n = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n)} = B(\alpha_1, \dots, \alpha_n) \quad (4.24)$$

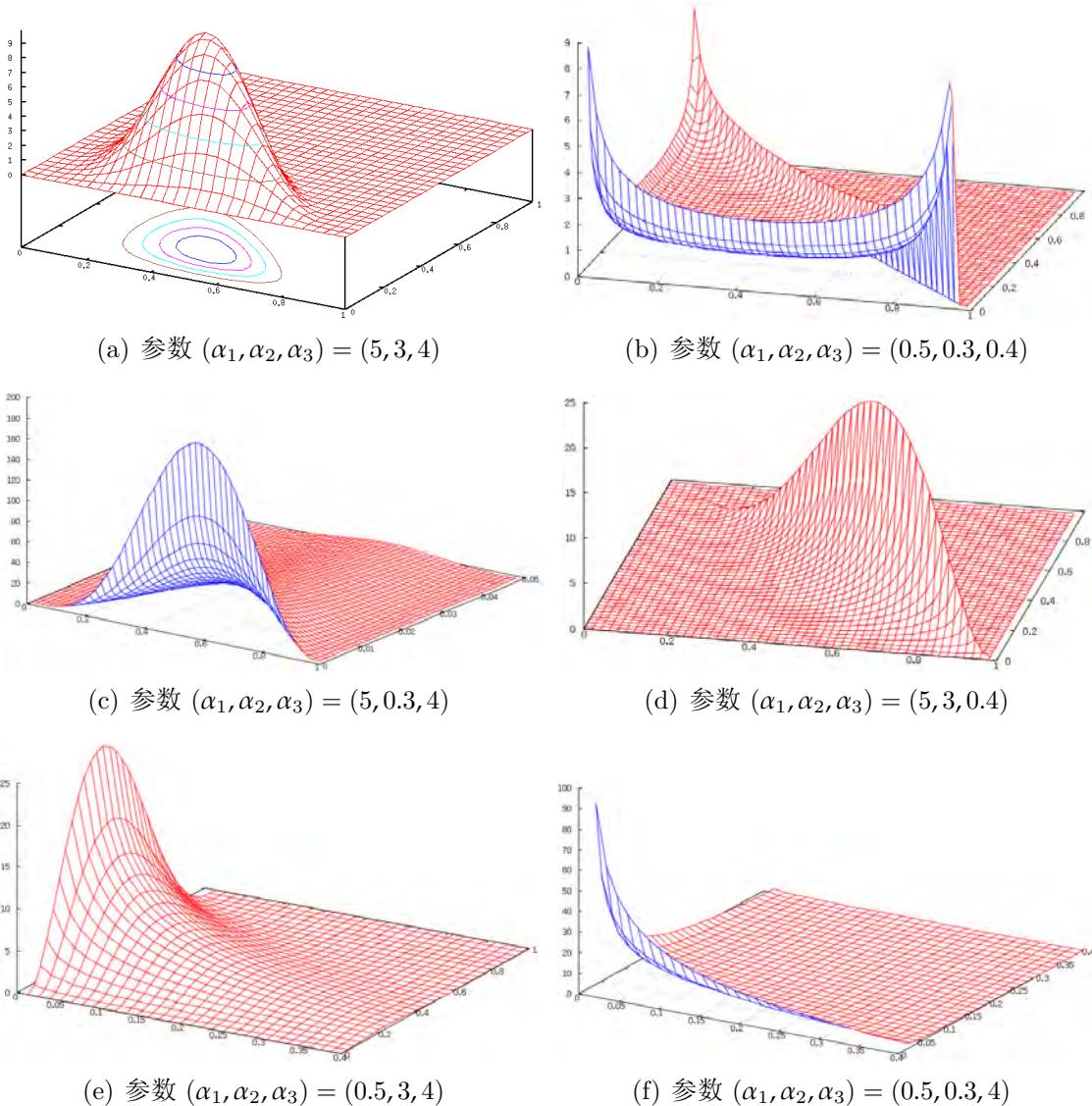


图 4.46: Dirichlet 分布 $(X, Y, 1 - X - Y)^T \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ 的密度函数曲面 $z = f(x, y, 1 - x - y)$ 。

下面, 我们不加证明地介绍 Dirichlet 分布的另一个重要性质, 该性质借用 Pólya 的球-盒子模型给 Dirichlet 分布一个直观的解释。

性质 4.42 (Pólya 的球-盒子模型). 将例 4.15 稍作推广: 盒子里有 n 种颜色的球, 其中第 j 种颜色的球的个数是 $\alpha_j, j = 1, 2, \dots, n$ 。现从盒子中随机地摸出一球并放回两个同颜色的球, 将此过程独立重复至无穷次, 盒子中各种颜色的球所占比例服从分

布 Dirichlet($\alpha_1, \alpha_2, \dots, \alpha_n$)。

定义 4.37. 已知随机向量 $\mathbf{p} = (p_1, \dots, p_k)^\top \sim \text{Dirichlet}(\boldsymbol{\alpha})$, 我们把复合分布 $\mathbf{X} \sim \text{Multin}(n; \mathbf{p})$ 称为 Dirichlet-多项分布, 记作 $\mathbf{X} \sim \text{Dirichlet-Multin}(n; \mathbf{p}; \boldsymbol{\alpha})$ 。特别地, 当 $k = 2$ 时, 也称为 Beta-二项分布, 记作 $\mathbf{X} \sim \text{Beta-B}(n, p; \alpha_1, \alpha_2)$ 。

性质 4.43. 已知 $\mathbf{X} \sim \text{Dirichlet-Multin}(n; p_1, \dots, p_k; \alpha_1, \dots, \alpha_k)$, 试证明:

$$\begin{aligned} P(X_1 = n_1, \dots, X_k = n_k | \boldsymbol{\alpha}) &= \binom{n}{n_1, \dots, n_k} \frac{B(\boldsymbol{\alpha} + \mathbf{n})}{B(\boldsymbol{\alpha})}, \quad \text{其中 } \begin{cases} \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)^\top \\ \mathbf{n} = (n_1, \dots, n_k)^\top \end{cases} \\ &= \frac{n! \Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1 + \dots + \alpha_k + n)} \prod_{j=1}^k \frac{\Gamma(\alpha_j + n_j)}{n_j! \Gamma(\alpha_j)} \end{aligned}$$

证明. 仿照第 276 页的例 4.22 并利用结果 (4.24) 和事实 $n_1 + \dots + n_k = n$,

$$\begin{aligned} P(X_1 = n_1, \dots, X_k = n_k | \boldsymbol{\alpha}) &= \int_{\Delta_{k-1}} P(X_1 = n_1, \dots, X_k = n_k | \mathbf{p}) f(\mathbf{p} | \boldsymbol{\alpha}) d\mathbf{p} \\ &= \binom{n}{n_1, \dots, n_k} \frac{1}{B(\boldsymbol{\alpha})} \int_{\Delta_{k-1}} \prod_{j=1}^k p_j^{\alpha_j + n_j - 1} d\mathbf{p} \\ &= \binom{n}{n_1, \dots, n_k} \frac{B(\boldsymbol{\alpha} + \mathbf{n})}{B(\boldsymbol{\alpha})} \end{aligned} \quad \square$$

在统计语言模型中, Dirichlet-多项分布常用来刻画不同文本类型之下词的计数分布, 详见第 670 页的例 12.33 所介绍的层级贝叶斯模型。

练习 4.68. 绘制 $X \sim \text{Beta-B}(n, p; \alpha, \beta)$ 概率密度函数的折线图, 其中参数 α, β 仿照图 4.26 中 Beta 分布的参数设置, 并与图 4.26 进行比较。

性质 4.44 (中立性). 已知随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_n)$, 我们有

① 随机向量 $(X_1, \dots, X_k)^\top / (X_1 + \dots + X_k) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$, 其中 $k < n$ 。

② 随机向量 $(X_1, \dots, X_k)^\top / (X_1 + \dots + X_k)$ 与 $(X_{k+1}, \dots, X_n)^\top$ 相互独立。

证明. 因为 $X_1 + \dots + X_n = 1$, 所以 $X_1 + \dots + X_k = 1 - X_{k+1} - \dots - X_n$ 。

$$\text{变换 } \begin{cases} Y_1 = \frac{X_1}{1-X_{k+1}-\dots-X_n} \\ \vdots \\ Y_k = \frac{X_k}{1-X_{k+1}-\dots-X_n} \\ Y_{k+1} = X_{k+1} \\ \vdots \\ Y_n = X_n \end{cases} \quad \text{的逆变换是} \quad \begin{cases} X_1 = Y_1(1 - Y_{k+1} - \dots - Y_n) \\ \vdots \\ X_k = Y_k(1 - Y_{k+1} - \dots - Y_n) \\ X_{k+1} = Y_{k+1} \\ \vdots \\ X_n = Y_n \end{cases}$$

其雅可比行列式为

$$\det J\left(\frac{x_1, \dots, x_k, x_{k+1}, \dots, x_n}{y_1, \dots, y_k, y_{k+1}, \dots, y_n}\right) = (1 - y_{k+1} - \dots - y_n)^k$$

由随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的密度函数 $f(x_1, \dots, x_n) \propto \prod_{j=1}^n x_j^{\alpha_j-1}$ 可以得到 $(Y_1, \dots, Y_k, Y_{k+1}, \dots, Y_n)^\top$ 的密度函数如下,

$$g(y_1, \dots, y_k, y_{k+1}, \dots, y_n) \propto \left[(1 - y_{k+1} - \dots - y_n)^{\alpha_1 + \dots + \alpha_k} \prod_{j=k+1}^n y_j^{\alpha_j-1} \right] \prod_{j=1}^k y_j^{\alpha_j-1}$$

由 $(Y_1, \dots, Y_k, Y_{k+1}, \dots, Y_n)^\top$ 密度函数的形式可知, 随机向量 $(Y_1, \dots, Y_k)^\top$ 与 $(Y_{k+1}, \dots, Y_n)^\top$ 相互独立, 并且 $(Y_1, \dots, Y_k)^\top$ 的密度函数正比于 $\prod_{j=1}^k y_j^{\alpha_j-1}$, 即

$$(Y_1, \dots, Y_k)^\top \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$$

□

性质 4.45. 已知随机变量 $Y_j \sim \text{Gamma}(\alpha_j, \beta), j = 1, 2, \dots, n$ 相互独立, 定义随机变量 $Y = \sum_{j=1}^n Y_j$ 和 $X_j = Y_j/Y$, 则随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n)$ 。

证明. 变换 $X_j = Y_j/Y, j = 1, 2, \dots, n$ 的逆变换是 $Y_1 = YX_1, \dots, Y_{n-1} = YX_{n-1}, Y_n = Y(1 - X_1 - X_2 - \dots - X_{n-1})$, 其雅可比行列式为

$$\det J\left(\frac{y_1, y_2, \dots, y_n}{y, x_1, \dots, x_{n-1}}\right) = \begin{vmatrix} x_1 & x_2 & \cdots & x_{n-1} & 1 - \sum_{j=1}^{n-1} x_j \\ y & 0 & \cdots & 0 & -y \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & y & -y \end{vmatrix} = y^{n-1}$$

由随机向量 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ 的密度函数 $f(y_1, y_2, \dots, y_n) \propto \prod_{j=1}^n y_j^{\alpha_j-1} e^{-\beta y_j}$ 可以得到 $(Y, X_1, X_2, \dots, X_{n-1})^\top$ 的密度函数如下,

$$g(y, x_1, x_2, \dots, x_{n-1}) \propto y^{\sum_{j=1}^n \alpha_j-1} e^{-\beta y} \prod_{j=1}^n x_j^{\alpha_j-1}, \text{ 其中 } x_n = 1 - \sum_{j=1}^{n-1} x_j$$

进而得到随机向量 $(X_1, X_2, \dots, X_{n-1})^\top$ 的密度函数为

$$h(x_1, x_2, \dots, x_{n-1}) = \int_0^{+\infty} g(y, x_1, x_2, \dots, x_{n-1}) dy \propto \prod_{j=1}^n x_j^{\alpha_j-1}$$

由分布 $\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n)$ 的密度函数得证。 □



性质 4.45 刻画了 Dirichlet 分布和 Gamma 分布的关系。由**性质 4.25** 可知，在**性质 4.45** 中， $Y \sim \text{Gamma}(\alpha_1 + \dots + \alpha_n, \beta)$ 。进而，**性质 4.45** 可粗略地表示为

$$\text{Dirichlet}(\alpha_1, \dots, \alpha_n) = \left(\frac{\text{Gamma}(\alpha_1, \beta)}{\text{Gamma}(\alpha_1 + \dots + \alpha_n, \beta)}, \dots, \frac{\text{Gamma}(\alpha_n, \beta)}{\text{Gamma}(\alpha_1 + \dots + \alpha_n, \beta)} \right)$$

简而言之，Dirichlet 分布可由一组独立的 Gamma 分布来表示。下面介绍**性质 4.45** 的两个重要应用：一是 Dirichlet 分布的整合性 (aggregation property)，二是 Dirichlet 分布的随机数产生算法。

→**性质 4.46** (整合性). 已知 n 维随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_n)$ ，将两个分量 X_i 和 X_j 合并成一个新的分量 $X_i + X_j$ ，所得到的 $(n - 1)$ 维随机向量 \mathbf{X}' 依然服从 Dirichlet 分布，即

$$\mathbf{X}' = (X_1, \dots, X_i + X_j, \dots, X_n)^\top \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_i + \alpha_j, \dots, \alpha_n)$$

证明. 由**性质 4.45**，存在相互独立的随机变量 $Y_k \sim \text{Gamma}(\alpha_k, \beta), k = 1, 2, \dots, n$ 使得 $X_k = Y_k / Y$ ，其中 $Y = \sum_{k=1}^n Y_k$ 。显然， \mathbf{X}' 的各分量之间相互独立，并且 $X_i + X_j \sim \text{Gamma}(\alpha_i + \alpha_j, \beta)$ 。进而，由**性质 4.45** 证得**性质 4.46**。□

推论 4.4. 已知随机向量 $(X_1, X_2, \dots, X_n)^\top \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n)$ ，则边缘分布 $X_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j), j = 1, 2, \dots, n$ ，其中 $\alpha_0 = \sum_{j=1}^n \alpha_j$ 。此外，

$$E(X_j) = \frac{\alpha_j}{\alpha_0}, \quad V(X_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}, \quad \text{并且} \quad \text{Cov}(X_j, X_k) = -\frac{\alpha_j \alpha_k}{\alpha_0^2(\alpha_0 + 1)}$$

证明. 由**性质 4.46**，二维随机向量 $(X_j, X_1 + \dots + X_{j-1} + X_{j+1} + \dots + X_n)^\top$ ，即 $(X_j, 1 - X_j)^\top$ 服从分布 $\text{Beta}(\alpha_j, \alpha_0 - \alpha_j)$ 。剩下的证明留作习题，请读者补全。□

推论 4.5. 已知 $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_n)$ ，如果 $k < m \leq n$ ，则有

$$\frac{X_1 + \dots + X_k}{X_1 + \dots + X_m} \sim \text{Beta}(\alpha_1 + \dots + \alpha_k, \alpha_{k+1} + \dots + \alpha_m)$$

证明. 由**性质 4.44** 可得

$$\frac{(X_1, \dots, X_m)^\top}{X_1 + \dots + X_m} \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_m)$$

再由**性质 4.46**，便可证得结果。

$$\frac{(X_1 + \dots + X_k, X_{k+1} + \dots, X_m)^\top}{X_1 + \dots + X_m} \sim \text{Beta}(\alpha_1 + \dots + \alpha_k, \alpha_{k+1} + \dots + \alpha_m) \quad \square$$

算法 4.23. 根据性质 4.45, 分布 $\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n)$ 的随机数 $\mathbf{x}^* = (x_1, x_2, \dots, x_n)^\top$ 通过 $x_j = y_j / (y_1 + y_2 + \dots + y_n)$ 产生, 其中 y_j 是独立地抽取自 $\text{Gamma}(\alpha_j, 1)$ 分布的随机数, $j = 1, 2, \dots, n$ 。例子见图 4.47。

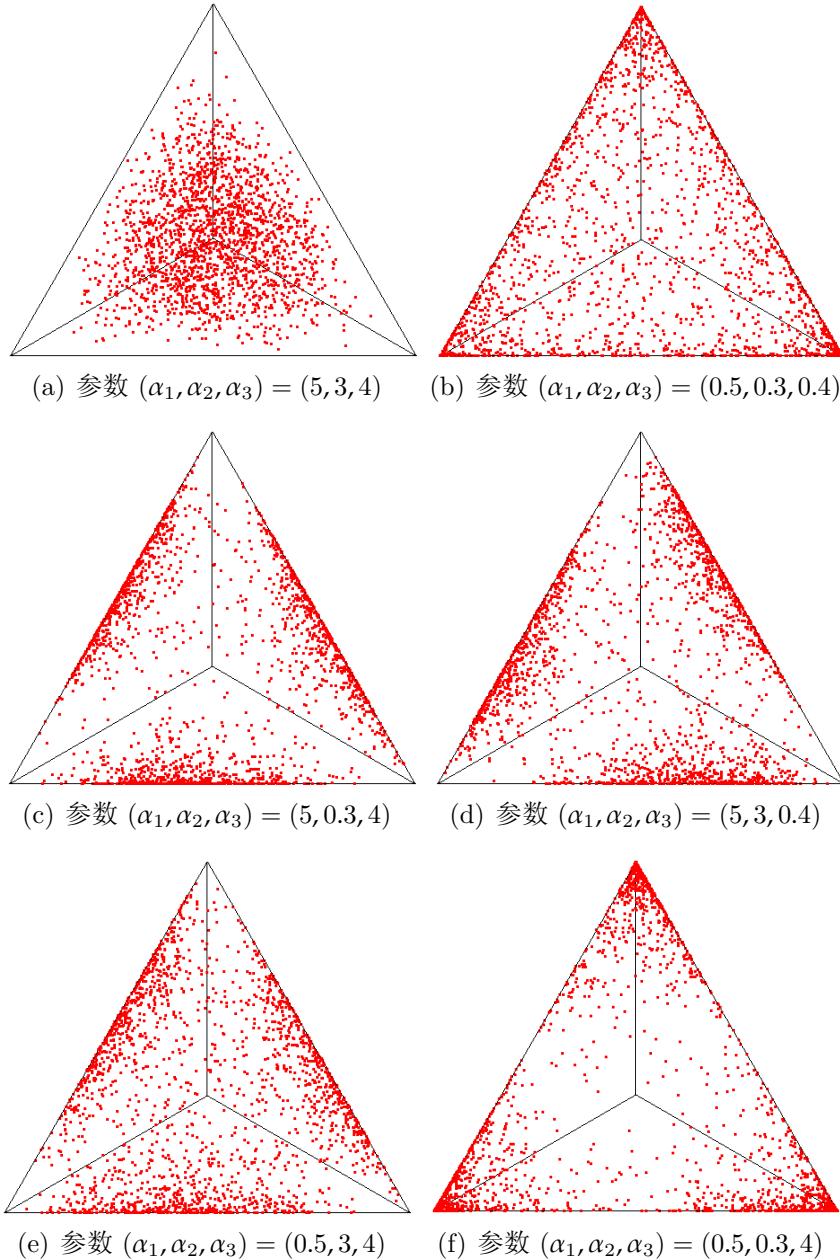


图 4.47: 利用算法 4.23 产生 2000 个图 4.46 所示分布 $\text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ 的随机数。

算法 4.24. $\mathbf{X} \sim \text{Dirichlet-Multin}(n; \mathbf{p}; \boldsymbol{\alpha})$ 的随机数是这样产生的:

- 利用算法 4.23 产生 $\text{Dirichlet}(\boldsymbol{\alpha})$ 的随机数 $\mathbf{p}_1, \dots, \mathbf{p}_m$;
- 利用算法 4.22 产生 $\text{Multin}(n; \mathbf{p}_i)$ 的一个随机数 \mathbf{x}_i , 其中 $i = 1, \dots, m$ 。

4.3.4 多元正态分布与多元 t 分布

定义 4.38 (多元正态分布). 如果连续型随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 使得内积 $\langle \mathbf{t}, \mathbf{X} \rangle = \mathbf{t}^\top \mathbf{X} = t_1 X_1 + t_2 X_2 + \dots + t_n X_n$ 是正态分布或常数, 其中 $\mathbf{t} = (t_1, t_2, \dots, t_n)^\top \in \mathbb{R}^n$ 是任意 n 维列向量, 则称 \mathbf{X} 服从 n 元正态分布或 n 维正态分布。

性质 4.47. 正态分布的随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的所有分量 $X_j, j = 1, 2, \dots, n$ 都服从正态分布, 即多元正态分布的边缘分布亦是正态分布。

若随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 服从 n 元正态分布, 通常记作 $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, 有时为强调维数也记作 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, 其中

$$\begin{aligned}\boldsymbol{\mu} &= (\mu_1, \mu_2, \dots, \mu_n)^\top = (E(X_1), E(X_2), \dots, E(X_n))^\top = E\mathbf{X} \\ \Sigma &= (\sigma_{ij})_{n \times n} = \text{Cov}(\mathbf{X}, \mathbf{X})\end{aligned}$$

由**定义 3.4** 和**例 3.6**, 以及式 (2.84), 得到 $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ 的特征函数为

$$\begin{aligned}\varphi(\mathbf{t}) &= E \exp\{i\mathbf{t}^\top \mathbf{X}\} \\ &= \exp\left\{iE(\mathbf{t}^\top \mathbf{X}) - \frac{1}{2}V(\mathbf{t}^\top \mathbf{X})\right\} \\ &= \exp\left\{i\mathbf{t}^\top E(\mathbf{X}) - \frac{1}{2}\mathbf{t}^\top \text{Cov}(\mathbf{X}, \mathbf{X})\mathbf{t}\right\} \\ &= \exp\left\{i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}\right\}\end{aligned}\tag{4.25}$$

当 Σ 为正定矩阵时, 行列式 $\det(\Sigma) > 0$ 。利用随机向量的 Lévy 反演公式 (3.20), 得到 n 元正态分布随机向量 \mathbf{X} 的密度函数如下:

$$\phi(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \text{ 其中 } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\tag{4.26}$$

式 (4.26) 即是第 215 页的式 (2.97), 我们已经证明它是概率密度函数。

练习 4.69. **例 2.22** 所描述的随机向量 $(X, Y)^\top \sim N(\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的密度函数 (2.22) 可表示为 (4.26) 的形式, 其中 $\boldsymbol{\mu} = (\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y)^\top$, 协方差矩阵 Σ 及其逆矩阵分别为

$$\begin{aligned}\Sigma &= \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \\ \Sigma^{-1} &= \frac{1}{1-\rho^2} \begin{pmatrix} \sigma_X^{-2} & -\rho\sigma_X^{-1}\sigma_Y^{-1} \\ -\rho\sigma_X^{-1}\sigma_Y^{-1} & \sigma_Y^{-2} \end{pmatrix}\end{aligned}$$

例 4.35. 性质 4.47 的逆命题不成立, 即随机向量 \mathbf{X} 的边缘分布都是正态分布并不能推导出 \mathbf{X} 也服从正态分布。例如, 下图所示的分布。

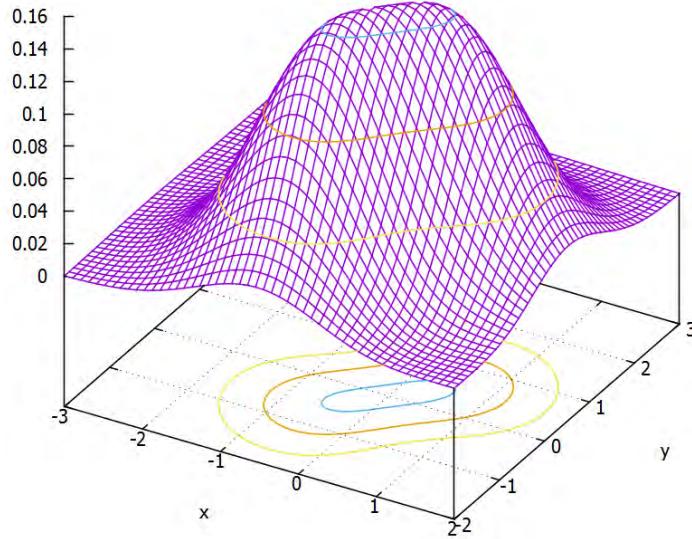


图 4.48: 密度函数为 $f(x, y) = \frac{1}{2\pi} \exp\left\{-\frac{x^2+y^2}{2}\right\}(1 + \sin x \sin y)$ 的随机向量 $(X, Y)^\top$ 非正态分布, 但它的任一边缘分布都是标准正态分布。

定理 4.10. 正态分布的随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top \sim N_n(\boldsymbol{\mu}, \Sigma)$ 经过线性变换所得的新随机向量 $A_{m \times n}\mathbf{X}$ 仍为正态分布, 其中 $m \leq n$ 。

$$\mathbf{Y} = A_{m \times n}\mathbf{X} \sim N_m(A\boldsymbol{\mu}, A\Sigma A^\top)$$

证明. 在式 (4.25) 中令 $\mathbf{t} = A^\top \mathbf{s}$ 便得到随机向量 \mathbf{Y} 的特征函数为

$$\begin{aligned}\varphi(s_1, s_2, \dots, s_m) &= E \exp\{is^\top \mathbf{Y}\} \\ &= E \exp\{is^\top A\mathbf{X}\} \\ &= E \exp\{i\mathbf{t}^\top \mathbf{X}\} \\ &= \exp\left\{i\mathbf{t}^\top \mathbf{X} - \frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}\right\} \\ &= \exp\left\{i\mathbf{s}^\top (A\boldsymbol{\mu}) - \frac{1}{2}\mathbf{s}^\top (A\Sigma A^\top)\mathbf{s}\right\}\end{aligned}$$

具有这个特征函数的分布只有多元正态分布 $N_m(A\boldsymbol{\mu}, A\Sigma A^\top)$ 。 □

练习 4.70. 已知 $\boldsymbol{\alpha} \in \mathbb{R}^m$ 是常向量, 在定理 4.10 的条件之下, 试证明:

$$A\mathbf{X} + \boldsymbol{\alpha} \sim N(A\boldsymbol{\mu} + \boldsymbol{\alpha}, A\Sigma A^\top)$$

练习 4.71. 利用定理 4.10 来解决第 138 页的例 2.23 的问题。

推论 4.6. 如果 Σ 为 $n \times n$ 阶对称正定矩阵, 设 Σ 有形如 $\Sigma = A^\top A$ 的分解*, 其中 A 是 $n \times n$ 阶矩阵, 则

$$\mathbf{X} = \boldsymbol{\mu} + A^\top \mathbf{Z} \sim N(\boldsymbol{\mu}, \Sigma)$$

其中随机向量 $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)^\top$ 满足 $Z_1, Z_2, \dots, Z_n \stackrel{\text{iid}}{\sim} N(0, 1)$ 。

算法 4.25. 利用**推论 4.6**, 构造多元正态分布 $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ 的随机数

$$\mathbf{x}^* = \boldsymbol{\mu} + A^\top \mathbf{z}^*$$

其中, 方阵 $A_{n \times n}$ 满足 $\Sigma = A^\top A$, 向量 $\mathbf{z}^* = (z_1, z_2, \dots, z_n)^\top$ 是独立地抽取自 $N(0, 1)$ 的 n 个随机数。在实践中, 考虑到算法的稳定性, 矩阵 A 常通过 Σ 的奇异值分解 (详见第 768 页的**定理 E.5**) 得到。

定理 4.11. 已知随机向量 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, 其中 $|\Sigma| > 0$, 则

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_n^2$$

证明. 因为矩阵 Σ 是对称正定矩阵, 所以存在奇异值分解 $\Sigma = U\Lambda U^\top$, 其中 $U = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ 是正交矩阵, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ 是 Σ 的特征值构成的对角矩阵, 满足 $\lambda_1 \geq \dots \geq \lambda_n > 0$ 。矩阵 Σ^{-1} 具有奇异值分解

$$\Sigma^{-1} = U\Lambda^{-1}U^\top, \text{ 进而有}$$

$$\begin{aligned} (\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) &= (\mathbf{X} - \boldsymbol{\mu})^\top U\Lambda^{-1}U^\top (\mathbf{X} - \boldsymbol{\mu}) \\ &= [A(\mathbf{X} - \boldsymbol{\mu})]^\top [A(\mathbf{X} - \boldsymbol{\mu})], \text{ 其中 } A = \Lambda^{-1/2}U^\top \end{aligned}$$

令 $\mathbf{Z} = A(\mathbf{X} - \boldsymbol{\mu})$, 则 $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)^\top$ 服从正态分布, 显然其均值为零, 其协方差阵分别是

$$\begin{aligned} \text{Cov}(\mathbf{Z}, \mathbf{Z}) &= A \text{Cov}(\mathbf{X} - \boldsymbol{\mu}, \mathbf{X} - \boldsymbol{\mu}) A^\top \\ &= A \Sigma A^\top \\ &= \Lambda^{-1/2} U^\top U \Lambda U^\top U \Lambda^{-1/2} \\ &= I \end{aligned}$$

因此, $\mathbf{Z} \sim N(\mathbf{0}, I)$, 即 $Z_1, Z_2, \dots, Z_n \stackrel{\text{iid}}{\sim} N(0, 1)$ 。由**性质 4.26**, 显然 $\|\mathbf{Z}\|^2 = \mathbf{Z}^\top \mathbf{Z} = \sum_{j=1}^n Z_j^2 \sim \chi_n^2$, 即 $(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_n^2$ 得证。 \square

*譬如, Cholesky 分解 $\Sigma = Q^\top Q$, 其中 Q 为上三角矩阵, 见第 767 页的**定理 E.3**。

\nwarrow 定理 4.12. 已知 $A_{n \times n}$ 是对称幂等矩阵, 其秩为 r 。若随机向量 $\mathbf{X} \sim N_n(\mathbf{0}, I_n)$, 则

$$\mathbf{X}^\top A \mathbf{X} \sim \chi_r^2$$

证明. 由矩阵 A 的已知条件, 存在分解 $A = UU^\top$, 其中 $U_{n \times r}$ 是秩为 r 的正交矩阵 (见附录 E 的性质 E.2)。于是

$$\mathbf{X}^\top A \mathbf{X} = \mathbf{X}^\top UU^\top \mathbf{X} = \|U^\top \mathbf{X}\|^2$$

与定理 4.11 的证明类似, 只需说明随机向量 $U^\top \mathbf{X} \sim N_r(\mathbf{0}, I_r)$ 。显然, $U^\top \mathbf{X}$ 的均值为零。另外, 其协方差阵为

$$\text{Cov}(U^\top \mathbf{X}, U^\top \mathbf{X}) = U^\top I_n U = I_r$$

例 4.36. 设 n 维随机向量 $\mathbf{X} = (X_1, \dots, X_m, X_{m+1}, \dots, X_n)^\top \sim N(\boldsymbol{\mu}, \Sigma)$ 且参数 $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_1 & O \\ O & \Sigma_2 \end{pmatrix}$, 其中 $\boldsymbol{\mu}_{(1)}$ 是 m 维列向量, $\boldsymbol{\mu}_{(2)}$ 是 $n-m$ 维列向量, Σ_1 是 m 阶方阵, Σ_2 是 $n-m$ 阶方阵, O 表示零矩阵。试证明:

$$\mathbf{X}_{(1)} = (X_1, \dots, X_m)^\top \sim N(\boldsymbol{\mu}_{(1)}, \Sigma_1)$$

$$\mathbf{X}_{(2)} = (X_{m+1}, \dots, X_n)^\top \sim N(\boldsymbol{\mu}_{(2)}, \Sigma_2)$$

证明. 不妨设 $\mathbf{x} = (x_1, \dots, x_n)^\top$, $\mathbf{x}_{(1)} = (x_1, \dots, x_m)^\top$, $\mathbf{x}_{(2)} = (x_{m+1}, \dots, x_n)^\top$ 。由已知条件, 随机向量 $\mathbf{X} = (X_1, \dots, X_m, X_{m+1}, \dots, X_n)^\top$ 的概率密度函数为

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^m |\Sigma_1|}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})^\top \Sigma_1^{-1} (\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}) \right\} \times \\ &\quad \frac{1}{\sqrt{(2\pi)^{n-m} |\Sigma_2|}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)})^\top \Sigma_2^{-1} (\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}) \right\} \end{aligned}$$

于是, 随机向量 $\mathbf{X}_{(1)} = (X_1, \dots, X_m)^\top$ 的密度函数为

$$f(\mathbf{x}_{(1)}) = \int_{\mathbb{R}^{n-m}} f(\mathbf{x}) d\mathbf{x}_{(2)} = \frac{1}{\sqrt{(2\pi)^m |\Sigma_1|}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})^\top \Sigma_1^{-1} (\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}) \right\}$$

即 $(X_1, \dots, X_m)^\top \sim N(\boldsymbol{\mu}_{(1)}, \Sigma_1)$, 同理 $(X_{m+1}, \dots, X_n)^\top \sim N(\boldsymbol{\mu}_{(2)}, \Sigma_2)$ 。 \square

性质 4.48. 设 $\mathbf{X} = (X_1, \dots, X_m, \dots, X_n)^\top \sim N(\boldsymbol{\mu}, \Sigma)$, 试证明: 对于 $1 < m < n$, 皆有

$$(X_1, \dots, X_m)^\top \sim N(\boldsymbol{\mu}_{(1)}, \Sigma_{11})$$

其中, $\boldsymbol{\mu}_{(1)}$ 是 $\boldsymbol{\mu}$ 的前 m 个分量构成的列向量, Σ_{11} 是 Σ 的左上角的 m 阶子矩阵。

证明. 不妨设 $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$ 且 $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ 。令 $(Y_1, \dots, Y_n)^\top = \begin{pmatrix} I_m & O \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix} \mathbf{X}$, 其中 I_m 是 m 阶单位阵, 则 $(Y_1, \dots, Y_m) = (X_1, \dots, X_m)$ 。因为

$$\begin{pmatrix} I_m & O \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_m & O \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix}^\top = \begin{pmatrix} \Sigma_{11} & O \\ O & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix}$$

由定理 4.10 和例 4.36 可证得。 \square

定理 4.13. 已知随机向量 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, 将 $\mathbf{X}, \boldsymbol{\mu}, \Sigma$ 作如下相应的分块: $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$ 且 $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, 其中 $\mathbf{X}_{(1)}, \boldsymbol{\mu}_{(1)}$ 都是 m 维列向量, Σ_{11} 是 $m \times m$ 阶矩阵。若 Σ 正定且 Σ_{22} 可逆, 则

① 在给定 $\mathbf{X}_{(2)} = \mathbf{x}_{(2)}$ 的条件下, $\mathbf{X}_{(1)}$ 的条件分布

$$\mathbf{X}_{(1)} | \mathbf{X}_{(2)} = \mathbf{x}_{(2)} \sim N_m \left(\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$$

② 随机向量 $\mathbf{X}_{(1)}$ 与 $\mathbf{X}_{(2)}$ 相互独立当且仅当 $\Sigma_{12} = O_{m \times (n-m)}$ (第 204 页的例 2.85 是二元正态分布时的特例)。

③ 随机向量 $\mathbf{X}_{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{X}_{(2)}$ 与 $\mathbf{X}_{(2)}$ 相互独立。

证明. 因为分块矩阵 Σ 正定, 所以它的逆矩阵存在, 按照附录 E 中的定理 E.8 不妨设 $\Sigma^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$, 其中 $S_{21} = S_{12}^\top$ 。于是 $\Delta = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})$ 可进一步分解为

$$\begin{aligned} \Delta &= -\frac{1}{2}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})^\top S_{11}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}) - \frac{1}{2}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})^\top S_{12}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}) \\ &\quad - \frac{1}{2}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)})^\top S_{21}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}) - \frac{1}{2}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)})^\top S_{22}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}) \\ &= -\frac{1}{2}\mathbf{x}_{(1)}^\top S_{11}\mathbf{x}_{(1)} + \mathbf{x}_{(1)}^\top \{S_{11}\boldsymbol{\mu}_{(1)} - S_{12}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)})\} + \text{不含 } \mathbf{x}_{(1)} \text{ 的项} \end{aligned}$$

(1) 把 Δ 视作有关 $\mathbf{x}_{(1)}$ 的函数, 不难看出 $\mathbf{X}_{(1)} | \mathbf{X}_{(2)} = \mathbf{x}_{(2)}$ 依然是正态分布, 不妨设为 $N_m(\boldsymbol{\mu}_{1|2}, \Sigma_{1|2})$ 。

$$-\frac{1}{2}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{1|2})^\top \Sigma_{1|2}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{1|2}) = -\frac{1}{2}\mathbf{x}_{(1)}^\top \Sigma_{1|2}^{-1}\mathbf{x}_{(1)} + \mathbf{x}_{(1)}^\top \Sigma_{1|2}^{-1}\boldsymbol{\mu}_{1|2} + \text{常数}$$

对比上式和 Δ 中 $\mathbf{x}_{(1)}$ 的二次项, 由定理 E.8 得到

$$\Sigma_{1|2} = S_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

即矩阵 $\Sigma = (\Sigma_{11}, \Sigma_{12}; \Sigma_{21}, \Sigma_{22})$ 关于 Σ_{22} 的 Schur 补 (见第 770 页)。类似地, 对

比 $\boldsymbol{x}_{(1)}$ 的一次项，不难得到

$$\begin{aligned}\boldsymbol{\mu}_{1|2} &= \Sigma_{1|2}\{S_{11}\boldsymbol{\mu}_{(1)} - S_{12}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)})\} \\ &= \boldsymbol{\mu}_{(1)} - S_{11}^{-1}S_{12}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)}) \\ &= \boldsymbol{\mu}_{(1)} - S_{11}^{-1}S_{11}\Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)}) \\ &= \boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)})\end{aligned}$$

(2) 若 $\Sigma_{12} = O_{m \times (n-m)}$ ，则 $\mathbf{X}_{(1)}|\mathbf{X}_{(2)} = \boldsymbol{x}_{(2)} \sim N_m(\boldsymbol{\mu}_{(1)}, \Sigma_{11})$ ，即 $\mathbf{X}_{(1)}$ 与 $\mathbf{X}_{(2)}$ 相互独立。

(3) 随机向量 $\begin{pmatrix} \mathbf{X}_{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{X}_{(2)} \\ \mathbf{X}_{(2)} \end{pmatrix}$ 服从正态分布，这是因为

$$\begin{pmatrix} \mathbf{X}_{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{X}_{(2)} \\ \mathbf{X}_{(2)} \end{pmatrix} = \begin{pmatrix} I_{m \times m} & -\Sigma_{12}\Sigma_{22}^{-1} \\ O & I_{(n-m) \times (n-m)} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{pmatrix}$$

经过简单的验证（请读者来完成它），随机向量 $\mathbf{X}_{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{X}_{(2)}$ 和 $\mathbf{X}_{(2)}$ 的协方差矩阵为 $\begin{pmatrix} \Sigma_{11} & O \\ O & \Sigma_{22} \end{pmatrix}$ ，由本定理第二个结论，这两个随机向量相互独立。□

定义 4.39. 已知 $\mathbf{X} \sim N_d(\mathbf{0}, \Sigma)$ 与 $Y \sim \chi_n^2$ 相互独立，令 $\boldsymbol{\mu}$ 为 d 维常向量，随机向量 $\mathbf{T} = \boldsymbol{\mu} + \mathbf{X}/\sqrt{Y/n}$ 的分布称为自由度为 n 的多元 t 分布，记作 $\mathbf{T} \sim t_n(\boldsymbol{\mu}, \Sigma)$ 。特别地，当 $n = 1$ 时，称为多元 Cauchy 分布。 $\mathbf{T} \sim t_n(\boldsymbol{\mu}, \Sigma)$ 的概率密度函数（留作习题）是

$$f_n(\mathbf{t}) = \frac{\Gamma\left(\frac{n+d}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n^d\pi^d|\Sigma|}} \left[\frac{1}{n}(\mathbf{t} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{t} - \boldsymbol{\mu}) + 1 \right]^{-\frac{n+d}{2}} \quad (4.27)$$

多元 t 分布可应用于贝叶斯多元分析、聚类分析、判别分析、回归分析、缺失数据分析等 [99]。

4.3.5 随机矩阵与 Wishart 分布

与随机向量的定义类似，随机矩阵就是取值为矩阵的随机变量，我们同样可以谈论随机矩阵的分布。例如，

例 4.37. 考虑 $n \times n$ 对称矩阵 $A = (a_{ij})_{n \times n}$ 的全体，其元素独立同分布于某一给定的分布 $p(x)$ ，则观察到矩阵 $A = (a_{ij})_{n \times n}$ 的概率是

$$p(A) = \prod_{1 \leq i \leq j \leq n} p(a_{ij})$$

例 4.38. 已知 $n \times n$ 随机矩阵 A 的元素独立同分布于 $N(0, 1)$ ，定义新的随机矩阵 $W = (A + A^\top)/2$ ，则 W 的对角线元素独立同分布于 $N(0, 1)$ ，非对角线元素独立同分布于 $N(0, 1/2)$ 。随机矩阵 W 被称为 Wigner 矩阵，其特征值的分布参见**例 4.26**。

1928 年，英国统计学家 John Wishart (1898-1956) 首次提出了一类随机矩阵，其分布被后人称为 Wishart 分布。Wishart 分布由多元正态分布导出，可视作 χ_n^2 分布的矩阵推广。因为 Wishart 分布恰是多元正态分布协方差矩阵最大似然估计的概率分布 [83, 129]，所以它在多元统计学中非常重要，该分布的发现揭开了多元统计学的篇章。

定义 4.40. 已知 d 维随机向量 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \stackrel{\text{iid}}{\sim} N_d(\mathbf{0}, \Sigma)$ ，其中 d, n 是正整数， Σ 是 $d \times d$ 半正定矩阵，则称 $d \times d$ 半正定的随机矩阵 $W = \sum_{j=1}^n \mathbf{X}_j \mathbf{X}_j^\top$ 服从自由度为 n 的 Wishart 分布^{*}，记作 $W \sim \text{Wishart}_n(\Sigma, d)$ ，当无需强调维数 d 时也简记作 $W \sim \text{Wishart}_n(\Sigma)$ 。显然，当 Σ 退化为 1 时， $W \sim \chi_n^2$ 。当 $n \geq d$ 且 Σ 正定时，称分布 $\text{Wishart}_n(\Sigma, d)$ 是非退化的，其密度函数为

$$p(W) = \frac{|\Sigma|^{-n/2} |W|^{(n-d-1)/2} \exp\{-\frac{1}{2}\text{tr}(\Sigma^{-1}W)\}}{2^{nd/2} \pi^{d(d-1)/4} \prod_{j=1}^d \Gamma\left(\frac{n+1-j}{2}\right)}, \text{ 其中 } W \text{ 是正定矩阵}$$



性质 4.49. 如果随机矩阵 $W_1 \sim \text{Wishart}_{n_1}(\Sigma)$ 与 $W_2 \sim \text{Wishart}_{n_2}(\Sigma)$ 相互独立，则

$$W_1 + W_2 \sim \text{Wishart}_{n_1+n_2}(\Sigma)$$

定理 4.14. 已知 $W \sim \text{Wishart}_n(\Sigma, d)$ 非退化且 C 是一个 $p \times d$ 矩阵，则

$$CWC^\top \sim \text{Wishart}_n(C\Sigma C^\top, p)$$

特别地，令 $\mathbf{c} \in \mathbb{R}^d$ 使得 $\sigma^2 = \mathbf{c}^\top \Sigma \mathbf{c} \neq 0$ ，则 $\mathbf{c}^\top W \mathbf{c} / \sigma^2 \sim \chi_n^2$ 。

^{*}模仿 χ_n^2 分布的记法，我们在此书中把 Wishart 分布的自由度 n 也放在脚标的位置。有些文献把维数 d 放在脚标的位置，请读者阅读时注意区分。

证明. $CWC^\top = \sum_{j=1}^n C\mathbf{X}_j(C\mathbf{X}_j)^\top$, 因为 $C\mathbf{X}_j \sim N_p(\mathbf{0}, C\Sigma C^\top)$ 得证。 \square

性质 4.50. 令 I_d 是 d 阶单位阵, 分布 $W \sim \text{Wishart}_n(\Sigma, d)$ 的特征函数为

$$\varphi(T_{d \times d}) = E\{i\langle T, W \rangle\} = E\{i \cdot \text{tr}(TW)\} = |I_d - 2iT\Sigma|^{-n/2}$$

4.4 习题

- 4.1. 有一个程序可以产生 $p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 的随机数, 仅仅利用该程序如何产生 $0.5\langle 1 \rangle + 0.5\langle 0 \rangle$ 的随机数?
- 4.2. 试求第 258 页的例 4.9 所定义的离散随机变量 Y 的期望和方差。
- 4.3. 试求离散均匀分布 $X \sim \frac{1}{n}\langle 1 \rangle + \frac{1}{n}\langle 2 \rangle + \cdots + \frac{1}{n}\langle n \rangle$ 的期望、方差、偏度系数、峰度系数、变异系数。
- 4.4. 已知 $X_n \sim \text{Poisson}(r^n/n)$, $n = 1, 2, \dots$ 相互独立, 其中 $r \in (0, 1)$ 。求证:

$$Y = \sum_{n=1}^{\infty} nX_n \sim \text{Geom}(r)$$

- 4.5. 设随机变量 $X \sim U(0, 1)$ 。现有常数 $0 < a < 1$, 如果任取 4 个随机数, 至少有一个大于 a 的概率为 0.9, 问: a 为多少?
- 4.6. 已知连续型随机变量 X 的某些观测值的直方图, 请问如何从该直方图产生出更多 X 的随机数?
- 4.7. 设连续型随机变量 X 的密度函数是 $p\phi(x|\mu_1, \sigma_1^2) + (1-p)\phi(x|\mu_2, \sigma_2^2)$, 其中常数 $0 < p < 1$, 如何产生 X 的随机数 x^* ?
- ☆ 4.8. 若随机变量 X 的密度函数是 $f(x) = \begin{cases} \phi(x)/[1 - \Phi(r)] & \text{当 } x > r \\ 0 & \text{其他} \end{cases}$
则称 X 服从截尾正态分布。求 X 的特征函数、期望和方差。
- 4.9. 设随机变量 $X \sim U[-1/2, 1/2]$, 令 $g(x) = \begin{cases} \ln x & \text{当 } x > 0 \\ 0 & \text{当 } x \leq 0 \end{cases}$
求随机变量 $Y = g(X)$ 的期望与方差。
- 4.10. 设 $X \sim U[0, 1]$, 求单调增函数 $h(x)$ 使得 $Y = h(X) \sim \text{Expon}(\beta)$ 。
- 4.11. 已知 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(1)$, 试证明 $2(X_1 + \cdots + X_n) \sim \chi_{2n}^2$ 。
- 4.12. 已知 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} U(0, 1)$, 试证明 $-2 \ln(X_1 X_2 \cdots X_n) \sim \chi_{2n}^2$ 。
- 4.13. 设随机变量 $X \sim N(0, 1)$, 求 X^n 的数学期望和方差, 其中 $n \in \mathbb{N}$ 。
- ☆ 4.14. 设随机变量 $X, Y \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 求 $E[\max(X, Y)]$ 和 $E[\min(X, Y)]$ 。
- 4.15. 已知随机向量 $(X, Y)^\top \sim N(0, 0, 1, 1, \rho)$, 求 $E[\max(X, Y)]$ 。

- 4.16. 验证图 4.48 中给出的联合密度 $f(x, y)$ 使得边缘分布为正态分布。
- 4.17. 设随机变量 X 的密度函数为 $f(x) = e^{-|x|}/2$, 其中 $-\infty < x < +\infty$ 。求 $E(|X|)$ $V(|X|)$ 和 $\text{Cov}(X, |X|)$, 并判定 X 与 $|X|$ 是否独立。
- ☆ 4.18. 若 $X \sim \text{Cauchy}(0, 1)$, 试证明 $Y = 2X/(1 - X^2) \sim \text{Cauchy}(0, 1)$ 。
- ☆ 4.19. 已知随机变量 $X \sim \text{Gamma}(\alpha, \beta)$ 和条件分布 $Y|X = x \sim \text{Poisson}(x)$, 求 Y 的分布列。
- 4.20. 已知随机变量 $X \sim \text{Gamma}(\alpha_1, \beta)$ 与 $Y \sim \text{Gamma}(\alpha_2, \beta)$ 相互独立, 求随机变量 $Z = X/Y$ 的密度函数。
- 4.21. 已知 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(\beta)$, 令 $U = \max(X_1, \dots, X_n)$ 和 $V = \min(X_1, \dots, X_n)$,
(1) 试求 U 的密度函数; (2) 试证 $V \sim \text{Expon}(n\beta)$ 。
- 4.22. 已知 $X \sim \text{Expon}(\beta)$, 如果 $P\{X \geq 1\} = P\{X \leq 1\}$, 求 $\sum_{k=1}^{\infty} P\{X \geq k\}$ 。
- 4.23. 利用逆 CDF 法给出分布 $\text{Pareto}(\alpha, \mu)$ 的随机数产生算法。
- 4.24. 试证明: Cauchy 分布、对数正态分布、 t 分布、 F 分布、 $0 < \alpha < 1$ 时的 Weibull(λ, α) 都是重尾分布。
- 4.25. 设随机变量 $X \sim \text{Rayleigh}(\sigma)$, 求 $E(1/X)$ 。
- ☆ 4.26. 已知随机向量 $\mathbf{X} = (X_1, \dots, X_k)^T \sim \text{Multin}(n; p_1, \dots, p_k)$ 且 $m_1 < m_2 < k$, 令 $Y_1 = X_1 + \dots + X_{m_1}, Y_2 = X_1 + \dots + X_{m_2}$, 则
- $$Y_1|Y_2 = y_2 \sim \text{B}\left(y_2, \frac{p_1 + \dots + p_{m_1}}{p_1 + \dots + p_{m_2}}\right)$$
- 4.27. 试证明第 326 页的推论 4.3。
- 4.28. 请读者补全第 332 页的推论 4.4 的证明。
- 4.29. 已知 $(X_1, X_2, \dots, X_n)^T \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n)$ 且 $k < n$, 试证明:
(a) 随机向量 $(X_1, \dots, X_k)^T$ 和 $(X_{k+1}, \dots, X_n)^T / (1 - X_1 - \dots - X_k)$ 相互独立。
(b) 随机变量 $X_1, X_2 / (1 - X_1), \dots, X_{k-1} / (1 - X_1 - \dots - X_{k-2})$ 相互独立, 并且
- $$\frac{X_k}{1 - X_1 - \dots - X_{k-1}} \sim \text{Beta}(\alpha_k, \alpha_1 + \dots + \alpha_{k-1})$$
- ☆ 4.30. 独立进行 n 次 k -分类试验 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} p_1\langle 1 \rangle + \dots + p_k\langle k \rangle$, 其中未知参数 $\mathbf{p} = (p_1, \dots, p_k)^T \sim \text{Dirichlet}(\boldsymbol{\alpha})$, 试计算 $P(X_1 = x_1, \dots, X_n = x_n | \boldsymbol{\alpha})$ 。

- ☆ 4.31. 设 $(X, Y)^\top$ 服从二元正态分布, 且有 $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$ 。证明: 当 $a^2 = \sigma_X^2/\sigma_Y^2$ 时, $W = X - aY$ 与 $V = X + aY$ 相互独立。
- 4.32. 设随机向量 $(X, Y, Z)^\top \sim N(\mu, \Sigma)$, 其中均值为 $\mu = (3, 5, 7)^\top$, 协方差矩阵为 $\Sigma = (8, 3, 2; 3, 4, 1; 2, 1, 2)$, 求 $X + Y$ 的密度函数。
- 4.33. 如果 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top \sim N(\mu, \Sigma)$, 则 X_1, X_2, \dots, X_n 相互独立的充分必要条件是 $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$, 其中 σ_j^2 是 X_j 的方差, $j = 1, 2, \dots, n$ 。
- ☆ 4.34. 试证明多元 t 分布的密度函数 (4.27)。
- 4.35. 求分布 $W \sim \text{Wishart}_n(\Sigma, d)$ 的期望。

第五章

大数律与中心极限定理

孤帆远影碧空尽，唯见长江天际流。

李白《黄鹤楼送孟浩然之广陵》

在十七世纪末，瑞士数学家 Jacob Bernoulli (1654-1705) 发现了 Bernoulli 弱大数律（见第 62 页的定理 1.1），首次严格地给出了概率的频率解释。通过例 1.48 的模拟试验，读者对结果 (1.16) 的含义也有了直观的理解：随机事件 A 的概率是在 n 次独立的重复试验中，事件 A 发生的频率 m/n 的“稳定值”。1837 年，法国数学家 S. D. Poisson 把 Bernoulli 弱大数律推广为 Poisson 弱大数律。

定理 5.1 (Poisson 弱大数律，1837). 随机事件 A 在 n 次独立的试验中出现了 m 次，令 p_j 是事件 A 在第 j 次试验中出现的概率，则 $\forall \epsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{m}{n} - \frac{p_1 + \dots + p_n}{n} \right| \leq \epsilon \right\} = 1$$

Poisson 把这一结果定名为“大数律”，而对它的严格证明是 Chebyshev 于 1846 年给出的。定义随机变量 X_j 如下：

$$X_j = \begin{cases} 1 & \text{若事件 } A \text{ 在第 } j \text{ 次试验中出现} \\ 0 & \text{若事件 } A \text{ 在第 } j \text{ 次试验中不出现} \end{cases} \quad (5.1)$$

在 Bernoulli 和 Poisson 弱大数律中， $m = \sum_{j=1}^n X_j$ ，进而它们有了“随机变量版”的表达形式：

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n EX_j \right| \leq \epsilon \right\} = 1 \quad (5.2)$$

1867 年, Chebyshev 把 Bernoulli 和 Poisson 弱大数律做了推广。后来 A. A. Markov 又进一步推广了 Chebyshev 的结果, 并建议把 Bernoulli 弱大数律的所有推广统称为“大数律”, 于是提炼出下面的概念。

定义 5.1 (弱大数律). 如果 $\forall \epsilon > 0$, 随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 满足条件 (5.2), 则称随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足弱大数律 (weak law of large numbers) 或大数律 (law of large numbers, LLN)。

定义 5.2 (依概率收敛). 随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 依概率收敛 (converge in probability) 于随机变量 X , 记为 $X_n \xrightarrow{P} X$, 当且仅当 $\forall \epsilon > 0$,

$$\lim_{n \rightarrow \infty} P\{|X_n - X| \leq \epsilon\} = 1 \quad (5.3)$$

它意味着 X_n 与 X 差距大于 ϵ 的机会随 n 的增加而趋于 0。譬如, $\{X_j\}_{j=1}^{\infty}$ 满足弱大数律意味着

$$Y_n = \frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n E(X_j) \xrightarrow{P} 0$$

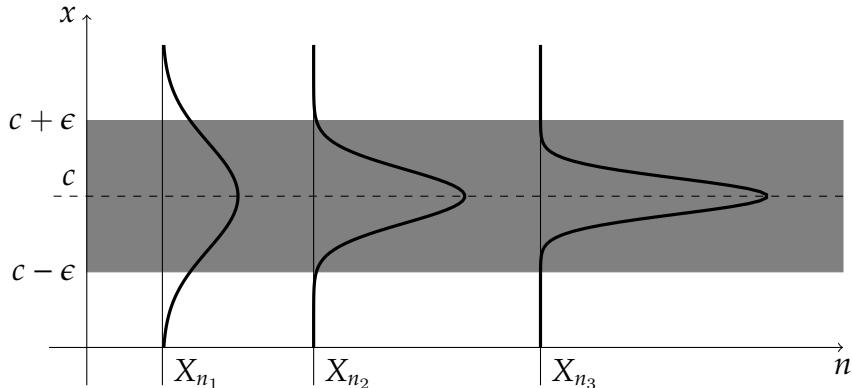


图 5.1: 依概率收敛 $X_n \xrightarrow{P} c$ 的含义: $\forall \epsilon > 0$, 随着 n 的增大, X_n 落在 $[c - \epsilon, c + \epsilon]$ 的概率越来越大。请回顧序列 $\{x_n\}$ 收敛的几何含义, 讨论它与依概率收敛的区别。

例 5.1. 若 $X_n \sim \left(1 - \frac{1}{n}\right)\langle 0 \rangle + \frac{1}{n}\langle n \rangle$, 则 $X_n \xrightarrow{P} 0$ 。然而, 当 $n \rightarrow \infty$ 时,

$$\begin{aligned} E(X_n) &\rightarrow 1 \\ E(X_n^2) &\rightarrow \infty \end{aligned}$$

练习 5.1. $X_n \xrightarrow{P} X$ 当且仅当 $X_n - X \xrightarrow{P} 0$ 。提示: 都满足式 (5.3)。

例 5.2. 若随机变量 X_n 的分布列为 $P(X_n = k/n - 1/2) = 2^{-n}C_n^k$, 其中 $k = 0, 1, \dots, n-1, n$ 。试证明: $X_n \xrightarrow{P} 0$ 。

证明. 令 $Y_n \sim B(n, 1/2)$, 则 $P(X_n = k/n - 1/2) = P(Y_n = k)$, 进而

$$P(|X_n| \leq \epsilon) = P(|Y_n - n/2| \leq n\epsilon)$$

仿照例 2.78, 利用 Hoeffding 不等式 (2.81) 可证得 i

$$\lim_{n \rightarrow \infty} P(|Y_n - n/2| \leq n\epsilon) = 1$$

进而, $\lim_{n \rightarrow \infty} P(|X_n| \leq \epsilon) = 1$ 得证。 \square

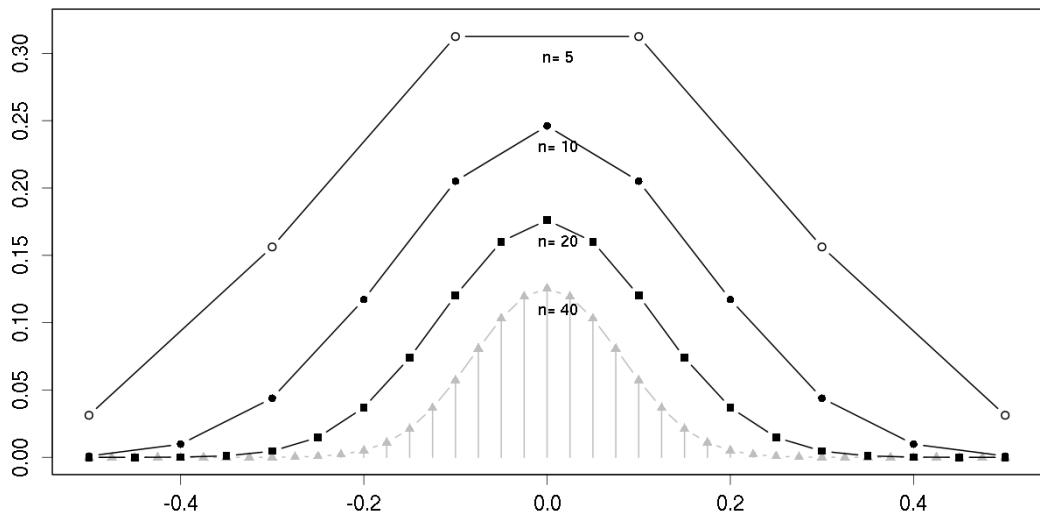


图 5.2: 例 5.2 中定义的随机变量 X_n 的分布列的折线图: 随着 n 的增加, X_n 落在区间 $[-\epsilon, \epsilon]$ 上的概率越来越接近 1, 即 $\{X_n\}$ 依概率收敛到单点分布 $\langle 0 \rangle$ 。

例 5.3. 回顾式 (2.5), 利用简单随机变量序列 $\{X_n\}$ 来逼近有界随机变量 $X \in (a, b]$, 显然 $X_n \xrightarrow{P} X$ 。

例 5.4. 接着第 236 页的例 3.25, 对于任意的 $\epsilon > 0$, 当 $n > 1/\epsilon$ 时,

$$P(|X_n - X| \geq \epsilon) = P(|X_n| \geq \epsilon) = 0$$

于是, $X_n \xrightarrow{P} X$, 也记作 $X_n \xrightarrow{P} 0$ 。另外, 由例 3.25 知, $X_n \xrightarrow{L} 0$ 。

例 5.5. 如果随机变量序列 $X_1, X_2, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 则随机变量

$$Y_n = \frac{1}{n} \sum_{j=1}^n X_j \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

显然, n 越大 Y_n 的方差越小, 而 $P\{|Y_n - \mu| \leq \epsilon\}$ 越接近 1。于是 $Y_n \xrightarrow{P} \mu$, 即 $\{Y_n\}$ 依概率收敛到单点分布 $Y \sim \langle \mu \rangle$ 。另外, 由例 3.27 知, $Y_n \xrightarrow{L} \mu$ 。

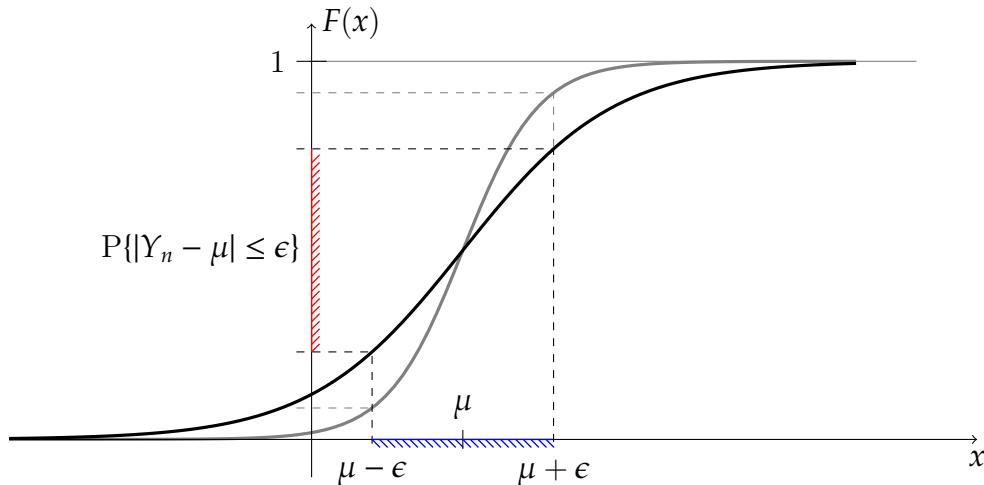
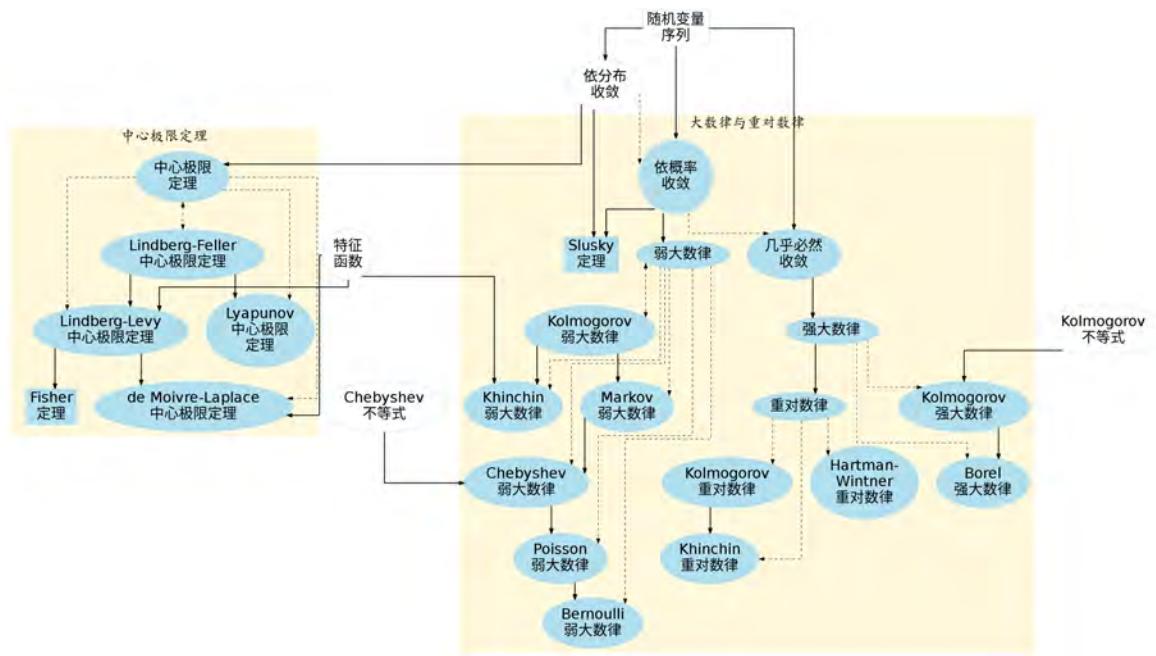


图 5.3: 随着 n 的增大, Y_n 的分布函数越来越“陡”。直观上, 随机变量序列 $Y_n \sim N(\mu, \sigma^2/n), n = 1, 2, \dots$ 依分布收敛, 同时依概率收敛到单点分布 $Y \sim \langle \mu \rangle$ 。

通过上面的两个例子, 人们自然会问: 依分布收敛和依概率收敛这两种收敛方式之间有怎样的关系呢? 后者是否一定能推出前者? [性质 5.1](#) 将回答这些问题。

第五章的主要内容及其关系



5.1 大数律

从大量的随机现象中寻找必然的规律,是概率论的研究目标。Bernoulli 和 Poisson 弱大数律就是从足够多次的独立 Bernoulli 试验中发掘随机事件频率的规律,最终以极限定理^{*}的形式给出概率的频率解释。再如,气压是由于空气分子运动而在器壁单位面积上所产生的压力——虽然单个分子的运动是随机的,并且对器壁的撞击是瞬间的,但大量分子对器壁撞击的总体效果却表现出持续而均匀的压力。

在大量分子的随机运动中,个体的偶然性在一定程度上相互消解或补偿,以至于宏观上的平均效果呈现出必然法则,这类“均等化”的物理现象是普遍存在的。一些问题在数学上可归结为论证随机变量序列 $\{X_j\}_{j=1}^\infty$ 满足弱大数律,即证明

$$Y_n = \frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n \mathbb{E}X_j \xrightarrow{P} 0$$

通过例 5.4、例 5.5 和例 5.2,读者对依概率收敛有了直观的了解,但要证明 $Y_n \xrightarrow{P} 0$ 也非易事。对依概率收敛缺乏必要研究手段的时候,人们很容易想到去考察它同依分布收敛的内在联系——“他山之石,可以攻玉”,毕竟依分布收敛有 Lévy 连续性定理和特征函数等工具。下面,我们将说明 $Y_n \xrightarrow{P} 0$ 和 $Y_n \xrightarrow{L} 0$ 是一回事儿。

例 5.6. 如果 $Y_n \xrightarrow{P} 0$, 则 $Y_n \xrightarrow{L} 0$, 即 $\forall y \in \mathbb{R} \setminus \{0\}$ 有

$$\lim_{n \rightarrow \infty} F_n(y) = \begin{cases} 0 & \text{当 } y < 0 \\ 1 & \text{当 } y > 0 \end{cases}$$

证明. 由 $Y_n \xrightarrow{P} 0$ 可知, 当 $n \rightarrow \infty$ 时, $P(|Y_n| < \epsilon) \rightarrow 1$, 其中 ϵ 是任意正数。因此,

$$F_n(y) = P(Y_n \leq y) \begin{cases} \leq 1 - P(|Y_n| < -y) \rightarrow 0 & \text{当 } y < 0 \\ \geq P(|Y_n| \leq y) \rightarrow 1 & \text{当 } y > 0 \end{cases} \quad \square$$

人们自然要问,例 5.6 的结果是否可以推广到一般情形,即依概率收敛蕴含依分布收敛?我们将论证这个猜想(见性质 5.1 及其证明),并举例说明反之不成立(见例 5.7)。然而,这两种收敛方式在收敛于常数时是等价的。因此,要证明 $Y_n = \frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n \mathbb{E}X_j \xrightarrow{P} 0$,只需证明 $Y_n \xrightarrow{L} 0$ 即可。

有关依概率收敛和依分布收敛的关系,下述结果给出了一个基本描述。

□ 若 $X_n \xrightarrow{P} X$, 则 $X_n \xrightarrow{L} X$ 。

*极限定理的研究内容很广泛,其中大数律和中心极限定理是最重要的两类。

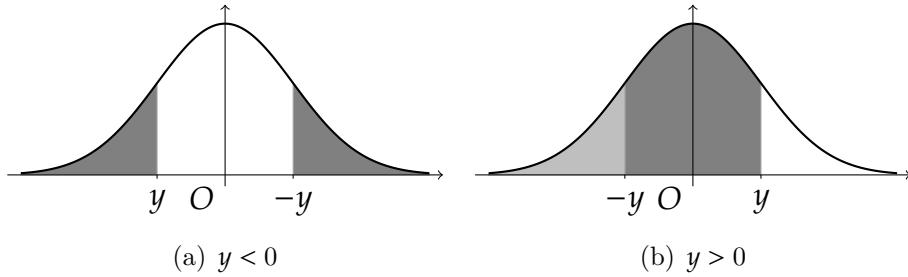


图 5.4: 例 5.6 的直观证明: 当 $y < 0$ 时, $P(Y_n \leq y) \leq 1 - P(|Y_n| < -y)$ 。当 $y > 0$ 时, $P(Y_n \leq y) \geq P(|Y_n| \leq y)$ 。

□ 若 $X_n \xrightarrow{L} x_0$, 其中 x_0 是常数, 则 $X_n \xrightarrow{P} x_0$ 。

※证明. 往证第一款: 假设 X 的分布函数 $F_X(x)$ 在点 $x = a$ 处连续, 对于任意 $\epsilon > 0$, 利用第 212 页的不等式 (2.96),

$$P(X_n \leq a) \leq P(X \leq a + \epsilon) + P(|X_n - X| > \epsilon)$$

由上面的两式，于是

$$\begin{aligned} \mathbb{P}(X \leq a - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon) &\leq \mathbb{P}(X_n \leq a) \\ &\leq \mathbb{P}(X \leq a + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \end{aligned}$$

令 $n \rightarrow \infty$, 便得到

$$F_X(a - \epsilon) \leq \lim_{n \rightarrow \infty} P(X_n \leq a) \leq F_X(a + \epsilon)$$

在上式中，令 $\epsilon \rightarrow 0$ ，立即得到

$$\lim_{n \rightarrow \infty} P(X_n \leq a) = \lim_{n \rightarrow \infty} F_{X_n}(a) = F_X(a), \text{ 即 } X_n \xrightarrow{L} X$$

往证第二款：对于任意 $\epsilon > 0$ ，令开集 $B_\epsilon = (x_0 - \epsilon, x_0 + \epsilon)$ 。因为 $X_n \xrightarrow{P} x_0$ ，由性质 3.2，不难得得到

$$\limsup_{n \rightarrow \infty} P(X_n \in B_\epsilon^c) \leq P(x_0 \in B_\epsilon^c) = 0$$

进而, $P(|X_n - x_0| \geq \epsilon) = P(X_n \in B_\epsilon^c) \rightarrow 0$, 得证。

例 5.7. 性质 5.1 第一款的逆命题不成立，即依分布收敛并不蕴含依概率收敛。举例说明：设随机变量 $X, X_1, X_2, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} \frac{1}{2}\langle 1 \rangle + \frac{1}{2}\langle 0 \rangle$ ，因此 $X_n \xrightarrow{L} X$ 。下面说明 $\{X_n\}$

不依概率收敛于 X : 这是因为 $\forall \epsilon \in (0, 1)$ 皆有

$$\begin{aligned} P(|X_n - X| \geq \epsilon) &= P(X_n = 1, X = 0) + P(X_n = 0, X = 1) \\ &= P(X_n = 1)P(X = 0) + P(X_n = 0)P(X = 1) = \frac{1}{2} \end{aligned}$$

或者, 直接从 $X_n - X \sim \frac{1}{4}\langle -1 \rangle + \frac{1}{2}\langle 0 \rangle + \frac{1}{4}\langle 1 \rangle$ 这一事实也能得到上述结论。由性质 5.1 和例 5.7 不难看出, 依分布收敛要弱于依概率收敛。

\nwarrow 定理 5.2. 已知 $Y_n \xrightarrow{L} X$ 和 $X_n - Y_n \xrightarrow{L} 0$, 则 $X_n \xrightarrow{L} X$ 。

\nwarrow 证明. 对于任意闭集 F , 定义 y 到 F 的距离 $\rho(y, F) = \min\{|y - x| : x \in F\}$ 。令闭集 $F_\epsilon = \{y : \rho(y, F) \leq \epsilon\}$, 其中 $\epsilon > 0$ 。显然, $F \subseteq F_\epsilon$ 。不难得到,

$$P(X_n \in F) \leq P(|X_n - Y_n| \geq \epsilon) + P(Y_n \in F_\epsilon)$$

由已知条件 $X_n - Y_n \xrightarrow{L} 0$ 可得 $X_n - Y_n \xrightarrow{P} 0$, 即 $\lim_{n \rightarrow \infty} P(|X_n - Y_n| \geq \epsilon) = 0$ 。根据性质 3.2, 我们有

$$\limsup_{n \rightarrow \infty} P(X_n \in F) \leq \limsup_{n \rightarrow \infty} P(Y_n \in F_\epsilon) \leq P(Y \in F_\epsilon)$$

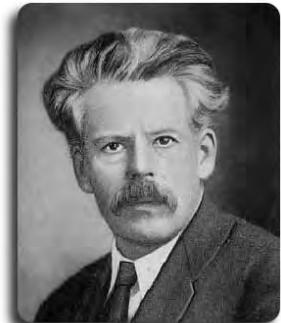
令 $\epsilon \rightarrow 0$, 则有 $F_\epsilon \rightarrow F$ 。再次利用性质 3.2, 证得 $X_n \xrightarrow{L} X$ 。 \square

推论 5.1. 如果 $X_n - X \xrightarrow{L} 0$, 则 $X_n \xrightarrow{L} X$ 。

练习 5.2. 请读者将推论 5.1 与练习 5.1 所述的结论进行比较, 举例说明推论 5.1 的逆命题不成立。提示: 见例 5.7。

有关依分布收敛和依概率收敛, 下面不加证明地介绍一个经典的结果——Slutsky 定理。该定理非常实用, 例如, 由它可直接推得定理 5.2, 以及 t_n 分布的极限是 $N(0, 1)$ (见第 357 页的例 5.10)。Slutsky 定理是苏联统计学家兼经济学家 Eugen E. Slutsky (1880-1948) 于 1925 年证得的。

\nwarrow 定理 5.3 (Slutsky, 1925). 若 $X_n \xrightarrow{L} X$ 且 $Y_n \xrightarrow{P} y_0$ (常数), 则有



$$\begin{aligned} X_n + Y_n &\xrightarrow{L} X + y_0 \\ Y_n X_n &\xrightarrow{L} y_0 X \\ \frac{X_n}{Y_n} &\xrightarrow{L} \frac{X}{y_0}, \text{ 如果 } y_0 \neq 0 \end{aligned}$$

※证明. 详见 A. N. Shirayev 的《概率论》[145] 第二章第十节。 □

练习 5.3. 设 $\{X_j\}_{j=1}^{\infty}, \{Y_j\}_{j=1}^{\infty}$ 是两个满足弱大数律的随机变量序列, 令 $Z_j = X_j + Y_j, j = 1, 2, \dots$ 。试证明: $\{Z_j\}_{j=1}^{\infty}$ 也满足弱大数律。提示: 根据性质 5.1 和 Slutsky 定理 5.3。

定理 5.4. 已知实数序列 $\{a_n\}$ 满足 $\lim_{n \rightarrow \infty} a_n = \infty$, 并且 $a_n(X_n - c) \xrightarrow{L} Y$, 其中 c 为常数。如果函数 $g: \mathbb{R} \rightarrow \mathbb{R}$ 在 c 处可导, 则

$$a_n(g(X_n) - g(c)) \xrightarrow{L} g'(c)Y$$

证明. 严格的证明详见 P. J. Bickel 和 K. A. Doksum 的《数理统计 基本概念及专题》[12] 第 311-312 页。下面给出一个不严格的“证明”, 它是启发性的: 将 $g(X_n)$ 在点 c 处做 Taylor 展开, 得到

$$g(X_n) - g(c) = g'(c)(X_n - c) + o(a_n^{-1})$$

$$\text{进而, } a_n(g(X_n) - g(c)) = g'(c)a_n(X_n - c) + o(1) \xrightarrow{L} g'(c)Y \quad \square$$

例 5.8. 如果 $\sqrt{n}(X_n - c) \xrightarrow{L} N(0, 1)$, 令 $Y_n = \sqrt{n}(X_n - c)$, 则 $Y_n \xrightarrow{L} Y$, 其中 $Y \sim N(0, 1)$ 。因为 $\frac{1}{\sqrt{n}} \rightarrow 0$, 于是 $X_n - c \xrightarrow{L} 0$ 。由 Slutsky 定理 5.3,

$$\sqrt{n}(X_n^2 - c^2) = Y_n(X_n + c) \xrightarrow{L} 2cY$$

根据定理 5.4, 可以直接得到 $\sqrt{n}(X_n^2 - c^2) \xrightarrow{L} N(0, 4c^2)$ 。

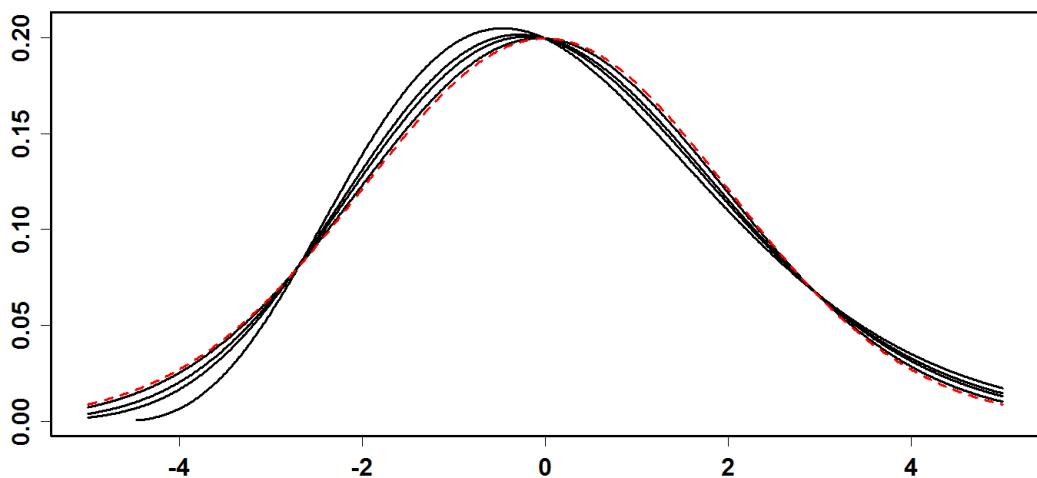


图 5.5: 已知独立同分布的随机变量序列 $Z_1, Z_2, \dots, Z_n, \dots \stackrel{\text{iid}}{\sim} N(c, 1)$, 则 $X_n = \frac{1}{n} \sum_{j=1}^n Z_j \sim N(c, 1/n)$, 进而利用例 2.17 的结果可求得随机变量 $V_n = \sqrt{n}(X_n^2 - c^2)$ 的密度函数。取 $c = 1, n = 20, 50, 100, 1000$, 图中实线是随机变量 V_n 的密度函数曲线, 随着 n 的增大, 越来越接近 $N(0, 4c^2)$ (虚线)。

本节内容

弱大数律是一组对 Bernoulli 弱大数律的推广，第一小节依次介绍了 Chebyshev、Markov、Khinchin、Kolmogorov 等数学家发现的弱大数律，其中 Kolmogorov 弱大数律是一个充分必要条件。第二小节是 Borel 强大数律和 Kolmogorov 的两个强大数律的简介。随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足弱（或强）大数律可用依概率（或几乎必然）收敛来描述。作为补充，第二小节的最后介绍了比强大数律更精细的重对数律。

关键知识

- (1) 掌握依分布收敛、依概率收敛、几乎必然收敛的定义以及性质；(2) 理解弱大数律和强大数律的概率意义；(3) 了解大数律的特征函数证明方法。

5.1.1 弱大数律

1866 年, Chebyshev 在其论文《论均值》中得到如下结论 [165]: “当试验次数趋于无穷时, 试验中事件发生的概率的算术平均与事件发生的次数与试验次数之比的差不超过任意给定值的概率趋于 1。一个特殊情形是, 试验中每次试验事件发生的概率为常数, 这便是 Bernoulli 定理。”

\nwarrow 定理 5.5 (Chebyshev 弱大数律). 如果独立的随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足方差一致有界, 即存在正的常数 c 使得

$$V(X_j) \leq c, \text{ 其中 } j = 1, 2, \dots, n, \dots \quad (5.4)$$



则该随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足弱大数律。

证明. 令 $Y_n = \frac{1}{n} \sum_{j=1}^n X_j$, 由定理的前提假设, 我们有

$$V(Y_n) = V\left(\frac{1}{n} \sum_{j=1}^n X_j\right) = \frac{1}{n^2} \sum_{j=1}^n V(X_j) \leq \frac{c}{n}$$

对任意正实数 ϵ , 由 Chebyshev 不等式不难得到

$$P\{|Y_n - E(Y_n)| \leq \epsilon\} \geq 1 - \frac{V(Y_n)}{\epsilon^2} \geq 1 - \frac{c}{n\epsilon^2}$$

在上式中, 令 $n \rightarrow \infty$ 即证得 $P\{|Y_n - E(Y_n)| \leq \epsilon\} \rightarrow 1$ 。 \square

从定理 5.5 的证明不难看出, 要使得 $\{X_j\}_{j=1}^{\infty}$ 满足弱大数律, 只要保证 $n \rightarrow \infty$ 时, $V(Y_n) \rightarrow 0$ 即可。于是, 便有了下面更强的结果。

\nwarrow 定理 5.6 (Markov 弱大数律). 如果随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足下述所谓的“Markov 条件”, 则它满足弱大数律。

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} V\left(\sum_{j=1}^n X_j\right) = 0 \quad (5.5)$$

练习 5.4. 已知 $\{X_n\}_{n=1}^{\infty}$ 是独立同分布的随机变量序列, X_n 的方差存在。设级数 $\sum_{n=1}^{\infty} a_n$ 绝对收敛, 令随机变量 $Y_n = a_n \sum_{j=1}^n X_j$, 试证明: $\{Y_n\}_{n=1}^{\infty}$ 满足弱大数律。

提示：验证 $\{Y_n\}_{n=1}^{\infty}$ 满足 Markov 条件 (5.5)，即当 $n \rightarrow \infty$ 时，

$$\frac{1}{n^2}V\left(\sum_{j=1}^n Y_j\right) \leq \frac{V(X_1)}{n} \sum_{j=1}^n |a_j| \rightarrow 0$$

例 5.9. 设随机变量序列 $\{X_k\}_{k=1}^{\infty}$ 满足方差一致有界的条件 (5.4)，并且当 $|k-j| \geq 2$ 时 X_k 和 X_j 不相关。试证明： $\{X_k\}_{k=1}^{\infty}$ 满足弱大数律。

证明. 由已知条件，当 $|k-j| \geq 2$ 时， X_k 和 X_j 不相关，所以

$$\begin{aligned} \text{Cov}(X_k, X_j) &= 0 \\ |\text{Cov}(X_k, X_{k+1})| &= |\rho(X_k, X_{k+1})| \sqrt{V(X_k)V(X_{k+1})} \leq c \end{aligned}$$

进而，Markov 条件 (5.5) 成立，即当 $n \rightarrow \infty$ 时，

$$\begin{aligned} \frac{1}{n^2}V\left(\sum_{k=1}^n X_k\right) &= \frac{1}{n^2}\left[\sum_{k=1}^n V(X_k) + 2 \sum_{k=1}^{n-1} \text{Cov}(X_k, X_{k+1})\right] \\ &\leq \frac{1}{n^2}[nc + 2(n-1)c] \leq \frac{3c}{n} \rightarrow 0 \end{aligned}$$

根据 Markov 弱大数律，随机变量序列 $\{X_k\}_{k=1}^{\infty}$ 满足弱大数律。 \square

练习 5.5. 在 Bernoulli 试验中，事件 A 出现的概率为 p ，令随机变量

$$X_k = \begin{cases} 1 & \text{若在第 } k \text{ 次及第 } k+1 \text{ 次试验中 } A \text{ 出现} \\ 0 & \text{其他} \end{cases}$$

试证明：随机变量序列 $\{X_k\}_{k=1}^{\infty}$ 满足弱大数律。

提示：利用上例的结果。

由 Lévy 连续性定理，人们可以轻易证得下面的 Khinchin 弱大数律，所用的 Lyapunov 特征函数方法具有一定的代表性，也可用于其他极限定理的证明，如 de Moivre-Laplace、Lindeberg-Lévy 中心极限定理等。Khinchin 弱大数律由苏联数学家 A. Ya. Khinchin (1894-1959, 照片见右) 于 1929 年证得，Khinchin 是概率论苏联学派的代表人物之一。

定理 5.7 (Khinchin 弱大数律, 1929). 已知 $\{X_j\}_{j=1}^{\infty}$ 是独立同分布的随机变量序列，满足 $EX_j = \mu < \infty$ ，则随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足弱大数律。



证明. 随机变量 X_j 的特征函数 $\varphi(t)$ 在 $t = 0$ 处有 Taylor 级数展开

$$\varphi(t) = \varphi(0) + \varphi'(0)t + o(t) = 1 + \mu it + o(t)$$

于是随机变量 $Y_n = \frac{1}{n} \sum_{j=1}^n X_j$ 的特征函数为

$$[\varphi(t/n)]^n = [1 + \mu it/n + o(t/n)]^n$$

对每个暂时固定的 t , 总存在足够大的 n 使得 $|\mu it/n + o(t/n)| < 1$, 利用式 (3.13) 进而得到

$$\begin{aligned} \ln[\varphi(t/n)]^n &= n \ln \left[1 + \frac{\mu it}{n} + o(t/n) \right] \\ &= n \left\{ \left[\frac{\mu it}{n} + o(t/n) \right] - \frac{1}{2} \left[\frac{\mu it}{n} + o(t/n) \right]^2 + \dots \right\} \\ &= \mu it + no(t/n) \end{aligned}$$

于是, $\lim_{n \rightarrow \infty} [\varphi(t/n)]^n = e^{\mu it}$, 显然它是单点分布 $Y \sim \langle \mu \rangle$ 的特征函数。根据 Lévy 连续性定理, 有 $Y_n \xrightarrow{L} \mu$, 进而 $Y_n \xrightarrow{P} \mu$, 得证。 \square

练习 5.6. 设独立同分布的随机变量序列 $X_1, X_2, \dots, X_n, \dots$ 满足

$$P \left\{ X_n = \frac{2^k}{k^2} \right\} = \frac{1}{2^k}, \text{ 其中 } k = 1, 2, \dots$$

试证明: 随机变量序列 $\{X_n\}_{n=1}^\infty$ 满足弱大数律。提示: $\sum_{k=1}^\infty k^{-2} = \pi^2/6$ 。

例 5.10. 如果随机变量 $T_n \sim t_n$, 试证明: $T_n \xrightarrow{L} N(0, 1)$ 。

证明. 由 t_n 分布的定义 4.21, $T_n = X_n / \sqrt{Y_n/n}$, 其中 $X_n \sim N(0, 1)$ 与 $Y_n \sim \chi_n^2$ 相互独立。而 $Y_n = \sum_{j=1}^n Z_j$, 其中 $Z_1, \dots, Z_n \stackrel{iid}{\sim} \chi_1^2$, 由定理 5.7,

$$\frac{Y_n}{n} = \frac{1}{n} \sum_{j=1}^n Z_j \xrightarrow{P} 1$$

利用 Slutsky 定理 5.3 有 $T_n \xrightarrow{L} N(0, 1)$ 。 \square

练习 5.7. 如果随机变量 $X_n \sim \chi_n^2$, 试证明: $(X_n - n) / \sqrt{2n} \xrightarrow{L} N(0, 1)$ 。

例 5.11. 设 $X_1, X_2, \dots, X_n, \dots$ 独立同分布, 且 $E(X_j) = 0, V(X_j) = \sigma^2 < \infty, j = 1, 2, \dots, n, \dots$ 。试证明: 随机变量序列 $\{X_j^2\}_{j=1}^\infty$ 满足弱大数律。

证明. 显然 $X_1^2, X_2^2, \dots, X_n^2, \dots$ 也独立同分布。由于 $E(X_j) = 0$, 因此 $E(X_j^2) = \sigma^2, j = 1, 2, \dots$ 。由 Khinchin 弱大数律即可证得。 \square

人们对 Bernoulli 弱大数律容易形成误解, 认为事件的频率随着试验次数的增加而趋于该事件的概率。事实上, Bernoulli 弱大数律仅仅断言, 对于事件 $E = “|m/n - p| \leq \epsilon”$, 只要实验次数 n 充分地大, 就能保证 E 发生的概率不小于 $1 - \eta$, 其中 $\eta < 1$ 是一给定的正数。换句话说, 只要 n 足够地大, 事件 E 便以接近 1 的概率发生, 对其他弱大数律的解释亦是如此。

定理 5.8 (Kolmogorov 弱大数律, 1926). $\{X_j\}_{j=1}^\infty$ 满足弱大数律当且仅当

$$\lim_{n \rightarrow \infty} E \left\{ \frac{\left[\sum_{j=1}^n (X_j - EX_j) \right]^2}{n^2 + \left[\sum_{j=1}^n (X_j - EX_j) \right]^2} \right\} = 0 \quad (5.6)$$

证明. 往证 “ \Leftarrow ”: 令 $G_n(x)$ 为 $Y_n = \frac{1}{n} \sum_{j=1}^n (X_j - EX_j)$ 的分布函数。

$$\begin{aligned} P(|Y_n| \geq \epsilon) &= \int_{|x| \geq \epsilon} dG_n(x) \leq \frac{1 + \epsilon^2}{\epsilon^2} \int_{|x| \geq \epsilon} \frac{x^2}{1 + x^2} dG_n(x) \\ &\leq \frac{1 + \epsilon^2}{\epsilon^2} \int_{\mathbb{R}} \frac{x^2}{1 + x^2} dG_n(x) \\ &= \frac{1 + \epsilon^2}{\epsilon^2} E \left\{ \frac{Y_n^2}{1 + Y_n^2} \right\} \rightarrow 0 \end{aligned}$$

往证 “ \Rightarrow ”: 当 $n \rightarrow \infty$ 时,

$$\begin{aligned} P(|Y_n| \geq \epsilon) &= \int_{|x| \geq \epsilon} dG_n(x) \geq \int_{|x| \geq \epsilon} \frac{x^2}{1 + x^2} dG_n(x) \\ &= \int_{\mathbb{R}} \frac{x^2}{1 + x^2} dG_n(x) - \int_{|x| < \epsilon} \frac{x^2}{1 + x^2} dG_n(x) \\ &\geq E \left\{ \frac{Y_n^2}{1 + Y_n^2} \right\} - \frac{\epsilon^2}{1 + \epsilon^2} \int_{\mathbb{R}} dG_n(x) \\ &\geq E \left\{ \frac{Y_n^2}{1 + Y_n^2} \right\} - \epsilon^2 \rightarrow 0 \quad \square \end{aligned}$$

 Kolmogorov 弱大数律保证了 $\forall \epsilon, \eta > 0$, 存在 $N \in \mathbb{N}$ 使得 $\forall n > N$ 皆有 $P(|Y_n| \geq \epsilon) < \eta$, 但不保证 $P\{(|Y_{N+1}| \geq \epsilon) \cup (|Y_{N+2}| \geq \epsilon) \cup \dots\} < \eta$ 。Khinchin 弱大数律不要求有限方差, Markov 弱大数律在条件中不要求独立性, 它们都是

Kolmogorov 弱大数律的特例，这是因为

$$\frac{Y_n^2}{1+Y_n^2} \leq Y_n^2 = \left[\frac{1}{n} \sum_{j=1}^n (X_j - \mathbb{E}X_j) \right]^2$$

于是 $\mathbb{E} \left\{ \frac{Y_n^2}{1+Y_n^2} \right\} \leq \frac{1}{n^2} V \left(\sum_{j=1}^n X_j \right)$

练习 5.8. 请说明 Kolmogorov 弱大数律 \Rightarrow Markov 弱大数律 \Rightarrow Chebyshev 弱大数律 \Rightarrow Poisson 弱大数律 \Rightarrow Bernoulli 弱大数律。

5.1.2 强大数律与重对数律

1902 年, 法国数学家 Émile Borel (1871-1956) 有一个重大的发现: 抛一枚均匀的硬币 n 次, 出现正面的频率 m/n 以概率 1 趋向 $1/2$ 。后来这个结果被整理成下述一般情形, 称为 Borel 强大数律。

定理 5.9 (Borel 强大数律, 1909). 随机事件 A 在 n 重 Bernoulli 试验中出现的频率 m/n 以概率 1 趋向 $P(A)$, 即

$$P\left\{\lim_{n \rightarrow \infty} \left[\frac{m}{n} - P(A)\right] = 0\right\} = 1$$

按照式 (5.1) 的定义可给出 Borel 强大数律的“随机变量版”的表达形式:

$$P\left\{\lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n EX_j\right) = 0\right\} = 1 \quad (5.7)$$

上式比弱大数律的条件 (5.2) 要求得更强些, 于是有了下面的概念。

定义 5.3 (强大数律). 如果随机变量序列 $\{X_j\}_{j=1}^\infty$ 满足条件 (5.7), 则称 $\{X_j\}_{j=1}^\infty$ 满足强大数律 (strong law of large numbers), 它由苏联数学家 Khinchin 于 1927-1928 年定名。

定义 5.4 (几乎必然收敛). 随机变量序列 $\{X_n\}_{n=1}^\infty$ 几乎必然收敛 (converge almost surely) 于随机变量 X , 记为 $X_n \xrightarrow{a.s.} X$, 当且仅当

$$P\left\{\lim_{n \rightarrow \infty} X_n - X = 0\right\} = 1$$

譬如, $\{X_j\}_{j=1}^\infty$ 满足强大数律意味着

$$\frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n EX_j \xrightarrow{a.s.} 0$$

性质 5.2. 若 $X_n \xrightarrow{a.s.} X$, 则 $X_n \xrightarrow{P} X$ 。

证明. 如果对于每个 $\omega \in \Omega$, 皆有 $X_n(\omega) \rightarrow X(\omega)$, 则对于任意 $\epsilon > 0$ 有

$$A_n = \bigcup_{m \geq n} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| > \epsilon\} \downarrow \emptyset$$

如果 $X_n \xrightarrow{a.s.} X$, 则 $P\{|X_n - X| > \epsilon\} \leq P(A_n) \rightarrow 0$, 得证。 \square

由**性质 5.1**和**性质 5.2**，随机变量序列的收敛性从强到弱的次序是几乎必然收敛、依概率收敛、依分布收敛，据此强大数律可推出弱大数律。以概率 1 发生的事件，在现实中常被视作“必然事件”，毕竟它已经非常接近真正的必然事件了。强大数律是数理统计学的基石之一，它们为“多次独立重复观测的结果的算术平均为总体期望的强相合估计”提供了理论依据。

定理 5.10 (Mann-Wald*, 1943). 已知 \mathbb{R} 上的随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 和随机变量 X ，假设函数 $g : \mathbb{R} \rightarrow \mathbb{R}$ 的不连续点集 D_g 满足 $P\{X \in D_g\} = 0$ ，则

$$\square X_n \xrightarrow{L} X \Rightarrow g(X_n) \xrightarrow{L} g(X);$$

$$\square X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X);$$

$$\square X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X).$$

证明. 第一款的证明见 [153] 的第二章的定理 2.3，本书不作要求。下面先往证第二款：对任意 $\epsilon > 0$ ，设 $\delta > 0$ ，定义集合 $B_{\epsilon}(\delta)$ 如下，

$$B_{\epsilon}(\delta) = \{x \in \mathbb{R} : x \notin D_g, \text{ 并且 } \exists y \text{ 满足 } |x - y| < \delta \text{ 使得 } |g(x) - g(y)| > \epsilon\}$$

在 g 的连续点集 D_g^c 上，有 $\lim_{\delta \rightarrow 0} B_{\epsilon}(\delta) = \emptyset$ 。如果 $|g(X_n) - g(X)| > \epsilon$ ，则 $|X_n - X| \geq \delta$ 或 $X \in B_{\epsilon}(\delta)$ 或 $X \in D_g$ ，于是当 $n \rightarrow \infty$ 且 $\delta \rightarrow 0$ 有

$$P\{|g(X_n) - g(X)| > \epsilon\} \leq P\{|X_n - X| \geq \delta\} + P\{X \in B_{\epsilon}(\delta)\} + P\{X \in D_g\} \rightarrow 0$$

即 $\lim_{n \rightarrow \infty} P\{|g(X_n) - g(X)| > \epsilon\} = 0$ ，第二款得证。

下面往证第三款：设 $X(\omega) \in \mathbb{R}$ 是函数 g 的连续点，则由 $\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)$ 可推导出 $\lim_{n \rightarrow \infty} g(X_n(\omega)) = g(X(\omega))$ ，进而

$$\begin{aligned} P\left\{\lim_{n \rightarrow \infty} g(X_n) = g(X)\right\} &\geq P\left\{\lim_{n \rightarrow \infty} g(X_n) = g(X) \text{ 且 } X \notin D_g\right\} \\ &\geq P\left\{\lim_{n \rightarrow \infty} X_n = X \text{ 且 } X \notin D_g\right\} \\ &\geq P\left\{\lim_{n \rightarrow \infty} X_n = X\right\} - P\{X \notin D_g\} \\ &= 1 - 0 = 1 \end{aligned}$$

□

*1943 年，美籍奥地利裔数学家 Henry Berthold Mann (1905-2000) 和美籍罗马尼亚裔统计学家 Abraham Wald (1902-1950) 证得此定理。

由**定义 5.3**, 随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足强大数律当且仅当对任意 $\epsilon > 0$, 存在充分大的 $N \in \mathbb{N}$ 使得下面的事件以概率 1 发生。

$$\left| \frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n \mathbb{E}X_j \right| \leq \epsilon, \text{ 其中 } n = N, N+1, \dots$$

利用 Borel-Cantelli 引理 (见第 74 页的**引理 1.1** 或 Borel 0-1 律), 若能证得下面的结果,

$$\sum_{n=1}^{\infty} P \left\{ \left| \frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{n} \sum_{j=1}^n \mathbb{E}X_j \right| \geq \epsilon \right\} < \infty$$

便证得了 $\{X_j\}_{j=1}^{\infty}$ 满足强大数律。利用这一工具和 Kolmogorov 不等式 (2.76) 可以证明下面的 Kolmogorov 强大数律。

定理 5.11 (Kolmogorov 强大数律, 1930). 如果独立的随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足下面的条件, 则 $\{X_j\}_{j=1}^{\infty}$ 满足强大数律。

$$\sum_{j=1}^{\infty} \frac{V(X_j)}{j^2} < \infty$$

证明. 令 $Z_n = \sum_{j=1}^n X_j - \mathbb{E}X_j$ 且 $Y_n = \frac{1}{n}Z_n$, 由 Kolmogorov 不等式可得

$$\begin{aligned} p_m &= P(\max |Y_n| \geq \epsilon, 2^m \leq n < 2^{m+1}) \\ &\leq P(\max |Z_n| \geq 2^m \epsilon, 2^m \leq n < 2^{m+1}) \\ &\leq \frac{1}{(2^m \epsilon)^2} \sum_{j<2^{m+1}} V(X_j) \end{aligned}$$

下面往证 $\sum_{m=1}^{\infty} p_m < \infty$, 再利用 Borel-Cantelli 引理可证得一组事件 $A_m = \{\max |Y_n| \geq \epsilon, 2^m \leq n < 2^{m+1}\}$, $m = 1, 2, \dots$ 中有无穷多个发生的概率为 0。事实上,

$$\begin{aligned} \sum_{m=1}^{\infty} p_m &\leq \sum_{m=1}^{\infty} \frac{1}{(2^m \epsilon)^2} \sum_{j<2^{m+1}} V(X_j) \\ &= \frac{1}{\epsilon^2} \sum_{j=1}^{\infty} V(X_j) \sum_{\{m: j<2^{m+1}\}} 2^{-2m} \\ &\leq \frac{16}{3\epsilon^2} \sum_{j=1}^{\infty} \frac{V(X_j)}{j^2} < \infty \end{aligned} \quad \square$$

练习 5.9. 如果存在正实数 c 使得独立的随机变量序列 $X_j, j = 1, 2, \dots$ 满足 $V(X_j) \leq c$,

即 $\{X_j\}_{j=1}^{\infty}$ 的方差一致有界，则随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足强大数律。提示：因为级数 $\sum_{j=1}^{\infty} j^{-2}$ 收敛，利用定理 5.11 可证。

\curvearrowleft 定理 5.12 (Kolmogorov 强大数律, 1933). 与 Khinchin 弱大数律 (5.7) 的条件相同，随机变量 $X_j, j = 1, 2, \dots$ 独立同分布且 $E(X_j) = \mu < \infty$ ，则 $\{X_j\}_{j=1}^{\infty}$ 满足强大数律。

※证明. 详见 W. Feller 的《概率论及其应用》下卷第七章第八节。 □

练习 5.10. 试说明 Borel 强大数律是 Kolmogorov 强大数律的特例。

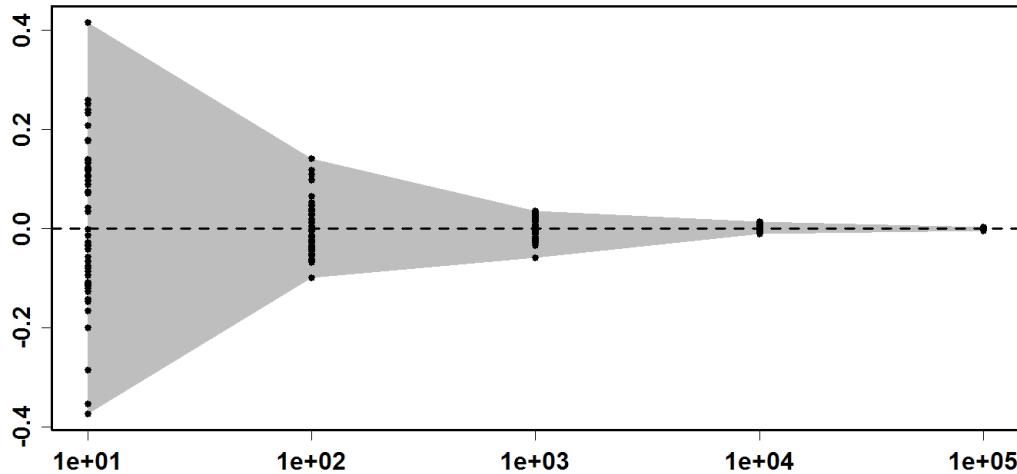


图 5.6: 考察 $Y_n = \frac{1}{n} \sum_{j=1}^n X_j$ 随 n 增加的变化情况，其中随机变量 $X_j, j = 1, 2, \dots$ 独立同分布于 $U[-1, 1]$ 。分别取 $n = 10, 10^2, 10^3, 10^4, 10^5$ ，对每个 Y_n 取 50 个随机数，读者不难发现散点图与强大数律断言的 $P(\lim_{n \rightarrow \infty} Y_n = 0) = 1$ 是吻合的。

例 5.12. 已知随机变量 $X \sim f(x)$ 满足 $E_f[h(X)] < \infty$ ，设 $X_1, \dots, X_n, \dots \stackrel{iid}{\sim} f(x)$ ，则 $h(X_1), \dots, h(X_n), \dots$ 独立同分布。由定理 5.12 可知，

$$\frac{1}{n} \sum_{j=1}^n [h(X_j)] \xrightarrow{a.s.} E_f[h(X)] \quad (5.8)$$

进一步，设概率/密度函数 $p(x) \neq 0$ ，则有

$$E_f[h(X)] = \begin{cases} \sum_j h(x_j) f(x_j) = \sum_j h(x_j) \frac{f(x_j)}{p(x_j)} p(x_j) \\ \int_{\mathbb{R}} h(x) f(x) dx = \int_{\mathbb{R}} h(x) \frac{f(x)}{p(x)} p(x) dx \\ = E_p[h(X)w(X)], \text{ 其中 } w(x) = \frac{f(x)}{p(x)} \end{cases}$$

与结果 (5.8) 类似, 设 $X_1, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} p(x)$, 利用强大数律有

$$\frac{1}{n} \sum_{j=1}^n [h(X_j)w(X_j)] \xrightarrow{a.s.} E_p[h(X)w(X)] = E_f[h(X)] \quad (5.9)$$

结果 (5.8) 和 (5.9) 都可以用于近似地计算 $E_f[h(X)]$: 当易于从 $f(x)$ 产生随机数的时候, 直接利用结果 (5.8); 否则, 找一个易于抽样的 $p(x) \neq 0$, 利用结果 (5.9)。具体实例分别见第 712 页的例 15.1 和第 721 页的例 15.9。

例 5.13. 设 $X_1, X_2, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 令 $S_n = \sum_{j=1}^n X_j$, 强大数律说 $P\{\lim_{n \rightarrow \infty} \frac{1}{n} S_n = p\} = 1$, 即 $\forall \epsilon > 0$, 不等式 $|\frac{1}{n} S_n - p| \leq \epsilon$ 除了有限个 n 外以概率 1 成立。1924 年, Khinchin 证得了一个比该例更强的结果。

定理 5.13 (Khinchin 重对数律, 1924). 若 $X_1, X_2, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 则随机变量 $S_n = X_1 + X_2 + \dots + X_n$ 满足下面的性质。

$$P \left\{ \limsup_{n \rightarrow \infty} \frac{|S_n - np|}{\sqrt{2p(1-p)n \ln \ln n}} = 1 \right\} = 1$$

也就是说, $\forall \epsilon > 0$, 以下不等式除了有限个 n 外以概率 1 成立。

$$\left| \frac{1}{n} S_n - p \right| \leq (1 + \epsilon) \sqrt{\frac{2p(1-p)}{n} \ln \ln n}$$

这个结果被称为 Khinchin 重对数律 (law of the iterated logarithm) 或迭对数律, 它利用重对数函数 $\ln \ln n$ 描述了 $\frac{1}{n} S_n$ 向其期望收敛的速度, 比 Borel 强大数律更精细些。重对数律的实例参见第 70 页的例 1.53。

定义 5.5 (重对数律). 由随机变量序列 X_1, \dots, X_n, \dots 构造新的随机变量序列 $S_n = X_1 + \dots + X_n$, 其中 $n = 1, 2, \dots$ 。不失一般性, 假设 $E(S_n) = 0$ 。如果存在数列 $\{c_n : n = 1, 2, \dots\}$ 满足以下条件, 则称随机变量序列 X_1, \dots, X_n, \dots 满足重对数律。

$$P \left\{ \limsup_{n \rightarrow \infty} \frac{|S_n|}{c_n} = 1 \right\} = 1$$

下面不加证明地介绍 Kolmogorov (1929)、P. Hartman 和 A. Wintner (1941) 在更一般的条件下发现的独立随机变量序列的重对数律。这些重对数律的证明都比较复杂, 感兴趣的读者可以参阅 V. V. Petrov 的著作《独立随机变量之和的极限定理》[121] 第七章或《独立随机变量之和》[120] 第十章。

定理 5.14 (Kolmogorov 重对数律, 1929). 若 $\{X_n\}_{n=1}^{\infty}$ 为独立随机变量序列, 且 $E(X_n) = 0, V(X_n) = \sigma_n^2$ 。记 $\tau_n^2 = \sum_{j=1}^n \sigma_j^2$, 如果存在某个趋于 0 的正数序列 $\{c_n\}$ 以概率 1 使

得 $|X_n| \leq c_n \sqrt{\tau_n^2 (\ln \ln \tau_n^2)^{-1}}$ 成立, 则 $S_n = X_1 + \cdots + X_n$ 满足

$$P \left\{ \limsup_{n \rightarrow \infty} \frac{|S_n|}{\sqrt{2\tau_n^2 \ln \ln \tau_n^2}} = 1 \right\} = 1 \quad (5.10)$$

练习 5.11. 试说明 Khinchin 重对数律是 Kolmogorv 重对数律的推论。

定理 5.15 (Hartman-Wintner 重对数律, 1941). 若随机变量序列 $\{X_n\}_{n=1}^\infty$ 独立同分布, 且 $E(X_n) = 0, V(X_n) = \sigma^2$, 则 $S_n = X_1 + \cdots + X_n$ 满足

$$P \left\{ \limsup_{n \rightarrow \infty} \frac{|S_n|}{\sqrt{2\sigma^2 n \ln \ln n}} = 1 \right\} = 1 \text{ 当且仅当 } \sigma^2 < \infty$$

例 5.14. 为了直观了解 Hartman-Wintner 重对数律, 从 $N(0, 1)$ 产生 $N = 10^5$ 个随机数, 观察 $\frac{1}{n}|S_n| = \frac{1}{n}|\sum_{j=1}^n X_j|, n = 1, 2, \dots, N$ 。不难发现当 n 增大时, $\frac{1}{n}|S_n|$ 被 $\sqrt{(2/n) \ln \ln n}$ “控制” 着趋向于 0 (见图 5.7)。

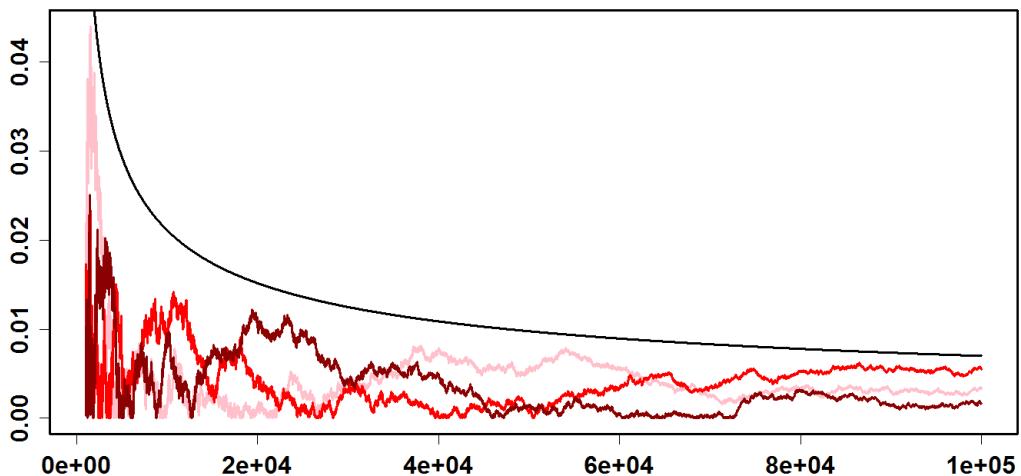


图 5.7: 若 $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} N(0, 1)$, 则 $\frac{1}{n}|\sum_{j=1}^n X_j|$ 被 $\sqrt{(2/n) \ln \ln n}$ “控制” 着趋向于 0。

5.2 中心极限定理

中心极限定理揭示了在一定条件下为何正态分布是一个普遍存在的分布，也给出了正态分布之所以重要的理由（见附录 A）。在实践中，人们经常遇到这样的随机现象，它受许多独立的随机因素的影响，而每一个因素对该现象的影响都是微小的，所有因素的集体作用才是我们真正关心的，而不是那些细枝末节的单个随机因素。

例如，测量不可避免地有误差，有些误差是因为测量仪器受空气湿度、大气压力、地球磁场等因素影响而产生的，有些则可能由测量者的心灵或生理情况的变化而引起的。这些不可控的微小因素使得随机误差可视为是众多独立随机变量之和，每项对总和的影响都很小，虽然每个组成部分的随机变量的分布是未知的，但它们的总体效应却明显地呈现出规律性——正态分布*。

例 5.15 (Galton 的正态漏斗). 1874 年，英国学者 Francis Galton 设计了一个叫 quincunx 的装置能够直观地揭示 de Moivre-Laplace 中心极限定理：板子上钉着一行一行的小钉儿，总共有 n 行，而且行与行之间是交错的，如图 5.8 中的 (b) 所示 (Galton 绘制的原图)。

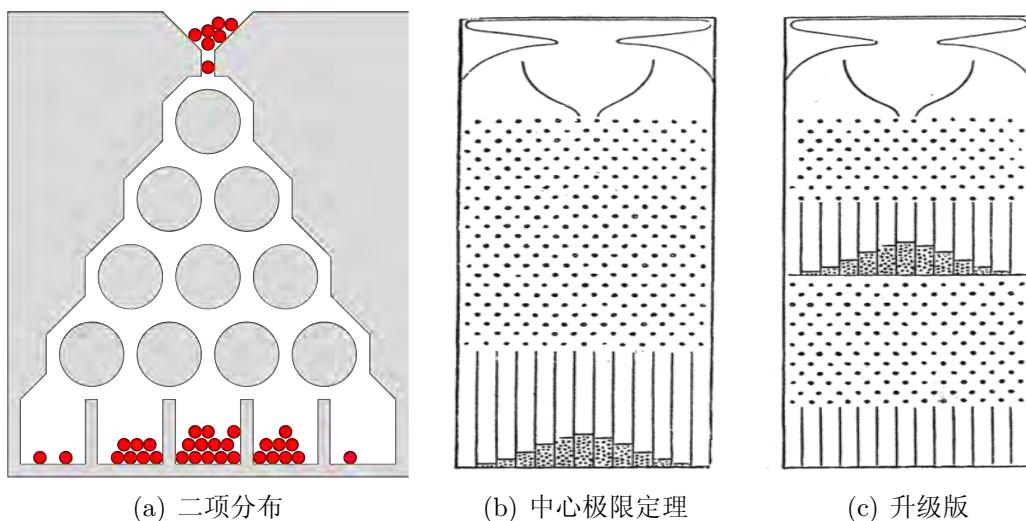


图 5.8: Galton 的正态漏斗可用于演示生成二项分布的随机数 (a)，以及 de Moivre-Laplace 中心极限定理 (b)。若按 (c) 所示设置一个临时的中间挡板，珠子在中间挡板上的槽内呈现出正态分布后抽掉中间挡板让珠子继续滑落，珠子在底部的槽内也将呈现出正态分布，这一现象说明了什么？

让珠子从第一行的中点处滚落，假设珠子碰撞到小钉儿后向左和向右滑落的概率相等，若 n 足够大，珠子落于底部槽中将呈现出正态分布。

* 独立随机变量之和不一定趋向正态分布。Lindeberg-Feller 中心极限定理断言，如果随机变量 X 是一些独立的“非本质”的随机变量之和，那么 X 服从正态分布。

图 5.8 中的 (a) 是这个游戏的简化版, 效果与 (b) 相同, 各槽内的珠子数服从二项分布 $B(n, 1/2)$, 当 n 足够大时, 可用正态分布 $N(n/2, n/4)$ 来近似。

例 5.16. 第 258 页的例 4.9 所定义的离散型随机变量 $Y = X_1 + \dots + X_n$, 当 n 足够大时, 其分布可以用 $N(kn/2, k(k+2)n/12)$ 来近似。

特别地, 第 25 页的例 1.16 中, $n = 6, k = 9$, 图 1.8 可由 $N(27, 49.5)$ 近似。有放回地随机抽取 6 次所得标号之和以大概率介于 12 和 42 之间。

正态分布如此之重要, 研究独立随机变量之和在什么条件下趋向于正态分布曾是概率论的核心问题, 传统把这一类命题统称为中心极限定理, 以突显它们在独立随机变量之和的极限定理中的地位。2000 年是世界数学年, 正态分布再一次被印在邮票上。

历史上, 中心极限定理的证明方法也经历过一个发展演变的过程, 首个系统的方法是由 Chebyshev 提出而经 Markov 完善化的“矩方法”。目前, 证明中心极限定理较多采用的是 Lyapunov 的特征函数方法, 该方法对其他极限定理的证明也是非常有效的, 是一件好用的数学工具。

de Moivre-Laplace 中心极限定理是古典概率论的巅峰, 附录 A 曾基于 Stirling 公式给出过证明, 下面的证法基于特征函数。

定理 5.16 (de Moivre-Laplace, 1733, 1801). 已知随机变量 $Y_n \sim B(n, p)$, 其中 $n = 1, 2, \dots$, 于是

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}} = \frac{\frac{1}{n}Y_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{L} N(0, 1) \quad (5.11)$$

证明. Y_n 和 Z_n 的特征函数分别是 $\varphi_n(t) = (q + pe^{it})^n$, 其中 $q = 1 - p$, 和

$$\begin{aligned} \tilde{\varphi}_n(t) &= \exp\left\{-\frac{npit}{\sqrt{npq}}\right\} \left[q + p \exp\left\{\frac{it}{\sqrt{npq}}\right\}\right]^n \\ &= \left[q \exp\left\{-\frac{pit}{\sqrt{npq}}\right\} + p \exp\left\{\frac{qit}{\sqrt{npq}}\right\}\right]^n \end{aligned}$$

利用解析函数 e^z 在 $z = 0$ 处的 Taylor 级数展开 $e^z = \sum_{j=0}^{\infty} z^j / j!$ 有,

$$\begin{aligned} q \exp\left\{-\frac{pit}{\sqrt{npq}}\right\} &= q - it \sqrt{\frac{pq}{n}} - \frac{pt^2}{2n} + o(t^2/n), \text{ 并且} \\ p \exp\left\{\frac{qit}{\sqrt{npq}}\right\} &= p + it \sqrt{\frac{pq}{n}} - \frac{qt^2}{2n} + o(t^2/n) \end{aligned}$$



仿照仿照 Khinchin 弱大数律（[定理 5.7](#)）的证明，我们有

$$\begin{aligned}\ln \tilde{\varphi}_n(t) &= n \ln \left[1 - \frac{t^2}{2n} + o(t^2/n) \right] \\ &= -\frac{t^2}{2} + no(t^2/n)\end{aligned}$$

因此， $\lim_{n \rightarrow \infty} \tilde{\varphi}_n(t) = \exp(-t^2/2)$ ，它是标准正态分布的特征函数。根据 Lévy 连续性定理， $Z_n \xrightarrow{L} N(0, 1)$ 得证。□

 回顾图 [1.7](#) 所示 n 很大时的二项分布 $B(n, p)$ 并阅读附录 A，二项分布与正态分布的关系就更明晰了。de Moivre-Laplace 中心极限定理也可以表述为：已知随机变量序列 $X_1, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ ，令 $Y_n = \sum_{j=1}^n X_j$ ，则式 [\(5.11\)](#) 成立。当 n 很大时，每个 $X_j, j = 1, 2, \dots, n$ 对 Y_n 的影响都不是至关重要的。1920 年，G. Pólya 将随机变量序列部分和的分布渐近于正态分布的这一类定理统称为中心极限定理。

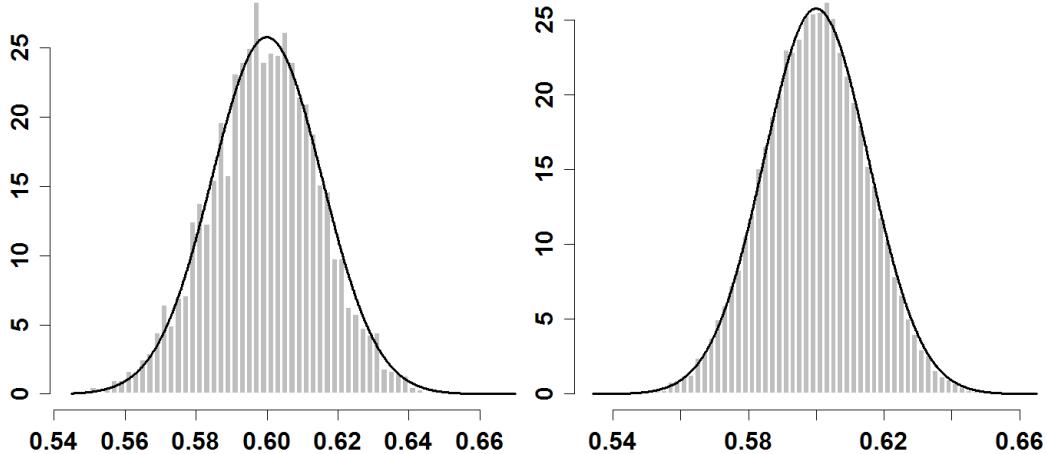


图 5.9：通过模拟试验了解 de Moivre-Laplace 中心极限定理：再次重复[例 1.48](#)中的随机试验，抛那枚不均匀的硬币（正面出现的概率为 $p = 0.6$ ） $n = 1000$ 次，记录下频率并重复此过程 3000 遍（左图）和 30000 遍（右图），得到频率的直方图，将之与正态分布 $N(p, p(1-p)/n)$ 的密度函数进行比较，发现重复遍数越多拟合效果越好。

定义 5.6. 随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立且每个 $X_j, j = 1, 2, \dots$ 有有限的期望 $\mu_j = EX_j$ 和方差 $\sigma_j^2 = V(X_j)$ 。随机变量序列 $\{X_n\}$ 被称为满足中心极限定理当且仅当

$$Z_n = \frac{1}{\tau_n} \sum_{j=1}^n (X_j - \mu_j) \xrightarrow{L} N(0, 1), \quad \text{其中 } \tau_n = \sqrt{\sum_{j=1}^n \sigma_j^2} \quad (5.12)$$

即, $Y_n = \sum_{j=1}^n X_j$ 标准化之后的随机变量序列依分布收敛到标准正态分布。有时, 条件 (5.12) 也等价地写为

$$Y_n = \sum_{j=1}^n X_j \xrightarrow{L} N\left(\sum_{j=1}^n \mu_j, \sum_{j=1}^n \sigma_j^2\right)$$

例 5.17. 随机变量序列 $X_1, X_2, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 满足中心极限定理。这是因为 $Y_n = \sum_{j=1}^n X_j \sim B(n, p)$, 由定理 5.16 可知,

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow{L} N(0, 1)$$

例 5.18. 设单词 w 和 w' 共同出现在一个句子里的概率为 p , 在随机抽取的大规模语料中, w, w' 共同出现在一个句子里的次数 $N_{w,w'}$ 服从二项分布 $B(n, p)$, 其中 n 是语料中句子的个数。当 n 足够地大, 近似地我们有

$$\frac{N_{w,w'} - np}{\sqrt{np(1-p)}} \sim N(0, 1)$$

本节内容

第一小节依次介绍了 Lindeberg-Lévy、Lindeberg-Feller、Lyapunov 中心极限定理, 并利用特征函数的方法证明了第一个结果。Lindeberg-Feller 中心极限定理给出了随机变量序列满足中心极限定理的充分必要条件, 即 Lindeberg 条件, 由此不难推导出 Lyapunov 中心极限定理。第二小节是中心极限定理在近似计算等方面的应用。

关键知识

(1) 通过模拟试验理解中心极限定理的概率意义; (2) 掌握中心极限定理的特征函数证法; (3) 了解 Lindeberg-Feller 中心极限定理中 Lindeberg 条件的含义; (4) 会利用中心极限定理做近似计算, 特别是 Lindeberg-Lévy 中心极限定理的应用。

5.2.1 Lindeberg-Feller 中心极限定理

中心极限定理是概率论的“无冕之王”，它在大量“混乱”之中揭示规律。Galton 甚至盛赞它所蕴涵的宇宙秩序之美妙无可比拟，可见其对人类认知的影响力是多么深远。本节所介绍的几个中心极限定理都是概率论的精品之作。

1920 年，芬兰数学家 Jarl Waldemar Lindeberg (1876-1932，照片见右) 发表中心极限定理的研究论文，以不同的方法独立重复了 Lyapunov 的某些工作。两年后，Lindeberg 得到了一个更好的结果，即中心极限定理成立的 Lindeberg 条件。1935 年，美国数学家 W. Feller 证明 Lindeberg 条件也是必要的。二人工作合称为 Lindeberg-Feller 中心极限定理，以前发现的诸多中心极限定理都是它的推论。



定理 5.17 (Lindeberg-Lévy 中心极限定理). 如果随机变量

$X_1, X_2, \dots, X_n, \dots$ 独立同分布，具有有限期望和方差 $E(X_n) = \mu, V(X_n) = \sigma^2 > 0$ ，则随机变量序列 $\{X_n\}$ 满足中心极限定理，即

$$Y_n = \frac{\frac{1}{n} \sum_{j=1}^n X_j - \mu}{\sigma / \sqrt{n}} = \frac{\sum_{j=1}^n (X_j - \mu)}{\sigma \sqrt{n}} \xrightarrow{L} N(0, 1)$$

证明. 随机变量 $X_n - \mu$ 和 Y_n 的特征函数分别是

$$\begin{aligned} \varphi(t) &= \varphi(0) + \frac{\varphi'(0)}{1!} t + \frac{\varphi''(0)}{2!} t^2 + o(t^2) = 1 - \frac{1}{2} \sigma^2 t^2 + o(t^2) \\ \tilde{\varphi}_n(t) &= \left[\varphi\left(\frac{t}{\sigma \sqrt{n}}\right) \right]^n = \left[1 - \frac{t^2}{2n} + o(t^2/n) \right]^n \end{aligned}$$

仿照定理 5.7 的证明，于是 $\ln \tilde{\varphi}_n(t) = -t^2/2 + no(t^2/n)$ ，显然 $\lim_{n \rightarrow \infty} \tilde{\varphi}_n(t) = \exp(-t^2/2)$ 。根据 Lévy 连续性定理， $Y_n \xrightarrow{L} N(0, 1)$ 得证。□

推论 5.2. 在 Lindeberg-Lévy 中心极限定理的条件之下，如果 n 充分大， $\frac{1}{n} \sum_{j=1}^n X_j$ 近似地服从如下的正态分布：

$$\frac{1}{n} \sum_{j=1}^n X_j \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (5.13)$$

 Lindeberg-Lévy 中心极限定理中方差 $V(X_n) < \infty$ 的条件必不可少，譬如随机变量序列 $X_1, \dots, X_n, \dots \stackrel{iid}{\sim} \text{Cauchy}(0, 1)$ ，每个随机变量的特征函数都是 $\varphi(t) = e^{-|t|}$ ，故 $\frac{1}{n} \sum_{j=1}^n X_j \sim \text{Cauchy}(0, 1)$ 。Lindeberg-Lévy 中心极限定理 5.17 也可轻易地推广到高维的情形。

定理 5.18. 已知 d 维随机向量 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots$ 独立同分布, 期望 μ 的每个分量都有限, 协方差矩阵 Σ 正定且 $|\Sigma| < \infty$, 则

$$\frac{\sum_{j=1}^n \mathbf{X}_j - \mu}{\sqrt{n}} \xrightarrow{L} N_d(\mathbf{0}, \Sigma)$$

例 5.19. 根据 Lindeberg-Lévy 中心极限定理, 图 4.14 所示的由 $U[0, 1]$ 的随机数构造正态分布随机数的方法也就显得很自然了。下图模拟了 $Y_n = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n (X_j - \mu)$ 随着 n 增加的演变情况, 其中 $\{X_n\}_{n=1}^\infty$ 满足 Lindeberg-Lévy 中心极限定理的条件, 但 X_n 的分布不是常见的。

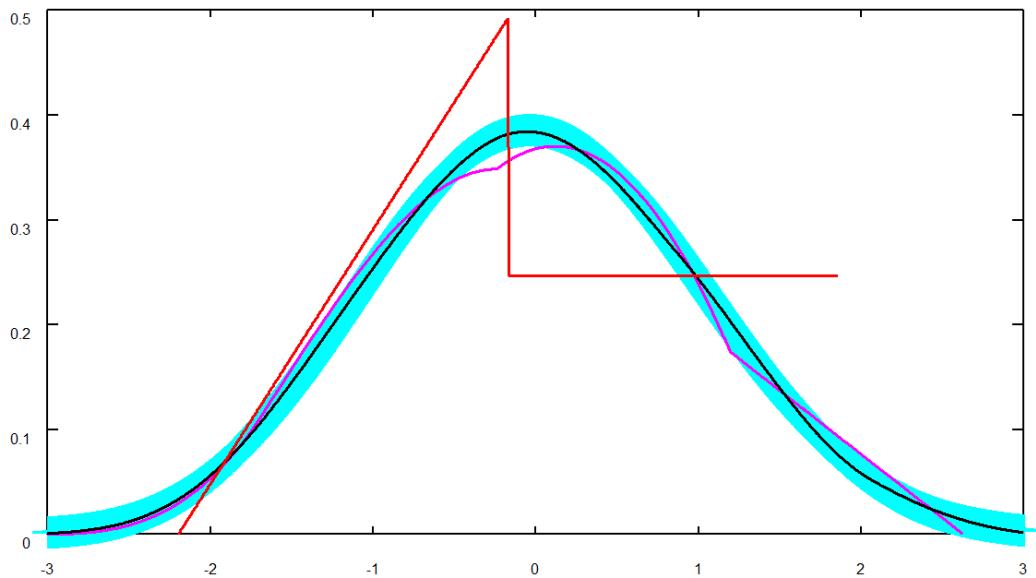


图 5.10: 独立同分布的随机变量 X_1, X_2, X_3 的密度函数是图中的折线, 期望为 0, 方差为 1。随机变量 $\frac{1}{\sqrt{2}}(X_1 + X_2)$ 的密度函数已基本进入 $\phi(x)$ 的一条窄带之中, 更不用说 $\frac{1}{\sqrt{3}}(X_1 + X_2 + X_3)$ 可视为服从标准正态分布。

给定一个随机变量序列 $\{X_n\}_{n=1}^\infty$, 大数律和中心极限定理都为问题“当 $n \rightarrow \infty$ 时, $S_n = \sum_{j=1}^n X_j$ 的极限状态是什么?”提供了部分答案, 它们之间有怎样的关系呢? 若 $\{X_n\}_{n=1}^\infty$ 满足 Lindeberg-Lévy 中心极限定理的条件 (即 X_1, X_2, \dots 独立同分布于一个有有限期望 μ 和非零有限方差 σ^2 的分布), 大数律只是说 $\frac{1}{n}S_n \xrightarrow{P} \mu$, 并没回答 $P\left\{ \left| \frac{1}{n}S_n - \mu \right| \leq \epsilon \right\}$ 究竟多大。而 Lindeberg-Lévy 中心极限定理则进一步描述这个收敛是按照式 (5.13) 的方式进行的, 并且指出 n 足够大时,

$$P\left\{ \left| \frac{1}{n}S_n - \mu \right| \leq \epsilon \right\} \approx 2\Phi\left(\frac{\epsilon\sqrt{n}}{\sigma} \right) - 1$$

定理 5.19 (Lindeberg-Feller 中心极限定理, 1922, 1935). 已知独立随机变量序列 $\{X_n\}$

满足 $E(X_n) = \mu_n$ 且 $V(X_n) = \sigma_n^2 > 0$ 。

$$\text{令 } Y_n = \sum_{j=1}^n \frac{X_j - \mu_j}{\tau_n} \text{ 和 } \tau_n = \sqrt{\sum_{j=1}^n \sigma_j^2}$$

$$\lim_{n \rightarrow \infty} \frac{\max_{1 \leq j \leq n} \sigma_j}{\tau_n} = 0 \text{ 且 } Y_n \xrightarrow{L} N(0, 1) \text{ 当且仅当 } \forall \epsilon > 0$$

$$\lim_{n \rightarrow \infty} \frac{1}{\tau_n^2} \sum_{j=1}^n \int_{|x-\mu_j|>\epsilon\tau_n} (x - \mu_j)^2 dF_j(x) = 0 \quad (5.14)$$

条件式 (5.14) 通常称为 Lindeberg 条件。

※证明. 详见 A. N. Shirayev 的《概率论》[145] 第三章第四节。 □

1935 年, 美籍克罗地亚裔数学家、二十世纪最杰出的概率论学者之一 W. Feller (1906-1970) 证得了 Lindeberg-Feller 中心极限定理的必要性。Lindeberg-Feller 中心极限定理的价值在于它的广泛性, 应用起来却非易事, 这是因为分布函数 $F_j(x), j = 1, 2, \dots$ 往往是未知的, 即便 $F_j(x)$ 已知, Lindeberg 条件式 (5.14) 中求极限的过程也是非常复杂的。下面给出的例 5.20 以及 Lyapunov 定理 5.20 和例 5.21 是 Lindeberg-Feller 中心极限定理的应用。

虽然应用起来很困难, Lindeberg-Feller 中心极限定理却蕴藏着深刻的思想, 如何理解其内在含义呢? 定义随机事件 $A_j = \{|X_j - \mu_j| > \epsilon\tau_n\}, j = 1, 2, \dots, n$, 则

$$\begin{aligned} P \left\{ \max_{1 \leq j \leq n} |X_j - \mu_j| > \epsilon\tau_n \right\} &= P \left\{ \bigcup_{j=1}^n A_j \right\} \\ &\leq \sum_{j=1}^n P(A_j) = \sum_{j=1}^n \int_{|x-\mu_j|>\epsilon\tau_n} dF_j(x) \\ &\leq \frac{1}{(\epsilon\tau_n)^2} \sum_{j=1}^n \int_{|x-\mu_j|>\epsilon\tau_n} (x - \mu_j)^2 dF_j(x) \end{aligned}$$



由 Lindeberg 条件式 (5.14), 于是

$$\lim_{n \rightarrow \infty} P \left\{ \max_{1 \leq j \leq n} \left| \frac{X_j - \mu_j}{\tau_n} \right| > \epsilon \right\} = 0$$

这说明每个随机变量 $X_j, j = 1, 2, \dots, n$ 在 $Y_n = \sum_{j=1}^n (X_j - \mu_j)/\tau_n$ 中所起的作用都是微不足道的，这是 Lindeberg 条件蕴含的结论。足够多这样“非本质”的独立随机变量的共同作用使得 Y_n 渐近于标准正态分布。Lindeberg-Feller 中心极限定理并不要求随机变量序列同分布，甚至可以把条件 $\sigma_j^2 > 0, j = 1, 2, \dots$ 减弱为 $\sigma_1^2, \sigma_2^2, \dots$ 不全为零。

练习 5.12. 设独立随机变量序列 $\{X_n\}$ 满足：(1) 存在常数 $c > 0$ 使得 $|X_n| < c, n = 1, 2, \dots$ ，即 $\{X_n\}$ 一致有界；(2) 方差 $V(X_n) = \sigma_n^2$ 存在，但是 $\sum_{n=1}^{\infty} V(X_n) = \infty$ 。试问强大数律和中心极限定理是否成立？为什么？

答案：都成立。 $\{X_n\}$ 满足强大数律是因为 $V(X_n) \leq c^2, n = 1, 2, \dots$ （参见第 362 页的练习 5.9）。 $\{X_n\}$ 满足中心极限定理是因为当 $n \rightarrow \infty$ 时， $\tau_n^2 = \sum_{j=1}^n \sigma_j^2 \rightarrow \infty$ ，于是对任意 $\epsilon > 0$ ，总存在 N 使得当 $n > N$ 时 Lindeberg 条件 (5.14) 中每个积分项的积分区域 $D_j = \{x : |x - \mu_j| > \epsilon \tau_n\}$ 满足 $P(X_j \in D_j) = 0$ ，从而 Lindeberg 条件 (5.14) 成立。

例 5.20. 试证明：Lindeberg-Feller 中心极限定理 \Rightarrow Lindeberg-Lévy 中心极限定理 \Rightarrow de Moivre-Laplace 中心极限定理。

证明. 只需往证 Lindeberg-Lévy 中心极限定理 5.17 的条件使得 Lindeberg 条件 (5.14) 成立，即

$$\lim_{n \rightarrow \infty} \int_{|\frac{x-\mu}{\sigma}| > \epsilon \sqrt{n}} \left(\frac{x-\mu}{\sigma} \right)^2 dF(x) = 1 - \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{\sigma} \right)^2 dF(x) = 1 - 1 = 0$$

如果 $Z_1, Z_2, \dots, Z_j, \dots \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ ，则 $X_n = \sum_{j=1}^n Z_j \sim B(n, p)$ ，由 Lindeberg-Lévy 中心极限定理 5.17 便可推导出结果 (5.11)，即

$$\frac{\frac{1}{n} \sum_{j=1}^n Z_j - p}{\sqrt{p(1-p)/n}} = \frac{\frac{1}{n} X_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{L} N(0, 1) \quad \square$$

例 5.21. 设 $\{X_n\}_{n=1}^{\infty}$ 是独立随机变量序列，并且 $X_n \sim \frac{1}{2}\langle -n^{\alpha} \rangle + \frac{1}{2}\langle n^{\alpha} \rangle, n = 1, 2, \dots$ ，其中 $\alpha > -1/2$ ，试证明： $\{X_n\}$ 满足中心极限定理。

证明. 算得 $E(X_j) = 0$ 且 $V(X_j) = j^{2\alpha}$ ，于是

$$\tau_n^2 = \sum_{j=1}^n j^{2\alpha} > \sum_{j=1}^n \int_{j-1}^j z^{2\alpha} dz = \frac{n^{2\alpha+1}}{2\alpha+1}$$

$\forall \epsilon > 0$ ，当 $\epsilon \tau_n > n^{\alpha}$ 时 Lindeberg 条件式 (5.14) 成立，因此只需 $\epsilon^2 n^{2\alpha+1} / (2\alpha+1) > n^{2\alpha}$ ，即 $n > (2\alpha+1)/\epsilon^2$ 。 \square

在概率论方面, Lyapunov 发展了 Chebyshev 和 Markov 的工作, 开创性地在中心极限定理的证明中使用了特征函数方法, 使得该方法在其后半个世纪里大放异彩。下面的中心极限定理即是 Lyapunov 于 1901 年发现的, 已经非常接近 Lindeberg 和 Feller 的结果了。

定理 5.20 (Lyapunov 中心极限定理, 1901). 对**定义 5.6** 描述的独立随机变量序列 $\{X_n\}$, 如果能找到正数 $\delta > 0$ 使得下述 Lyapunov 条件成立, 则随机变量序列 $\{X_n\}$ 满足中心极限定理。

$$\lim_{n \rightarrow \infty} \frac{1}{\tau_n^{2+\delta}} \sum_{j=1}^n E|X_j - \mu_j|^{2+\delta} = 0, \text{ 其中 } \tau_n^2 = \sum_{j=1}^n V(X_j) \quad (5.15)$$

证明. 只需验证 Lyapunov 条件 (5.15) 是 Lindeberg 条件 (5.14) 的特款即可。

$$\begin{aligned} \frac{1}{\tau_n^2} \sum_{j=1}^n \int_{|x-\mu_j|>\epsilon\tau_n} (x-\mu_j)^2 dF_j(x) &\leq \frac{1}{\tau_n^2(\epsilon\tau_n)^\delta} \sum_{j=1}^n \int_{|x-\mu_j|>\epsilon\tau_n} |x-\mu_j|^{2+\delta} dF_j(x) \\ &= \frac{1}{\epsilon^\delta} \left[\frac{1}{\tau_n^{2+\delta}} \sum_{j=1}^n E|X_j - \mu_j|^{2+\delta} \right] \end{aligned} \quad \square$$

例 5.22. 设 $\{X_n\}_{n=1}^\infty$ 是独立的离散型随机变量的序列, 其中

$$X_n \sim \frac{1}{2\sqrt{n}} \langle -n \rangle + (1 - \frac{1}{\sqrt{n}}) \langle 0 \rangle + \frac{1}{2\sqrt{n}} \langle n \rangle$$

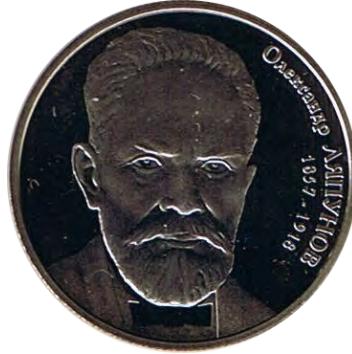
试证明: $\{X_n\}_{n=1}^\infty$ 满足中心极限定理。

证明. 利用 Lyapunov 条件 (5.15): 首先算得 $E(X_n) = 0, V(X_n) = E(X_n^2) = n^{\frac{3}{2}}$, 进而 $\tau_n^2 = \sum_{k=1}^n k^{\frac{3}{2}}$ 。对任意的 $\delta > 0$, 皆有

$$E|X_n|^{2+\delta} = 2 \cdot n^{2+\delta} \cdot \frac{1}{2\sqrt{n}} = n^{\frac{3}{2}+\delta} \text{ 和 } \tau_n^{2+\delta} = \left(\sum_{k=1}^n k^{\frac{3}{2}} \right)^{\frac{2+\delta}{2}}$$

对于任意 $\alpha > 0$, 下面的不等式总是成立的,

$$\begin{aligned} \frac{n^{\alpha+1}}{\alpha+1} &= \int_0^n x^\alpha dx \leq \sum_{k=1}^n k^\alpha \leq \int_0^{n+1} x^\alpha dx = \frac{(n+1)^{\alpha+1}}{\alpha+1} \\ \text{因此, } \sum_{k=1}^n E|X_k|^{2+\delta} &\leq \frac{(n+1)^{\frac{5}{2}+\delta}}{\frac{5}{2}+\delta} \text{ 并且 } \frac{n^{\frac{5}{4}(2+\delta)}}{\left(\frac{5}{2}\right)^{\frac{2+\delta}{2}}} \leq \tau_n^{2+\delta} \end{aligned}$$



由此可见, $\lim_{n \rightarrow \infty} \frac{1}{\tau_n^{2+\delta}} \sum_{k=1}^n \mathbb{E}|X_k|^{2+\delta} = 0$, 即 Lyapunov 条件 (5.15) 成立, 由定理 5.20, 于是 $\{X_n\}_{n=1}^\infty$ 满足中心极限定理。□

Lindeberg-Feller 中心极限定理是一个分水岭, 既涵盖了先前的结果, 又使得其后对中心极限定理的研究转向为: (1) 减弱对随机变量独立性的要求; (2) 与收敛速度有关的问题; (3) 与其他应用挂钩, 如 2007 年 Imre Bárány 和 Van Vu 证得的高斯多面体的中心极限定理。下面不加证明地介绍两个结果。

定理 5.21 (Bárány-Vu, 2007). 已知 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \stackrel{\text{iid}}{\sim} N_d(\mathbf{0}, I)$, 其中 I 是 $d \times d$ 的单位阵。 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ 的凸包 K_n 称作高斯随机多面体 (Gaussian random polytope), 令 X_n 是 K_n 的体积或面的个数, 则

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{V(X_n)}} \xrightarrow{L} N(0, 1)$$

定理 5.22. 对于多项分布 $\mathbf{X} = (X_1, X_2, \dots, X_k)^\top \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$, 当 $n \rightarrow \infty$ 时, \mathbf{X} 的每个分量经过标准化 $Y_j = (X_j - np_j)/\sqrt{np_j(1-p_j)}$ 后所得随机向量 \mathbf{Y} 依分布收敛于一个多元正态分布。同时也有

$$\sum_{j=1}^k (1-p_j) Y_j^2 \xrightarrow{L} \chi_{k-1}^2$$

5.2.2 中心极限定理的应用

定理 5.23. 对于二项分布 $X \sim B(n, p)$, 当 n 很大时, 有近似计算公式

$$P(a < X \leq b) \approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right)$$

例 5.23. 抛硬币出现正面的概率是 0.5, 抛此硬币 100 次, 出现正面的次数记为 X_{100} 。试求: $P(50 < X_{100} \leq 60)$ 。

解. 由**定理 5.23**, $P(50 < X_{100} \leq 60) \approx \Phi(2) - \Phi(0) \approx 0.4772499$ 。

例 5.24. 对一批产品进行抽样检查, 若发现次品数多于 10 件时, 则认为这批产品不合格。求应检查多少件产品, 才能使次品率为 10% 的一批产品以 90% 的概率被认为不合格?

解. 设应检查 n 件产品, 则次品数 $X \sim B(n, 0.1)$ 。由**定理 5.23**, 这批产品被认为不合格的概率为

$$P\{10 < X \leq n\} \approx \Phi(3\sqrt{n}) - \Phi\left(\frac{10 - 0.1n}{0.3\sqrt{n}}\right)$$

依题意知 $n > 10$, 故 $\Phi(3\sqrt{n}) \approx 1$ 且 $10 - 0.1n < 0$, 故 $P\{10 < X \leq n\} \approx \Phi[(0.1n - 10)/(0.3\sqrt{n})] \geq 0.9$, 求得 $n \geq 147$ 。

练习 5.13. 某汽车制造厂生产汽车发动机的合格率为 0.8, 为了能以 0.997 的概率保证每月组装的 10000 辆汽车都装上合格发动机, 问该厂每月应生产多少台发动机? (答案: 12655)

例 5.25. 若事件 A 发生的概率为 $p = 0.7$, 在 n 重 Bernoulli 试验中, 要使 A 出现的频率在 0.68 与 0.72 间的概率至少为 0.9, 问至少要做多少次试验? 如果进行 1000 次试验, 事件 A 发生的次数在 650 至 750 次之间的概率是多少? 请分别利用 (i) Chebyshev 不等式; (ii) 中心极限定理来估计, 并比较不同方法所得的结果。

解. 令随机变量 $X_n \sim B(n, p)$ 表示在 n 重 Bernoulli 试验中事件 A 出现的次数, 下面分别用 Chebyshev 不等式和中心极限定理来估计 $P\{0.68 < X_n/n < 0.72\}$ 。

□ 利用 Chebyshev 不等式 (2.75), 我们有

$$\begin{aligned} P\left\{0.68 < \frac{X_n}{n} < 0.72\right\} &= P\{|X_n - np| < 0.02n\} \\ &\geq 1 - \frac{0.3 \times 0.7n}{(0.02n)^2} = 0.9 \end{aligned}$$

解之得 $n = 5250$, 即用 Chebyshev 不等式求得至少需要 5250 次试验。下面用 Chebyshev 不等式估计 $P(650 < X_{1000} < 750)$ 的下界,

$$\begin{aligned} P(650 < X_{1000} < 750) &= P(-50 < X_{1000} - 1000p < 50) \\ &\geq 1 - \frac{0.3 \times 0.7 \times 1000}{50^2} = 0.916 \end{aligned}$$

□ 由定理 5.23 (该定理由 de Moivre-Laplace 中心极限推得),

$$\begin{aligned} P\left\{0.68 < \frac{X_n}{n} < 0.72\right\} &= P\{0.68n < X_n < 0.72n\} \\ &\approx \Phi\left(\frac{0.72n - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{0.68n - np}{\sqrt{np(1-p)}}\right) \\ &= 2\Phi\left(\frac{0.02n}{\sqrt{0.21n}}\right) - 1 = 0.9 \end{aligned}$$

即 $\Phi(0.02n/\sqrt{0.21n}) = 0.95$ 。利用 R 语言的命令 qnorm(0.95) 求得 $0.02n/\sqrt{0.21n} \approx 1.644854$, 解之得 $n \approx 1420.41$, 即利用定理 5.23 求得至少需要 1421 次试验。

下面用定理 5.23 估计 $P(650 < X_{1000} < 750)$,

$$\begin{aligned} P(650 < X_{1000} < 750) &\approx \Phi\left(\frac{749 - 0.7 \times 1000}{\sqrt{0.21 \times 1000}}\right) - \Phi\left(\frac{650 - 0.7 \times 1000}{\sqrt{0.21 \times 1000}}\right) \\ &\approx 0.9994 \end{aligned}$$



该例再一次印证了第 195 页的例 2.77 之后的讨论, 即 Chebyshev 不等式不适用于精确地估计概率, 它只能粗糙地给出所估概率的某个下界或上界, 对付类似该例的问题还是中心极限定理管用些。

例 5.26. 在某路边报亭经过的人购买报纸的概率是 $1/3$, 令随机变量 X 表示出售了 100 份报纸时过路人的总数, 试求 $P(280 < X \leq 320)$ 。

解. 设随机变量 X_k 表示第 $k-1$ 份报纸被买走后, 过报亭的第 X_k 个人购买第 k 份报纸, 则 $X = \sum_{k=1}^{100} X_k$ 。显然 $X_1, X_2, \dots, X_k, \dots, X_{100}$ 是独立同分布的, 其中 X_k 的分布列为

$$P(X_k = j) = \left(\frac{2}{3}\right)^{j-1} \times \frac{1}{3}, \text{ 其中 } j = 1, 2, \dots$$

求得 $\mu = E(X_k) = 3$ 和 $\sigma^2 = V(X_k) = 6$ 。由 Lindeberg-Lévy 中心极限定理 5.17,

近似地有 $\frac{X-100\mu}{10\sigma} \sim N(0, 1)$, 因此

$$P(280 < X \leq 320) = 2\Phi(\sqrt{2/3}) - 1 = 0.5857838$$

例 5.27. 某测量值 $X_j \sim N(\mu, \sigma^2), j = 1, 2, \dots$, 其中 $\sigma^2 = 10^{-2}$ 。需要测量多少次才能使得

$$P\left\{\left|\frac{1}{n} \sum_{j=1}^n X_j - \mu\right| < 10^{-4}\right\} = 0.95$$

解. 由 Lindeberg-Lévy 中心极限定理知, 当 n 足够大时, 近似地

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n X_j - \mu &\sim N(0, \sigma^2/n) \\ \text{于是, } P\left\{\left|\frac{1}{n} \sum_{j=1}^n X_j - \mu\right| < 10^{-4}\right\} &= \Phi(10^{-4} \sqrt{n}/\sigma) - \Phi(-10^{-4} \sqrt{n}/\sigma) \\ &= 2\Phi(\sqrt{n}/100) - 1 \end{aligned}$$

所以 $2\Phi(\sqrt{n}/100) - 1 = 0.95$ 或 $\Phi(\sqrt{n}/100) = 0.975$ 。进而, 求得 $n \geq 38415$ 次测量。

~定理 5.24 (Fisher, 1925). 如果随机变量 $X \sim \chi_n^2$, 则

$$\lim_{n \rightarrow \infty} P\{\sqrt{2X} - \sqrt{2n-1} \leq z\} = \Phi(z)$$

证明. 因为 X 是 n 个独立同分布的 χ_1^2 随机变量之和, 由 Lindeberg-Lévy 中心极限定理 5.17, 当 $n \rightarrow \infty$ 时, 渐近地有 $X \sim N(n, 2n)$ 或 $Z_n = \frac{X-n}{\sqrt{2n}} \sim N(0, 1)$, 即

$$\begin{aligned} \lim_{n \rightarrow \infty} P\left\{\frac{X-n}{\sqrt{2n}} \leq z\right\} &= \Phi(z), \text{ 同时我们还有} \\ \lim_{n \rightarrow \infty} P\left\{\frac{X-n}{\sqrt{2n}} \leq z\right\} &= \lim_{n \rightarrow \infty} P\left\{\frac{X-n}{\sqrt{2n}} \leq z + \frac{z^2-1}{2\sqrt{2n-1}}\right\} \\ &= \lim_{n \rightarrow \infty} P\left\{\frac{X-n}{\sqrt{2n-1}} \leq z + \frac{z^2-1}{2\sqrt{2n-1}}\right\} \\ &= \lim_{n \rightarrow \infty} P\left\{X \leq n + z\sqrt{2n-1} + \frac{z^2-1}{2}\right\} \\ &= \lim_{n \rightarrow \infty} P\{\sqrt{2X} \leq z + \sqrt{2n-1}\} \end{aligned} \quad \square$$

在实践中, 定理 5.24 当 $n \geq 30$ 时便有 $P\{\sqrt{2X} - \sqrt{2n-1} \leq z\} \approx \Phi(z)$, 于是定

理 5.24 有如下在近似计算上的应用。

$$\begin{aligned} P(X \leq z) &= P\{\sqrt{2X} - \sqrt{2n-1} \leq \sqrt{2z} - \sqrt{2n-1}\} \\ &\approx \Phi(\sqrt{2z} - \sqrt{2n-1}) \end{aligned}$$

于是, χ_n^2 分布的密度函数 (见定义 4.17) 可近似为

$$g(z) = \frac{\phi(\sqrt{2z} - \sqrt{2n-1})}{\sqrt{2z}}, \text{ 其中 } z > 0 \quad (5.16)$$

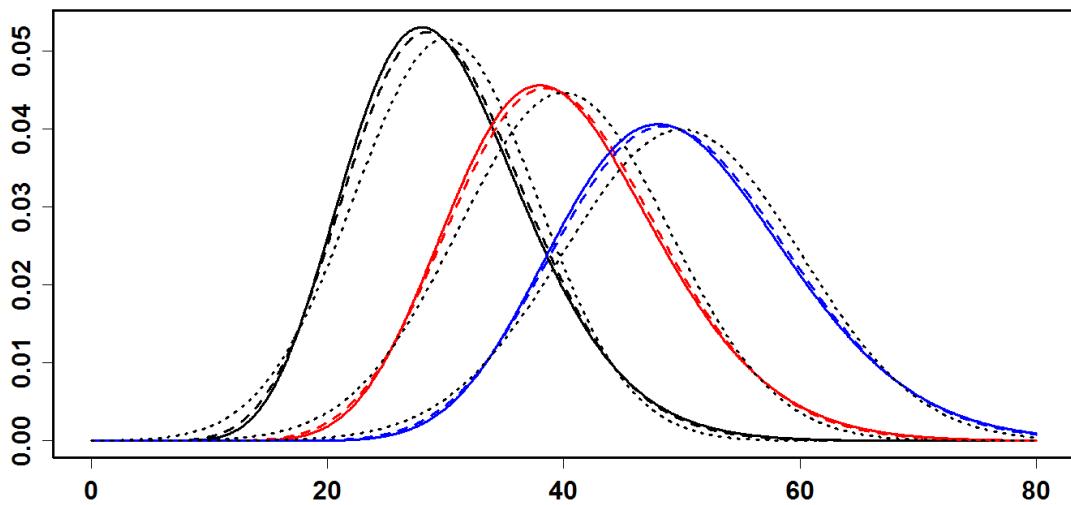


图 5.11: 实线从左至右依次是 $n = 30, 40, 50$ 时 χ_n^2 分布的密度函数曲线, 虚线是由式 (5.16) 定义的曲线, 点线是 $N(n, 2n)$ 的密度函数曲线, 它们相应地都是对 χ_n^2 分布的近似。

5.3 习题

- 5.1. 已知 $\{X_k\}_{k=1}^{\infty}$ 为独立随机变量序列，并且 $X_k \sim \frac{1}{k+1}\langle -\sqrt{k+1} \rangle + (1 - \frac{2}{k+1})\langle 0 \rangle + \frac{1}{k+1}\langle \sqrt{k+1} \rangle, k = 1, 2, \dots$ 。试证明： $\{X_k\}$ 满足弱大数律。
- 5.2. 已知 $\{X_k\}$ 为独立随机变量序列，并且 $X_k \sim \frac{1}{2}\langle k^s \rangle + \frac{1}{2}\langle -k^s \rangle, k = 1, 2, \dots, n, \dots$ 。
试证明：当 $s < 1/2$ 时， $\{X_k\}$ 满足弱大数律。
- ☆ 5.3. 将标号为 $1, 2, \dots, n$ 的 n 个球随机放入标号为 $1, 2, \dots, n$ 的盒中，每盒放一个球，设 S_n 为球与盒子的号码相同的个数。试证明： $\forall \epsilon > 0$, 有 $\lim_{n \rightarrow \infty} P\{|\frac{1}{n}S_n - \frac{1}{n}| \geq \epsilon\} = 0$ 。
- ★ 5.4. 设随机变量 X_1, X_2, \dots 的方差都不超过 $c > 0$ ，当 $|k - j| \rightarrow \infty$ 时 X_k 和 X_j 的相关系数 $\rho_{kj} \rightarrow 0$ 。试证明： $\{X_i\}_{i=1}^{\infty}$ 满足弱大数律。
- ☆ 5.5. 设 $\{X_k\}$ 为独立随机变量序列， X_k 具有有限方差 $V(X_k), k = 1, 2, \dots$ 且 $\sum_{k=1}^{\infty} V(X_k)/k^2 < \infty$ ，试证明： $\{X_k\}$ 满足弱大数律。
- 5.6. 设随机变量序列 $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} U[0, 1]$ ，令 $Y_n = (\prod_{k=1}^n X_k)^{1/n}$ 。
(1) 试证明： $Y_n \xrightarrow{P} c$ ，其中 c 是某常数；
(2) 试求 c 。
- ☆ 5.7. 设 $h(x)$ 是在 $(0, \infty)$ 上的连续单调增函数，且 $\lim_{x \rightarrow 0} h(x) = 0, \sup h(x) < \infty$ ，求证：
随机变量序列 $X_n \xrightarrow{P} 0$ 的充要条件是 $\lim_{n \rightarrow \infty} E[h(|X_n|)] = 0$ 。
- 5.8. 设 $X_1, X_2, \dots, X_n, \dots$ 为独立同分布的随机变量序列， $E(X_n) = \mu, V(X_n) = \sigma^2$ ，
试证明： $\frac{2}{n(n+1)} \sum_{k=1}^n kX_k \xrightarrow{P} \mu$ 。
- ☆ 5.9. 已知独立随机变量序列 $X_1, X_2, \dots, X_n, \dots$ 满足 $X_n \sim \frac{1}{2}n^{-1/3}\langle -\sqrt{n} \rangle + \frac{1}{2}(1 - n^{-1/3})\langle -1 \rangle + \frac{1}{2}(1 - n^{-1/3})\langle 1 \rangle + \frac{1}{2}n^{-1/3}\langle \sqrt{n} \rangle$ ，问随机变量序列 $\{X_n\}$ 是否满足中心极限定理？
- 5.10. 设独立同分布的随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 满足 $E(X_n) = 0, V(X_n) = 1$ 。试证明：当 $n \rightarrow \infty$ 时，随机变量 $Y_n = \frac{X_1 + \dots + X_n}{\sqrt{X_1^2 + \dots + X_n^2}} \sqrt{n}$ 和 $Z_n = \frac{X_1 + \dots + X_n}{\sqrt{X_1^2 + \dots + X_n^2}}$ 的极限分布都是标准正态分布。
- 5.11. 已知随机变量 X_1, X_2, \dots, X_{100} 独立同分布且 $E(X_1) = 1, V(X_1) = 2.4$ ，计算 $P\{\sum_{i=1}^{100} X_i \geq 90\}$ 。
- 5.12. 设事件 A 在随机试验中发生的概率为 $1/4$ ，独立重复 400 次这样的试验，利用中心极限定理计算事件 A 发生的次数在 50 到 150 之间的概率为多少？
- 5.13. 某厂产品中优等品率为 20%，现从该厂的产品中随机地抽出 100 件，设优等品的个数为 X ，求概率 $P(18 < X \leq 25)$ ？

- 5.14. 为 n 个正实数的加法 $x_1 + x_2 + \dots + x_n$ 设计一个算法：先对每个加数四舍五入取整，再将所得到的这些整数相加。按此算法将任意 1200 个正实数相加，求总误差的绝对值不超过 15 的概率？
- ☆ 5.15. 已知连续型随机变量序列 $X_1, X_2, \dots, X_n, \dots$ 独立同分布，密度函数都为 $f(x)$ ，分布函数 $F(x)$ 是 x 的严格增函数，试求 $\lim_{n \rightarrow \infty} P\{\frac{1}{n} \sum_{i=1}^n F(X_i) \leq \frac{1}{2}\}$ 。
- 5.16. 已知随机变量 $X_1, X_2, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} \text{Expon}(\lambda)$ ，令 $Y_n = \frac{1}{n} \sum_{j=1}^n X_j^2$ 。（1）试证明： $Y_n \xrightarrow{P} 2/\lambda^2$ ；（2）问当 n 充分大时， Y_n 服从什么分布？
- 5.17. 设随机变量 X_1, X_2, \dots, X_n 独立同分布， $E(X_1^k) = m_k, k = 1, 2, 3, 4$ 都存在且 $m_4 - m_2^2 > 0$ 。问当 n 充分大时，随机变量 $Y_n = \frac{1}{n} \sum_{i=1}^n X_i^2$ 近似地服从什么分布？
- 5.18. 设 $\{X_n\}_{n=1}^\infty$ 是独立同分布的随机变量序列且 $E(X_n) = V(X_n) = 1$ ，求常数 c 使得 $\lim_{n \rightarrow \infty} P\{\frac{c}{\sqrt{n}} \sum_{j=1}^n (X_{2j} - X_{2j-1}) \leq x\} = \phi(x)$ 。
- ☆ 5.19. 利用中心极限定理证明： $\lim_{n \rightarrow \infty} e^{-n} \sum_{k=0}^n n^k / k! = 1/2$ 。

第六章

随机过程简介

前不见古人，后不见来者。念天地之悠悠，独怆然而涕下。

陈子昂《登幽州台歌》

以时间为轴将随机现象“串起来”而得到的自然过程也带有随机性，在数学里被抽象为随机过程。例如，某地的年降雨量受很多随机因素的影响而成为随机现象，以时间为指标这些降雨量的序列 $\{X_n : n = 1, 2, \dots\}$ 就是一个随机过程。再如，某时间段内股票价格或成交量的波动。因为现实世界中大多数过程都多多少少带有随机性，随机过程的数学理论对自然科学、工程技术中这类问题的研究都极为重要，另外它与微分方程、复变函数、位势论等众多数学分支都相互渗透，成为随机分析中的主要组成部分 [85, 138, 139]。

随机过程的研究始于对一组不相互独立的随机事件的概率模型的探讨——俄国圣彼得堡学派的代表人物 A. A. Markov (1856-1922) 于 1907 年提出的 Markov 链和 Markov 过程。Markov 过程是一类特殊的随机过程，它具有所谓的“Markov 性”，即在当前的条件之下，未来的演化独立于过去的状态，像传染病的受感染人数、液体中悬浮微粒的布朗运动等都可视为 Markov 过程 [86]。

例 6.1. 某人有 100 元赌资，抛一枚均匀的硬币，出现正面赚一元，出现反面赔一元，一直赌下去直至赌资为零。

令 X_n 表示抛 n 次硬币后的赌资，则序列 $\{X_0 = 100, X_1, \dots\}$ 是一个 Markov 过程。

定义 6.1 (随机过程). 设 (Ω, \mathcal{S}, P) 为一个概率空间， T 为某个连续的或离散的时间指标集，一般为实数集合 $\mathbb{R} = (-\infty, \infty)$ 或非负整数集合 $\{0, 1, 2, \dots\}$ 或自然数集合 $\{1, 2, \dots\}$ 。一个随机过程 (stochastic process)，简称过程，就是一族随机变量



$X = \{X(t) : t \in T\}$, 其中每个 $X(t)$ 都是定义在 Ω 上的随机变量。在不引起歧义的前提下, $X(t)$ 有时记作 X_t , 随机过程 X 记作 $\{X_t\}$ 或者 X_t 。

定义 6.2. 随机过程 $X = \{X(t) : t \in T\}$ 中所有随机变量的所有可能取值的全体被称为状态空间, 记作 \mathbb{S} , 其中每个元素称作一个状态 (state)。状态空间有离散和连续之分。

定义 6.3. 若 T 为连续统, 则称 X 为连续时间过程; 若 T 为可数集, 则称 X 为离散时间过程, 也称为随机序列 (random sequence) 或时间序列 (time series)。当 $T \subseteq \mathbb{R}^d$ 时, 其中 d 为大于 1 的自然数, 则称随机过程 X 为多指标随机过程。

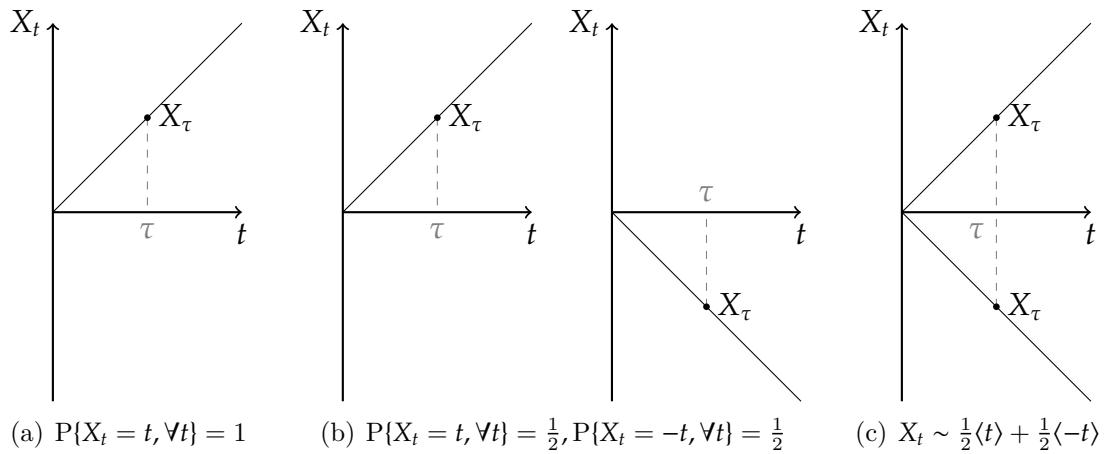


图 6.1: 随机过程关注变量间的依赖关系和过程的长期行为特征; 随机过程 (a) 是可预测的。(b) 一旦经过二选一, 也是可预测的。但是, (c) 在任何时刻 τ , 都不是可预测的 $X_\tau \sim \frac{1}{2}\langle \tau \rangle + \frac{1}{2}\langle -\tau \rangle$ 。

表 6.1: 随机过程的四种类型: discrete-time (DT), continuous-time (CT) 和 discrete-valued (DV), continuous-valued (CV) 的四个组合。

时间	离散	连续
取值		
离散	DTDV	CTDV
连续	DTCV	CTCV

离散时间过程很常见, 我们可以在相邻的时间点上考察随机变量间的关系。DTCV 过程如 $X_t \stackrel{\text{iid}}{\sim} N(0, 1)$, 其中 $t = 0, 1, \dots$ 。DTDV 过程如下面的例子。

例 6.2 (Bernoulli 过程). 多重 Bernoulli 试验 $X_1, X_2, \dots, X_t, \dots \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 是一个随机过程, 称作 参数为 p 的 Bernoulli 过程 (Bernoulli process with parameter p), 状态空间是 $\mathbb{S} = \{0, 1\}$, 其中 1 表示 “成功”。下面三种离散分布通过 Bernoulli 过程得到新的解释, 有助于我们搞清楚它们的研究背景。

- n 次试验中成功的次数 $S_n = X_1 + \dots + X_n \sim \text{B}(n, p)$ 。 $\{S_n\}$ 也是一个随机过程，被称为（离散时间）二项过程 (binomial process)，二项过程是一类计数过程 (counting process)，显然 $S_n \leq S_{n+1}$ 。
- 出现一次成功所需的试验次数服从几何分布 $\text{Geom}(p)$ 。
- 获得 k 次成功所需要的试验次数服从负二项分布，即

$$Y_k = \min\{n : S_n = k\} \sim \text{NegB}(k, p)$$

证明. 下面验证 Y_k 服从负二项分布：

$$\begin{aligned} P(Y_k = t) &= P(S_{t-1} = k-1, X_t = 1) \\ &= P(S_{t-1} = k-1)P(X_t = 1) \\ &= C_{t-1}^{k-1} p^{k-1} (1-p)^{t-k} \cdot p \\ &= C_{t-1}^{k-1} p^k (1-p)^{t-k} \end{aligned}$$

□

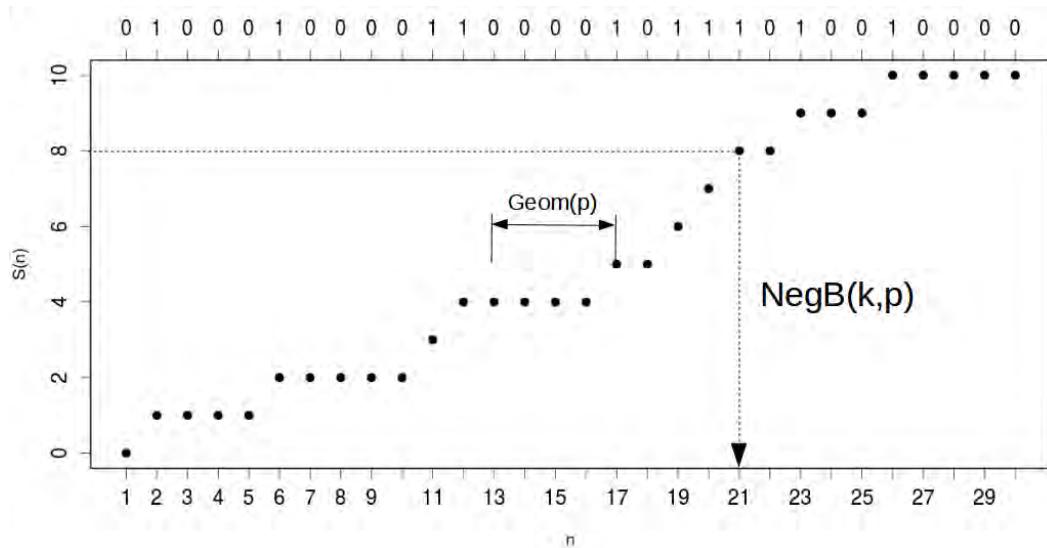


图 6.2: Bernoulli 过程：相邻两次成功之间所间隔的试验次数服从几何分布 $\text{Geom}(p)$ ，获得 k 次成功所需要的试验次数服从负二项分布 $\text{NegB}(k, p)$ 。

例 6.3 (简单随机游动). 已知 $X_1, X_2, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle -1 \rangle$ ，如下定义的随机过程被称为简单随机游动 (simple random walk)，其状态空间是整数集合 \mathbb{Z} 。

$$S_0 = 0$$

$$S_n = X_1 + \dots + X_n, \text{ 其中 } n = 1, 2, \dots$$

例 6.4. 由 Bernoulli 过程 $X_0, X_1, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} 0.3\langle 1 \rangle + 0.7\langle 0 \rangle$ 构造 CTDV 随机过程 Y_t , 其状态空间是非负整数集合。

$$Y_t = \sum_{j=0}^{\lfloor t \rfloor} X_j, \text{ 其中 } t \in [0, +\infty)$$

定义 6.4. 就如同对随机变量进行抽样产生随机数, 对随机过程 X_t 的一次抽样 $\{X_t(\omega) : \omega \in \Omega\}$ 被称为该随机过程的一个实现 (realization) 或者一个样本路径 (sample path)。随机过程研究的对象是所有可能的样本路径上的概率分布。

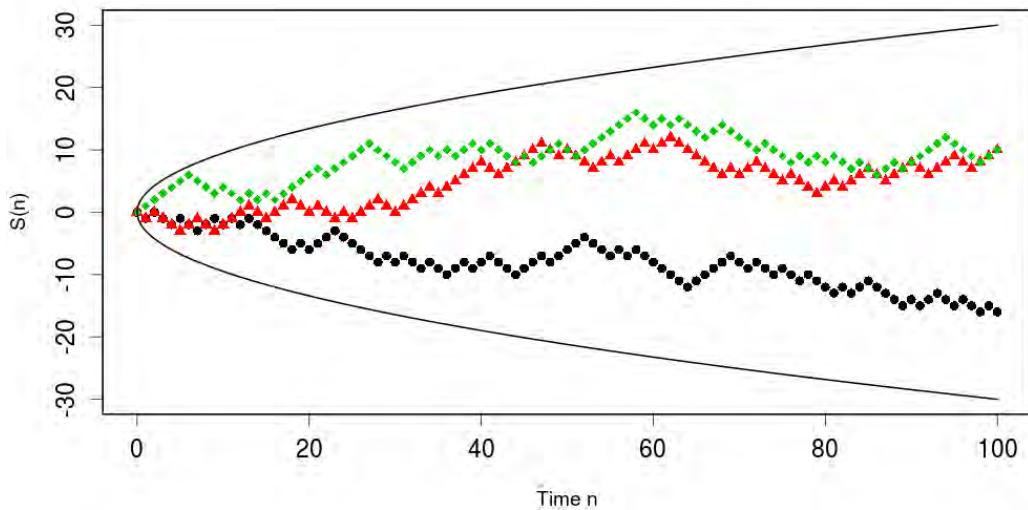


图 6.3: 在例 6.3 中, $p = 0.5$ 定义的简单随机游动的三个样本路径。请读者验证: $E(S_n) = 0, V(S_n) = n$ 。想一想 S_n 服从什么样的分布?

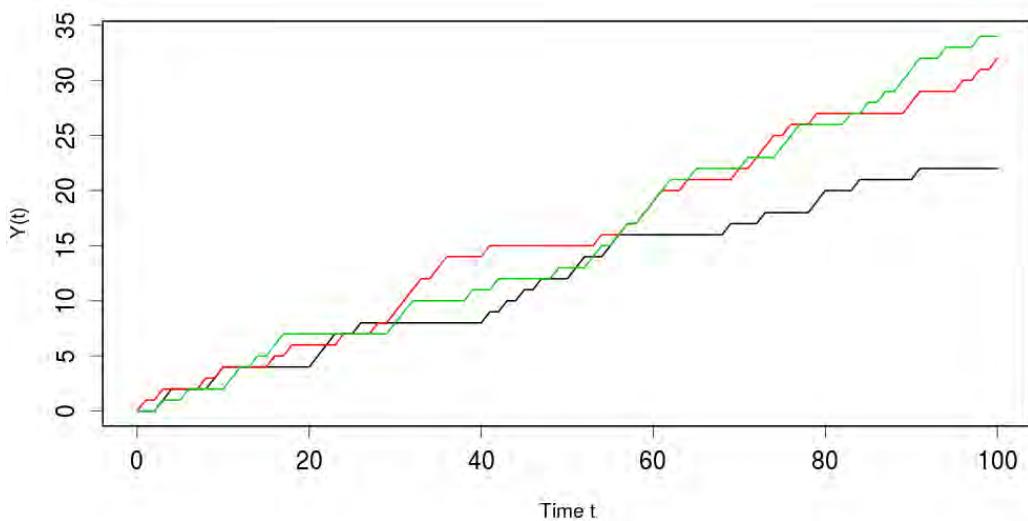


图 6.4: 例 6.4 定义的连续时间二项过程 Y_t 的三个样本路径。

1827 年，英国植物学家 Robert Brown (1773-1858) 在显微镜下发现悬浮在水面上的花粉迸出的细微颗粒有不规则的运动，这种现象被称为布朗运动 (Brownian motion)。后来发现，灰尘等细微颗粒也有类似的现象，Brown 并没有从理论上给出该现象的合理解释。布朗运动的数学建模几乎迟到了一百年，其物理解释也是在二十世纪初才出现（即，由水中的水分子对细微颗粒的碰撞造成）。人们看清布朗运动的数学本质，非得借助随机过程这一高级工具。回顾历史，布朗运动的数学模型是二十世纪最重要的数学发现之一，深刻地影响着随机分析的发展。



图 6.5: 布朗运动可用于描述自然界的地形地貌。

布朗运动同时引起了数学家和物理学家的兴趣，虽然在二十世纪初它的严格理论还没有建立起来，但以它为工具的应用已经开始了。在数学史中，先用起来后补基础的理论也不罕见，如数学分析。



1900 年，金融数学的先驱、法国数学家 Louis Bachelier (1870-1946) 在他的博士论文里首次提出利用布朗运动模型来研究股票价格的变化，被视为金融数学的开山之作。Bachelier 的研究工作虽然得到他的导师 Henri Poincaré 的赞扬和支持，但在当时并未得到应有的认可，以至于 Bachelier 的职业生涯充满坎坷，包括 Paul Lévy 对他的误解和不公正的评价。尽管 Lévy 后来为此道歉并赢得和解，然而像 Lévy 这样杰出的概率论大师始终低估了 Bachelier 对金融数学的巨大贡献。令 Lévy 惊奇的是，他眼中的数学奇才 Kolmogorov 却视 Bachelier 的工作为珍宝。显然，Kolmogorov 在扩散过程的工作受到 Bachelier 的影响。为纪念 Bachelier，Feller 在他的名著《概率论及其应用》中把布朗运动称为 Wiener-Bachelier 过程。

1905 年是 Einstein 的奇迹年，是年他发表了六篇学术论文，几乎篇篇在物理学里都是划时代的。其中，论文《热的分子运动论所要求的静液体中悬浮粒子的运动》建立了布朗运动的物理模型。Einstein 发现布朗粒子在时刻 t 处于位置 x 的密度函数 $\rho(x, t)$ 满足以下扩散方程，其中 D 是质量扩散系数。



$$\frac{\partial \rho}{\partial t} = D \frac{\partial^2 \rho}{\partial x^2}$$

若初始时刻 $t = 0$ 粒子处于原点，上述扩散方程的解是

$$\rho(x, t) = \frac{\rho_0}{\sqrt{4\pi Dt}} \exp\left\{-\frac{x^2}{4Dt}\right\}$$



1923 年，美国数学家、控制论之父 Norbert Wiener (1894-1964) 给出了布朗运动的数学定义。布朗运动是 CTCV 过程，譬如一维的布朗运动 $\{B_t\}$ ，状态空间是 \mathbb{R} ，对于任意的时间列 $t_1 \leq t_2 \leq \dots \leq t_n$ ， $(B_{t_1}, B_{t_2}, \dots, B_{t_n})^\top$ 都服从多元正态分布。我们将在 §6.2.3 专门讨论布朗运动，它的数学和物理意义都是非凡的。

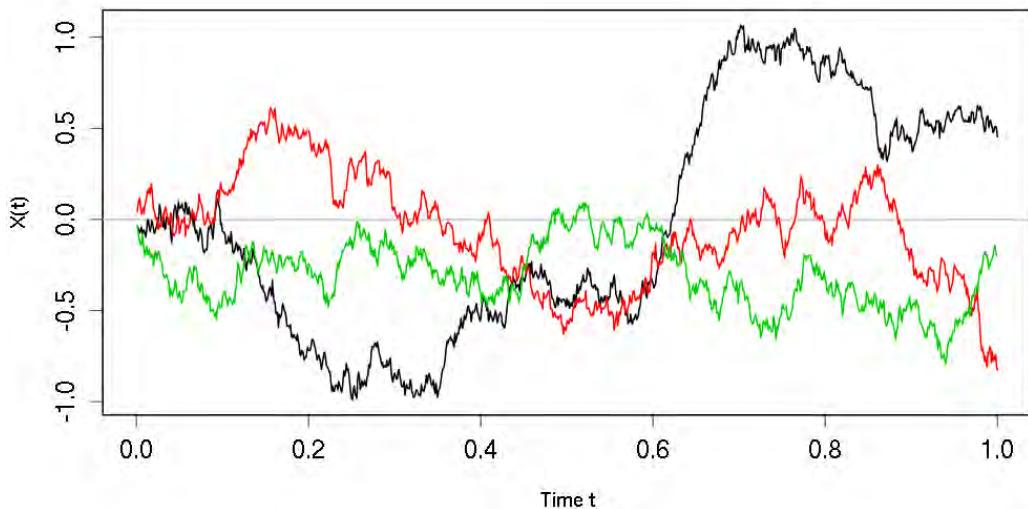


图 6.6: 一维布朗运动的三个样本路径，状态空间是 \mathbb{R} 。

十八世纪的法国流行一种加倍投注的赌博策略：赌徒先报赌多少钱，譬如 M 元，若抛硬币出现正面则赌徒赢得 M 元，若出现反面则输掉 M 元。在输了的情况下，赌徒为了赢回上次输掉的钱，同时又能取得上次赢钱的效果，往往把赌注加倍为 $2M$

元（这次若赢了就算白赢，但总体效果是上次赢了）。除非赌资无穷，否则加倍投注的赌徒很快就会破产。

为证明不存在成功的赌博策略，法国概率大师 Paul Lévy 于 1934 年引入一类特殊的随机过程——鞅 (martingale)，后续由美国数学家 Joseph Leo Doob (1910-2004，照片见右) 系统发展为鞅论。从上世纪七十年代起，鞅论广泛应用于金融数学、数学物理等分支。

据说，martingale 这个词来源于法国的一个小镇 Martique。这个小镇的居民很小气，明天将花的钱的期望值就是今天花的钱。这个习性用数学的语言描述就是鞅。



定义 6.5 (鞅). 离散时间随机过程 X_1, X_2, \dots ，如果满足以下条件，则分别称之为鞅 (martingale)、下鞅 (submartingale) 和上鞅 (supermartingale)。

鞅: $E(X_{n+1}|X_1 = x_1, \dots, X_n = x_n) = x_n$, 其中 $E(|X_n|) < \infty$

下鞅: $E(X_{n+1}|X_1 = x_1, \dots, X_n = x_n) \geq x_n$

上鞅: $E(X_{n+1}|X_1 = x_1, \dots, X_n = x_n) \leq x_n$

房价是一个下鞅，随时间均值增高。生活是一个上鞅，随时间期望降低。

对随机变量的研究可以转嫁到对它的分布函数的研究上，那么，对于一个随机过程而言情况又如何呢？

定义 6.6. 随机过程 X 实质上就是两变元 (t, ω) 的函数：当 $\omega \in \Omega$ 固定时， X 称为对应于 ω 的轨道；当 $t \in T$ 固定时， X 是一个随机变量 $X(t)$ ，它的分布函数称为 X 的一维分布函数，即 $F_t(x) = P\{X(t) \leq x\}$ 。

定义 6.7 (有限维分布族). 对任意 n 个相异的时间点 $t_1, t_2, \dots, t_n \in T$ ，随机变量 $X(t_1), X(t_2), \dots, X(t_n)$ 的联合分布为

$$F_{t_1, t_2, \dots, t_n}(x_1, x_2, \dots, x_n) = P\{X(t_1) \leq x_1, X(t_2) \leq x_2, \dots, X(t_n) \leq x_n\}$$

随机过程 X 的有限维分布族就是这些联合分布函数的全体，即

$$\mathcal{F} = \{F_{t_1, t_2, \dots, t_n}(x_1, x_2, \dots, x_n) : n \in \mathbb{N} \text{ 且 } t_1, t_2, \dots, t_n \in T\}$$

例 6.5. 有限维分布都是正态分布的随机过程称为正态过程，或者高斯过程 (Gaussian process)。当它是 DTCV 过程时，我们称之为离散高斯过程；当它是 CTCV 过程时，我们称之为连续高斯过程。

定理 6.1 (Kolmogorov, 1931). 随机过程 $X = \{X(t) : t \in T\}$ 的有限分布族 \mathcal{F} 满足以下两个性质。反之，如果对于指标集 T ，一族分布函数 $\mathcal{F} = \{F_{t_1, t_2, \dots, t_n}(x_1, x_2, \dots, x_n) : n \in \mathbb{N} \text{ 且 } t_1, t_2, \dots, t_n \in T\}$ 满足这两个性质，则存在概率空间 (Ω, \mathcal{S}, P) 和定义于该概率空间上的随机过程 X 使得 \mathcal{F} 即是 X 的有限分布族。

□ 对称性：对 $(1, 2, \dots, n)$ 的任意排列 $(\pi_1, \pi_2, \dots, \pi_n)$ 皆有

$$F_{t_1, t_2, \dots, t_n}(x_1, x_2, \dots, x_n) = F_{t_{\pi_1}, t_{\pi_2}, \dots, t_{\pi_n}}(x_{\pi_1}, x_{\pi_2}, \dots, x_{\pi_n})$$

□ 相容性：若 $m < n$ ，则分布函数 $F_{t_1, t_2, \dots, t_m}(x_1, x_2, \dots, x_m)$ 可由 $F_{t_1, t_2, \dots, t_n}(x_1, x_2, \dots, x_n)$ 按下面的方式唯一决定。

$$F_{t_1, t_2, \dots, t_m}(x_1, x_2, \dots, x_m) = \lim_{\substack{x_{m+1} \rightarrow \infty \\ \vdots \\ x_n \rightarrow \infty}} F_{t_1, t_2, \dots, t_n}(x_1, x_2, \dots, x_n)$$

④ **定义 6.8.** 随机过程 X 最常见的数字特征包括：均值函数、方差函数和协方差函数，它们都是关于时间的函数。

□ 均值函数 $E(X(t))$ 和方差函数 $V(X(t))$ ，分别简记作 $E_X(t)$ 和 $V_X(t)$ ，有时也记作 $\mu_X(t)$ 和 $\sigma_X^2(t)$ ，简记作 $\mu(t)$ 和 $\sigma^2(t)$ 。

□ 协方差函数 (covariance function) $Cov_X(t, s) = Cov[X(t), X(s)] = E[(X(t) - E_X(t))(X(s) - E_X(s))]$ ，其中 $t, s \in T$ 。协方差函数有时也记作 $\gamma_X(t, s)$ ，简记作 $Cov(t, s)$ 或 $\gamma(t, s)$ 。

□ 相关函数 (correlation function) 定义为

$$\rho_X(t_1, t_2) = \frac{\gamma_X(t_1, t_2)}{\sigma_X(t_1)\sigma_X(t_2)}$$

例 6.6. 假设一个加油站一天有 N 辆车来加油，其中 N 是随机变量，所加的油量 X_1, X_2, \dots, X_N 不妨设为独立同分布，其概率母函数为 $G_X(s)$ 。令 $Y_j = X_1 + \dots + X_j$ ，

其中 $j = 1, \dots, N$, 则

$$\begin{aligned}
G_{Y_N}(s) &= \sum_{k=0}^{\infty} P(Y_N = k)s^k \\
&= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} P(Y_N = k|N = j)P(N = j)s^k \\
&= \sum_{j=0}^{\infty} P(N = j) \sum_{k=0}^{\infty} P(Y_j = k)s^k, \text{ 根据性质 4.2 我们有} \\
&= \sum_{j=0}^{\infty} P(N = j)G_X^j(s) \\
&= G_N(G_X(s))
\end{aligned}$$

进而利用性质 4.2, 不难得到

$$\begin{aligned}
E(Y_N) &= G'_N(G_X(s))G'_X(s)|_{s=1} \\
&= \mu_N \mu_X \\
V(Y_N) &= G''_N(G_X(s))[G'_X(s)]^2 + G'_N(G_X(s))G''_X(s)|_{s=1} + E(Y_N) - [E(Y_N)]^2 \\
&= \sigma_N^2 \mu_X^2 + \sigma_X^2 \mu_X
\end{aligned}$$

高斯过程的协方差函数有很好的性质, 譬如, 用于产生再生核 Hilbert 空间 (reproducing kernel Hilbert space, RKHS), 进而有 Karhunen-Loeve 展开。由于篇幅的限制, 这些漂亮的结果在本书中不能一一展现, 感兴趣的读者可参考 [7]。这里, 我们只是不加证明地介绍下面的定理。

定理 6.2. 高斯过程 $X(t)$ 由均值函数和协方差函数唯一决定。即, 给定任意连续函数 $A(t)$ 和任意正定的连续函数 $B(t, s)$, 总存在高斯过程 $X(t)$ 使其均值函数和协方差函数恰为 $A(t)$ 和 $B(t, s)$ 。

定义 6.9. 随机过程 $X(t)$ 如果满足下面两个条件, 则被称为弱平稳的 (weakly stationary), 或简称平稳的。平稳过程具有良好的性质, 是被重点研究过的。

- 期望函数 $\mu_X(t)$ 与 t 无关, 是个常数。
- 对每个整数 h , 自协方差函数 $\gamma_X(t + h, t)$ 都与 t 都无关, 记作 $\gamma_X(h)$ 。特别地, 方差函数 $\sigma_X^2(t)$ 是与 t 无关的常数 $\gamma_X(0)$ 。

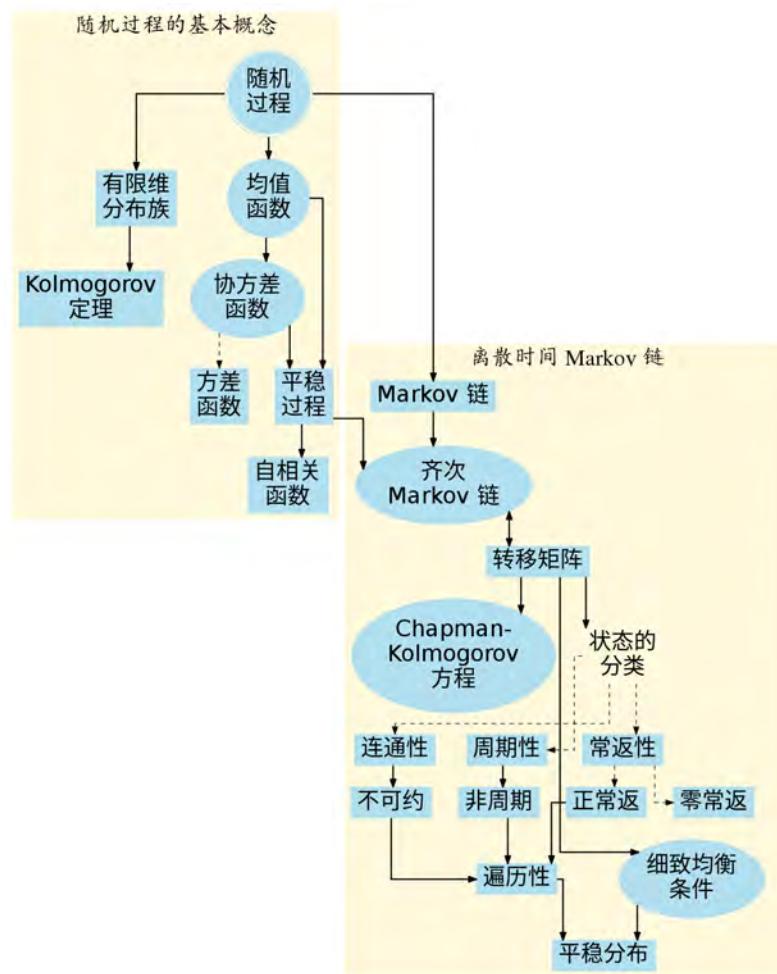
如果随机过程 $X(t)$ 在任何时间段 $[t_1, t_2]$ 上的联合分布满足平移不变性, 即在时间段 $[t_1 + \Delta t, t_2 + \Delta t]$ 也具有相同的联合分布, 则称该随机过程是强平稳的 (strongly stationary)。显然, 强平稳过程一定是弱平稳的。

性质 6.1. 平稳随机过程 X 的自相关函数 (autocorrelation function, ACF) 为

$$\rho_X(h) = \rho(X(t+h), X(t)) = \frac{\gamma_X(h)}{\gamma_X(0)}$$

1931 年, Kolmogorov 发表论文《概率论的解析方法》, 首次将微分方程等分析方法用于 Markov 过程的研究。1934 年, Khinchin 的论文《平稳过程的相关理论》开始了平稳过程的研究。1951 年, 日本数学家伊藤清 (Kiyoshi Itô, 1915-2008) 建立了随机微分方程理论。1953 年, 美国数学家 Joseph Leo Doob (1910-2004) 出版名著《随机过程论》[36], 系统论述了随机过程的基本理论。另外, Doob 在概率位势理论也颇有建树。对一些特殊随机过程, 如 Markov 过程、独立增量过程、平稳过程、鞅、点过程、分支过程等, 已进行过系统的研究。近年来, 人们关注多指标随机过程、流形上的随机过程与随机微分方程、无穷质点 Markov 过程、关于半鞅的随机微分方程等话题。

第六章的主要内容及其关系



6.1 离散时间 Markov 链

考虑离散时间过程 $\{X_n : n = 1, 2, \dots\}$, 其状态空间 \mathbb{S} 是有限的或可数的, 不妨设为 $\{1, 2, \dots\}$ 。表达式 “ $X_n = k$ ” 意味着过程在时刻 n 处于状态 $k \in \mathbb{S}$ 。1907 年, Markov 首次考虑了下面一类常见的离散时间过程。

定义 6.10 (Markov 链). 具有至多可数状态空间的离散时间过程 $\{X_n : n = 1, 2, \dots\}$ 如果满足下面的条件, 则称之为离散时间 Markov 链 (discrete-time Markov chain, DTMC), 简称 Markov 链: 对任意非负整数 n , 对任意状态 $j, i, i_{n-1}, \dots, i_1 \in \mathbb{S}$ 皆有

$$P(X_{n+1} = j | X_n = i) = P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1) \quad (6.1)$$

条件 (6.1) 称为 Markov 性 (Markov property), 它表示在已知当前的状态 X_n 的条件下, 将来的状态 X_{n+1} 与过去的状态 X_{n-1}, \dots, X_1 条件独立。

离散时间 Markov 过程具有广泛的实用价值, 譬如, 用于隐 Markov 模型和 Markov 链 Monte Carlo (MCMC) 方法。

定义 6.11 (齐性 Markov 链). 若对于任意的 n 皆有相同的 $P(X_{n+1} = j | X_n = i) = p_{ij}$, 则称 Markov 链 $\{X_n : n = 1, 2, \dots\}$ 为齐性的 (homogeneous) 或时齐的 (time-homogeneous), 它意味着从状态 i 转移到状态 j 的概率是固定的某个值, 与所在的时刻无关。

显然, $\sum_{j=1}^{\infty} p_{ij} = 1$, 即从状态 i 转移到下个状态的概率为 1。所有的转移概率可整理成如下的转移矩阵, 每行之和都为 1。

$$\begin{matrix} & 1 & 2 & \cdots & j & \cdots \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ i \\ \vdots \end{matrix} & \left(\begin{matrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{matrix} \right) & = (p_{ij}) = P \end{matrix}$$

在下面的内容中, 如果没有特殊的说明, 约定所涉及的离散时间 Markov 链都是指齐性的。Markov 链有个形象的比喻: 一只无记忆的青蛙在标号为 $1, 2, \dots, j, \dots$ 的荷叶间跳跃, 标号就是状态。青蛙从当前状态 i 跳到下一个状态 j 与它过去走过的路径无关。

令 $T_1^{(j)}$ 是从状态 j 出发第一次返回该状态的步数, 令 $T_2^{(j)}$ 是第一次返回状态 j 后第二次返回该状态的步数, ……。显然,

性质 6.2. 随机变量 $T_1^{(j)}, T_2^{(j)}, \dots$ 是独立同分布的, 其分布可仿照下述结果算得。

$$\begin{aligned} P(T_1^{(j)} = 2) &= \sum_{i \neq j} p_{ij} p_{ji} \\ P(T_1^{(j)} = 3) &= \sum_{i,k \neq j} p_{kj} p_{ik} p_{ji} \end{aligned}$$

例 6.7. n 重 Bernoulli 试验就是一个离散时间的 Markov 过程, 只有两个状态 $\mathbb{S} = \{1, 2\}$, 状态转移矩阵是 $P = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$ 。

例 6.8. 设一个微观粒子有 $N+1$ 个状态 $0, 1, \dots, N$, 粒子随机游动的状态转移矩阵 $P = (p_{ij})$ 定义为: $p_{00} = p_{NN} = 1$, 即粒子一旦转入状态 0 或 N , 下一步转移将以概率 1 返回, 称这样的状态为吸收壁*。对于任意的 $1 \leq i \leq N-1$,

$$p_{ij} = \begin{cases} p & \text{当 } j = i+1 \\ 1-p & \text{当 } j = i-1 \\ 0 & \text{其他} \end{cases}$$

该 Markov 链的状态转移矩阵 P 如下, 其中 $q = 1 - p, 0 < p < 1$ 。

$$P = \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots & N-3 & N-2 & N-1 & N \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & q & 0 & p & 0 & \cdots & 0 & 0 & 0 & 0 \\ 2 & 0 & q & 0 & p & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ N-1 & 0 & 0 & 0 & 0 & \cdots & 0 & q & 0 & p \\ N & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{pmatrix}_{(N+1) \times (N+1)}$$

有时我们用下面的状态转移图来表示 Markov 链 (的转移矩阵), 其中节点表示状态, 转移概率标记在有向弧上, 满足“从每个节点转移出去的概率之和等于 1”。

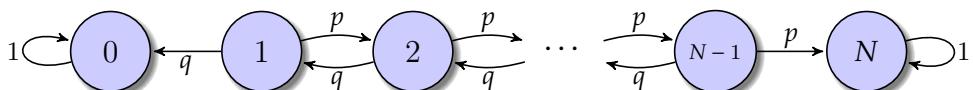


图 6.7: 例 6.8 的 Markov 链的状态转移图, 其中 $q = 1 - p, 0 < p < 1$ 。

*公子王孙逐后尘, 绿珠垂泪滴罗巾。侯门一入深似海, 从此萧郎是路人。

崔郊《赠婢》

练习 6.1. 孔乙己和阿 Q 赌博，赢者得到一元，输者失去一元。孔乙己赢的概率是 p ，拥有赌资 a 元；阿 Q 赢的概率是 $q = 1 - p$ ，拥有赌资 b 元。他们二人谁也不服谁，一直赌下去直到某人的赌资为零。显然，孔乙己和阿 Q 的赌资状态都是 $0, 1, \dots, N = a + b$ ，请读者写出孔乙己（和阿 Q）赌资的状态转移矩阵并画出状态转移图。

例 6.9. 奥地利生物学家 Gregor Johann Mendel (1822-1884) 通过豌豆实验发现了生物遗传的三个基本规律——基因的分离定律、自由组合定律和连锁交换定律。

Mendel 揭示了生物的性状都是由遗传基因控制，控制性状的基因总是成对的，显性和隐性基因分别控制显性和隐性性状（分别用大写和小写字母表示）。当体细胞中同时含有显性与隐性基因时，体细胞只表现显性基因所表达的性状。

考虑基因类型 AA, Aa, aa ，对于有 N 个个体的一代，总共有 $2N$ 个基因，其中显性基因 A 的个数有 $2N+1$ 个可能的状态： $0, 1, \dots, 2N$ 。显然，状态 $0, 2N$ 是吸收壁。假设子代按照 Bernoulli 试验的方式从父代继承基因，则从状态 i 到状态 j 的转移概率为



$$p_{ij} = C_{2N}^j \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}, \text{ 其中 } i, j = 0, 1, \dots, 2N$$

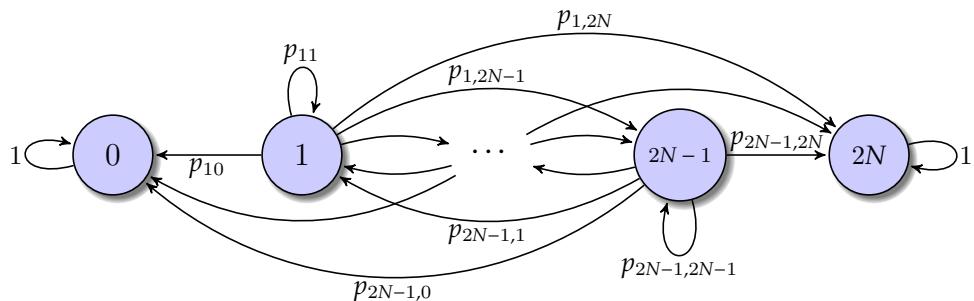


图 6.8: 例 6.9 的 Markov 链的状态转移图。

定义 6.12. 已知某 Markov 链的状态转移矩阵 $P = (p_{ij})$ 。令 $p_{ij}^{(n)}$ 表示从状态 i 经过 n 步转移到状态 j 的概率，称矩阵 $P_n = (p_{ij}^{(n)})$ 为 n -步转移矩阵 (n -step transition matrix)。我们常把 $\lim_{n \rightarrow \infty} P_n$ 简记作 P_∞ 或 P^∞ ，其 (i, j) 元素 $\lim_{n \rightarrow \infty} p_{ij}^{(n)}$ 简记作 $p_{ij}^{(\infty)}$ 。

例 6.10. 例 6.8 中, 2-步转移矩阵 P_2 为

$$P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q & pq & 0 & p^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & pq & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}_{(N+1) \times (N+1)}$$

练习 6.2. 计算例 6.8 中 Markov 链的 3-步转移矩阵。

定理 6.3 (Chapman-Kolmogorov 方程). 给定 Markov 链的状态转移矩阵 $P = (p_{ij})$, 则对于任意的 $1 \leq m < n$, 皆有

$$p_{ij}^{(n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n-m)}, \text{ 或等价地 } P_n = P_m P_{n-m} \quad (6.2)$$

证明. 从状态 i 经过 n 步转移到状态 j , 等价于先经过 m 步转移到到达某一状态 $k = 1, 2, \dots$, 然后再从状态 k 经过 $n - m$ 步转移到状态 j 。由 Markov 性, 我们有

$$\begin{aligned} p_{ij}^{(n)} &= P(X_n = j | X_0 = i) \\ &= \sum_{k \in S} P(X_n = j, X_m = k | X_0 = i) \\ &= \sum_{k \in S} P(X_m = k | X_0 = i) P(X_n = j | X_m = k, X_0 = i) \\ &= \sum_{k \in S} P(X_m = k | X_0 = i) P(X_n = j | X_m = k) \\ &= \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n-m)} \end{aligned}$$

□

式 (6.2) 称为 Chapman-Kolmogorov 方程。该方程于 1928 年由英国数学家和地球物理学家 Sydney Chapman (1888-1970, 照片见右) 首次发现, 1931 年由 Kolmogorov 重新发现并深入研究。对于连续时间离散状态的 Markov 过程, 也有类似式 (6.2) 的 Chapman-Kolmogorov 方程, 见 (6.17)。



练习 6.3. 试证明 n -步转移矩阵 $P_n = P^n$, 即方阵 P 的 n 次幂*。

本节内容

第一小节是 Markov 链中状态的分类, 我们着重考虑连通性、周期性和常返性这几个等价关系。第二小节是关于 Markov 链的遍历性和平稳分布。

*矩阵幂的算法具体见 [164] 的第八章第三节《Cayley-Hamilton 定理及其应用》。

关键知识

(1) (齐性) Markov 链; (2) 状态转移矩阵; (3) n -步转移矩阵和 Chapman-Kolmogorov 方程; (4) 不可约 Markov 链; (5) 常返的状态, 包括正常返和零常返; (6) 遍历性与平稳分布; (7) 细致均衡条件。

6.1.1 状态的分类

对于一个 Markov 链，有的状态一旦进入就再也逃不出来，如例 6.8 中的吸收壁；有的一旦离开后还可以无穷次地“故地重游”，……。本小节着重考虑在不同标准之下 Markov 链中状态的分类。

定义 6.13. 状态 i 经过若干步可转移到状态 j ，即意味着存在 $m \geq 0$ 使得 $p_{ij}^{(m)} > 0$ ，记作 $i \rightarrow j$ ，否则记作 $i \not\rightarrow j$ 。其中 $p_{ij}^{(0)}$ 定义为

$$p_{ij}^{(0)} = \begin{cases} 1 & \text{当 } i = j \\ 0 & \text{当 } i \neq j \end{cases}$$

“往而不来，非礼也”。若 $i \rightarrow j$ 且 $j \rightarrow i$ ，即从状态 i 或 j 出发可相互到达对方，则称它们是连通的或互达的，记作 $i \leftrightarrow j$ 。

练习 6.4. 请读者验证“连通的”是一个等价关系，即满足 (i) 自反性 $i \leftrightarrow i$; (ii) 对称性 $i \leftrightarrow j \Rightarrow j \leftrightarrow i$; (iii) 传递性 $i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k$ 。

定义 6.14. 等价关系 “ \leftrightarrow ” 决定了 Markov 链的状态空间 \mathbb{S} 上的一个分类。如果 $\forall i, j \in \mathbb{S}, i \leftrightarrow j$ ，则该 Markov 链称为不可约的 (irreducible)。它意味着，按照连通性，这些状态都在同一个类，即从任何状态出发，都有可能到达指定的状态。显然，Markov 链不可约当且仅当它的状态转移图是一个有向连通图。

例 6.11. 例 6.8 中，对任意的状态 $1 \leq i \leq N - 1$ 皆有 $i \rightarrow 0$ ，但是 $0 \not\rightarrow i$ 。另外， $0 \leftrightarrow N$ 。因此，该 Markov 链不是不可约的。按照连通性，该 Markov 链的状态有三个等价类： $\{0\}$ ， $\{1, 2, \dots, N - 1\}$ 和 $\{N\}$ 。

定义 6.15. 状态 i 的周期定义为满足下述条件的最大自然数 d_i ：

$$p_{ii}^{(n)} > 0 \text{ 仅当 } d_i | n, \text{ 即 } n \text{ 被 } d_i \text{ 整除}$$

即，从状态 i 出发，经过 d_i 的整倍的步数后，都有可能回到状态 i 。若步数不是 d_i 的整倍，则回到 i 的概率为零。不难看出，周期刻画了系统从状态 i 出发并重回该状态的概率规律。

□ 若对于任意的 $n \geq 1$ 皆有 $p_{ii}^{(n)} = 0$ ，则定义 $d_i = 0$ ，它意味着系统一旦离开状态 i 就不再回来（严格地说，回来的概率为零）。可谓，

昔人已乘黄鹤去，此地空余黄鹤楼。
黄鹤一去不复返，白云千载空悠悠。

崔浩《黄鹤楼》

□ 如果一个状态的周期为 1，则称它为非周期的 (aperiodic)。即，对于任意的 $n \geq 1$ 皆有 $p_{ii}^{(n)} > 0$ ，意味着系统离开状态 i 随时都有可能重新回来。

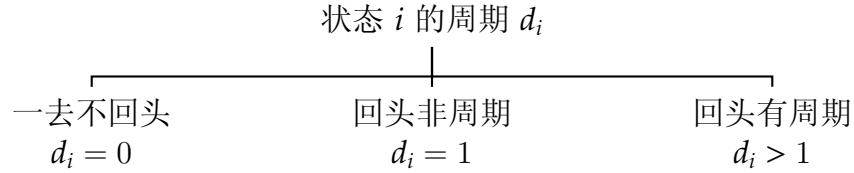


图 6.9: Markov 链的状态分类: 零周期 + 非周期 + 周期。

定义 6.16. 如果一个 Markov 链中所有满足 $p_{ii}^{(n)} > 0$ 的 n 没有大于 1 的公因子，则称该 Markov 链为非周期的。

例 6.12. 对于下面转移矩阵定义的非周期的 Markov 链，状态 1,2 都是非周期的，状态 3,4,5 的周期都是 2。该 Markov 链不是不可约的，因为状态 1,2 与状态 3,4,5 不连通。

$$\begin{matrix} & 1 & 2 & 3 & 4 & 5 \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \left(\begin{array}{ccccc} \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \end{array} \right) = P \end{matrix}$$

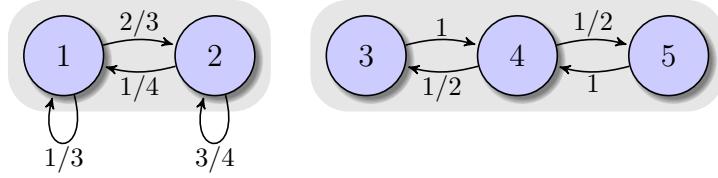
该 Markov 链的状态 1,2 是连通的，状态 3,4,5 是连通的。经过计算， ∞ -步转移矩阵 P_∞ 如下所示，状态转移图的拓扑结构发生了变化。

性质 6.3. 从状态 i 出发，经过 k 步首次到达状态 j 的概率记为 $f_{ij}^{(k)} = P(X_k = j, X_s \neq j, 1 \leq s \leq k-1 | X_0 = i)$ ，则经过 n 步从状态 i 达到状态 j 的概率为

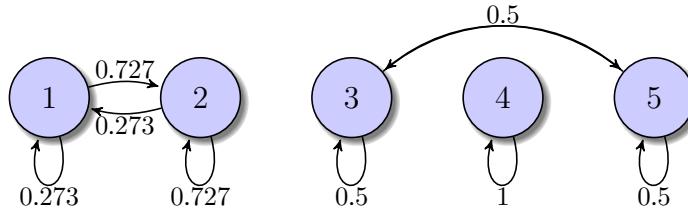
$$p_{ij}^{(n)} = \sum_{k=1}^n f_{ij}^{(k)} p_{jj}^{(n-k)}$$

并且，从状态 i 出发迟早到达状态 j 的概率为

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$



(a) 从状态 3 出发总要经过偶数步才能回到起点, 故其周期为 2。



(b) 所有状态都是非周期的。

图 6.10: (a) 例 6.12 中 P 的状态转移图。 (b) P_∞ 的状态转移图。

特别地, 从状态 i 出发迟早返回自身的概率为

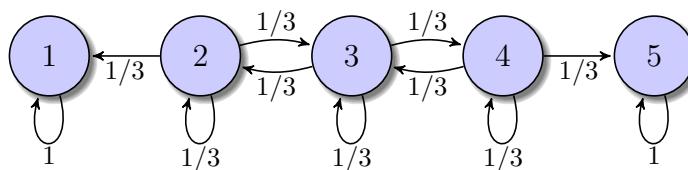
$$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)} \quad (6.3)$$

证明. 从状态 i 经过 n 步到达状态 j , 等价于经过 k 步首次达到状态 j 后, 再从状态 j 出发经过 $n-k$ 步回到状态 j , 其中 $k = 1, 2, \dots, n$ 。 \square

定义 6.17. 如果从状态 i 出发, 以概率 1 能够返回状态 i , 即式 (6.3) 满足 $f_{ii} = 1$, 则称状态 i 为常返的 (recurrent)。否则 (即 $f_{ii} < 1$), 则称状态 i 为非常返的或滑过的 (transient)。

练习 6.5. 请读者说明例 6.12 中所有状态都是常返的。

练习 6.6. 请根据下面的状态转移图写出转移矩阵, 并说明状态 2, 3, 4 是非常返的。



性质 6.4. Markov 链中某状态 i 是常返的当且仅当

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$$

证明. 由**定义 6.13** 的约定, $p_{ii}^{(0)} = 1$ 。 $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$ 当且仅当 $f_{ii} = 1$, 这是因为

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^n f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \sum_{n=k}^{\infty} p_{ii}^{(n-k)} = f_{ii} \sum_{n=0}^{\infty} p_{ii}^{(n)} = f_{ii} \left(1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}\right) \quad \square$$

性质 6.5. 定义随机变量 $I_{n,j} = \begin{cases} 1 & \text{当 } X_n = j \\ 0 & \text{当 } X_n \neq j \end{cases}$, 则随机变量 $\sum_{n=0}^{\infty} I_{n,j}$ 刻画了到达状态 j 的次数, 并且

$$E\left(\sum_{n=0}^{\infty} I_{n,j} \mid X_0 = j\right) = \sum_{n=0}^{\infty} E(I_{n,j} \mid X_0 = j) = \sum_{n=0}^{\infty} p_{jj}^{(n)}$$

推论 6.1. Markov 链中某状态 j 是非常返的当且仅当从 j 出发至多有限多次返回 j 。

 若状态 i 是常返的, 则从 i 出发以概率 1 回到 i , 然后再从 i 出发以概率 1 又回到 i , 如此下去就能够无穷次地回到状态 i , 这是“常返的”的直观含义。显然, 吸收壁一定是常返的。若状态 j 是非常返的, 则它只能有限次地被造访。

人生不相见，动如参与商。今夕复何夕，共此灯烛光。
少壮能几时，鬢发各已苍。訪旧半为鬼，惊呼热中肠。
焉知二十载，重上君子堂。昔別君未婚，儿女忽成行。
怡然敬父執，问我來何方。問答乃未已，儿女羅酒漿。
夜雨剪春韭，新炊間黃粱。主稱會面難，一舉累十觴。
十觴亦不醉，感子故意長。明日隔山岳，世事兩茫茫。

杜甫《贈卫八處士》

性质 6.6. 如果状态 j 是常返的并且 $i \leftrightarrow j$, 则状态 i 也是常返的。

证明. 不妨设 $p_{ij}^{(a)} > 0, p_{ji}^{(b)} > 0$, 显然有

$$p_{ii}^{(n+a+b)} \geq p_{ij}^{(a)} p_{jj}^{(n)} p_{ji}^{(b)}$$

如果状态 j 是常返的, 由**性质 6.4** 知 $\sum_{n=1}^{\infty} p_{jj}^{(n)} = \infty$, 进而有 $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$, 即状态 i 也是常返的。 \square

性质 6.7. 若状态 j 是非常返的, 则对任意状态 i 皆有 $p_{ij}^{(\infty)} = 0$, 即

$$\lim_{n \rightarrow \infty} p_{ij}^{(n)} = 0$$

证明. 类似于**性质 6.4** 的证明,

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^n f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=k}^{\infty} p_{jj}^{(n-k)} = f_{ij} \sum_{n=0}^{\infty} p_{jj}^{(n)} \leq \sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty \quad \square$$

例 6.13. [例 6.8](#) 中, 状态 $0, N$ 都是非周期的且是常返的, 而状态 $1, 2, \dots, N-1$ 都是周期的 (周期为 2) 且是非常返的。

※例 6.14 (赌徒输光问题). 接着第 395 页的练习 6.1, 问孔乙己的赌资达到 N 元或 0 元 (即赌博停止) 的概率?

解. 设孔乙己在 t 时刻的赌资为 X_t , 则过程 $\{X_t : t = 0, 1, 2, \dots\}$ 是 Markov 链, 转移概率 P 为**例 6.8** 所描述。

- 令孔乙己的赌资从 j 元出发, 到达 N 元之前到达 0 的概率是 f_j , 则 $f_0 = 1, f_N = 0$ 并且满足下面的递归关系。

$$f_j = pf_{j+1} + qf_{j-1}, \text{ 其中 } 1 < j < N-1$$

求解该递归式, 得到

$$f_j = \begin{cases} (r^j - r^N)/(1 - r^N) & \text{当 } p \neq 1/2, \text{ 其中 } r = q/p \\ 1 - j/N & \text{当 } p = 1/2 \end{cases}$$

- 令孔乙己的赌资从 j 元出发, 到达 0 元之前到达 N 的概率是 g_j , 则

$$g_j = \begin{cases} (1 - r^j)/(1 - r^N) & \text{当 } p \neq 1/2 \\ j/N & \text{当 } p = 1/2 \end{cases}$$

不难看出, $f_j + g_j = 1$, 其中 $j = 0, 1, \dots, N$ 。换句话说, 如果一直赌下去, 总有一人要彻底输光, 即赌博不可能永远进行下去。那种输赢交替拉锯式的赌局理论上存在, 但发生的概率为零。

另外, 当孔乙己与赌资无穷的阿 Q 赌博时, 则孔乙己彻底输光的概率是

$$\lim_{N \rightarrow \infty} f_j = \begin{cases} r^j & \text{当 } r < 1 \\ 1 & \text{当 } r \geq 1 \end{cases}$$

若孔乙己赌技不佳, 即 $p \leq 1/2$, 他将以概率 1 全部输光。若孔乙己赌技高超, 即 $p > 1/2$, 则面对不会破产的阿 Q, 即使只有一元赌资, 孔乙己依然有 $1 - r$ 的机会逃避彻底输光。

练习 6.7. 在练习 6.1 里, 若赌博是公平的, 即 $p = 1/2$, 求孔乙己彻底输光的概率?

答案: $b/(a+b)$, 即阿 Q 的赌资越多, 孔乙己输光的概率越大。

练习 6.8. 有限状态的不可约 Markov 链的所有状态都是常返的。提示: 由习题 6.5 知, 总存在某个状态是常返的, 再由性质 6.6 可证。

性质 6.8. 设 R 是由所有常返的状态构成的类, 则对于常返状态 $i \in R$ 和非常返状态 $j \notin R$, 有 $p_{ij} = 0$ 。

证明. 假设 $p_{ij} > 0$, 则 $i \rightarrow j$, 而 $j \not\rightarrow i$, 于是 $p_{ji}^{(n)} = 0$ 。从 i 出发至少有正概率 p_{ij} 使得过程回不到 i , 与 i 是常返的矛盾, 因此 $p_{ij} = 0$ 。 \square

例 6.15. 接着考虑第 34 页的例 1.26, 构造 Markov 链如下, 其中 $P(H) = p, q = 1 - p$, 状态空间为 $\mathbb{S} = \{0, 1, 2, \dots, t\}$, 表示当前连续正面的长度。我们约定: 一旦进入状态 t , 不管以后的抛掷情况如何, 状态都不再变。从状态 i 出发到达状态 t 的概率 f_{it} 简记作 f_i , 下面我们证明 $f_0 = 1$ 。

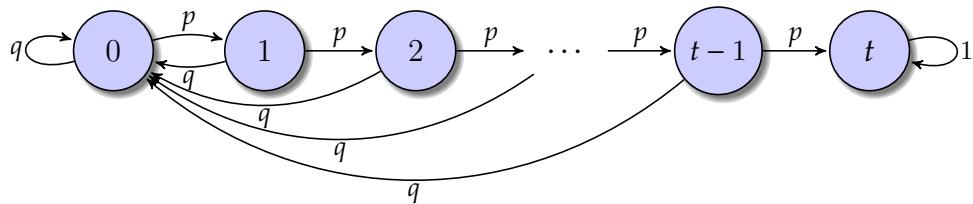


图 6.11: 该 Markov 链所有状态都是非周期的、常返的, 状态 t 是吸收壁。

显然, $f_t = 1$ 。不难得到递归关系如下,

$$f_k = pf_{k+1} + (1-p)f_0, \text{ 其中 } k = 0, 1, \dots, t-1$$

解之, 得 $f_0 = f_1 = \dots = f_t = 1$ 。牛刀小试, Markov 链是一个很实用的工具。读者可以对比第 58 页的例 1.46 的解法, 看哪个更容易理解。

对于非常返的状态, 根据性质 6.8, 长时间后系统造访它们的概率趋于零。再加上性质 6.8, 非常返的状态愈显得不重要, 因此我们只需把注意力放在常返的状态及其性质上。

定义 6.18. 已知某状态 j 是常返的, 如果从状态 j 出发返回自身的期望步数 $t_j = E(T_1^{(j)})$ 有限, 即 $t_j = \sum_{n=1}^{\infty} nf_{jj}^{(n)} < \infty$, 则称状态 j 为正常返的 (positive recurrent); 如果 $t_j = \sum_{n=1}^{\infty} nf_{jj}^{(n)} = \infty$, 则称状态 j 为零常返的 (null recurrent)。

例 6.16. 在图 6.11 中, 状态 $0, 1, 2, \dots, t-1$ 是正常返的, 状态 t 是零常返的。

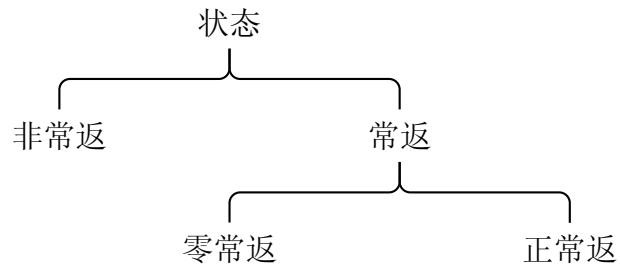


图 6.12: Markov 链的状态分类: 非常返 + 零常返 + 正常返。

性质 6.9. 作为**性质 6.6** 的扩充: 若 Markov 链中状态 $i \leftrightarrow j$, 则它们或都是非常返的, 或都是零常返的, 或都是正常返的。

6.1.2 Markov 链的遍历性与平稳分布

对于 Markov 链 $P = (p_{ij})$, 人们关注的是 $p_{ij}^{(n)}$ 的极限情况, 即经过足够多步从状态 i 转移到状态 j 的概率。其中一种极端的情形是当 $n \rightarrow \infty$, $p_{ij}^{(n)}$ 趋向一个和初始状态无关的概率 p_j , 这就是“遍历性”的由来。

定义 6.19 (遍历性). 称 Markov 链 $P = (p_{ij})$ 是遍历的 (ergodic), 如果对于任意的 $i, j \in \mathbb{S}$, 极限 $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = p_j$ 皆存在且不依赖于 i 。

换句话说, 从一个遍历的 Markov 链的任何状态出发, 经过足够多的 n 步之后到达状态 j 的概率都是一样的。简而言之, 矩阵 P_∞ 每行都一样, 即

$$\lim_{n \rightarrow \infty} P^n = \lim_{n \rightarrow \infty} (p_{ij}^{(n)}) = \begin{pmatrix} p_1 & p_2 & \cdots & p_j & \cdots \\ p_1 & p_2 & \cdots & p_j & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

例 6.17. 已知 Markov 链 $P = \begin{pmatrix} 1/3 & 2/3 \\ 1/4 & 3/4 \end{pmatrix}$, 考察 n 很大时的 P^n , 猜测极限 $\lim_{n \rightarrow \infty} p_{ij}^{(n)}$ 不依赖于 i , 即该 Markov 链是遍历的。事实上,

$$\lim_{n \rightarrow \infty} P^n = \begin{pmatrix} 3/11 & 8/11 \\ 3/11 & 8/11 \end{pmatrix}$$

练习 6.9. 请读者验证**例 6.12** 中的 Markov 链不是遍历的。提示: $p_{22}^{(\infty)} \neq p_{32}^{(\infty)}$ 。

练习 6.10. 图 6.11 所示的 Markov 链是遍历的。

例 6.18. 已知 Markov 链的状态转移图如下, 状态空间 $\mathbb{S} = \{0, 1, 2, \dots, t, \dots\}$ 。不难看出, 它是不可约的、非周期的

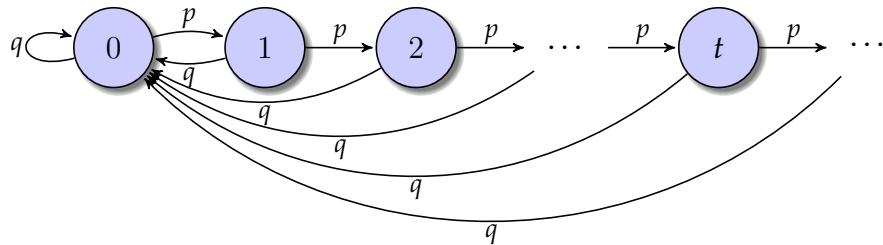


图 6.13: 该 Markov 链是不可约的, 所有状态都是非周期的、零正常返的。

性质 6.10. 遍历的 Markov 链当且仅当它是非周期的、不可约的和正常返的。即

遍历的 = 非周期的 + 不可约的 + 正常返的

定义 6.20 (平稳分布). 对于 Markov 链 $P = (p_{ij})$, 如果存在概率分布 $\pi_1\langle 1 \rangle + \pi_2\langle 2 \rangle + \cdots + \pi_j\langle j \rangle + \cdots$ 满足关系

$$\pi_j = \sum_{i=1}^{\infty} \pi_i p_{ij} \quad (6.4)$$

则称此分布为该 Markov 链的一个平稳分布, 即系统经过一定时间的演化后进入平衡态——系统的宏观状态不再随时间而变化。

性质 6.11. 若 Markov 链所有状态都是正常返的, 则平稳分布 $\pi_1\langle 1 \rangle + \pi_2\langle 2 \rangle + \cdots + \pi_j\langle j \rangle + \cdots$ 存在且唯一, 其中

$$\pi_j = \frac{1}{t_j}, \text{ 此处 } t_j = E(T_1^{(j)})$$

定理 6.4 (遍历性). 对于不可约、非周期的 Markov 链, 极限 $\lim_{n \rightarrow \infty} P_{ij}^{(n)} = p_j$ 存在, 并且只有以下两种情况:

- 所有 p_j 都为零, 此时不存在平稳分布;
- 所有 p_j 大于零, 此时 $p_1\langle 1 \rangle + p_2\langle 2 \rangle + \cdots + p_j\langle j \rangle + \cdots$ 是该 Markov 链唯一的平稳分布。

练习 6.11. 设 Markov 链有 s 个状态, 如果存在 $n \in \mathbb{N}$ 使得 $\forall i, j = 1, 2, \dots, s$ 皆有 $p_{ij}^{(n)} > 0$, 则该 Markov 链是遍历的。

性质 6.12. 如果有限状态的 Markov 链 $P_{s \times s} = (p_{ij})$ 有平稳分布 $x_1\langle 1 \rangle + \cdots + x_s\langle s \rangle$, 则 $\mathbf{x} = (x_1, x_2, \dots, x_s)^\top$ 一定 是以下方程的非零解。

$$P^\top \mathbf{x} = \mathbf{x} \quad (6.5)$$

即, \mathbf{x} 是 P^\top 的非零不动点, 或 $P^\top - I_s$ 的非平凡零点。

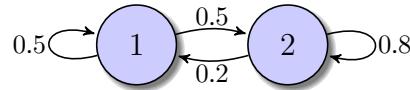
性质 6.13. 若 Markov 链的所有状态都是非常返的或零常返的, 则不存在平稳分布。

例 6.19. 已知 Markov 链的转移矩阵 P 为

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}, \text{ 借助 Maxima 猜测 } \lim_{n \rightarrow \infty} P^n = \begin{pmatrix} \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \end{pmatrix}$$

不难验证, $x_1 = \frac{2}{5}, x_2 = \frac{1}{5}, x_3 = \frac{2}{5}$ 恰是方程 $P^\top \mathbf{x} = \mathbf{x}$ 的非零解。由定理 6.4 和性质 6.12 知, 该 Markov 链有平稳分布 $\frac{2}{5}\langle 1 \rangle + \frac{1}{5}\langle 2 \rangle + \frac{2}{5}\langle 3 \rangle$ 。

练习 6.12. 求下面 Markov 链的平稳分布，并计算概率： $P\{X_1, X_{100} = 1 | X_0 = 1\}$, $P\{X_{100} = 1, X_{101} = 2\}$, $P\{X_{100} = 1, X_{200} = 1\}$ 。



答案： $\frac{2}{7}\langle 1 \rangle + \frac{5}{7}\langle 2 \rangle$, $P\{X_1, X_{100} = 1 | X_0 = 1\} = \pi_1 \cdot p_{11} = \frac{1}{7}$, $P\{X_{100} = 1, X_{101} = 2\} = p_{12} \cdot \pi_1 = \frac{1}{7}$, $P\{X_{100} = 1, X_{200} = 1\} = \pi_1 \cdot \pi_1 = \frac{4}{49}$ 。

练习 6.13. 求转移矩阵为 $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{pmatrix}$ 的 Markov 链的平稳分布。

答案： $0.2213\langle 1 \rangle + 0.4098\langle 2 \rangle + 0.3689\langle 3 \rangle$ 。

例 6.20 (PageRank 算法^{*}). 设有向图 $G = (V, E)$ 是由 n 个页面之间的链接关系定义的：顶点 $v_i \in V$ 表示第 i 个页面，顶点 v_i 和 v_j 之间有边 $(i, j) \in E$ 当且仅当页面 i 有链接指向页面 j 。

令 l_i 表示离开顶点 v_i 的边的个数。若 $l_i = 0$ ，则按照如下方法更新有向图 G ：增加从顶点 v_i 出发到各个顶点的边，总共 n 条。总而言之，保证每个顶点都边出来。

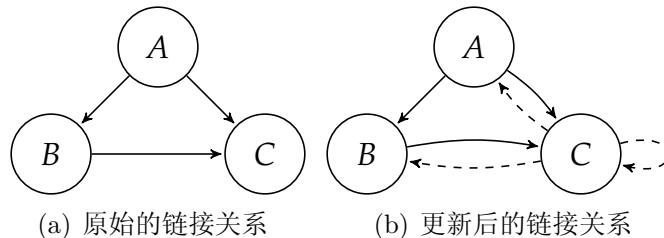


图 6.14: 有向图 G 经过更新后，所有顶点 v_i 的 l_i 都大于零。

对于更新后的有向图 G ，定义矩阵 $P_{n \times n} = (p_{ij})$ 如下，

$$p_{ij} = \begin{cases} 1/l_i & \text{如果 } (i, j) \in E \\ 0 & \text{如果 } (i, j) \notin E \end{cases}$$

显然， P 是一个状态转移矩阵。设顶点 v_i 的 PageRank 值是 $r_i \in [0, 1]$ （满足 $\sum_{i=1}^n r_i = 1$ ），可视为顶点 v_i 拥有的票数。于是，在顶点 v_i 每条出来的边上，都承载

^{*}1998 年，Stanford 大学博士研究生 Larry Page 和 Sergey Brin 发表 PageRank 算法，并由此创立了 Google 公司。简而言之，PageRank 算法就是如何由页面的链接关系构造一个 Markov 链，其平稳分布就是页面的 PageRank 值。

了 r_i/l_i 的票数投给他人。顶点 v_j 的 PageRank 值是

$$r_j = \sum_{(i,j) \in E} \frac{r_i}{l_i} \quad (6.6)$$

令 $\mathbf{r} = (r_1, \dots, r_i, \dots, r_n)^\top$ 表示 PageRank 向量, 不难验证式 (6.6) 就是 $P^\top \mathbf{r} = \mathbf{r}$ 。根据性质 6.12, PageRank 向量 \mathbf{r} 就是 Markov 链 P 的平稳分布。例如, 图 6.14 中更新后的链接关系所对应的状态转移矩阵是

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

进而算出页面 A, B, C 的 PageRank 向量是 $(0.1818, 0.2727, 0.5455)^\top$ 。实际应用中, 转移矩阵 P 有时经过平滑, 变为下面的转移矩阵 Q 。

$$Q = dP + \frac{1-d}{n}, \text{ 其中 } 0 < d < 1$$

一般地, $d = 0.85$ 。例如, 图 6.14 中的 PageRank 值大约是 $0.1976, 0.2815, 0.5209$ 。不论采用哪个转移矩阵, 页面重要性由高至低总是 C, B, A 。PageRank 算法的变种体现在如何定义转移矩阵上。

定义 6.21. 如果存在概率分布 $\pi_1\langle 1 \rangle + \pi_2\langle 2 \rangle + \dots + \pi_s\langle s \rangle$, 使得有限的 Markov 链 $P_{s \times s} = (p_{ij})$ 满足细致均衡条件 (detailed balance condition),

$$\pi_i p_{ij} = \pi_j p_{ji}, \text{ 其中 } i, j = 1, 2, \dots, s \quad (6.7)$$

则称 Markov 链 $P = (p_{ij})$ 是可逆的 (reversible)。

性质 6.14. 满足细致均衡条件 (6.7) 的可逆 Markov 链 $P_{s \times s} = (p_{ij})$ 具有平稳分布 $\pi_1\langle 1 \rangle + \pi_2\langle 2 \rangle + \dots + \pi_s\langle s \rangle$ 。

证明. 直接验证 $\pi_1, \pi_2, \dots, \pi_s$ 满足条件 (6.4), 事实上,

$$\sum_{i=1}^s \pi_i p_{ij} = \sum_{i=1}^s \pi_j p_{ji} = \pi_j \sum_{i=1}^s p_{ji} = \pi_j \quad \square$$



假设时间可以逆转, 利用 Bayes 公式计算条件概率 $P(X_n = i | X_{n+1} = j)$ 便产生了定义 6.21 所描述的“可逆性”。

$$P(X_n = i | X_{n+1} = j) = \frac{P(X_n = i)P(X_{n+1} = j | X_n = i)}{P(X_{n+1} = j)}$$

未完成。

6.1.3 分支过程

在父系社会里，家族姓氏由父亲传给儿子，如果某一代没有男性子嗣，这个姓氏就会消亡。1873-1874 年，英国统计学家、遗传学家 Francis Galton 和数学家 H. W. Watson (1827-1903) 发表论文《关于家族灭绝的可能性》，提出一个数学模型来研究姓氏消亡的问题，它更深的背景是遗传学中 Y 染色体的传播。

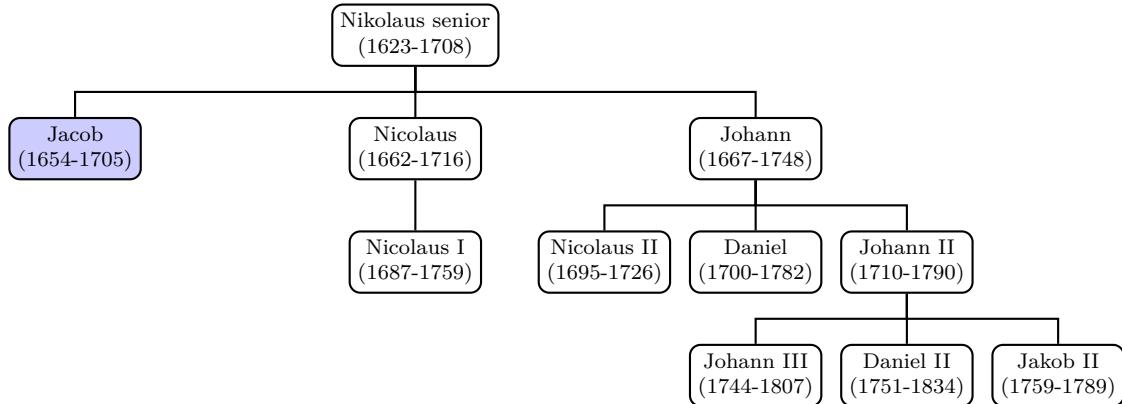


图 6.15: Bernoulli 家族中的数学家们，其中 Jacob Bernoulli (1654-1705) 是概率论的奠基者、Bernoulli 弱大数律的提出者、《猜度术》一书的作者。

每个男性在死之前留下若干儿子，数目为 $0, 1, 2, \dots$ ，一代一代下去，男性人口要么爆炸要么消亡。初始人数 $Y_0 = 1$ ，第一代 Y_1 个男性，则第二代男性人数 Y_2 为

$$Y_2 = X_1^{(1)} + X_2^{(1)} + \cdots + X_{Y_1}^{(1)}$$

其中 $X_1^{(1)}, X_2^{(1)}, \dots, X_{Y_1}^{(1)}$ 独立同分布，分别表示第一代每个男性所生儿子的个数，其概率母函数为 $G(s)$ 。例如，图 6.15 中，第一代 3 人，第二代 4 人，第三代 3 人。

练习 6.14. 仿照例 6.6 验证： Y_2 的概率母函数是 $G(G(s))$ ，简记作 $G_{(2)}(s)$ 。以此类推，第 n 代男性人数 Y_n 的概率母函数为

$$G_{(n)}(s) = G(G_{(n-1)})(s) = \underbrace{G(G(\cdots G(s) \cdots))}_{n \text{ 个 } G} \quad (6.8)$$

定义 6.22 (Galton-Watson 过程). 满足 $Y_0 = 1$ 及以下递归条件的随机过程 $\{Y_n\}$ 被称为 Galton-Watson 过程。

$$Y_{n+1} = \sum_{j=1}^{Y_n} X_j^{(n)} \quad (6.9)$$

其中, 对于所有 $n \in \mathbb{N}$, 取值自然数的随机变量 $X_1^{(n)}, X_2^{(n)}, \dots, X_j^{(n)}, \dots \stackrel{\text{iid}}{\sim} X$ 。

练习 6.15. 在**定义 6.22** 中, 若 $X \sim \text{Geom}(p)$, 求 Y_n 的概率母函数 $G_{(n)}(s)$ 。

$$G_{(n)}(s) = \frac{(p^2 - p)s + p}{(p - 1)s + p^2 - p + 1}$$

一般地, 很难得到 Y_n 的概率母函数的显示表达, 但我们可以利用**性质 4.2** 计算它的数字特征。

性质 6.15. 在**定义 6.22** 中, 若 $\mathbb{E}X = \mu, \mathbb{V}X = \sigma^2$, 则

$$\mathbb{E}(Y_n) = \mu^n \quad \mathbb{V}(Y_n) = \frac{\mu^{n-1}(1 - \mu^n)}{1 - \mu} \sigma^2$$

证明. 由**练习 6.14** 知,

$$G'_{(n)}(s) = G'(G_{(n-1)}(s))G'_{(n-1)}(s) \quad (6.10)$$

将 $s = 1$ 代入上式, 根据**性质 4.2** 和 $G_{(n-1)}(1) = 1$, 上式蕴含

$$\mathbb{E}(Y_n) = \mu \mathbb{E}(Y_{n-1})$$

这个结果也可以用下面的方法证得:

$$\begin{aligned} \mathbb{E}(Y_n) &= \mathbb{E}[\mathbb{E}(Y_n|Y_{n-1})] \\ &= \sum_{k=0}^{\infty} \mathbb{P}(Y_{n-1} = k) \mathbb{E}(Y_n|Y_{n-1} = k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(Y_{n-1} = k) \mathbb{E}\left(\sum_{j=1}^k X_j^{(n-1)}\right) \\ &= \mu \sum_{k=0}^{\infty} k \mathbb{P}(Y_{n-1} = k) \\ &= \mu \mathbb{E}(Y_{n-1}) \end{aligned}$$

由 $\mathbb{E}(Y_1) = \mu$ 知 $\mathbb{E}(Y_n) = \mu^n$ 。式 (6.10) 两边对 s 再求一阶导, 得到

$$\begin{aligned} G''_{(n)}(s) &= G''(G_{(n-1)}(s))[G'_{(n-1)}(s)]^2 + G'(G_{(n-1)}(s))G''_{(n-1)}(s) \\ G''_{(n)}(1) &= G''(1)(\mathbb{E}Y_{n-1})^2 + \mu G''_{(n-1)}(1) \\ &= (\sigma^2 - \mu + \mu^2)\mu^{2(n-1)} + \mu G''_{(n-1)}(1) \end{aligned}$$

求解上面的递归关系式，得到

$$G''_{(n)}(1) = \frac{\sigma^2 - \mu + \mu^2}{\mu(1-\mu)}(\mu^n - \mu^{2n})$$

利用性质 4.2 不难求得 $V(Y_n)$ ，请读者补全。 \square

一般地， $\lim_{n \rightarrow \infty} G_{(n)}(0)$ 刻画了 Galton-Watson 过程消亡的概率，这个概率是序列 $x_1 = G(0), x_2 = G(x_1), \dots, x_n = G(x_{n-1}), \dots$ 的极限，即方程 $G(x) = x$ 的解。显然 $x = 1$ 是一个解，除此之外， $[0, 1)$ 里还可能有一个解。

练习 6.16. 练习 6.15 中 $G(x) = x$ 有两个根 1 和 $p/(1-p)$ 。若 $p < 1/2$ ，则消亡的概率是 $p/(1-p)$ ，否则消亡是几乎必然的。

性质 6.16. 考虑第 0 代到第 n 代总体人数 Z_n ，它是以第 0 代为根的一棵的所有节点个数，也等于以第一代为根的 Y_1 棵子树的所有节点个数之和加 1，即

$$\begin{aligned} Z_n &= Y_0 + Y_1 + \dots + Y_n, \text{ 其中 } Y_0 = 1 \\ &= Z_{n-1}^{(1)} + Z_{n-1}^{(2)} + \dots + Z_{n-1}^{(Y_1)} + 1 \end{aligned}$$

Z_n 的期望和方差分别为

$$\begin{aligned} E(Z_n) &= \frac{1 - \mu^{n+1}}{1 - \mu} \\ V(Z_n) &= \begin{cases} \left[\frac{1 - \mu^{2n+1}}{(1 - \mu)^3} - \frac{(2n+1)\mu^n}{(1 - \mu)^2} \right] \sigma^2 & \text{当 } \mu \neq 1 \\ \frac{1}{3}n\left(n + \frac{1}{2}\right)(n+1)\sigma^2 & \text{当 } \mu = 1 \end{cases} \end{aligned}$$

证明. 令 $H_n(s)$ 为 Z_n 的概率母函数，由性质 4.2 和例 6.6 不难得到

$$\begin{aligned} H_0(s) &= s \\ H_n(s) &= sG(H_{n-1}(s)), \text{ 其中 } G \text{ 是 } Y_1 \text{ 的概率母函数} \end{aligned} \tag{6.11}$$

上式两边对 s 求一阶导，令 $s = 1$ 得到递归关系

$$\begin{aligned} E(Z_0) &= 1 \\ E(Z_n) &= 1 + \mu E(Z_{n-1}) \\ \text{于是, } E(Z_n) &= 1 + \mu + \mu^2 + \dots + \mu^n = \frac{1 - \mu^{n+1}}{1 - \mu} \end{aligned}$$

仿照性质 6.15 的证明,

$$\begin{aligned} H_n''(s) &= 2G'(H_{n-1}(s))H'_{n-1}(s) + G''(H_{n-1}(s))[H_{n-1}(s)]^2 + sG'(H_{n-1}(s))H''_{n-1}(s) \\ H_n''(1) &= 2\mu E(Z_{n-1}) + (\sigma^2 - \mu + \mu^2)[E(Z_{n-1})]^2 + \mu H''_{n-1}(1) \end{aligned}$$

利用 $H_n''(1) = V(Z_n) + [E(Z_n)]^2 - E(Z_n)$ 以及上式得到

$$\begin{aligned} V(Z_n) - \mu V(Z_{n-1}) &= \mu E(Z_{n-1}) + (\mu^2 + \sigma^2)[E(Z_{n-1})]^2 - [E(Z_n)]^2 + E(Z_n) \\ &= \left(\frac{1-\mu^n}{1-\mu}\sigma\right)^2 \\ V(Z_0) &= 0 \end{aligned}$$

求解上面的递归关系式, 即得到 $V(Z_n)$ 的表达式。 \square

性质 6.17. 序列 $H_n(s)$ 的极限, 记作 $H_\infty(s)$, 它是家族所有男性人数 Z_∞ 的概率母函数, 满足如下关系

$$\begin{aligned} H_\infty(s) &= sG(H_\infty(s)) \\ H_\infty(0) &= 0 \end{aligned}$$

证明. 对式 (6.11) 两边取极限 $n \rightarrow \infty$ 即得。 \square

例 6.21. 一般地, 很难得到 $H_\infty(s)$ 的显示表达, 但有些例子除外。接着练习 6.15, 利用性质 6.17, 我们有

$$H_\infty(s) = \frac{sp}{1 - (1-p)H_\infty(s)}$$

解上面的方程, 有两个解, 其中只有下面的满足 $H_\infty(0) = 0$ 。

$$H_\infty(s) = \frac{1 - \sqrt{1 - 4p(1-p)s}}{2(1-p)}$$

练习 6.17. 请读者验证

$$H_\infty(1) = \begin{cases} 1 & \text{如果 } p \geq \frac{1}{2} \\ \frac{p}{1-p} & \text{如果 } p < \frac{1}{2} \end{cases}$$

定义 6.23. 一个分支过程就是状态为 $0, 1, 2, \dots$ 的 Markov 过程 $X(t)$, 其转移概率

$p_{ij}(t) = \text{P}(X(t + t_0) = j | X(t_0) = i)$ 满足以下条件

$$p_{ij}(t) = \sum_{j_1+\dots+j_i=j} p_{ij_1}(t) \cdots p_{1j_i}(t) \quad (6.12)$$

分支过程 $X(t)$ 的概率母函数定义为

$$G_t(s) = \sum_{k=0}^{\infty} \text{P}(X(t) = k | X(0) = 1) s^k$$

性质 6.18. 分支过程 $X(t)$ 的概率母函数满足

$$\begin{aligned} G_0(s) &= s \\ G_{t+\tau}(s) &= G_t(G_\tau(s)) \end{aligned} \quad (6.13)$$

证明. 由结果 (6.13), □



特别地, 离散时间分支过程 $X(t)$ 当 t 取非负整数时, 式 (6.13) 就是式 (6.8), 其中 $G(s) = G_1(s)$ 。这样的分支过程就是 Galton-Watson 过程。

6.2 连续时间过程

连续时间过程 $\{X_t : t \geq 0\}$ 不是一系列随机试验的结果，而是时间的连续函数。我们可以将之离散化，但这样做一般是复杂的，也会丧失很多跟时间有关的属性，莫不如直接对连续时间建模。根据定理 6.1， $\{X_t : t \geq 0\}$ 由 X_{t_1}, \dots, X_{t_n} 的联合分布完全决定，其中 $t_1 < \dots < t_n$ 是任意有限个时间点。我们先介绍应用广泛的连续时间 Markov 过程（也称作 Markov 链）。

定义 6.24 (Markov 链). 对于任意非降的时间点的序列 $0 = t_0 < t_1 < \dots < t_{n-1} < s < t$ ，若取值为整数的连续时间过程 $X(t)$ 满足

$$\begin{aligned} & P\{X(t) = j | X(s) = i, X(t_{n-1}) = i_{n-1}, \dots, X(t_1) = i_1, X(t_0) = i_0\} \\ &= P\{X(t) = j | X(s) = i\} \end{aligned} \quad (6.14)$$

则称 $X(t)$ 是连续时间 Markov 链 (continuous-time Markov chain, CTMC)，其中整数值表示状态。式 (6.14) 被称为 Markov 性 (Markov property)，即当前状态只与所考虑的上一个时间点的状态有关，与之前的状态无关。

本书我们重点考虑一类特殊的 Markov 链——齐性 Markov 链，也称为时齐 Makrov 链 (time-homogeneous Markov chain)，即

$$P\{X(t) = j | X(s) = i\} = P\{X(t-s) = j | X(0) = i\} \quad (6.15)$$

换句话说，对于任意的 $t, \tau \geq 0$ 和状态 i, j ，无论进入状态 i 的时刻 τ 是何时，在时间 t 之后从状态 i 到状态 j 的转移概率总是不变的，称作平稳转移概率 (stationary transition probability)，记作 $p_t(j|i)$ 。我们有

$$p_t(j|i) = P(X_{t+\tau} = j | X_\tau = i), \forall t \geq 0$$

性质 6.19. 给定转移概率 $p_t(j|i) \geq 0, \forall t \geq 0$ 和初始分布 $\pi(j) = P(X_0 = j)$ ，则唯一确定一个齐性 Markov 链。并且，

$$\sum_j p_t(j|i) = 1 \quad (6.16)$$

$$p_{t+\tau}(j|i) = \sum_k p_t(j|k)p_\tau(k|i) \quad (6.17)$$

式 (6.16) 被称为归一性；式 (6.17) 被称为（连续时间）Chapman-Kolmogorov 方程，它与离散时间 Chapman-Kolmogorov 方程 (6.2) 是类似的。

证明. 对于任意时间点的序列 $0 = t_0 < t_1 < \dots < t_n$ ，随机变量 $X_{t_n}, X_{t_{n-1}}, \dots, X_0$ 的联

合分布为

$$\begin{aligned} \mathrm{P}(X_{t_n} = i_n, X_{t_{n-1}} = i_{n-1}, \dots, X_0 = i_0) &= \mathrm{P}(X_0 = i_0) \prod_{k=1}^n \mathrm{P}(X_{t_k} = i_k | X_{t_{k-1}} = i_{k-1}) \\ &= \pi(i_0) \prod_{k=1}^n p_{t_k-t_{k-1}}(i_k | i_{k-1}) \end{aligned}$$

由定理 6.1 知，此齐性 Markov 链唯一被确定。式 (6.17) 成立是因为

$$\begin{aligned} \mathrm{P}(X_{t+\tau} = j, X_\tau = k | X_0 = i) &= \mathrm{P}(X_{t+\tau} = j | X_\tau = k) \mathrm{P}(X_\tau = k | X_0 = i) \\ &= p_t(j|k) p_\tau(k|i) \end{aligned}$$

上式两边对 k 求和，则 τ 时刻是啥状态无所谓了，左边即是 $p_{t+\tau}(j|i)$ ，得证。□

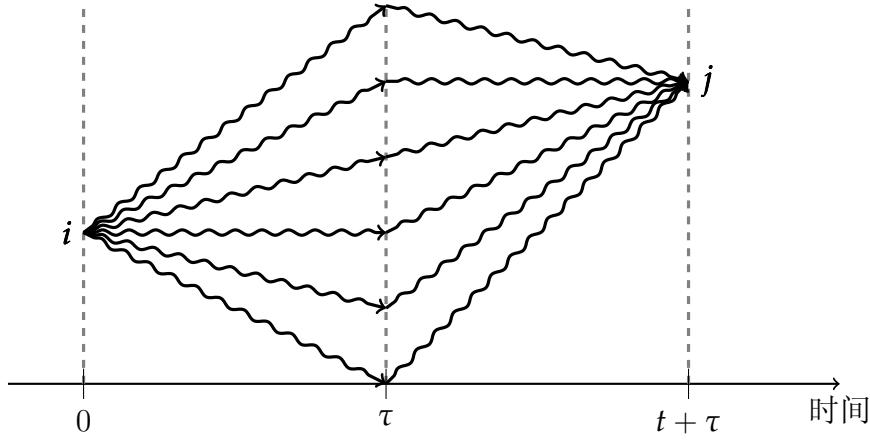


图 6.16: Chapman-Kolmogorov 方程的直观解释：在 $0, t + \tau$ 时刻分别处于状态 i, j 的所有可能的路径的概率，就是在中间某时刻 τ 遍历所有可能的状态而连接起来的两段路径之概率的总和。

由式 (6.16)，我们有

$$\begin{aligned} p_{\Delta t}(i|i) &= 1 - \sum_{j \neq i} p_{\Delta t}(j|i) \\ &= 1 - \sum_{j \neq i} q(j|i) \Delta t + o(\Delta t) \\ &= 1 - Q(i) \Delta t + o(\Delta t) \end{aligned}$$

$$\text{其中, } Q(i) = \sum_{j \neq i} q(j|i)$$

除了可以用平稳转移概率 $p_t(j|i)$ ，也可以用 $Q(i)$ 来刻画连续时间 Markov 过程，例如即将介绍的两类连续时间 Markov 过程：Poisson 过程、生灭过程。

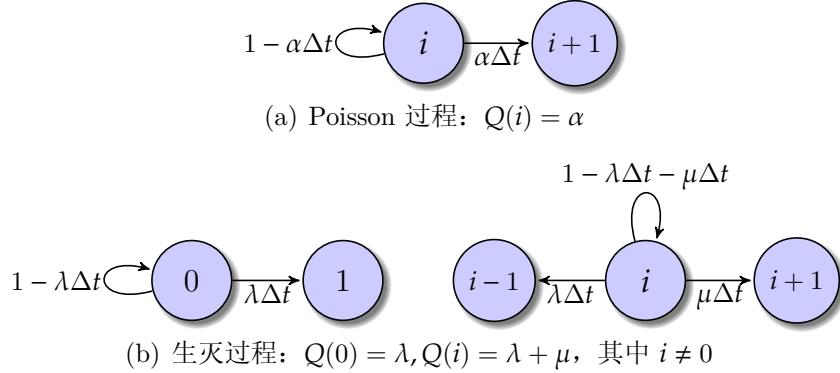


图 6.17: 两类特殊的连续时间 Markov 过程: Poisson 过程和生灭过程。

为了简化连续时间随机过程的数学模型, 本书我们特别关注以下两种类型的过程: 独立增量过程和平稳增量过程, 定义如下。

定义 6.25. 连续时间过程 $X = \{X(t) : t \in T\}$,

- 如果对任意 $t_1 < t_2 < \dots < t_{n-1} < t_n$ 皆有 $X(t_2) - X(t_1), \dots, X(t_n) - X(t_{n-1})$ 相互独立, 则称 X 有独立增量 (independent increment), 或称 X 是独立增量过程。
- 如果 $X(t) - X(s)$ 的分布只依赖于时间区间的长度 $t - s$, 其中 $t > s$, 则称 X 有平稳增量 (stationary increment), 或称 X 是平稳增量过程。

例如, 布朗运动满足 $X_t - X_s \sim N(0, t - s)$, 即将介绍的 Poisson 过程满足 $X_t - X_s \sim \text{Poisson}(\alpha(t - s))$, 其中 $\alpha > 0$ 是一个常数。

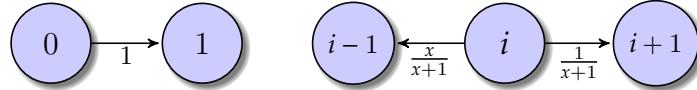
例 6.22. 考虑系统中某类部件的替换问题: 该部件一旦失效, 马上就会有一个新的将之替换。若部件个体的有效期和失效都独立于其他个体, 我们关心在时刻 t 该部件已被替换了多少次、当前部件使用了多久等问题。

不妨设部件的有效期服从分布 $X_1, X_2, X_3, \dots \stackrel{\text{iid}}{\sim} F(x)$, 定义 CTDV 过程 $N(t)$ 为 $X_1, X_1 + X_2, X_1 + X_2 + X_3, \dots$ 中不超过 t 的个数, 表示在时间段 $[0, t]$ 该部件的更新次数。该随机过程可用于解决排队论 (queueing theory)、库存论 (storage theory)、可靠性理论 (reliability theory) 等问题, 被称为更新过程 (renewal process)。更新过程既是独立增量过程, 也是平稳增量过程。

性质 6.20. (齐性) Markov 链与独立/平稳增量过程的关系是:

- 独立增量过程一定是 Markov 链。
- 具有独立且平稳增量的随机过程一定是齐性 Markov 链。反之不成立。

例 6.23. 齐性 Markov 链不一定有平稳增量。例如, 状态空间 $\{0, 1, 2, \dots\}$ 上的平稳转移概率



令初始分布 $\pi(0) = 1$, 则有

$$\begin{aligned} P(X_1 = 1) &= \sum_{i=0}^{\infty} P(X_1 = 1)P(X_0 = i) \\ &= P(X_1 = 1|0)\pi(0) \\ &= 1 \end{aligned}$$

于是, $P(X_1 - X_0 = 1) = 1$ 。但是,

$$\begin{aligned} P(X_2 - X_1 = 1) &= \sum_{i=0}^{\infty} P(X_2 - X_1 = 1)P(X_1 = i) \\ &= \sum_{i=0}^{\infty} P(X_2 = i+1)P(X_1 = i) \\ &= P(X_2 = 2|X_1 = 1)P(X_1 = 1) \\ &= \frac{1}{2} \end{aligned}$$

另外, 有平稳增量的 Markov 过程不一定是齐性的, 也不一定有独立增量。例如, 布朗桥 (稍后见例 6.31)。

6.2.1 Poisson 过程与更新过程

先让我们考虑一个特殊的 CTDV 过程：假设 $\{N(t)\}$ 具有独立增量和平稳增量。为直观起见，令 $N(t)$ 表示时间段 $[0, t]$ 内达到邮局（或银行、加油站、图书馆等公共场所）的人数，它是一个离散型随机变量，满足 $N(0) = 0$ 。

- 独立增量意味着在不相交时间段内到达邮局的人数是独立的。
- 平稳增量意味着在任意时间段内到达邮局人数的分布只与该时间段的长度有关，丝毫不依赖于历史情况。

根据这两个假设，我们有 $\forall t$,

$$P\{N(t + \Delta t) - N(t) = n\} = P\{N(\Delta t) = n\}, \text{ 其中 } n = 0, 1, 2, \dots$$

记 $P\{N(\Delta t) = n\}$ 为 $P_n(\Delta t)$ ，它表示在间隔时间 Δt 后，有 n 个人到达邮局的概率。在 $[0, t + \Delta t)$ 时间段有 n 人到达，等同于在 $[0, t)$ 时间段有 i 人到达，在 $[t, t + \Delta t)$ 有 $n - i$ 人到达。于是，

$$\begin{aligned} P_n(t + \Delta t) &= \sum_{i=0}^n P\{N(t) = i, N(t + \Delta t) - N(t) = n - i\} \\ &= \sum_{i=0}^n P\{N(t) = i\} P\{N(t + \Delta t) - N(t) = n - i\} \\ &= \sum_{i=0}^n P_i(t) P_{n-i}(\Delta t) \end{aligned} \tag{6.18}$$

显然， Δt 越小，概率 $P_1(\Delta t)$ 越小。不妨设在时间间隔 Δt 内有一人到达的概率是 $\alpha \Delta t$ ，其中 α 是一个常数，称为到达率。

$$P_1(\Delta t) = \alpha \Delta t + o(\Delta t), \text{ 并且 } \sum_{n=2}^{\infty} P_n(\Delta t) = o(\Delta t), \text{ 其中 } o(\Delta t) \text{ 是 } \Delta t \text{ 的高阶无穷小}$$

于是， $P_0(\Delta t) = 1 - \alpha \Delta t + o(\Delta t)$ 。令 $\Delta t \rightarrow 0$ ，由 (6.18) 可得

$$\begin{aligned} \frac{d}{dt} P_0(t) &= -\alpha P_0(t), \text{ 满足 } P_0(0) = 1 \\ \frac{d}{dt} P_n(t) &= -\alpha P_n(t) + \alpha P_{n-1}(t), \text{ 满足 } n \geq 1, P_n(0) = 0 \end{aligned}$$

解这些常微分方程，得到

$$P_0(t) = e^{-\alpha t}, \text{ 并且 } P_n(t) = \frac{(\alpha t)^n}{n!} e^{-\alpha t}$$

即，这个随机过程在任意两个时间点上的增量都是一个 Poisson 分布。我们把这个特殊的 CTDV 过程称为 Poisson 过程，正式定义如下。

定义 6.26 (Poisson 过程). 具有独立增量的连续时间过程 $\{N(t) : t \geq 0\}$ ，如果 $N(t_2) - N(t_1), t_2 > t_1$ 服从 Poisson 分布，则被称为 Poisson 过程 (Poisson process)。

定义 6.27 (齐性 Poisson 过程). Poisson 过程 $\{N(t)\}$ 若满足以下条件，则被称为齐性 Poisson 过程 (homogeneous Poisson process)。

$$P\{N(\tau + t) - N(\tau) = k\} = \frac{(\alpha t)^k}{k!} e^{-\alpha t} \quad (6.19)$$

其中，常数 $\alpha > 0$ 称为该 Poisson 过程的强度 (intensity) 或到达率 (arrival rate)，通常表示单位时间内事件的平均发生次数 (或平均到达人数)。齐性 Poisson 过程的实例见第 90 页的例 1.70。

显然，齐性 Poisson 过程是一个齐性 Markov 过程，其平稳转移概率是

$$p_t(j|i) = \begin{cases} 0 & \text{当 } j < i \\ \frac{(\alpha t)^{j-i}}{(j-i)!} e^{-\alpha t} & \text{当 } j \geq i \end{cases}$$

齐性 Poisson 过程的条件 (6.19) 要求概率 $P\{N(\tau + t) - N(\tau) = k\}$ 与 τ 无关，即时间段 $[\tau, \tau + t]$ 内到达邮局人数的概率分布不依赖于历史。说到这种对历史的无记忆性，第 270 页的定理 4.2 和第 298 页的定理 4.9 揭示，一个连续型 (或离散型) 随机变量满足无记忆性当且仅当它服从指数分布 (或几何分布)。Poisson 过程被视为连续版的 Bernoulli 过程。

性质 6.21. 对于 Poisson 过程 $N(t) \sim \text{Poisson}(\alpha t)$ ，令随机变量 T_n 表示 n 个人到达所需的时间，则 $T_n \sim \text{Erlang}(n, \alpha)$ ，其概率密度函数如下 (Erlang 分布的定义见第 296 页的定义 4.17)。

$$f_{T_n}(t) = \frac{\alpha^n t^{n-1}}{(n-1)!} e^{-\alpha t}, \text{ 其中 } t \geq 0, n = 1, 2, \dots$$

特别地，当 $n = 1$ 时， $T_1 \sim \text{Expon}(\alpha)$ 。即，相邻到达的时间间隔服从指数分布。

证明. 由第 297 页的练习 4.43 即得。下面提供另外一种证法：令 Δt 非常之小，第 n 个人在 $[t, t + \Delta t]$ 到达的概率是

$$\begin{aligned} f_{T_n}(t) \cdot \Delta t &= P(t \leq T_n \leq t + \Delta t) \\ &= P\{N(t) = n-1\} \cdot (\alpha \Delta t) \\ &= \frac{(\alpha t)^{n-1}}{(n-1)!} e^{-\alpha t} \cdot (\alpha \Delta t) \end{aligned}$$

上式两边约掉 Δt 即得证。 \square

定义 6.28 (更新过程). 已知独立同分布的非负随机变量 $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} F(x)$, 令 $T_0 = 0, T_n = X_1 + \dots + X_n$, 下面的 CTDV 过程 $N(t)$ 被称为更新过程 (见例 6.22)。

$$N(t) = \max\{n : T_n \leq t\} \quad (6.20)$$

定义 $U(t) = EN(t)$, 称之为更新函数 (renewal function)。更新理论很多结果都是关于更新函数的渐近性质的。

性质 6.22. 已知随机变量序列 $X_1, \dots, X_n, \dots \stackrel{\text{iid}}{\sim} \text{Expon}(\alpha)$, 令 $T_0 = 0, T_n = X_1 + \dots + X_n$, 定义 CTDV 过程 $N(t) = \max\{n : T_n \leq t\}$, 则 $N(t) \sim \text{Poisson}(\alpha t)$ 。

证明. 由第 297 页的练习 4.43, $T_n \sim \text{Gamma}(n, \alpha)$ 。另外, $N(t) = k$ 当且仅当 $T_k \leq t < T_{k+1}$ 。即, 若 $T_k = s \leq t$, 则 $X_{k+1} > t - s$ 。

$$\begin{aligned} P\{N(t) = k\} &= \int_0^t P(T_k = s, X_{k+1} > t - s) ds \\ &= \int_0^t g_{k,\alpha}(s) P(X_{k+1} > t - s) ds \\ &= \int_0^t \frac{\alpha^k}{(k-1)!} s^{k-1} e^{-\alpha s} \cdot e^{-\alpha(t-s)} ds \\ &= \frac{(\alpha t)^k}{k!} e^{-\alpha t} \end{aligned} \quad \square$$

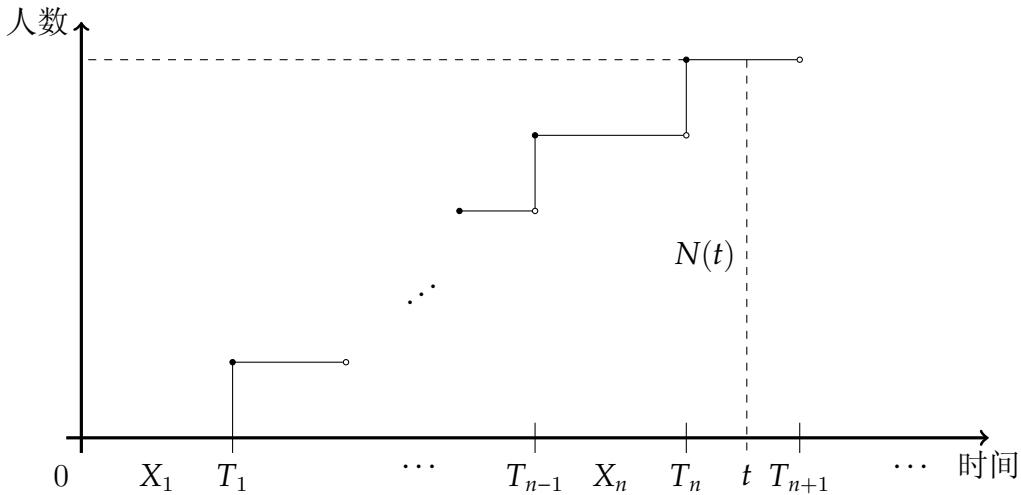


图 6.18: 性质 6.22 的直观理解: 两个相邻到达的间隔时间是独立的指数分布, $N(t) = \max\{n : X_1 + \dots + X_n \leq t\}$ 恰为时间段 $[0, t]$ 到达的人数。随机变量 $Z = T_{n+1} - t$ 具有分布函数 $F_Z(z) = \exp(-\alpha z)$ (见性质 4.27)。

由性质 6.21、性质 6.22 和式 (6.20) 易知: Poisson 过程是一类特殊的更新过程(部件的有效期服从指数分布)。

练习 6.18. 对于 Poisson 过程 $N_t \sim \text{Poisson}(\alpha t)$, $N_t - N_s$ 的分布与 N_{t-s} 的相同, 都是 $\text{Poisson}(\alpha(t-s))$ 。

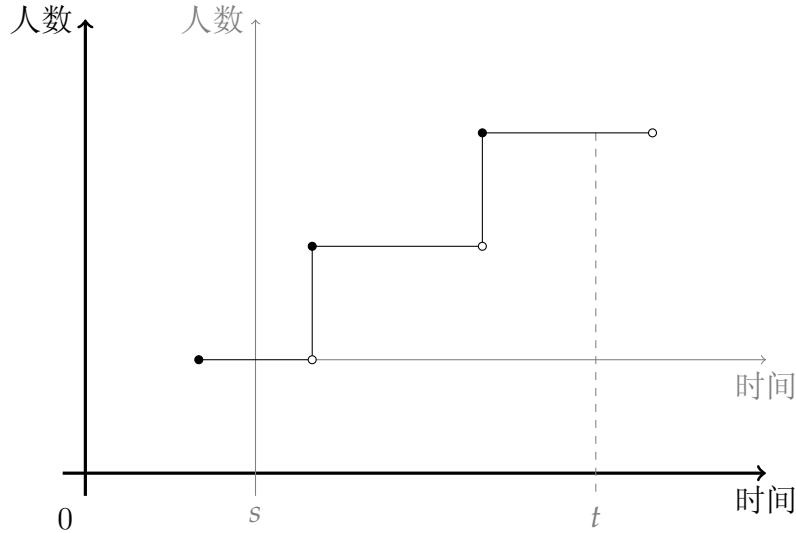


图 6.19: Poisson 过程是平稳增量过程: 以时间点 s 为起点, 以后仍是 Poisson 过程。

例 6.24. 为了更好地理解 Poisson 过程, 我们基于性质 6.22 来设计如下的随机试验: 令 $\alpha = 1, t = 4$ (或者 $t = 6$), 产生 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(\alpha)$ 的随机数, 进而求得 $N(t)$ 的随机数。重复该过程多次, 看这些随机数是否来自 $\text{Poisson}(\alpha t)$ 。

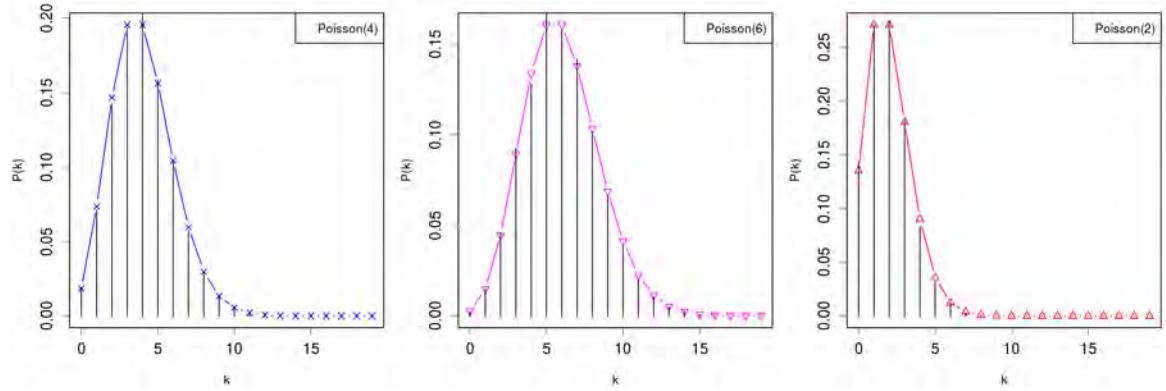


图 6.20: 按照例 6.24 产生 10^4 个 $N(t)$ 的随机数 (其频率见竖线图), 发现它们的确是来自 $\text{Poisson}(\alpha t)$ (图中折线)。另外, 练习 6.18 与试验结果是吻合的, 即 $N(6) - N(4)$ 的分布就是 $N(2)$ 的分布 $\text{Poisson}(2)$ 。

表 6.2: Bernoulli 过程和 Poisson 过程的性质对比。

随机变量	随机过程	Bernoulli	Poisson
相邻到达的间隔时间		几何分布	指数分布
第 k 次到达时间		负二项分布	Erlang 分布

性质 6.23. 对于 Poisson 过程 $N(t) \sim \text{Poisson}(\alpha t)$, 若 $s < t$, 则

$$\begin{aligned} P\{N(s) = k | N(t) = n\} &= C_n^k p^k (1-p)^{n-k}, \text{ 其中 } p = \frac{s}{t} \\ \rho(N(s), N(t)) &= \sqrt{\frac{s}{t}} \end{aligned}$$

证明. 在已知 $N(t) = n$ 的条件下, 利用 Bayes 公式计算 $N(s) = k$ 的概率。

$$\begin{aligned} P\{N(s) = k | N(t) = n\} &= \frac{P\{N(s) = k, N(t) = n\}}{P\{N(t) = n\}} \\ &= \frac{P\{N(s) = k, N(t) - N(s) = n - k\}}{P\{N(t) = n\}} \\ &= \frac{P\{N(s) = k\} P\{N(t-s) = n - k\}}{P\{N(t) = n\}} \\ &= \frac{\frac{(\alpha s)^k}{k!} e^{-\alpha s} \frac{(\alpha(t-s))^{n-k}}{(n-k)!} e^{-\alpha(t-s)}}{\frac{(\alpha t)^n}{n!} e^{-\alpha t}} \\ &= C_n^k \left(\frac{s}{t}\right)^k \left(1 - \frac{s}{t}\right)^{n-k} \end{aligned}$$

由 $E(N(t)) = V(N(t)) = \alpha t$ 知 $E(N^2(t)) = (\alpha t)^2 + \alpha t$ 。我们有

$$\begin{aligned} E(N(s)N(t)) &= E(N(s)(N(s) + N(t) - N(s))) \\ &= E(N^2(s)) + E(N(s))E(N(t) - N(s)) \\ &= \alpha s + \alpha^2 st \end{aligned}$$

$$\text{Cov}(N(s), N(t)) = E((N(s) - \alpha s)(N(t) - \alpha t)) = \alpha s$$

$$\rho(N(s), N(t)) = \frac{\text{Cov}(N(s), N(t))}{\sqrt{V(N(s))V(N(t))}} = \sqrt{\frac{s}{t}}$$

□

 我们知道: 几个独立的 Poisson 分布之和依然是 Poisson 分布 (见第 276 页的练习 4.17)。相反, 若两个独立随机变量之和是 Poisson 分布, 这两个随机变量也是 Poisson 分布 (见定理 4.3)。如果我们把强度为 α 的 Poisson 过程 $N(t)$ 拆成两个 Poisson 过程 $X(t)$ 和 $Y(t)$ 之和, 对应的强度分别为 $p\alpha$ 和 $q\alpha$, 其中 $q = 1 - p$, 则这

两个 Poisson 过程是独立的。事实上，

$$\begin{aligned} P\{X(t) = m, Y(t) = n\} &= P\{N(t) = m + n\} \\ &= \frac{(\alpha t)^{m+n}}{(m+n)!} e^{-\alpha t} \cdot \frac{(m+n)!}{m!n!} p^m q^n \\ &= \frac{(p\alpha t)^m}{m!} e^{-p\alpha t} \cdot \frac{(q\alpha t)^n}{n!} e^{-q\alpha t} \\ &= P\{X(t) = m\} P\{Y(t) = n\} \end{aligned}$$

例 6.25. 考虑强度是 α 的 Poisson 过程 $N(t)$ ，它描述是时间段 $[0, t]$ 逛商场的人数。令 X_j 表示第 j 个顾客的消费，不妨假设这些消费是独立同分布的，特征函数是 $\varphi_X(z)$ ，均值 μ_X ，方差 σ_X^2 。仿照**例 4.23**，我们从 Poisson 过程 $N(t)$ 构造新的随机过程 $Y(t)$ 如下，它表示时间段 $[0, t]$ 内的总消费。

$$Y(t) = \sum_{j=1}^{N(t)} X_j$$

于是， $Y(t)$ 的特征函数是

$$\varphi_{Y(t)}(z) = \exp\{-\alpha t(1 - \varphi_X(z))\}$$

请读者利用第 231 页的**定理 3.5** 验证

$$EY(t) = \alpha t \mu_X \quad VY(t) = \alpha t \mu_X^2 + \alpha t \sigma_X^2$$

例 6.26. 考虑 Poisson 过程 $N(t)$ 在某一随机时刻 $T \sim p(t)$ 时的均值和方差。

$$\begin{aligned} EN(T) &= \int_0^\infty E(N(T)|T=t)p(t)dt \\ &= \int_0^\infty \alpha t p(t)dt \\ &= \alpha ET \end{aligned}$$

请读者验证 $VN(T) = \alpha ET + \alpha^2 VT$

定理 6.5. 更新函数 $U(t)$ （见**定义 6.28**）满足下面的更新方程（renewal equation）。

$$U(t) = F(t) + \int_0^t U(t-x)dF(x) \tag{6.21}$$

证明. 根据双期望定理 2.18, 我们有

$$\begin{aligned}
 U(t) &= \mathbb{E}(N_t) = \mathbb{E}[\mathbb{E}(N_t|X_1)] \\
 &= \int_0^\infty \mathbb{E}(N_t|X_1 = x)dF(x) \\
 &= \int_0^\infty I_{\{t>x\}}[1 + \mathbb{E}(N_{t-x})]dF(x), \text{ 其中 } I_{\{t>x\}} \text{ 是指示函数} \\
 &= \int_0^t [1 + U(t-x)]dF(x) \\
 &= F(t) + \int_0^t U(t-x)dF(x)
 \end{aligned}
 \quad \square$$

练习 6.19. 验证 Poisson 过程的更新函数满足更新方程 (6.21)。

提示: $F(x) = 1 - e^{-\alpha t}$, $U(t) = \alpha t$, 并且

$$\int_0^t \alpha(t-x) \cdot \alpha e^{-\alpha x} dx = e^{-\alpha t} + \alpha - 1$$

\rightsquigarrow 定理 6.6 (W. Feller, 1941). 更新过程 N_t 和更新函数 $U(t)$ 满足下面的性质:

$$\lim_{t \rightarrow \infty} \frac{N_t}{t} = \frac{1}{m}, \text{ 其中 } m = \mathbb{E}(X_1) \quad (6.22)$$

$$\lim_{t \rightarrow \infty} \frac{U(t)}{t} = \frac{1}{m} \quad (6.23)$$

6.2.2 生灭过程

某服务窗口一次只能服务一个顾客，其他顾客排队等候。我们考虑以下状态：

状态 0：无服务顾客，无排队顾客。

状态 1：一个服务顾客，无排队顾客。

状态 2：一个服务顾客，一个排队顾客……

状态 n ：一个服务顾客， $n - 1$ 个排队顾客。

已知条件：顾客人数是到达率为 λ 的 Poisson 过程，

服务时间服从指数分布 $\text{Expon}(\mu)$ ，队列可以无限长，

客源数目无限多，服务规则是先到先服务。满足这些

已知条件的数学模型被称作 $M/M/1$ 排队模型，常用下图直观地描述。

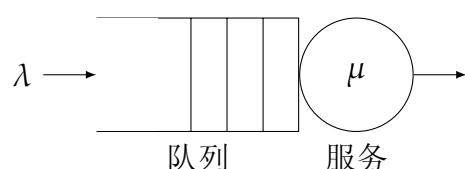


图 6.21: $M/M/1$ 排队模型：遵循先到先服务的单一服务，服务时间服从指数分布 $\text{Expon}(\mu)$ ，参与人数不限（到达率为 λ 的 Poisson 过程），队列长度不限。

状态之间的转移概率如下图所示，这样的一个 Markov 链被称为一个生灭过程 (birth-death process)。生灭过程是一类特殊的连续时间 Markov 链。

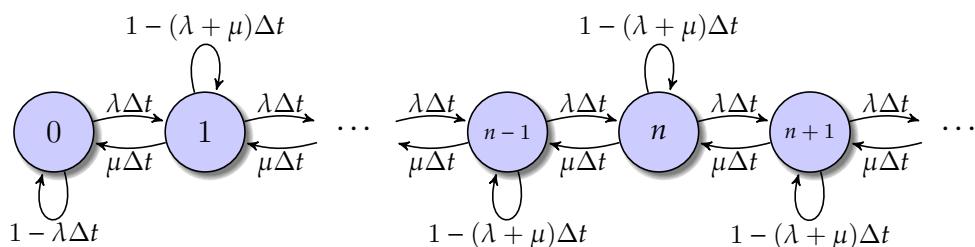


图 6.22: 当间隔时间 Δt 很小时，生灭过程在状态 $n > 0$ 只有三种状态转移：一生、一灭、无生无灭，转移概率分别是 $\lambda\Delta t + o(\Delta t)$, $\mu\Delta t + o(\Delta t)$, $1 - (\lambda + \mu)\Delta t + o(\Delta t)$ 。在单位时间内，排队的顾客人数（即生的数量） λ 如果小于可服务人数（灭的数量） μ ，则队列将变得越来越短。反之，队列将变得无限长。



最早思考这个问题的是丹麦数学家、统计学家、工程师 Agner Krarup Erlang (1878-1929)，他是排队论或随机服务系统理论的奠基者。1909 年，Erlang 发现 Poisson 分布可用于描述电话流量。1917 年，Erlang 利用概率论研究电话通讯中等待接线的平均人数、不同等待时间等问题。二十世纪三十年代，W. Feller 引入生灭过程，为排队论奠定了理论基础。

二十世纪中叶，英国数学家、统计学家 David George Kendall (1918-2007) 利用 Markov 链系统地研究了排队论。1951 年，Kendall 提出 Kendall 记号 $X/Y/Z/A/B$ 来表示排队系统，其中 X 表示相邻到达间隔时间的分布， Y 表示服务时间的分布， Z 表示服务窗口的个数， A 表示队列的长度， B 表示顾客总数。例如， $M/M/1/\infty/\infty$ 中的 M 表示无记忆 (memoryless)，简记作 $M/M/1$ 。如今，排队论已经发展成为运筹学的一个分支，通过到达人数的随机过程和服务时间研究系统空闲概率、等待时间、排队长度、逗留时间（即，等待时间+服务时间）等规律，以便花费最小的代价满足服务需求。



性质 6.24. 记 $P(X(t) = n)$ 为 $P_n(t)$ ，则由状态转移图 6.22 定义的生灭过程满足

$$\begin{aligned} P'_n(t) &= \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t) + \mu P_{n+1}(t), \text{ 其中 } n \geq 1 \\ P'_0(t) &= -\lambda P_0(t) + \mu P_1(t) \end{aligned}$$

证明. 当 Δt 很小时，我们考虑 $t + \Delta t$ 时刻处于状态 n 的概率 $P_n(t + \Delta t)$: 在 $[t, t + \Delta t]$ 内只有三种状态转移，即 (1) 从 t 时刻的状态 $n - 1$ “生”到 $t + \Delta t$ 时刻的状态 n ；(2) “无生无灭”仍处于状态 n ；(3) 从 t 时刻的状态 $n + 1$ “灭”到 $t + \Delta t$ 时刻的状态 n 。

$$\begin{aligned} P_n(t + \Delta t) &= \lambda \Delta t \cdot P_{n-1}(t) + (1 - \lambda \Delta t - \mu \Delta t) \cdot P_n(t) + \mu \Delta t \cdot P_{n+1}(t) + o(\Delta t) \\ P_0(t + \Delta t) &= (1 - \lambda \Delta t) \cdot P_0(t) + \mu \Delta t \cdot P_1(t) + o(\Delta t) \end{aligned}$$

分别将 $P_n(t)$ 和 $P_0(t)$ 移到左边，两边同时除以 Δt 便证得。 □

性质 6.25. 令 $t \rightarrow \infty$, $P_n(t) \rightarrow p_n$ 是生灭过程的平稳状态分布。若 $\rho = \lambda/\mu < 1$ ，有

$$\begin{aligned} p_0 &= \frac{1}{1 + \rho + \rho^2 + \dots} = 1 - \rho \\ p_n &= (1 - \rho)\rho^n, \text{ 其中 } n \geq 1 \end{aligned}$$

证明. 令 $t \rightarrow \infty$, 由性质 6.24 我们有

$$\begin{aligned} \lambda p_{n-1} - (\lambda + \mu)p_n + \mu p_{n+1} &= 0 \\ -\lambda p_0 + \mu p_1 &= 0 \end{aligned}$$

求解递归式，得到

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 = \rho^n p_0, \text{ 其中 } \rho = \frac{\lambda}{\mu}$$

若 $\rho < 1$, 根据 $p_0 + p_1 + \dots = 1$ 不难证得结果。 \square

练习 6.20. 在 $M/M/1$ 排队模型中, 平均顾客数目是 $\rho/(1 - \rho)$ 。

提示: 计算 $\sum_{n=0}^{\infty} np_n$ 。

练习 6.21. 如果生灭过程的状态转移图如下所示, 请证明:

$$P'_n(t) = \lambda_{n-1}P_{n-1}(t) - (\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t), \text{ 其中 } n \geq 1 \quad (6.24)$$

$$P'_0(t) = -\lambda_0P_0(t) + \mu_1P_1(t) \quad (6.25)$$

该生灭过程的平稳状态分布是

$$p_n = \frac{\lambda_0\lambda_1 \cdots \lambda_{n-1}}{\mu_1\mu_2 \cdots \mu_n} p_0, \text{ 其中 } p_0 = \left(1 + \sum_{n=1}^{\infty} \frac{\lambda_0\lambda_1 \cdots \lambda_{n-1}}{\mu_1\mu_2 \cdots \mu_n}\right)^{-1} \quad (6.26)$$

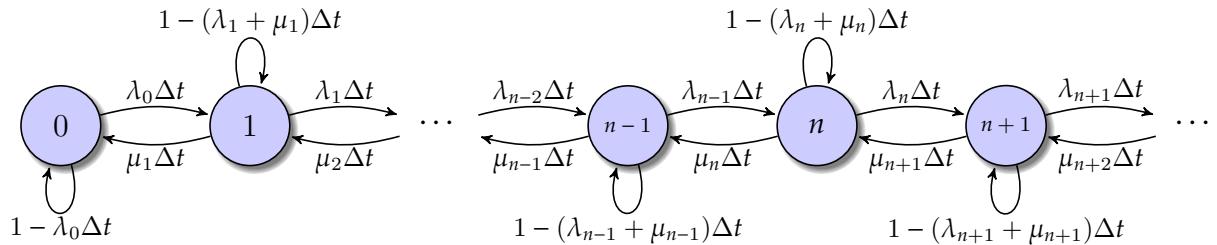


图 6.23: 与图 6.22 区别在于“生”、“灭”、“无生无灭”的概率因当前状态而异。

例 6.27. 考虑 $M/M/\infty$ 排队模型, 即有无穷个服务窗口, 每个顾客都能得到即时服务而无需等待。没有排队的顾客, 状态 n 就是有 n 个顾客正被服务。在图 6.23 中, $\lambda_n = \lambda$, $\mu_n = \mu n$ (即, 单位时间里可服务 μn 人)。令 $\rho = \lambda/\mu$, 根据练习 6.21 的结果, 平稳状态分布为 $\text{Poisson}(\rho)$, 即

$$p_0 = \left(\sum_{n=0}^{\infty} \frac{\rho^n}{n!} \right)^{-1} = e^{-\rho} \quad p_n = \frac{\rho^n}{n!} e^{-\rho}$$

※例 6.28. 考虑 $M/M/c$ 排队模型, 即有 c 个服务窗口。在图 6.23 中, $\lambda_n = \lambda$,

$$\mu_n = \begin{cases} \mu n & \text{如果 } n \leq c \\ \mu c & \text{如果 } n > c \end{cases}$$

令 $\rho = \lambda/\mu$, 根据练习 6.21 的结果, 平稳状态分布为

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \frac{1}{1 - \rho/c} \right)^{-1} \quad p_n = \begin{cases} \frac{\rho^n}{n!} p_0 & \text{如果 } 0 < n \leq c \\ \frac{\rho^n c^{c-n}}{c!} p_0 & \text{如果 } n > c \end{cases}$$

※例 6.29. 考虑 $M/M/1/k$ 排队模型, 即有 1 个服务窗口, 队列长度为 k 。在图 6.23 中,

$$\lambda_n = \begin{cases} \lambda & \text{如果 } 0 \leq n < k \\ 0 & \text{如果 } n \geq k \end{cases} \quad \mu_n = \begin{cases} \mu & \text{如果 } 1 \leq n \leq k \\ 0 & \text{如果 } n > k \end{cases}$$

令 $\rho = \lambda/\mu$, 根据练习 6.21 的结果,

$$P'_n(t) = 0, \text{ 其中 } n > k$$

$$P'_n(t) = \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t) + \mu_{n+1} P_{n+1}(t), \text{ 其中 } 1 \leq n \leq k$$

$$P'_0(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t)$$

仿照性质 6.25 的证明求解递归式, 得到平稳状态分布为

$$p_0 = \left(\sum_{n=0}^k \frac{\rho^n}{n!} \right)^{-1} \quad p_n = \rho^n p_0, \text{ 其中 } n = 1, 2, \dots, k$$

6.2.3 布朗运动

定义 6.29. 一维标准布朗运动 (standard Brownian motion), 或称为 Wiener-Bachelier 过程或 Wiener 过程, 就是满足下述条件的随机过程 $B_t, t \geq 0$:

① $B_0 \stackrel{a.s.}{=} 0$, 即在起始点几乎必然为零, $P(B_0 = 0) = 1$ 。

② 作为 t 的函数, $B(t)$ 几乎必然连续, 即

$$\lim_{\Delta t \downarrow 0} \frac{P\{|B(t + \Delta t) - B(t)| > \epsilon\}}{\Delta t} = 0, \forall \epsilon > 0$$

③ B_t 具有独立增量, 且 $B_t - B_s \sim N(0, t - s)$, 其中 $0 \leq s < t$ 。

缺省地, 我们用 B_t 或 W_t 表示标准布朗运动, 有时也简称为布朗运动。显然, 它是一个 CTCV 过程且 $B_t \sim N(0, t)$ 。

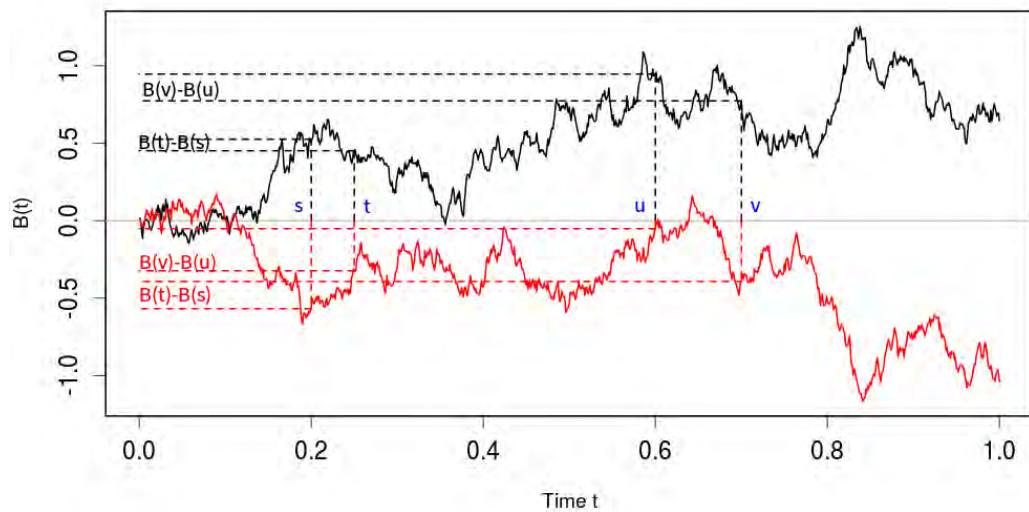
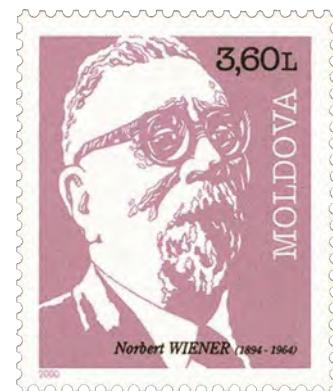


图 6.24: 标准布朗运动: (1) B_t 具有独立增量: 若 (s, t) 与 (u, v) 不相交, 则 $B_t - B_s$ 与 $B_v - B_u$ 独立。 (2) B_t 具有平稳增量: $B_0 = 0, B_t - B_s \sim N(0, t - s), 0 \leq s < t$ 。

1923 年, 美国数学家、控制论之父 Norbert Wiener (1894-1964) 首次证明标准布朗运动的存在性 (证明稍复杂, 从略)。他利用随机 Fourier 级数刻画区间 $[0, 1]$ 上的布朗运动,

$$B_t = \xi_0 t + \sqrt{2} \sum_{n=1}^{\infty} \xi_n \frac{\sin(n\pi t)}{n\pi}, \text{ 其中 } \xi_0, \xi_1, \dots \stackrel{\text{iid}}{\sim} N(0, 1)$$

另外, 区间 $[0, 1]$ 上的布朗运动还可以由简单随机游动 (见例 6.3) 来近似构造, 见下面的例子。



例 6.30. 将区间 $[0, 1]$ 划分为 n 等分, 在时间点 $t = \frac{k}{n}, k = 0, 1, \dots, n$ 上的随机变量是

$$W(t) = \sum_{i=1}^k X_i, \text{ 其中 } X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \frac{1}{2} \left\langle \sqrt{\frac{1}{n}} \right\rangle + \frac{1}{2} \left\langle -\sqrt{\frac{1}{n}} \right\rangle \quad (6.27)$$

我们用线段把离散的点连接起来, 当 $n \rightarrow \infty$ 时, 就是一维标准布朗运动。三维的布朗运动见第 783 页的**例 G.11**。

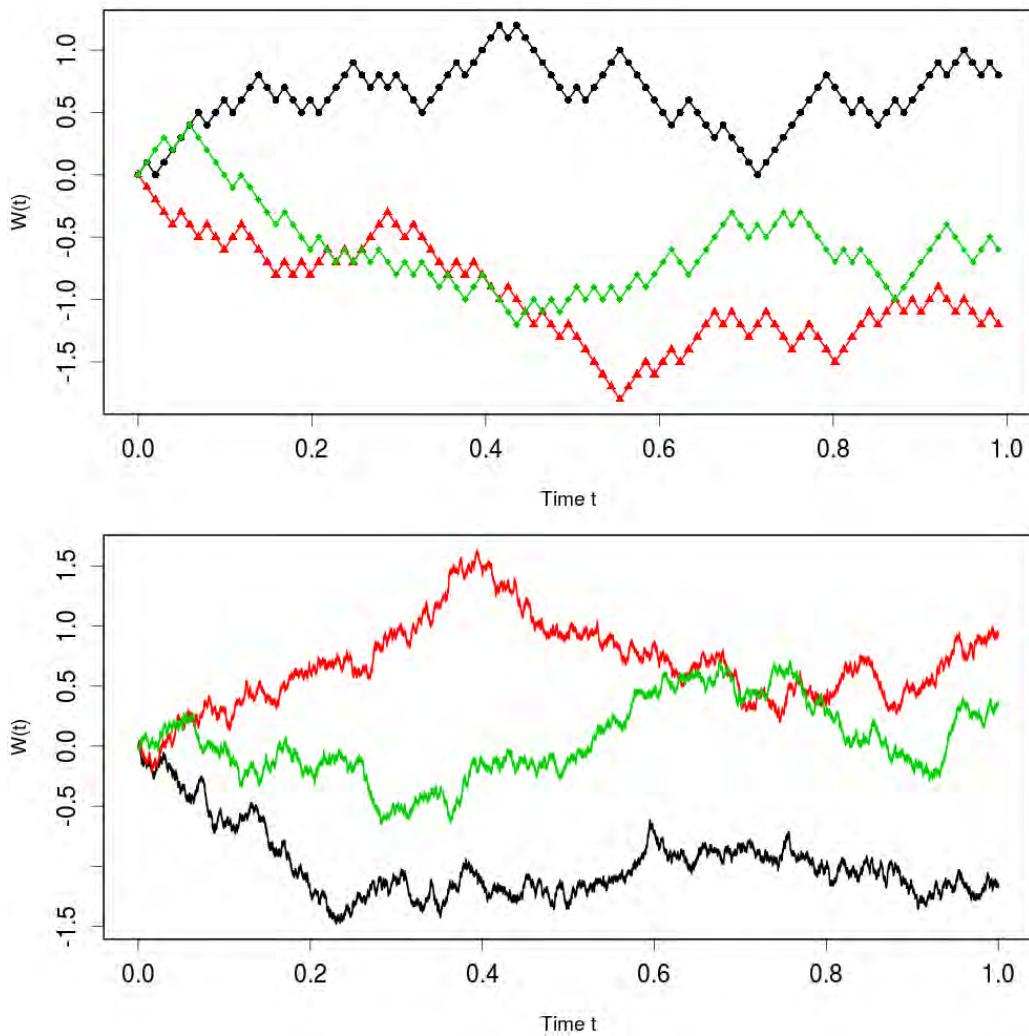


图 6.25: 由简单随机游动构造标准布朗运动: 在式 (6.27) 中分别取 $n = 10^2, n = 10^5$ 。

不难验证 $n \rightarrow \infty$ 随机过程 $W(t)$ 就是标准布朗运动。

$$\mathbb{E}(W(t)) = 0$$

$$\mathbb{V}(W(t)) = k\mathbb{V}(X_1) = k/n = t$$

由中心极限定理, $W(t)$ 服从正态分布 $N(0, t)$ 。另外, 考虑时间点 $s = \frac{k'}{n} < t = \frac{k}{n}$,

$$\begin{aligned}\mathbb{E}[W(t) - W(s)] &= 0 \\ \mathbb{V}[W(t) - W(s)] &= \sum_{i=k'+1}^k \mathbb{V}(X_i) = (k - k')/n = t - s\end{aligned}$$

于是, $W(t) - W(s) \sim N(0, t - s)$, $W(t)$ 是平稳增量过程。同时, 它显然也是独立增量过程。

\rightsquigarrow **性质 6.26.** 在固定时刻 t , $B_t \sim N(0, t)$, 并且

$$\begin{aligned}\text{Cov}(B_s, B_t) &= \min(s, t) \\ \rho(B_s, B_t) &= \frac{\min(s, t)}{st} = \sqrt{\frac{\min(s, t)}{\max(s, t)}}\end{aligned}$$

证明. 不妨设 $s < t$, 则 $B_t = (B_t - B_s) + B_s$, 进而

$$\begin{aligned}\text{Cov}(B_s, B_t) &= \mathbb{E}(B_s B_t) \\ &= \mathbb{E}[B_s(B_t - B_s) + B_s^2], \text{ 因为 } B_s - B_0 \text{ 与 } B_t - B_s \text{ 独立} \\ &= \mathbb{E}(B_s)\mathbb{E}(B_t - B_s) + \mathbb{E}(B_s^2) \\ &= s\end{aligned}$$

□

\rightsquigarrow **定理 6.7.** 若高斯过程 X_t 具有零均值且 $\text{Cov}(X_s, X_t) = \min(s, t)$, 则 X_t 是布朗运动。

性质 6.27. 基于布朗运动 B_t 可构造新的布朗运动 W_t 如下:

- ① 时间平移: $W_t = B_{t+h} - B_h$, 其中 h 是一个给定的正数
- ② 空间对称: $W_t = -B_t$
- ③ 尺度变换: $W_t = B_{\alpha t}/\sqrt{\alpha}$, 其中 $\alpha > 0$
- ④ 时间反转: $W_t = B_T - B_{T-t}$, 其中 $T > 0, 0 \leq t \leq T$
- ⑤ 时间反演: $W_0 = 0, W_t = tB_{\frac{1}{t}}$, 其中 $t > 0$

证明. 不妨设 $s < t$, 下面往证时间反演, 其他的留作习题。

$$\begin{aligned}\text{Cov}(W_s, W_t) &= \text{Cov}[sB_{\frac{1}{s}}, tB_{\frac{1}{t}}] \\ &= st \min\left(\frac{1}{s}, \frac{1}{t}\right) \\ &= s\end{aligned}$$

于是 $\text{Cov}(W_s, W_t - W_s) = \text{Cov}(W_s, W_t) - \text{V}(W_s) = s - s = 0$ 。因为 W_t, W_s 服从正态分布，所以 $\text{Cov}(W_s, W_t - W_s) = 0$ 意味着 $W_t - W_s, W_s$ 相互独立。 \square

定义 6.30. 基于布朗运动 B_s ，构造即时最大值 (running maximum) 过程 M_t 如下，它刻画了布朗运动在时间段 $[0, t]$ 的最大值。

$$M_t = \max_{0 \leq s \leq t} B_s \quad (6.28)$$

\curvearrowleft **性质 6.28.** 布朗运动的即时最大值 M_t 的分布函数和密度函数如下。

$$\begin{aligned} P(M_t > m) &= 2P(B_t > m) \\ F_{M_t}(m) &= 2\Phi\left(\frac{m}{\sqrt{t}}\right) - 1 \\ f_{M_t}(m) &= \frac{2}{\sqrt{t}}\phi\left(\frac{m}{\sqrt{t}}\right) = \sqrt{\frac{2}{\pi t}} \exp\left\{-\frac{m^2}{2t}\right\} \end{aligned}$$

证明. 只需往证 $P(M_t > m) = 2 - 2\Phi(m/\sqrt{t})$ 。事实上，

$$\begin{aligned} P(M_t > m) &= P(\max_{0 \leq s \leq t} B_s > m \text{ 且 } B_t > m) + P(\max_{0 \leq s \leq t} B_s > m \text{ 且 } B_t < m) \\ &= 2P(\max_{0 \leq s \leq t} B_s > m \text{ 且 } B_t > m), \text{ 见图 6.26} \\ &= 2P(B_t > m), \text{ 因为 } B_t \sim N(0, t) \\ &= 2 - 2\Phi\left(\frac{m}{\sqrt{t}}\right) \end{aligned} \quad \square$$

图 6.26 解释了为何 M_t 触 m 线的机会是 B_t 的两倍。这个性质很重要，用于证明布朗运动的下述重要性质。

\curvearrowleft **定理 6.8.** 对于任意的 $T > 0$ ，皆有

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n \left[B\left(\frac{k}{n}T\right) - B\left(\frac{k-1}{n}T\right) \right]^2 \stackrel{a.s.}{=} T$$

证明. 根据布朗运动的定义，

$$X_k = B\left(\frac{k}{n}T\right) - B\left(\frac{k-1}{n}T\right) \sim N\left(0, \frac{T}{n}\right)$$

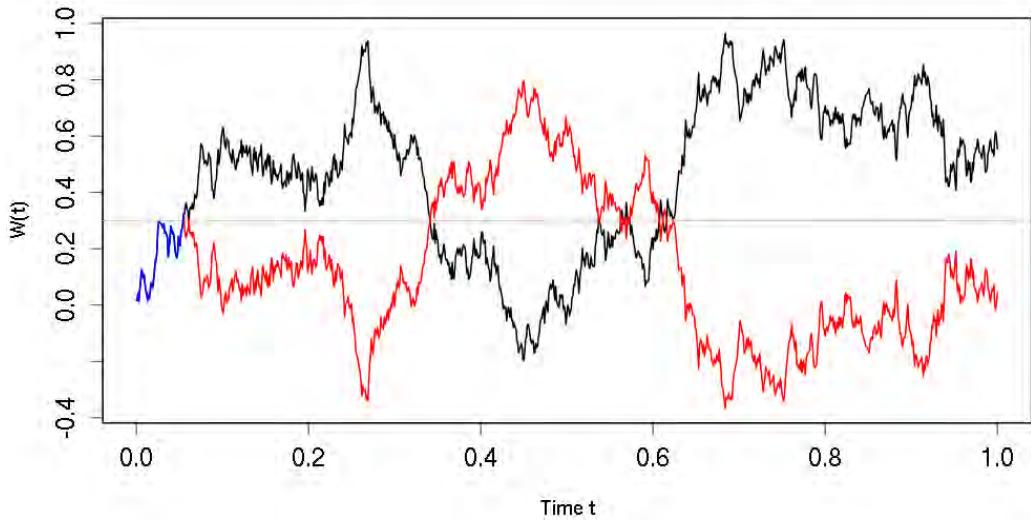


图 6.26: 对于随机过程 B_s 在时间段 $[0, t]$ 上到达状态 m 的任一样本路径, 不妨设首次到达时间是 t_m , 总存在另外一个样本路径, 在时刻 t_m 之后是原路径关于水平线 m 的镜像。

于是, $E(X_k^2) = T/n$, 进而利用强大数律,

$$\sum_{k=1}^n X_k^2 = n \left(\frac{1}{n} \sum_{k=1}^n X_k^2 \right) \xrightarrow{a.s.} n \cdot \frac{T}{n} = T \quad \square$$

上述结果暗示了布朗运动是不可导的。如果函数 $f(x)$ 在 $[0, T]$ 上可导, 令 $t_k = kT/n$, 则必有 $\lim_{n \rightarrow \infty} \sum_{k=1}^n [f(t_k) - f(t_{k-1})]^2 = 0$, 因为

$$\begin{aligned} \sum_{k=1}^n [f(t_k) - f(t_{k-1})]^2 &= \sum_{k=1}^n f'(\xi_k)(t_k - t_{k-1})^2, \text{ 其中 } \xi_k \in [t_{k-1}, t_k] \\ &\leq \max_{0 \leq t \leq T} f'(t) \frac{T^2}{n} \rightarrow 0, \text{ 当 } n \rightarrow \infty \end{aligned}$$

下面不加证明介绍布朗运动的一个重要性质, 感兴趣的读者可参考 Breiman 的《概率论》[22] 第 261 页定理 12.25 的证明。

性质 6.29 (Paley-Wiener-Zygmund, 1933^{*}). 布朗运动 $B(t)$ 路径几乎处处不可导。

例 6.31 (布朗桥). 由布朗运动 B_t 构造一个新的随机过程

$$X_t = B_t - tB_1, \text{ 其中 } 0 \leq t \leq 1$$

^{*}1933 年, 英国数学家 R. Paley (1907-1933), 美国数学家 N. Wiener 和波兰数学家 A. Zygmund (1900-1992) 证得此结果。遗憾的是论文尚未出版, Paley 滑雪时殒于雪崩。

于是, $X_0 = X_1 = 0$ 。该随机过程被称为布朗桥 (Brownian bridge)。随机变量 $X_{t+h} - X_t = B_{t+h} - B_t - hB_1$ 是正态的, 均值为零, 方差不依赖于 t , 且

$$\begin{aligned}\mathrm{E}(X_{t+h} - X_t)^2 &= \mathrm{E}(B_{t+h} - B_t)^2 + h^2\mathrm{E}(B_1^2) - 2h\mathrm{E}[(B_{t+h} - B_t)B_1] \\ &= h + h^2 - 2h^2 \\ &= h - h^2\end{aligned}$$

换句话说, Markov 过程 X_t 有平稳增量, 然而却不是齐性的, 并且也没有独立增量。

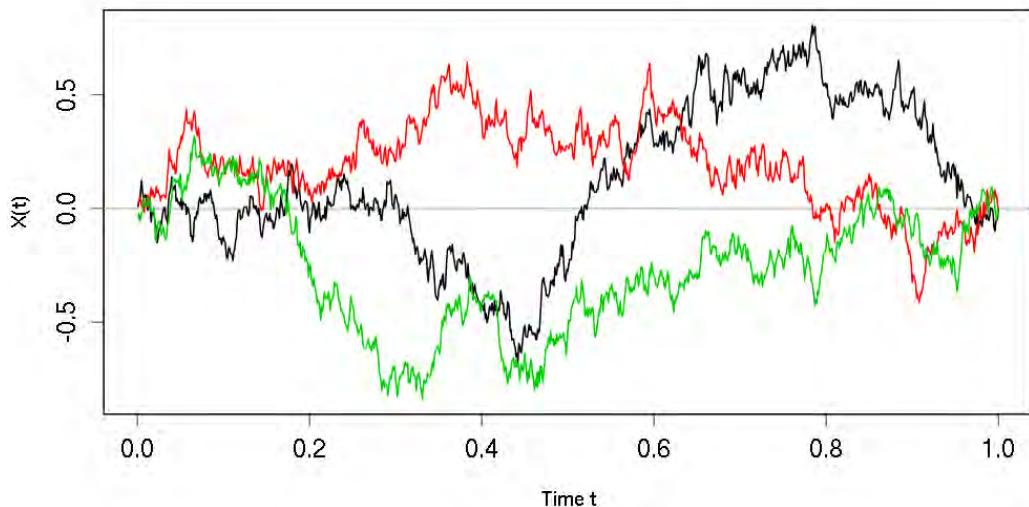


图 6.27: 布朗桥 $X_t = B_t - tB_1$ 的三个样本路径。

例 6.32 (带漂移的布朗运动). 给定常数 $\mu \in \mathbb{R}, \sigma > 0$, 由标准布朗运动 B_t 构造一个新的随机过程

$$X_t = \mu t + \sigma B_t, \text{ 其中 } t \geq 0 \quad (6.29)$$

显然, 该随机过程是一个高斯过程, 期望函数是 $\mu_X(t) = \mu t$, 协方差函数是 $\gamma_X(s, t) = \sigma^2 \min(s, t)$ 。随机过程 (6.29) 被称为带漂移的布朗运动, 期望函数 $\mu_X(t) = \mu t$ 基本决定了过程的样本路径。

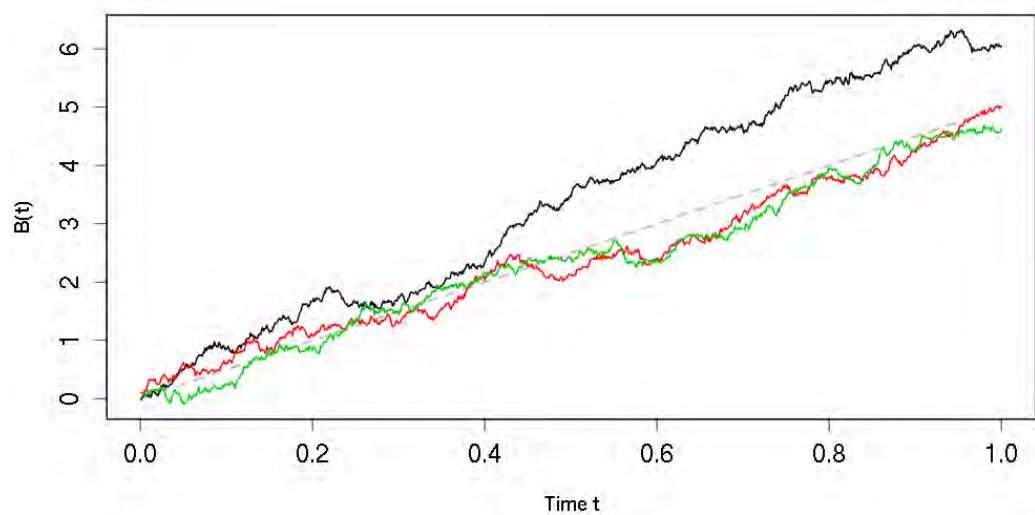


图 6.28: 带漂移的布朗运动 $X_t = 5t + B_t$ 的三个样本路径, 虚线是 $\mu_B(t) = 5t$ 。

6.3 随机分析

随机分析是概率论里以分析的方法研究随机过程的一个分支，包括 Itô 积分、随机微分方程等内容。日本数学家伊藤清 (Kiyoshi Itô, 1915-2008) 是随机分析的奠基者，被称为“随机分析之父”。

伊藤清在《我的六十年概率之路》回忆了他的学术生涯。“当时概率论的研究主要针对各种独立随机变量序列的特性，目的是为统计规律奠定数学基础。这个对应着分析基础里的级数理论。虽然对比级数理论统计规律的研究更难、更有变化，但在当时，我依然认为它们没有其他数学领域那么吸引人，并且不觉得自己投身于统计规律的研究之中。最终把我重新吸引回概率论世界的是阅读了法国数学家 Paul Lévy 的《独立随机变量之和的理论》(1937)。这是随机过程研究里的第一个巨大的进步，它把随机过程的概念视作某个对应到微积分里函数的东西。在 Lévy 的论文里我找到了新概率论的本质，于是我决定沿着 Lévy 铺设并以一抹微光照亮的道路走下去。”



6.3.1 伊藤积分

定义 6.31 (伊藤积分). 令 $B(t)$ 是布朗运动, 定义伊藤积分 (Itô integral) 为下面的 Riemann-Stieltjes 积分。

$$X_t = \int_0^t f(s) dB(s) = \lim_{n \rightarrow \infty} \sum_{j=1}^n f(t_{j-1})(B(t_j) - B(t_{j-1}))$$

其中, $t_0 = 0 < t_1 < \dots < t_{n-1} < t_n = t$ 。新定义的随机过程 X_t 简记作

$$dX_t = f(t) dB_t$$

定义 6.32 (伊藤过程). 如下定义的随机过程被称为伊藤过程,

$$X_t = X_0 + \int_0^t \mu(s) ds + \int_0^t \sigma(s) dB(s), \text{ 其中 } \int_0^t \{\sigma^2(s) + |\mu(s)|\} ds < \infty$$

伊藤过程 X_t 简记作

$$dX_t = \mu_t dt + \sigma_t dB_t \tag{6.30}$$

引理 6.1 (伊藤清). 令 $f(t, x)$ 是二次可微的函数, X_t 是伊藤过程 (6.30), 则 $Y_t = f(t, X_t)$ 仍然是一个伊藤过程, 并且

$$dY_t = \left(\frac{\partial f}{\partial t} + \mu_t \frac{\partial f}{\partial x} + \frac{\sigma_t^2}{2} \frac{\partial^2 f}{\partial x^2} \right) dt + \sigma_t \frac{\partial f}{\partial x} dB_t \tag{6.31}$$

证明. 下面是伊藤引理一个不严格的证明。由 $f(t, x)$ 的 Taylor 展开式可得

$$\begin{aligned} df &= f(t + dt, x + dx) - f(t, x) \\ &= \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dx + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} dx^2 + \dots \end{aligned}$$

将 x 和 dx 分别替换为 X_t 和 $\mu_t dt + \sigma_t dB_t$, 我们得到

$$df = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} (\mu_t dt + \sigma_t dB_t) + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} (\mu_t^2 dt^2 + 2\mu_t \sigma_t dt dB_t + \sigma_t^2 dB_t^2) + \dots$$

当 $dt \rightarrow 0$, $dt^2, dt dB_t$ 趋向 0 的速度大于 $dB_t^2 = dt$, 于是式 (6.31) 成立。 \square

例 6.33. 若随机过程 Y_t 满足随机微分方程 $dY_t = \mu dt + \sigma dB_t$, 其中 μ, σ 为常数, 则

$$Y_t = Y_0 + \mu t + \sigma B_t$$

定义 6.33 (几何布朗运动). 给定常数 μ, σ , 满足下面随机微分方程的随机过程 S_t 被称为几何布朗运动 (geometric Brownian motion, GBM)。

$$dS_t = \mu S_t dt + \sigma S_t dB_t \quad (6.32)$$

性质 6.30. 几何布朗运动 (6.32) 的解析表达式是

$$S_t = S_0 \exp \left\{ \left(\mu - \frac{\sigma^2}{2} \right) t + \sigma B_t \right\} \quad (6.33)$$

证明. 令函数 $f(t, x) = \ln x$, 则

$$\frac{\partial f}{\partial t} = 0, \frac{\partial f}{\partial x} = \frac{1}{x}, \frac{\partial^2 f}{\partial x^2} = -\frac{1}{x^2}$$

考虑随机过程 $Y_t = f(t, S_t)$, 将 $\mu_t = \mu S_t, \sigma_t = \sigma S_t$ 代入 Itô 公式 (6.31), 我们得到

$$dY_t = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma dB_t$$

于是, $Y_t = Y_0 + (\mu - \sigma^2/2)t + \sigma B_t$, 进而得到 (6.33)。 \square

推论 6.2. 几何布朗运动 S_t 对每个固定的 t 都服从对数正态分布, 即

$$S_t \sim \log N \left(\left\{ \mu - \frac{\sigma^2}{2} \right\} t + \ln S_0, \sigma^2 t \right)$$

练习 6.22. 给出几何布朗运动的期望函数和方差函数。提示: 利用练习 4.29 的结果。

$$E(S_t) = S_0 e^{\mu t}$$

$$V(S_t) = S_0^2 e^{2\mu t} (e^{\sigma^2 t} - 1)$$

6.3.2 随机微分方程

一个微分方程若有一项或多项是随机过程，其解也是一个随机过程，则称之为随机微分方程 (stochastic differential equation)。例如，

$$\begin{aligned} dX_t &= \mu(X_t, t) + \sigma(X_t, t)dB_t \\ dX_t &= \mu X_t + \sigma X_t dB_t \end{aligned}$$

6.4 习题

6.1. 已知两个 Bernoulli 过程 $\{X_n\}$ 和 $\{Y_n\}$ 相互独立, 参数分别为 p 和 q 。试证明:

- (a) $Z_n = \max\{X_n, Y_n\}$ 也是一个 Bernoulli 过程。
- (b) $U_n = X_n \cdot Y_n, V_n = X_n \cdot (1 - Y_n)$ 是两个 Bernoulli 过程, 二者并不独立。

6.2. 求即时最大值 (6.28) 和布朗运动 B_t 的联合密度函数, 并由此推导即时最大值的密度函数及其期望。

6.3. 请补充性质 6.27 的证明。

6.4.

6.5. 有限状态的 Markov 链总有某个状态是常返的。

6.6. 一个 Markov 链是遍历的当且仅当它的每个状态都是非周期的、正常返的。

- ☆ 6.7. 若 $i \leftrightarrow j$ 且 i 是常返的, 则 $f_{ij} = 1$ 。
- ☆ 6.8. 设 T 是由所有非常返的状态构成的类, 则从任意非常返状态 $i \in T$ 出发首次到达常返状态 j 的概率为

$$f_{ij} = \sum_{k \in T} p_{ik} f_{kj} + \sum_{k \in K} p_{ik}$$

其中 K 是与 j 相通的状态的集合。

6.9. 假设天气状态为“雨天”、“晴天”、“雪天”, 状态转移矩阵 P 如下, 求该 Markov 链的平稳分布。

$$P = \begin{matrix} & \text{雨天} & \text{晴天} & \text{雪天} \\ \text{雨天} & \left(\begin{matrix} 1/2 & 1/4 & 1/4 \end{matrix} \right) \\ \text{晴天} & \left(\begin{matrix} 1/2 & 0 & 1/2 \end{matrix} \right) \\ \text{雪天} & \left(\begin{matrix} 1/4 & 1/4 & 1/2 \end{matrix} \right) \end{matrix}$$

6.10. 一个具有有限状态 $1, 2, \dots, n$ 的不可约非周期的 Markov 链, 如果它的转移矩阵 $P = (p_{ij})$ 是对称的, 求该 Markov 链的平稳分布。

6.11. 接着性质 6.17, 利用已有结果证明:

$$\mathbb{E}(Z_\infty) = \frac{1}{1 - \mu} \quad \mathbb{V}(Z_\infty) = \frac{\sigma^2}{(1 - \mu)^3}$$

6.12. 设 n 个灯泡的寿命 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(\alpha)$, 令 $Y = \max\{X_1, \dots, X_n\}$, 求 EY 。

第二部分

数理统计学初步

数理统计学简史

人有悲欢离合，月有阴晴圆缺，此事古难全。

苏轼《水调歌头》

统 计 学（或数理统计学）是应用数学的一个分支，研究如何收集整理、分析探索带有随机性的数据，以便通过观察对研究的问题做出推断、预测甚至决策。这门学问由来已久，因为人类在社会生活中总是透过大量的随机现象总结经验或探究自然本质，不论是对经济、人口等国情的宏观了解，还是对天气、地震等自然现象的预测，都需要系统的科学方法的指导^{*}。在概率论成为统计学的基础之前，人们多采用描述性的统计方法，如 K. Pearson 引入的直方图等，并未考虑因随机抽样带来的随机性。

当概率论揭示了随机性的规律，统计学在某种意义上可看作是概率论的一个应用领域。譬如，人们使用统计方法来估算测量中的随机误差，对之进行分析并想方设法减小它。

另一方面，概率论的逆问题是统计学的一个主题：假定一个随机现象由概率空间 (Ω, \mathcal{S}, P) 来描述，研究者只了解 P 的部分信息（譬如， P 所在的概率分布族）和该随机现象的一些观察结果，他们所面对的一个经典统计问题就是寻求 P 的最优估计。

譬如，已知某测量结果服从正态分布 $N(\mu, \sigma^2)$ ，其中参数 μ, σ^2 未知，如何从有限的测量数据中估算出这些未知参数？这里面蕴藏着一个统计学的基本认识，即把数据视作来自具有一定概率分布的总体，只是这个概率分布对我们而言不是完全明确了的，其中有些信息缺失了。在这个认识之下，总体的分布是一个客观实在，理论上允许数据源源不断地从中产生，观察结果就是该分布的抽样结果，它们仅仅是表象。因此，统计学的真正研究对象是总体分布，而不是数据本身。

统计学伴随概率论的发展而变得更加数学化，它的发展历史一般被划分为以下三个阶段 [1, 167, 169]。



* “统计学是科学的语法。”

— 统计学之父 K. Pearson (1857-1936)

二十世纪前

几件重要的工作包括, (i) 直方图等描述方法。 (ii) 1763 年 T. Bayes 的论文《论有关机遇问题的求解》对统计思想产生巨大影响, 催生了贝叶斯学派。 (iii) Gauss、Legendre 基于最小二乘法的误差分析, 以及对统计学基本认识的确立。 (iv) 英国人类学家、遗传学家和统计学家 Francis Galton (1822-1911) 关于回归分析的先驱性的工作。 (v) χ^2 分布的发现以及对正态总体的研究。 (vi) 统计学之父 K. Pearson (1857-1936) 在研究曲线拟合时提出的矩方法成为参数点估计的经典方法之一。

二十世纪上半叶

统计学得到迅速发展, 诞生了许多新方法和新分支。 (i) 1900 年, K. Pearson 提出拟合优度的 χ^2 检验。 (ii) 1908 年, W. S. Gosset (笔名 Student) 提出 t 分布和正态总体均值的 t 检验。 (iii) 英国统计学家 R. A. Fisher 是一位在统计学发展史上举足轻重的天才人物, 以他的关键工作为标志数理统计学得以形成和发展^{*}: 1912-1925 年, 最大似然估计成为参数点估计的又一经典方法 [48]; 20 年代系统地发展了正态总体下各种统计量的抽样分布, 初步建立了相关分析、回归分析和多元分析等分支; 20-30 年代, 创立了试验设计与方差分析。另外, Fisher 提出了“信任推断”, 对一般统计思想也有很大的影响。 Fisher 认为统计学即是对总体、变异和数据简化的研究 [49]。 (iv) 1928-1938 年, 美籍波兰裔统计学家 Jerzy Neyman (1894-1981) 和 K. Pearson 之子、英国统计学家 Egon Sharpe Pearson (1895-1980) 创立了假设检验理论。 (v) 1934-1937 年, Neyman 建立了与 Neyman-Pearson 假设检验理论息息相关的置信区间估计理论。 (vi) 1925-1930 年, 英国统计学家 G. N. Yule (1871-1951) 奠定时间序列分析的基础。 (vii) 1928 年, 英国统计学家 John Wishart (1898-1956) 提出 Wishart 分布, 多元统计得以迅速发展。我国著名统计学家许宝騄于 1940 年前后对这一领域和线性模型的统计推断做出了奠基性的工作。 (viii) 美籍罗马尼亚裔统计学家 Abraham Wald (1902-1950) 于 1939 年开始发展统计决策理论, 引进了损失函数、风险函数、极小极大原则和最不利先验分布等重要概念。二战期间应军需品的检验工作而提出序贯概率比检验法并证明其最优性, 奠定了序贯分析的基础 [155]。 (ix) 1946 年, 瑞典统计学家 Harald Cramér (1893-1985) 发表著作《统计学数学方法》[29] 总结了当时数理统计学的成果, 标志着统计学走向成熟。

二十世纪下半叶

计算机和信息技术的发展对统计学产生了深远的影响, 它推动了统计学的应用发展, 这一时期的特点是注重实用效果。美国统计学家 B. Efron (1938-) 甚至说, “二十一世纪统计学中几乎所有的主题都是计算机相关的” [43]。

^{*}详见 C. R. Rao 的纪念文章《R. A. Fisher: 现代统计学的奠基人》[130]。



美国统计学家 J. Tukey (1915-2000) 在六十年代初就预测数据分析的未来是面向应用和计算的学问。七十年代末，Efron 的自助法 (bootstrap method) 和 Monte Carlo 方法借助计算机得以广泛应用于统计推断。九十年代，以色列统计学家 Y. Benjamini (1949-) 和 Y. Hochberg 提出错误发现率 (false-discovery rate, FDR)，即假设检验中第一类错误率的期望，并给出了在多重比较中对它的控制方法。

二十世纪末，统计学家开始关注机器学习、大数据分析等应用类数据科学。例如集成学习 (ensemble learning) 策略，美国统计学家 L. Breiman (1928-2005) 的自助聚集 (bootstrap aggregating) 方法和随机森林 (random forest) 成为机器学习的经典方法。还有以色列统计学家 Y. Freund 和 R. Schapire 的提升 (boosting) 方法，于 2003 年获得 Gödel 奖，也是集成学习的典型方法。



图 6.29: 统计学在国民经济生产、科学技术研究中有着广泛的应用。

(i) 在生物学、医学、金融数学、经济学、社会学以及工程技术上的应用越来越

普及，产生了一些新的应用分支，如生物统计、抽样检验、统计质量管理、排队论、库存论、可靠性与生存分析等。(ii) 非参数统计学的大样本理论得到发展，尤其是关于秩统计量和 U 统计量的大样本理论。(iii) 应小样本分析的需求，贝叶斯学派逐渐兴起，在很多具体应用上贝叶斯统计学（见第 12 章）已成为经典统计学的强有力竞争者。(iv) 随机模拟技术（见第 15 章）的发展令很多计算上的困难不复存在，一些复杂的抽样分布的推导变得不再需要，同时计算机处理海量数据的能力推动了理论模型的各种应用，也加剧了统计学中理论和应用逐渐分离的趋势。(v) 人工神经网络 [15]、模式识别 [39, 135]、机器学习 [16, 111]、数据挖掘等一些与数据分析和处理有关的边缘分支如雨后春笋般出现，它们模糊了统计学的边界。(vi) 各行各业对统计人员的需求越来越大，统计专业的教育和培训受到统计学发达国家的重视，甚至取得了与数学平起平坐的地位。

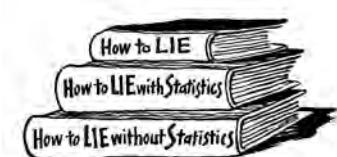
二十一世纪

统计学将深受大数据和人工智能的影响，向应用和计算的方向发展。

抽象地说，统计学透过数据研究这样的自然本质：对于输入变量 x ，自然本质这个黑箱有响应变量 y 输出。即，

$$x \rightarrow \boxed{\text{自然本质}} \rightarrow y$$

数据分析的目的无外乎（1）探索自然规律是如何把响应变量和输入变量关联起来的，（2）预测新的输入将会有怎样的输出。美国统计学家 Leo Breiman (1928-2005) 在论文 Statistical Modeling: The Two Cultures 里总结了两种不同的统计建模文化：一种是假设了数据的产生机制（如线性回归），另一种对数据产生的机制一无所知，把它当作黑箱来模拟（如决策树、神经网络）。经典统计学里第一种文化占主体，统计机器学习里第二种文化占主体。其实，这两种文化是互补的，共同构成数据科学 (data science) 的文化。



说到数学和统计的关系，很多人不把统计学列为数学的分支，而将之视为推断的艺术，原因是使用不同的统计方法、基于不同的观察数据有可能导致截然不同的结论——这是统计学固有的特点。也有人把统计学视为说谎的艺术，美国作家 Mark Twain 曾说过一句名言，“世上有三种谎言：谎言、该死的谎言和统计学。”他的话当然有些偏激，统计方法本身没有善恶，只有别有用心的使用者而没有说谎的统计学。

哲学家 Karl Popper (1902-1994) 认为“一个理论就是描述观测的数学模型。”统计理论可视作对观测数据进行数学建模 (mathematical modeling) 进而给出统计推断（包括预测、分析、解释等），其首要步骤是对问题的抽象。要得到合理的问题抽象并非易事，美国数学家、统计学家 J. W. Tukey (1915-2000) 曾说，“一个正确问题的近似解比一个似是而非问题的精确解更有价值得多。”意思是建模之重要甚于求解。建模是一门复杂的艺术，需要有一双看见本质的眼睛和足够多的经验积累。经

验包括对一些经典统计模型的了解和在真实数据分析上的实战。需要铭记的是统计学扎根于应用，靠实际效果来评价模型的优劣，而不是炫耀数学技巧搞一些华而不实的东西。

统计学经过百年的发展已经枝繁叶茂，本书第二部分仅是数理统计学的入门，对新近兴起的自助法 [30, 42]、刀切法 [144]、贝叶斯方法 [9, 10, 35, 55]、贝叶斯网络 [92] 等只能做蜻蜓点水式的介绍，这些有趣的内容大都未列入本科教学，感兴趣的读者只能去阅读相关的专著。为了更好地了解统计思想和方法，由易及难，向读者推荐以下几部专著作为课外读物。



- D. Freedman, R. Pisani, R. Purves, A. Adhikari 合著的《统计学》[51]（强调统计思想的入门书，像侦探小说一样引人入胜）和 J. L. Folks 的《统计思想》[50]（正如书名一样强调思想方法而不是技术细节）。
- L. Breiman 的《统计学：从应用的观点看》[21]，Breiman 是打通统计学和机器学习的杰出学者，他对统计学两种文化的见解 [23] 体现了实用主义统计学。
- R. A. Fisher 的《统计方法、试验设计和科学推断》[49]，是三本书的合订本，分别是《供研究人员用的统计方法》、《实验设计》、《统计方法与科学推断》。英国知名统计学家 Frank Yates (1902-1994) 为合订本写序，评价这位二十世纪最伟大的统计学家的工作。这部著作让我们重温数理统计初创的辉煌，不忘初心关注于统计的本质，即揭示数据背后的事，而不是单纯地玩弄数学技巧。
- G. Casella 和 R. L. Berger 的《统计推断》[24] 包含丰富的例子，预备知识仅需要数学分析和线性代数，适合初学者。
- P. J. Bickel 与 K. A. Doksum 合著的《数理统计——基本概念及专题》[11] 的内容非常紧凑。2001 年的新版较 1977 年的旧版改动很大，旧版更好一些。
- 我国著名统计学家陈希孺院士 (1934-2005) 的《高等数理统计学》[168] 是一部基于测度论的数理统计学基础教科书，有大量精心设计的习题，占了书的一半篇幅。

第七章

数理统计学的一些基本概念

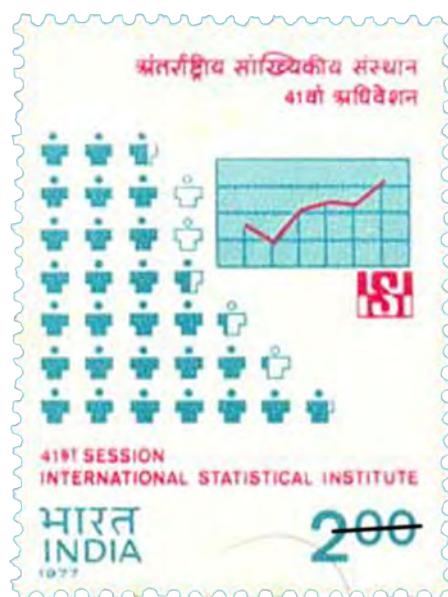
纸上得来终觉浅，绝知此事要躬行。

陆游《冬夜读书示子聿》

与所研究问题有关的全部个体的确定集合称作总体 (population)。若其中个体数目有限，则称之为有限总体。例如，调查一个班级学生的身高状况，总体就是这个班的所有学生；调研国民的年收入情况，总体包括所有国民。有时总体只是一种理想的存在，如测量给定地点某时刻的温度，总体就是所有可能的观测值，即实数集 \mathbb{R} ；再如，抛一枚硬币无穷多次所得结果的总体。像这种个体数目无穷多的总体被称为无限总体。如果有有限总体中个体数目足够大，也可近似地当作无限总体来处理，如某时间段内全球上网者的总体。

一元总体中的每个个体都可用度量总体某一属性的随机变量来描述。举个例子，如果关心上网者是否浏览股票信息，对每个上网者可联系一个 0-1 分布的随机变量 X ，若浏览股票信息则取 1，否则取 0。如果同时还对上网者的年龄特征感兴趣，就要用到两个随机变量来描述一个个体，这样的总体被称为二元总体。以此类推，也会有多元总体，多元统计学就是研究多元总体的统计学分支。

为了研究的方便，常把总体数量化，譬如与人均年收入问题有关的全部个体是一群人的集合，可以把这个集合简化为这群人的收入的集合。总体内各数值出现的可能性所形成的概率分布称为总体分布，例如，给定地点某时刻的温度测量值的总体分布就是以该时刻的真实温度为均值的正态分布 $N(\mu, \sigma^2)$ ，其中真实温度 μ 和



方差 σ^2 都是未知的。总体分布一旦完全明确，对统计学而言总体就是毫无神秘之处了，所以统计学仅对以下两种类型的总体分布感兴趣：

- 总体分布几乎是未知的，仅仅知道它是连续型的或离散型的，这种总体称为非参数总体。
- 总体分布 F 的数学形式已知，仅有若干参数 $\theta_1, \dots, \theta_k$ 未知，这样的总体被称为参数总体。本书重点考虑参数总体，其中正态总体因为相对常见和简单而被研究得较为透彻。

定义 7.1. 未知参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top$ 的所有可能取值称为参数空间，记作 Θ 。譬如已知总体是正态分布 $N(\mu, \sigma^2)$ ，但是参数 μ, σ^2 未知，参数空间是上半平面。

定义 7.2. 参数总体的所有可能分布的集合 $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ 称为一个分布族 (a family of distributions)^{*}。例如，正态分布族 $\mathcal{F} = \{\phi(x|\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$ ， k -参数指数族（见第 487 页的**定义 7.15**）等。

Fisher 在《供研究人员用的统计方法》说，“无关信息和相关信息之间的区别如下。即使在最简单的情况下，我们面对的数值（或数值集合）也被解释为在相同的情况下可能出现的所有数值构成的一个假想无限总体的一个随机样本。这个总体的分布能够有某种数学形式，公式中包含一定数目，通常是少数几个参数或“常数”。这些参数是这个总体的特征。如果我们能知道参数的精确值，我们就会知道来自这个总体的任何样本所能告诉我们的一切（甚至更多）。事实上我们无法精确地知道参数，但我们可以估计它们的值，这估计或多或少地不准确。这些估计，术语是统计量，当然是从观测算出来的。如果我们能为总体找到一个可充分表示观察数据的数学形式，然后从数据中计算出所需参数的最好的可能估计，那么数据将看起来只能告诉我们这么多；我们将从中提取所有可用的相关信息。”

在概率问题中，总体的分布是已知的；在统计问题中，总体的分布是未知的。在观察数据的基础上，统计的任务就是在圈定好的一些可能的总体分布中找出最合理的目标。要达到这个目标，必须回答下面的问题：

- 基于什么假设圈定这些分布？
- 模型的合理性体现在哪些方面？
- 如果假设不成立，会出现什么后果？
- 怎么衡量模型的效果？
- 还有其他的模型吗？

^{*}当谈到未知参数而无需强调它是单个参数还是向量参数时，我们也常用非粗体的小写字母来表示未知参数，分布族记作 $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ 。

- 与其他模型相比，你的模型有怎样的优点和缺点？
- 如何改进当前的模型？

不难看出，统计建模 (statistical modeling)^{*}是首要步骤，它就是在现实问题和统计理论之间构筑桥梁。统计建模是统计学家的日常工作，十之八九的精力耗费于此。譬如，可能的分布族圈小了会导致错误的结果，圈大了会增加搜索的难度，也有可能导致无意义的结果。统计建模既依赖于数学推理，也要靠一些不那么严谨的经验。书本里多强调前者，后者多是从数据分析的实战中磨练（譬如，探索性数据分析直观地考察数据的特点，见图 7.1）。Poincaré 说，“对发现来说，直觉比逻辑更重要。”统计发现亦是如此，所以统计既是一门科学，也是一门艺术。英国统计学家 George E. P. Box (1919-2013) 开玩笑说，“统计学家就像艺术家一样有个爱上他们的 models 的坏习惯。”



图 7.1: 柱状图 (bar chart, 或称条形图) 和折线图用高度表示重要性、比例、数量等，常用于直观比较大小、展示差异……。

参数模型和非参数模型所用的方法还是有很多差异的。前者关注于一类可能的分布，在参数空间里寻找一个最优化问题的解；后者所考虑的可能分布族的范围更广泛。

频率派认为参数是未知的固定值，参数估计有两种方式：一是点估计（即猜它等于多少），二是区间估计（即以某个概率被由样本构造的区间覆盖住）。与频率派不同，贝叶斯学派认为参数是随机变量，参数估计就是算该参数的后验分布。这两套理论无所谓对错，它们的基本理念是不同的，作为工具都有各自适合的应用场景，我们在后续的章节里会逐一给予介绍。

另外，对于有关总体分布的假设，基于观察样本统计学也有一些叫作“假设检验”的手段来评价它的真假。

^{*}John von Neumann 曾说，“科学不试图阐释，甚至几乎不试图说明，而只是构造模型。一个模型就是一个数学构造，稍加一点文字解释便能描述观察到的现象。对此数学构造合理性的证明恰好是该模型被期许做的事情，即正确解释一个相当广泛的领域中的现象。”

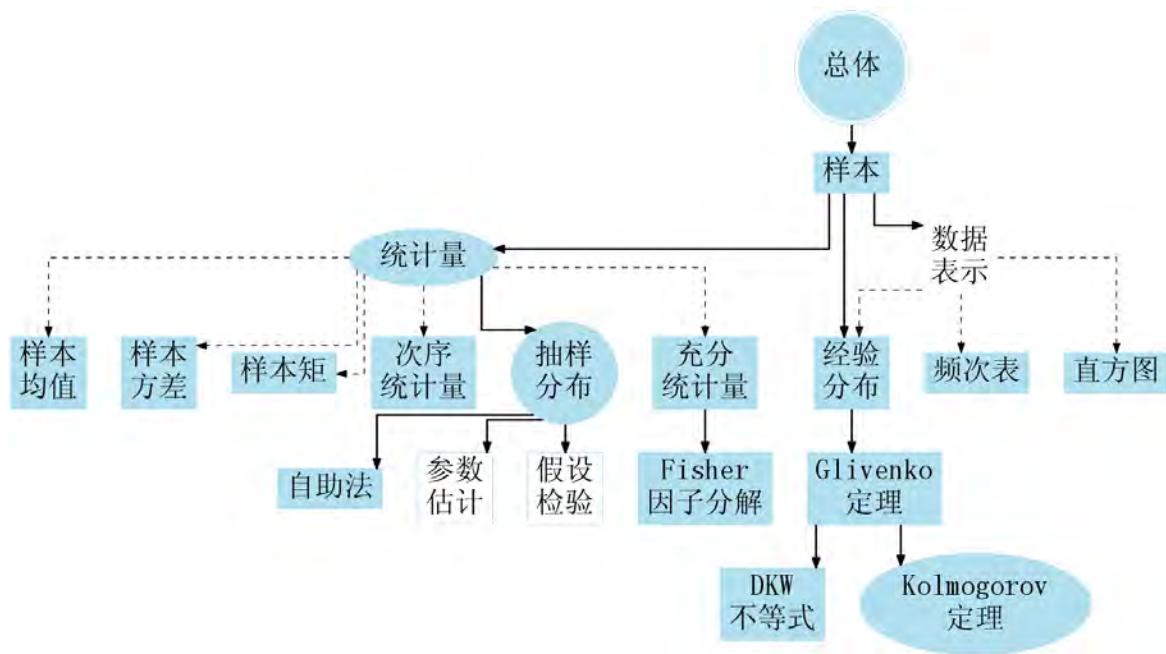
除了上述跟总体有关的统计学，我们经常还需发掘变量之间隐藏的函数关系。在输入变量 x 和响应变量 y 之间，我们假定有某个待定的函数关系可被参数化表示为 f_θ ，即

$$y = f_\theta(x) + \epsilon, \text{ 其中误差 } \epsilon \text{ 服从某个带参数的分布} \quad (7.1)$$

式 (7.1) 中的未知参数 θ 是待定的，它需要通过输入变量和响应变量的观测得到估计。频率派常把这个参数估计的问题转化为一个最优化的问题。

要深入了解这些统计分支，我们必须掌握一些基本的统计概念和它们的性质，如经验分布、统计量及其抽样分布等。我们还要专门研究 Fisher 提出的一类特殊的统计量——充分统计量的性质。

第七章的主要内容及其关系



7.1 样本的特征

将 总体中的每个个体逐一列举加以研究是不可行的或不经济的。为得到对总体的宏观了解，一个可行的策略是以一定的方式从总体中抽出若干个体（这些被抽取的个体称为样本点^{*}，它们构成的集合称为一个样本，其中所含个体的数目称为样本容量或样本量）进行考察，进而做出有关总体的结论。

譬如，调研去年省内人均年收入情况可根据职业比例，从工人、农民、个体商户、公司职员、政府机构公职人员等人群中随机抽取一定规模的样本，这样做的代价是结论可能因样本的差异而不同。根据样本获得正确的结论并指出其不可靠的范围是统计推断的主要研究内容之一。

对于数量化了的总体，样本的观察结果也是数值，称为样本值。由于抽样的随机性，在观察到样本值之前，频率派把每个样本点都视为一个服从总体分布的随机变量，所以样本量为 n 的样本就是一个由样本点构成的 n 维随机向量，记作 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ ，其分布称为样本分布，取决于总体分布、抽样方式[†]和样本量。



图 7.2: 探索性数据分析：用各式各样的直观方式去揭示数据的特点和背后的规律。

一方面，利用样本值可以直观地探索总体的特性，例如数据可视化，从而帮助我们画出总体的“素描”。每个数据分析人员都要有探索性数据分析的素质，计算机是很好的辅助工具。

另一方面，统计学更关注如何通过有限样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 尽可能准确地“猜出”总体分布的一般方法。所以，我们可以抽象地研究随机向量 \mathbf{X} 与总体分布之间的关系，甚至无需有具体的样本值。

允许样本量趋向无穷，即允许从总体中不断抽样的统计问题称为大样本问题，由此发展起来的大样本理论以概率论的极限理论为研究工具，以统计量的渐近性质及针对这些性质的统计方法为研究对象。大样本理论起源于 1900 年 K. Pearson 对

^{*}有时候在不引起歧义的情况下也把样本点简称为样本。如何获取样本是抽样调查和试验设计的任务，它们都是统计学的重要分支。

[†]有限总体的抽样有“有放回”和“无放回”之分。如果总体中个体的数目远大于样本量，无放回的抽样也可近似地看作有放回的抽样。

用于拟合优度检验的 χ^2 统计量渐近于 χ^2 分布的证明, 如今该理论已得到充分的发展 [102], 后续章节将介绍它的一些经典结果, 如最大似然估计、似然比检验等。

具有固定样本容量的统计问题称为小样本问题 (small sample problem), 大样本时的结论不再适用。小样本理论起源于 1908 年 W. S. Gosset 提出 t 分布并将之用于正态总体均值的小样本估计。请读者切记“大样本”和“小样本”并不是指样本容量的大小, 而是指是否允许样本量无限大。

定义 7.3 (统计量). 设 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 是从某总体抽得的样本, 若除了 \mathbf{X} 之外, Borel 可测函数 $T = T(\mathbf{X})$ 不依赖于其他任何未知量, 则称 T 为样本统计量或统计量^{*}(statistic)。简而言之, 统计量就是由样本构造的新的随机变量, 是为了反映了样本的某个特性而对样本的“深加工”。

为方便记述, 一般情况下统计量用大写字母 (如 T) 表示, 它的观察结果约定用相应的小写字母表示 (如 t)。常见的统计量包括

$$\text{样本均值: } \bar{X} = \frac{1}{n} \sum_{j=1}^n X_j, \text{ 有时记作 } \mu_n \quad (7.2)$$

$$\begin{aligned} \text{样本方差: } S^2 &= \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2, \text{ 有时记作 } \sigma_n^2 \\ &= \frac{1}{n-1} \left(\sum_{j=1}^n X_j^2 - n\bar{X}^2 \right) \end{aligned} \quad (7.3)$$

$$\text{样本 } k \text{ 阶矩: } A_k = \frac{1}{n} \sum_{j=1}^n X_j^k \quad (7.4)$$

$$\text{样本 } k \text{ 阶中心矩: } B_k = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^k, \text{ 其中 } B_2 \text{ 也常记作 } S_n^2 \quad (7.5)$$

性质 7.1. 样本均值 μ_k 和样本方差 σ_n^2 满足下面递归关系:

$$\begin{aligned} \mu_n &= \frac{1}{n} [X_n + (n-1)\mu_{n-1}] \\ &= \mu_{n-1} + \frac{1}{n}(X_n - \mu_{n-1}) \\ \sigma_n^2 &= \sigma_{n-1}^2 + \frac{1}{n-1} \left[\frac{n-1}{n} (X_n - \mu_{n-1})^2 - \sigma_{n-1}^2 \right] \end{aligned}$$

当 n 个样本不是一次性给定, 而是陆陆续续得到, 我们可以依据该性质不断更新样本均值和方差, 勿须重新计算。在线学习 (online learning) 的很多方法都是利用

*搞清楚统计量的分布是统计学的一个基本问题: $T(X_1, \dots, X_n)$ 的精确分布对小样本问题很重要, 而 $n \rightarrow \infty$ 时 $T(X_1, \dots, X_n)$ 的极限分布对大样本问题是至关重要的。

类似的递归关系去更新学习机，有效地降低计算复杂度。

练习 7.1. 请分别给出样本 k 阶矩 (7.4) 和 k 阶中心矩 (7.5) 的递归关系。

例 7.1. 若总体 $X \sim N(\mu, \sigma^2)$ 的期望 μ 已知，方差 σ^2 未知，则 $\bar{X} - \mu$ 是统计量，而 $\sum_{j=1}^n X_j / \sigma^2$ 不是统计量。

例 7.2. 对于任意实数 c ，样本方差 (7.3) 具有如下的分解。

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n \frac{1}{n^2} \left[n(X_i - c) - \sum_{j=1}^n (X_j - c) \right]^2 \\ &= \frac{1}{n^2(n-1)} \sum_{i=1}^n \left\{ n^2(X_i - c)^2 + \left[\sum_{j=1}^n (X_j - c) \right]^2 - 2n(X_i - c) \sum_{j=1}^n (X_j - c) \right\} \\ &= \frac{1}{n^2(n-1)} \left\{ n^2 \sum_{i=1}^n (X_i - c)^2 + n \left[\sum_{j=1}^n (X_j - c) \right]^2 - 2n \sum_{i=1}^n (X_i - c) \sum_{j=1}^n (X_j - c) \right\} \\ &= \frac{1}{n^2(n-1)} \left\{ (n^2 - n) \sum_{i=1}^n (X_i - c)^2 - 2n \sum_{i < j} (X_i - c)(X_j - c) \right\} \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - c)^2 - \frac{2}{n(n-1)} \sum_{i < j} (X_i - c)(X_j - c) \end{aligned}$$

练习 7.2. 请读者验证 $B_2 = \frac{n-1}{n} S^2 = A_2 - A_1^2$ 。

定义 7.4 (简单随机样本). 如果样本 X_1, X_2, \dots, X_n 独立同分布于总体分布 $X \sim F_\theta(x)$ ，则称样本 X_1, X_2, \dots, X_n 为独立同分布样本或简单随机样本，简记作

$$X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} F_\theta(x)$$

例如， $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ 表示简单随机样本来自正态总体。

性质 7.2. 设总体 X 的密度函数为 $f_\theta(x)$ ，则简单随机样本 X_1, X_2, \dots, X_n 的分布函数 $\hat{F}_\theta(x_1, x_2, \dots, x_n)$ 和密度函数 $\hat{f}_\theta(x_1, x_2, \dots, x_n)$ 分别为

$$\begin{aligned} \hat{F}_\theta(x_1, x_2, \dots, x_n) &= \prod_{j=1}^n F_\theta(x_j) \\ \hat{f}_\theta(x_1, x_2, \dots, x_n) &= \prod_{j=1}^n f_\theta(x_j) \end{aligned}$$

如果没有特殊声明，后文中所提的“样本”大多是指简单随机样本，即每个样本点都是从总体中独立抽得，对有限总体而言抽样要求是有放回的，这样才不至于改变总体的分布。

定义 7.5. 通过频次表（或频率表）、茎叶图（stem-and-leaf plot）、直方图等可对这些样本值有一个直观的了解。把样本值 x_1, x_2, \dots, x_n 中所有不同的值 $x_{(1)} < x_{(2)} < \dots < x_{(k)}$ 都列出来并标出它们出现的频次 n_1, n_2, \dots, n_k 或频率 $f_1 = n_1/n, f_2 = n_2/n, \dots, f_k = n_k/n$ （其中 $\sum_{j=1}^k n_j = n$ ），这样得到的如下列表称为频次表或频率表。

不同的样本值	$x_{(1)}$	$x_{(2)}$	\cdots	$x_{(k)}$
出现的频次	n_1	n_2	\cdots	n_k

不同的样本值	$x_{(1)}$	$x_{(2)}$	\cdots	$x_{(k)}$
出现的频率	f_1	f_2	\cdots	f_k

显然频次表是无损失的数据表示，频率表丢失了样本量的信息。若把 $x_{(1)}, x_{(2)}, \dots, x_{(k)}$ 按位数进行比较，将基本不变或变化不大的位作为“茎”，将变化大的位作为“叶”列在“茎”的后面，如此得到的图可以毫无损失地直观显示样本值，称为茎叶图。

例 7.3. Iris 数据是美国植物学家 Edgar Anderson (1897-1969) 在 1935 年收集的 150 组鸢尾花萼片和花瓣的长度与宽度，共分为三个类：setosa、virginica 和 versicolor，每个类都包含 50 组数据。1936 年，Fisher 在一篇有关判别分析的论文中使用了该数据而使之成为多元统计方法的一个公开测试数据，并被冠以“Fisher 的 Iris 数据”。考虑 setosa 类的花瓣长度的样本值，分别给出它们的频次表和茎叶图。



样本值	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.9
频次	1	1	2	7	13	13	7	4	2

10 | 0
 11 | 0
 12 | 00
 13 | 0000000
 14 | 00000000000000
 15 | 00000000000000
 16 | 0000000
 17 | 0000
 18 |
 19 | 00

直方图是直观显示数据聚散情况的最常见方法，已下放至中小学数学教育。第 63 页的例 1.48 曾利用直方图来揭示频率和概率的关系。绘制直方图最关键的步骤是区间的划分，区间个数可由用户指定。

茎叶图保留了（样本值）数据的原始信息而且便于更新。如果用矩形面积表示观测数据落于矩形底边区间的百分数，直方图中所有矩形的面积之和就等于 1，这样的直方图对了解连续型总体的密度函数很有用^{*}。但是，直方图丢失了数据的很多原始信息，更新起来也不太容易。



图 7.3: 直方图用面积而非柱高表示数量，这是它有别于条形图之处。直方图描述了德国人口的年龄结构（不难看出老龄化的趋势）。

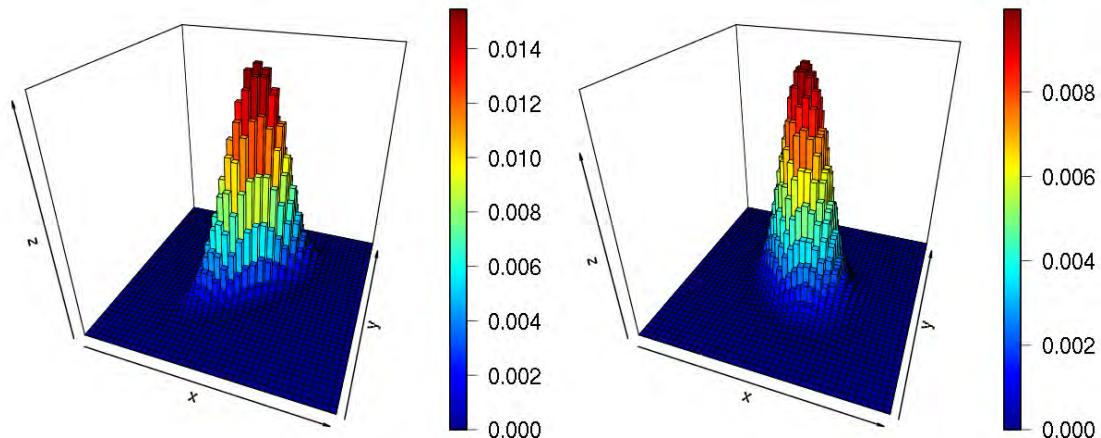


图 7.4: 三维直方图：分别从二元正态分布 $N(0, 0, 1, 4, 0.8)$ 和 $N(0, 0, 1, 4, -0.4)$ 抽取 $n = 10^6$ 个随机数，利用计算机绘制三维直方图。在每个三维直方图里，所有小长方体的体积之和等于 1。

本节内容

第一小节讨论经验分布及其性质，Glivenko 定理确保了经验分布可以任意地接近总体分布，DKW 不等式刻画了收敛速度，Kolmogorov 定理和 Rényi

^{*}在实践中，往往要通过核密度估计的方法得到密度函数，感兴趣的读者可参阅 [146]。

定理则揭示了二者接近程度的极限分布。第二小节利用大数律和中心极限定理说明样本矩依概率收敛于相应的总体矩，并给出了样本矩的极限分布。

关键知识

- (1) 统计量；(2) 数据表示的常见方法，如茎叶图、直方图等；(3) 经验分布函数的基本性质；(4) Glivenko 定理、Kolmogorov 定理和 DKW 不等式；(5) 样本矩的基本性质。

7.1.1 次序统计量

令函数 $h_j(x_1, x_2, \dots, x_n)$ 给出实数 x_1, x_2, \dots, x_n 按升序排列后第 j 个位置的数, 记作 $x_{(j)} = h_j(x_1, x_2, \dots, x_n)$, 其中 $1 \leq j \leq n$ 。显然, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 。

定义 7.6 (次序统计量). 由样本空间 (Ω, \mathcal{S}) 上的样本 X_1, X_2, \dots, X_n 定义一个新的随机变量 $X_{(j)}$, 满足以下条件

$$X_{(j)}(\omega) = h_j(X_1(\omega), X_2(\omega), \dots, X_n(\omega)), \text{ 其中 } \omega \in \Omega \quad (7.6)$$

$X_{(j)}$ 被称为第 j 个次序统计量 (order statistic)。其中, 统计量 $X_{(1)}$ 和 $X_{(n)}$ 被称为极值, 二者之差 $X_{(n)} - X_{(1)}$ 被称为极差 (range)。次序统计量在非参数统计学 [72, 100] 中是最基本的概念。

显然, 作为函数, $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ 。即, 如果 $i < j$, 则 $\forall \omega \in \Omega$ 有 $X_{(i)}(\omega) \leq X_{(j)}(\omega)$ 。因为样本 X_1, X_2, \dots, X_n 是定义在样本空间 (Ω, \mathcal{S}) 上的可测函数, 我们也常把 $X_{(1)}$ 记作 $\min(X_1, X_2, \dots, X_n)$, 把 $X_{(n)}$ 记作 $\max(X_1, X_2, \dots, X_n)$, 请读者按照式 (7.6) 来理解。

例 7.4. 接着**例 2.3**, 如果样本 X_1, X_2 满足 $X_1(k) = k, X_2(k) = 7-k$, 其中 $k = 1, 2, \dots, 6$ 。按照**定义 7.6**, 次序统计量 $X_{(1)}$ 和 $X_{(2)}$ 分别为

$$\begin{aligned} X_{(1)}(k) &= \begin{cases} k & \text{当 } k = 1, 2, 3 \\ 7-k & \text{当 } k = 4, 5, 6 \end{cases} \\ X_{(2)}(k) &= \begin{cases} 7-k & \text{当 } k = 1, 2, 3 \\ k & \text{当 } k = 4, 5, 6 \end{cases} \end{aligned}$$

性质 7.3. 设总体分布函数为 $F_X(x)$, 密度函数为 $f_X(x)$, 记第 i 个次序统计量 $X_{(i)}$ 的密度函数为 $f_{X_{(i)}}(x)$, 则

$$\begin{aligned} f_{X_{(i)}}(x) &= \frac{n!f_X(x)}{(i-1)!(n-i)!}[F_X(x)]^{i-1}[1-F_X(x)]^{n-i} \\ f_{X_{(i)}, X_{(j)}}(x, x') &= \frac{n!f_X(x)f_X(x')}{(i-1)!(j-i-1)!(n-j)!}[F_X(x)]^{i-1}[F_X(x')-F_X(x)]^{j-i-1}[1-F_X(x')]^{n-j} \end{aligned}$$

其中, $i < j, x \leq x'$

$$f_{X_{(1)}, \dots, X_{(n)}}(x_1, \dots, x_n) = n! \prod_{i=1}^n f_X(x_i)$$

性质 7.4. 已知样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} U[0, 1]$, 令 $X^{(1)} \leq X^{(2)} \leq \dots \leq X^{(n)}$ 是次序统计

量, 则对于 $1 \leq m_1 < m_2 < \cdots < m_{k-1}$,

$$\begin{aligned} & (X^{(m_1)}, X^{(m_2)} - X^{(m_1)}, \dots, X^{(m_{k-1})} - X^{(m_{k-2})}, 1 - X^{(m_k)})^\top \\ & \sim \text{Dirichlet}(m_1, m_2 - m_1, \dots, m_{k-1} - m_{k-2}, n - m_{k-1}) \end{aligned}$$

定义 7.7. 对于样本 X_1, X_2, \dots, X_n , 基于次序统计量 $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, 可以定义一个新的统计量——样本中位数 M 如下,

$$M = M(X_1, X_2, \dots, X_n) = \begin{cases} X_{(\frac{n+1}{2})} & \text{若 } n \text{ 为奇数} \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}] & \text{若 } n \text{ 为偶数} \end{cases}$$

样本值 x_1, \dots, x_n 是样本 X_1, X_2, \dots, X_n 的具体观察结果, 而这些样本值的均值 $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$ 和方差 $s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$ 则分别是样本均值 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ 和样本方差 $S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2$ 的观察结果; $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ 则分别是次序统计量 $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ 的观察结果。

定义 7.8. 设 X_1, X_2, \dots, X_n 是来自总体 $X \sim F(x)$ 的简单随机样本, 其中分布函数 $F(x)$ 连续。对任意 $j = 1, 2, \dots, n$, 由次序统计量 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ 来定义新的随机变量 R_j , 称为 X_j 的秩 (rank)。

$$R_j(\omega) = r \text{ 当且仅当 } X_j(\omega) = X_{(r)}(\omega), \text{ 其中 } \omega \in \Omega$$

由于 $F(x)$ 连续, R_j 不是唯一确定的概率为零。我们把随机向量 $\mathbf{R} = (R_1, R_2, \dots, R_n)^\top$ 称为秩统计量 (rank statistic)。显然, 对任意 $\omega \in \Omega$, $R_1(\omega), R_2(\omega), \dots, R_n(\omega)$ 都是 $1, 2, \dots, n$ 的某一排列。

R 语言中函数 sort 给出 x_1, x_2, \dots, x_n 的升序排列 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, 函数 rank 依次给出 x_1, x_2, \dots, x_n 在 $x_{(1)} \leq x_2 \leq \cdots \leq x_{(n)}$ 中的位置。

练习 7.3. 接着例 7.4, 请给出秩统计量 $(R_1, R_2)^\top$ 。

1948 年, 美籍芬兰裔统计学家, 非参数统计学的奠基者之一 Wassily Hoeffding (1914-1991) 提出了一类统计量—— U 统计量, 此概念在参数估计和非参数统计学中都十分重要。例如, 通过构造合适的 U 统计量, 能找到未知参数的最小方差无偏估计 (见 §8.1.3)。前面介绍的某些统计量也是 U 统计量。

定义 7.9 (U 统计量). 基于样本 X_1, X_2, \dots, X_n 和 m 元实值函数 h , 其中 $1 \leq m \leq n$, 来构造 U 统计量如下。



$$U_n = \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{i_1, \dots, i_m} h(X_{i_1}, \dots, X_{i_m}) \quad (7.7)$$

其中，上式中的求和项表示从 $\{1, 2, \dots, n\}$ 中任选出 m 个不同整数 i_1, \dots, i_m ，将所有可能的 $h(X_{i_1}, \dots, X_{i_m})$ 求和。函数 h 称为统计量 U_n 的核。特别地，当 h 是一个对称函数时，式 (7.7) 可简化为

$$U_n = \frac{1}{C_n^m} \sum_{1 \leq i_1 < \dots < i_m \leq n} h(X_{i_1}, \dots, X_{i_m}) \quad (7.8)$$

例 7.5. 若 $h(x) = x$ ，则 U 统计量 $U_n = \frac{1}{n} \sum_{j=1}^n X_j = \bar{X}$ 恰是样本均值。

例 7.6. 令 $n \geq 2$ ，若 $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$ ，则 h 是二元对称函数，由式 (7.8) 和**例 7.2** 的结果，

$$\begin{aligned} U_n &= \frac{1}{2C_n^2} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 \\ &= \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (X_i - c + c - X_j)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - c)^2 - \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} (X_i - c)(X_j - c) \\ &= S^2 (\text{即样本方差}) \end{aligned}$$

练习 7.4. 令 $n \geq 2$ ，若 $h(x_1, x_2) = x_1^2 - x_1 x_2$ ，请基于样本 X_1, X_2, \dots, X_n 和 h 构造 U 统计量。答案：

$$U_n = S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2$$

7.1.2 经验分布及其性质

德国哲学家 Immanuel Kant (1724-1804) 在《纯粹理性批判》(1781) 中提出这样的哲学目标：在认识之前必须首先确定认识能力，只有这样才能开始认识。譬如，数学问题的解的存在性总是走在求解之前的。

通过简单随机样本 X_1, X_2, \dots, X_n 对总体 $X \sim F(x)$ 进行研究，无非是为了搞清楚 X 的分布 $F(x)$ ，理论上有没有这个可能呢？在数学上，本节即将介绍的 Glivenko 定理保证了从样本到总体分布有一条通途，它便是经验分布。

样本 → 经验分布 → 总体分布

定义 7.10 (经验分布函数). 简单随机样本 X_1, X_2, \dots, X_n 的经验累积分布函数 (empirical cumulative distribution function, ECDF)，简称经验分布函数，定义为

$$\begin{aligned}\hat{F}_n(x) &= \frac{1}{n} \#\{X_j \leq x : j = 1, 2, \dots, n\} \\ &= \frac{1}{n} \sum_{j=1}^n J(x - X_j)\end{aligned}\tag{7.9}$$

其中， $\#\{X_j \leq x : j = 1, 2, \dots, n\}$ 表示 X_1, X_2, \dots, X_n 中不超过 x 的个数， $J(\cdot)$ 为式 (2.10) 定义的非负判定函数。

高维简单随机样本的经验累积分布函数的定义是类似的。譬如，二维简单随机样本 $(X_j, Y_j)^\top, j = 1, 2, \dots, n$ 的累积经验分布函数定义为

$$\hat{F}_n(x, y) = \frac{1}{n} \#\{X_j \leq x \text{ 且 } Y_j \leq y : j = 1, 2, \dots, n\}$$

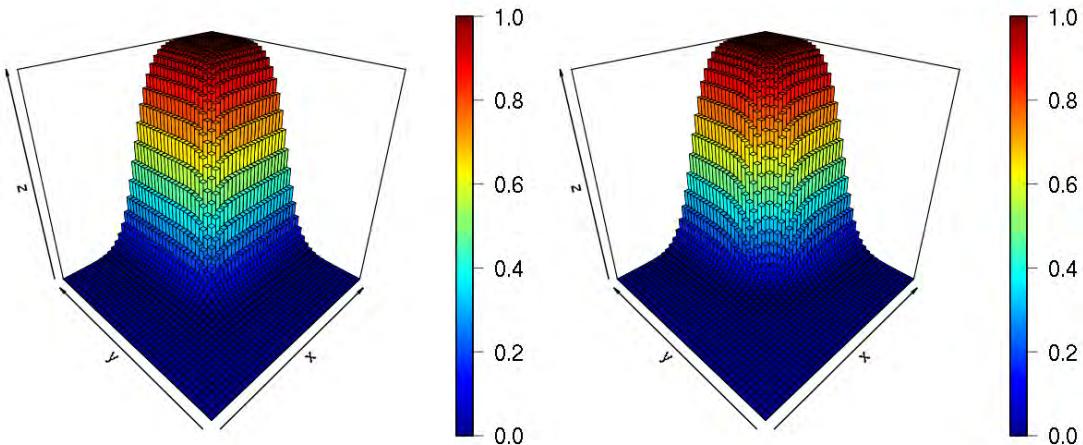


图 7.5: 分别从二元正态分布 $N(0, 0, 1, 4, 0.8)$ 和 $N(0, 0, 1, 4, -0.4)$ 抽取 $n = 10^6$ 个随机数，其经验累积分布函数如图所示。

性质 7.5. 总体分布函数 $F(x)$ 与经验分布函数 $\hat{F}_n(x)$ 有如下关系:

$$P\left\{\hat{F}_n(x) = \frac{k}{n}\right\} = C_n^k [F(x)]^k [1 - F(x)]^{n-k} \quad (7.10)$$

$$\hat{F}_n(x) \xrightarrow{P} F(x) \quad (7.11)$$

$$\frac{\sqrt{n}[\hat{F}_n(x) - F(x)]}{\sqrt{F(x)[1 - F(x)]}} \xrightarrow{L} N(0, 1) \quad (7.12)$$

证明. 显然, 对于每个固定的 x , 经验分布函数 $\hat{F}_n(x)$ 都是由样本 X_1, X_2, \dots, X_n 定义的一个统计量。并且, 随机变量 $J_j = J(x - X_j), j = 1, 2, \dots, n$ 独立同分布于 0-1 分布 $(1 - F(x))\langle 0 \rangle + F(x)\langle 1 \rangle$, 这是因为

$$P\{J_j = 1\} = P(x - X_j \geq 0) = F(x)$$

$$P\{J_j = 0\} = P(x - X_j < 0) = 1 - F(x)$$

由式 (7.9) 可看出 $n\hat{F}_n(x) \sim B(n, F(x))$, 结果 (7.10) 得证。由弱大数律和中心极限定理, 可证得结果 (7.11) 和结果 (7.12)。□

性质 7.6. 令 $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ 为样本 X_1, X_2, \dots, X_n 的次序统计量, 则经验分布函数 $\hat{F}_n(x)$ 可按如下方式构造:

$$\hat{F}_n(x) = \begin{cases} 0 & \text{当 } x < X_{(1)} \\ k/n & \text{当 } X_{(k)} \leq x < X_{(k+1)} \\ 1 & \text{当 } x \geq X_{(n)} \end{cases}$$

练习 7.5. 从某报刊中随机地抽出 5 篇文章, 统计每篇文章中的拼写错误数, 测得样本值为 0, 3, 2, 1, 1, 请写出经验分布函数。答案:

$$\hat{F}_5(x) = \begin{cases} 0 & \text{当 } x < 0 \\ 0.2 & \text{当 } 0 \leq x < 1 \\ 0.6 & \text{当 } 1 \leq x < 2 \\ 0.8 & \text{当 } 2 \leq x < 3 \\ 1 & \text{当 } x \geq 3 \end{cases}$$

例 7.7. 从正态总体 $N(0, 1)$ 抽取样本 x_1, \dots, x_n , 样本量分别为 $n = 10, 10^2, 10^3$ 。基于具体的样本值, 相对应的的经验分布函数 $\hat{F}_n(x)$ 如下图所示。不难发现, 样本量越大, 经验分布函数越接近总体分布。

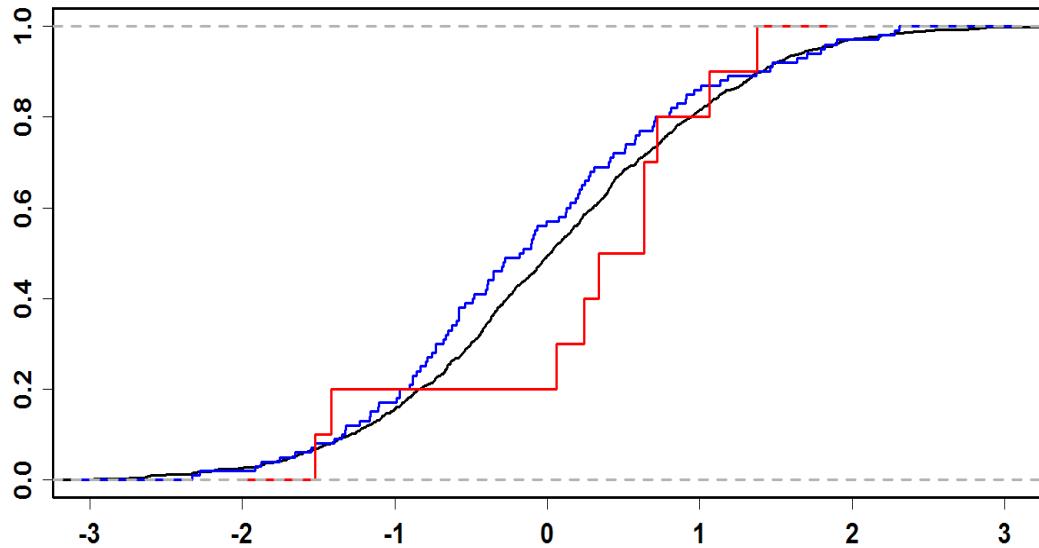


图 7.6: 经验分布函数是一个阶梯函数, 样本量越大越有可能接近总体分布。

算法 7.1. 由连续型随机变量 $X \sim F(x)$ 的简单随机样本构造经验分布函数 $\hat{F}_n(x)$, 不妨设其间断点为 $x_1 < x_2 < \dots < x_m$ 。显然, x_1, x_m 分别是样本里的最小值和最大值。根据第 280 页的算法 4.10, X 的随机数 x_* 可按下面的方法产生: 首先, 产生 $U[0, 1]$ 的随机数 y_* 。

- 若 $\hat{F}_n(x_1) \leq y_* < 1$, 寻找 i 使之满足 $y_* \in [\hat{F}_n(x_i), \hat{F}_n(x_{i+1})]$, 产生 $U[x_i, x_{i+1}]$ 的随机数 x_* 即为所求。
- 若 $y_* < \hat{F}_n(x_1)$, 则令 $x_* = x_1$; 若 $y_* = 1$, 则令 $x_* = x_m$ 。

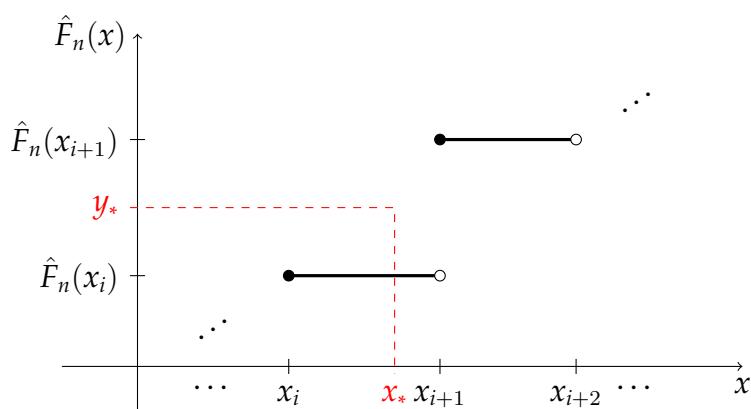


图 7.7: 当 n 很大时, $\hat{F}_n(x)$ 的跳跃很小, 它和 $F(x)$ 很接近。按照算法 7.1 产生的随机数可近似视为 $F(x)$ 的随机数。

例 7.8. 抽取 $X \sim 0.4N(-3, 1) + 0.6N(2, 0.64)$ 的 s 个样本，构造经验分布函数 $\hat{F}_s(x)$ 。按照算法 7.1 产生随机数 n 个，考察它们的经验分布函数 $\tilde{F}_n(x)$ 与 $F(x)$ 之间的 Kolmogorov 距离。

一个 Kolmogorov 距离说明不了什么，把上述随机试验重复 1000 次后，从这 1000 个 Kolmogorov 距离的分布不难看出： s, n 都很大时，算法 7.1 还是靠谱的。

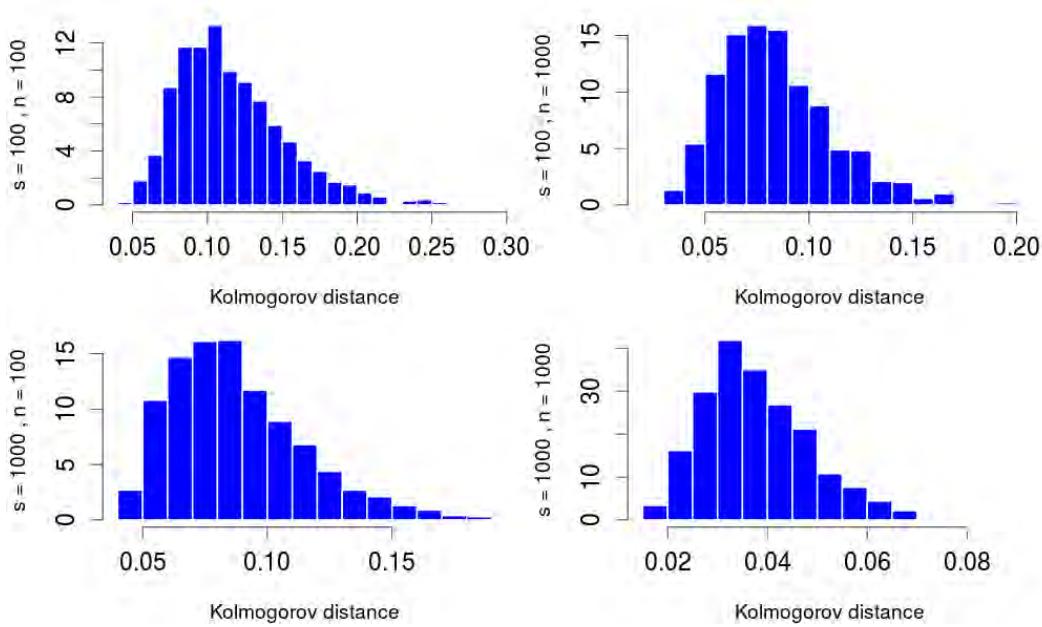


图 7.8: 例 7.8 中 Kolmogorov 距离的直方图： s, n 越大，越有可能取到小距离。

练习 7.6. 如果总体是离散型随机变量 X ，基于样本 x_1, \dots, x_n 如何抽样？

提示：参考第 258 页的算法 4.3。

例 7.9. 令 $\hat{F}_n(x)$ 是从样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F(x)$ 得到的经验分布函数，要使得 $\forall x \in \mathbb{R}$ 皆有 $P\{|\hat{F}_n(x) - F(x)| \geq \epsilon\} \leq \eta$ ，样本量 n 至少该多大？

解. 由式 (7.12)，当 n 很大时有

$$P\left\{\frac{\sqrt{n}|\hat{F}_n(x) - F(x)|}{\sqrt{F(x)[1 - F(x)]}} \leq z\right\} \approx 2\Phi(z) - 1$$

又因为 $F(x)[1 - F(x)] \leq 1/4$ ，所以

$$P\left\{|\hat{F}_n(x) - F(x)| \leq \frac{z}{2\sqrt{n}}\right\} \geq 2\Phi(z) - 1$$

或者等价地，

$$P\left\{|\hat{F}_n(x) - F(x)| \geq \frac{z}{2\sqrt{n}}\right\} \leq 2 - 2\Phi(z)$$

根据条件，从 $2 - 2\Phi(z) = \eta$ 解得 $z = z_*$ 。再从 $z_*/(2\sqrt{n}) = \epsilon$ 解得，

$$n = \left\lceil \frac{z_*^2}{4\epsilon^2} \right\rceil \quad (7.13)$$

表 7.1: 为满足 $P\{|\hat{F}_n(x) - F(x)| \geq \epsilon\} \leq \eta$ ，根据式 (7.13) 算得所需的样本量。

η	0.1	0.05	0.01	0.005	0.001
0.05	97	385	9604	38415	960365
0.01	166	664	16588	66349	1658725
0.005	197	788	19699	78795	1969860
0.001	271	1083	27069	108276	2706892

1933 年，苏联数学家 Valery Ivanovich Glivenko (1897-1940) 得到一个比式 (7.11) 更强的关键结果，被誉为“统计学基本定理”，它以概率 1 确保了只要样本量足够地大，经验分布 $\hat{F}_n(x)$ 就能以任何要求的精度逼近总体分布 $F(x)$ 。

~定理 7.1 (Glivenko, 1933). 设样本 X_1, X_2, \dots, X_n 来自分布函数为 $F(x)$ 的总体，我们约定经验分布函数 $\hat{F}_n(x)$ 与总体分布函数 $F(x)$ 的接近程度用 Kolmogorov 距离 $D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$ 来度量，则

$$P\left\{\lim_{n \rightarrow \infty} D_n = 0\right\} = 1$$

※证明. 见王梓坤的《概率论基础及其应用》第五章第一节。 □

Glivenko 定理只是说以概率 1 经验分布函数一致收敛于总体分布函数，并未揭示二者的 Kolmogorov 距离 $D_n \leq \epsilon$ 能以多大的概率发生。1933 年，Kolmogorov 发表著名的短文《论分布函数的经验测定》，给出了统计量 D_n 的极限分布，从而在大样本的情况下完美地解决了该问题*。Kolmogorov 定理 7.2 的一个应用是拟合优度的 Kolmogorov 检验，并导致了 Smirnov 检验的诞生（详见 §9.2.1）。

~定理 7.2 (Kolmogorov, 1933). 如果总体 X 的分布函数 $F(x)$ 是连续的，则有

$$\lim_{n \rightarrow \infty} P\{\sqrt{n}D_n \leq z\} = K(z)$$

*该论文被收录在《统计学中的重大突破》第二卷 [97] 中。

其中, Kolmogorov 分布函数 $K(z)$ 定义为 (见图 7.9)

$$K(z) = \begin{cases} 0 & \text{当 } z \leq 0 \\ \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 z^2) & \text{当 } z > 0 \end{cases} \quad (7.14)$$

定理 7.2 揭示了 D_n 的极限分布与总体分布 $F(x)$ 无关, 这使得在大样本前提下, 即便总体分布完全未知, 我们依然能给出合理的估计方法和实验设计方法。

表 7.2: Kolmogorov 分布函数 $K(z)$ 的取值。

z	$K(z)$	z	$K(z)$	z	$K(z)$
0.0	0.0000	1.0	0.7300	2.0	0.99932
0.2	0.0000	1.2	0.8877	2.2	0.99986
0.4	0.0028	1.4	0.9603	2.4	0.999973
0.6	0.1357	1.6	0.9880	2.6	0.9999964
0.8	0.4558	1.8	0.9969	2.8	0.99999966

Kolmogorov **定理 7.2** 考虑的是绝对误差 $|\hat{F}_n(x) - F(x)|$, 1953 年匈牙利数学家 A. Rényi (1921-1970) 研究了相对误差

$$D_{n,p} = \sup_{\substack{F(x) \geq p \\ F(p) > 0}} \frac{|\hat{F}_n(x) - F(x)|}{F(x)} \quad (7.15)$$

的渐近规律, 得到了下面的结果。

定理 7.3 (Rényi 定理, 1953). $\forall p \in (0, 1)$, 相对误差 (7.15) 渐近服从 Rényi 分布, 即

$$\lim_{n \rightarrow \infty} P\{\sqrt{n}D_{n,p} \leq z\} = R\left(z \sqrt{\frac{p}{1-p}}\right)$$

其中, Rényi 分布函数 $R(z)$ 定义为 (见图 7.9)

$$R(z) = \begin{cases} 0 & \text{当 } z \leq 0 \\ \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{-\frac{(2k+1)^2 \pi^2}{8z^2}\right\} & \text{当 } z > 0 \end{cases}$$

当 $z \geq 2$ 时, $R(z) \approx 4\Phi(z) - 3$ 的精度在 4×10^{-9} 以内。

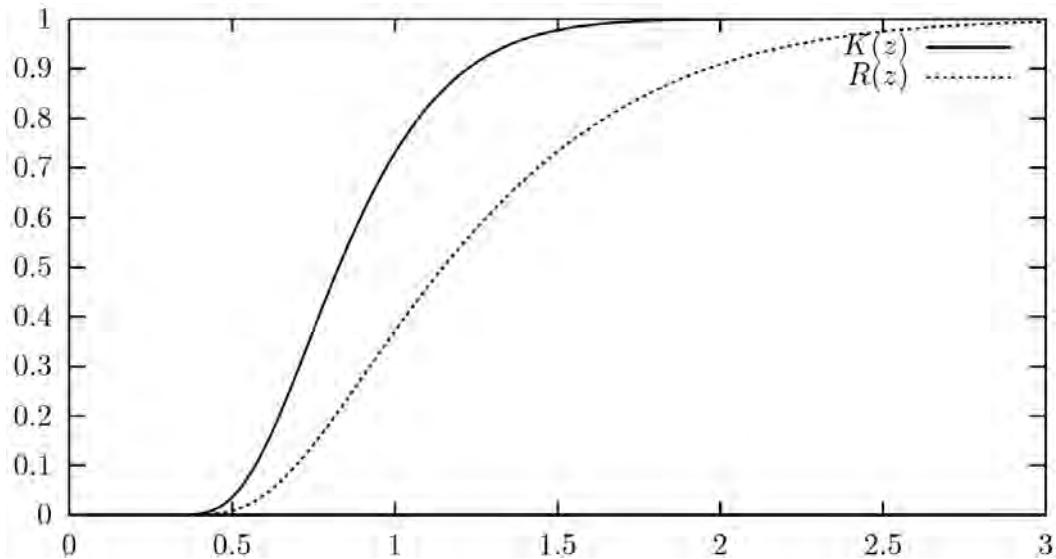


图 7.9: Kolmogorov 分布函数 (图中实线) 与 Rényi 分布函数 (图中虚线)。

1956 年, 以色列数学家 Aryeh Dvoretzky (1916-2008) 和美国统计学家 Jack Kiefer (1924-1981)、Jacob Wolfowitz (1910-1981) 在一般条件下, 对 $n \rightarrow \infty$ 时 $P\{D_n > \epsilon\}$ 收敛于零的速度进行了粗略的描述, 即下面的有关统计量 D_n 的小样本性质。

\nwarrow 定理 7.4 (DKW 不等式, 1956). 对于任意 $\epsilon > 0$, Kolmogorov 距离 D_n 满足

$$P\{D_n > \epsilon\} \leq 2 \exp\{-2n\epsilon^2\} \quad (7.16)$$

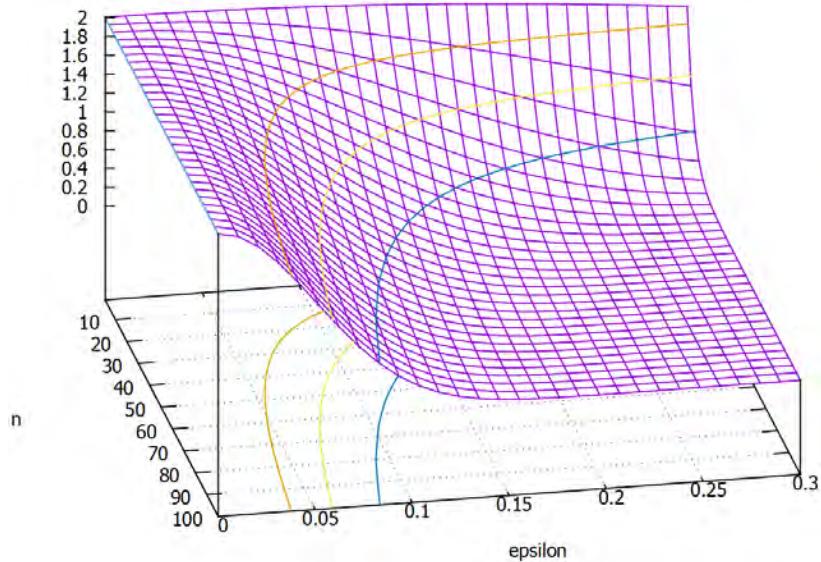


图 7.10: 曲面 $2 \exp\{-2n\epsilon^2\}$, 其中 $\epsilon \in (0, 0.3]$, $n = 1, 2, \dots, 100$ 。

练习 7.7. 请利用 DKW 不等式来解决例 7.9 的问题，并比较哪个结果更精细。

$$n = \left\lceil \frac{\ln 2 - \ln \eta}{2\epsilon^2} \right\rceil$$

表 7.3: 为满足 $P\{D_n > \epsilon\} \leq \eta = 2 \exp\{-2n\epsilon^2\}$, 根据式 (7.16) 算得所需的样本量。

η	ϵ	0.1	0.05	0.01	0.005	0.001
0.05		185	738	18445	73778	1844440
0.01		265	1060	26492	105967	2649159
0.005		300	1199	29958	119830	2995733
0.001		381	1521	38005	152019	3800452

7.1.3 样本矩及其极限分布

在测量长度、重量、容积等实践中，难以避免随机误差，不妨设测量值 $X \sim N(\mu, \sigma^2)$ ，其中 μ 是所测的真实值， σ^2 刻画了每次测量的精度，它们对于测量者来说都是未知的。为获得令人满意的测量值，人们往往独立地进行多次测量，测量值 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ ，只要测量次数 n 足够地多，以样本均值 $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ 为测量结果就能满足预定的误差要求。早在十七世纪，人们已经懂得利用样本均值来估算真实值 μ ，但是这样做的理论依据何在？

更宽泛地，人们关注在仅仅知道某些总体矩的存在性的前提下，样本矩与总体矩（譬如，样本均值 \bar{X} 与总体期望 μ ）之间有什么样的关系？如果能从样本矩中探索出总体分布的数字特征，如同给总体素描，将有助于了解总体分布（详见 §8.1.5 参数点估计的矩方法）。

$$\begin{array}{ccc} \text{总体} & \xrightarrow{?} & \text{数字特征} \\ \downarrow & & \uparrow \\ \text{样本} & \longrightarrow & \text{样本矩} \end{array}$$

性质 7.7. 令简单随机样本 X_1, \dots, X_n 来自的总体 X 具有期望 $E(X) = \mu$ ，方差 $V(X) = \sigma^2$ ， k 阶矩 $E(X^k) = m_k$ 和 k 阶中心矩 $E(X - \mu)^k = \mu_k$ ，则

$$E(\bar{X}) = \mu \text{ 且 } V(\bar{X}) = \frac{\sigma^2}{n} \quad (7.17)$$

$$E(S^2) = \sigma^2 \text{ 且 } V(S^2) = \frac{\mu_4}{n} + \frac{3-n}{n(n-1)}\mu_2^2 \quad (7.18)$$

$$A_k = \frac{1}{n} \sum_{j=1}^n X_j^k \xrightarrow{a.s.} m_k \text{ 且样本量足够大时，渐近地有}$$

$$A_k \sim N\left(m_k, \frac{m_{2k} - m_k^2}{n}\right) \quad (7.19)$$

证明. 结果 (7.17) 很显然，其证明留给读者练习。往证结果 (7.18)：由例 7.2，样本方差 S^2 具有下面的分解。

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n(n-1)} \sum_{i < j} (X_i - \mu)(X_j - \mu) \end{aligned}$$

立即可得 $E(S^2) = \mu_2 = \sigma^2$ 。下面求解 $V(S^2)$:

$$\begin{aligned} V(S^2) &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n(n-1)} \sum_{i<j} (X_i - \mu)(X_j - \mu)\right]^2 - \mu_2^2 \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n (X_i - \mu)^2\right]^2 + \frac{4}{n^2(n-1)^2} E\left[\sum_{i<j} (X_i - \mu)(X_j - \mu)\right]^2 - \mu_2^2 \\ &= \frac{\mu_4}{n} + \frac{n-1}{n} \mu_2^2 + \frac{2}{n(n-1)} \mu_2^2 - \mu_2^2 \\ &= \frac{\mu_4}{n} + \frac{3-n}{n(n-1)} \mu_2^2 \end{aligned}$$

由 Kolmogorov 强大数律（第 363 页的定理 5.12）证得结果 (7.19) 的前半部分。又因为 X_1^k, \dots, X_n^k 是独立同分布的且 $V(X_1^k) = m_{2k} - m_k^2$ ，由 Lindeberg-Lévy 中心极限定理 5.17 证得结果 (7.19) 的后半部分。□

练习 7.8. 在性质 7.7 的条件下，样本均值 \bar{X} 渐近服从于 $N(\mu, \sigma^2/n)$ 。

例 7.10. 样本 X_1, \dots, X_{100} 来自总体 $0.3\langle 1 \rangle + 0.7\langle 0 \rangle$ ，求 $P(|\bar{X} - 0.3| \leq 0.02)$ 。

解. $m_1 = m_2 = 0.3$ ，于是 $\sqrt{(m_2 - m_1^2)/100} \approx 0.0458$ 。利用结果 (7.19) 有

$$\begin{aligned} P(|\bar{X} - 0.3| \leq 0.02) &= P\left(\frac{|\bar{X} - 0.3|}{0.0458} \leq 0.44\right) \\ &= 2\Phi(0.44) - 1 \approx 0.34 \end{aligned}$$

练习 7.9. 验证样本二阶中心矩为 $B_2 = \frac{n-1}{n} S^2 = A_2 - A_1^2$ 。

定义 7.11. 仿照随机变量的变异系数、偏度系数和峰度系数（见定义 2.46），样本变异系数定义为 $C_v = S/\bar{X}$ ，样本偏度系数定义为 $C_s = B_3/B_2^{3/2}$ ，样本峰度系数定义为 $C_k = B_4/B_2^2 - 3$ 。

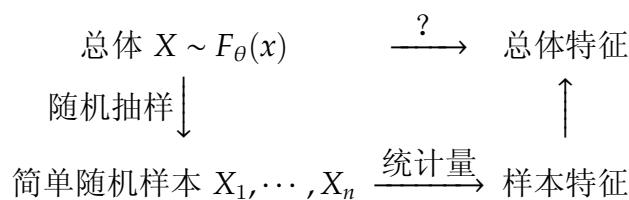
练习 7.10. 按照定义 7.11，利用 R 计算 Iris 数据中 setosa 类的花瓣长度数据（见第 457 页的例 7.3）的变异系数、偏度系数和峰度系数。

答案: $C_v = 0.1187852, C_s = 0.1031751, C_k = 0.8045921$ 。

7.2 样本统计量及其性质

样本是从总体中随机抽样而得，里面隐藏着总体分布中未知参数的信息。由样本构造而得的统计量是对样本的简化，Fisher 认为简化数据也是统计学的研究内容，即把样本中所含的未知参数的信息“压缩”到统计量中。他说，“现代统计学家都熟悉这样一个观念，任何有限数据只包含考察对象的限量信息；这个局限是由数据本身的性质决定的，不能通过统计研究中耗费的聪明才智得到延展：统计学家的任务实际上仅限于提取具体问题的所有可用信息。”

通常有关未知参数的统计推断（如参数估计、假设检验）是通过某些统计量实现的（这个过程的图示如下），而无需先通过经验分布来逼近总体分布后再研究这些未知参数。譬如，若总体期望 μ 未知，则可以通过样本均值 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ 对 μ 作出推断。在统计推断中，选择合适的统计量并搞清楚它的分布是非常关键的。



定义 7.12 (抽样分布). 统计量 $T = T(X_1, X_2, \dots, X_n)$ 的分布称作 T 的抽样分布 (sampling distribution)，它完全由样本 X_1, X_2, \dots, X_n 的分布唯一决定。

抽样分布就是随机变量 T 的分布，之所以冠以“抽样”这一限定词，无非是强调抽样分布可由随机抽样的方法得到：把第 k 次从总体抽得容量为 n 的样本记作 $X_1^{(k)}, X_2^{(k)}, \dots, X_n^{(k)}$ ，算出 $T^{(k)} = T(X_1^{(k)}, X_2^{(k)}, \dots, X_n^{(k)})$ ，则 $T^{(k)}, k = 1, 2, \dots, m$ 是总体 T 的简单随机样本，由 Glivenko 定理，只要 m 充分地大， T 的分布可通过 $T^{(1)}, T^{(2)}, \dots, T^{(m)}$ 的经验分布近似得到。

例 7.11 (样本均值的抽样分布). 下表给出了在若干不同总体分布之下，样本均值 $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ 或由它构造的新统计量的抽样分布。

表 7.4: 在不同的总体之下，由样本均值 \bar{X} 构造的统计量的抽样分布。

总体分布	统计量	抽样分布
$N(\mu, \sigma^2)$	\bar{X}	$N(\mu, \sigma^2/n)$
$B(m, p)$	$n\bar{X}$	$B(mn, p)$
Poisson(λ)	$n\bar{X}$	Poisson($n\lambda$)
Cauchy(μ, λ)	\bar{X}	Cauchy(μ, λ)
Expon(β)	$2n\beta\bar{X}$	χ^2_{2n}

证明. 例 7.2 中倒数第二行是因为 \bar{X} 的特征函数恰是 $\text{Cauchy}(\mu, \lambda)$ 分布的特征函数 $\exp\{i\mu t - \lambda|t|\}$ 。最后一行是因为 $2n\beta\bar{X}$ 与 χ^2_{2n} 的特征函数都是 $(1 - 2it)^{-n}$ 。 \square

算法 7.2. 样本均值 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ 是最常见的统计量。若再增加一个新的样本点 X_{n+1} , 新的样本均值可按下面的方式更新:

$$\bar{X}_{\text{new}} = \bar{X} + \frac{1}{n+1} (X_{n+1} - \bar{X})$$

※例 7.12. 容量为 n 的样本 X_1, X_2, \dots, X_n 来自从总体 $N(0, 1)$, 样本均值 $\bar{X} \sim N(0, 1/n)$ 。样本量 n 越大, 随机变量 \bar{X} 的取值越紧密围绕在总体均值周围。反复抽样的批次 m 越大, 样本均值 $\bar{X}^{(1)}, \bar{X}^{(2)}, \dots, \bar{X}^{(m)}$ 的经验分布函数越接近统计量 \bar{X} 的抽样分布。

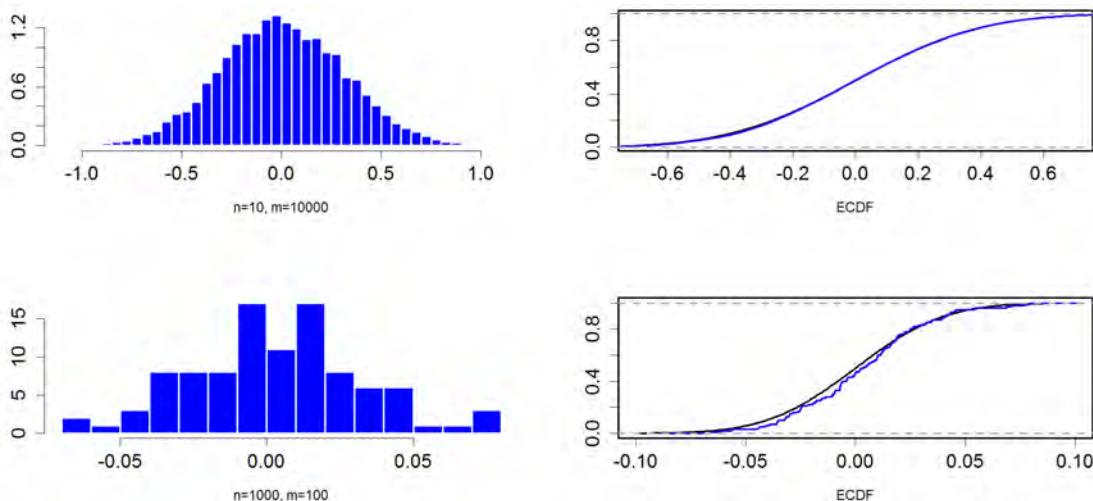


图 7.11: 样本均值 $\bar{X}^{(1)}, \bar{X}^{(2)}, \dots, \bar{X}^{(m)}$ 的直方图和经验分布函数: 第一行 $n = 10, m = 10^4$, 第二行 $n = 10^3, m = 10^2$ 。

即便实际情况不允许反复从总体中抽样, 利用已有样本通过“自助法”(bootstrap method)*依然可以得到统计量的经验分布(详见 §7.2.2)。

本节内容

第一小节总结了正态总体下几个常见统计量的抽样分布, Fisher-Geary 定理 7.5 是一个关键结果。在无法或很难得到抽样分布的确切表达式的时候, 第二小节介绍通过“自助法”可获得统计量的经验分布。第三小节引入了充分统计量的概念(它之所以重要是因为未丢失样本中所含的未知参数信息), 接着给出了充分统计量的判定方法——Fisher 因子分解定理。

关键知识

(1) 正态总体下几个常见统计量的抽样分布; (2) Fisher-Geary 定理的内容; (3) 自助法; (4) 统计量的充分性和 Fisher 因子分解定理。

* “bootstrap”一词来自习语 Pull yourself up by your bootstraps, 比喻不借助外部援助仅通过自身努力而改善状况或提升性能, 也暗指自立、自持、自助等性质的行为。

7.2.1 统计量的抽样分布

已知样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F(x)$, 要搞清楚任一统计量 $T(X_1, \dots, X_n)$ 的抽样分布并非易事, 绝大多数情况下很难找到具有简单形式的抽样分布。然而对于正态总体, 不难求得一些重要统计量的抽样分布, 因此正态总体之下的置信区间估计、假设检验等研究成果硕硕, 这些事实也抬高了正态分布在统计学中的地位——统计学在相当长的一段时间里把正态总体当作研究重点, 统计推断也往往是基于正态总体的。

性质 7.8. 如果样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 则有

$$\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)^2 \sim \chi_1^2 \quad (7.20)$$

证明. 因为 $Y = \sqrt{n}(\bar{X} - \mu) / \sigma \sim N(0, 1)$, 所以 $Y^2 \sim \chi_1^2$ 。 \square

定理 7.5 (Fisher-Geary, 1925, 1936). 样本 X_1, \dots, X_n 来自一个正态总体当且仅当样本均值 \bar{X} 与样本方差 S^2 独立。

证明. “ \Rightarrow ” 见第 245 页的定理 3.16。“ \Leftarrow ” 的证明由爱尔兰统计学家 Roy Charles Geary (1896-1983) 于 1936 年给出 [53], 已超出本书范围。 \square

性质 7.9. 已知样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 则有

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

证明. 我们知道

$$\sum_{j=1}^n \frac{(X_j - \mu)^2}{\sigma^2} \sim \chi_n^2, \text{ 并且 } \left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)^2 \sim \chi_1^2$$

根据 $\sum_{j=1}^n (X_j - \bar{X}) = 0$, 显然有

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{j=1}^n (X_j - \mu)^2 &= \frac{1}{\sigma^2} \sum_{j=1}^n (X_j - \bar{X} + \bar{X} - \mu)^2 \\ &= \left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)^2 + \frac{(n-1)S^2}{\sigma^2} \end{aligned}$$

因为总体是正态分布, 由定理 3.16 可知 \bar{X} 与 S^2 独立, 进而上式右侧两求和项独立。从上式的特征函数易得 $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ 。 \square

性质 7.10. 已知样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 样本方差为 S^2 , 则

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

证明. 由 $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$ 和 $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$, 我们有

$$\frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

□

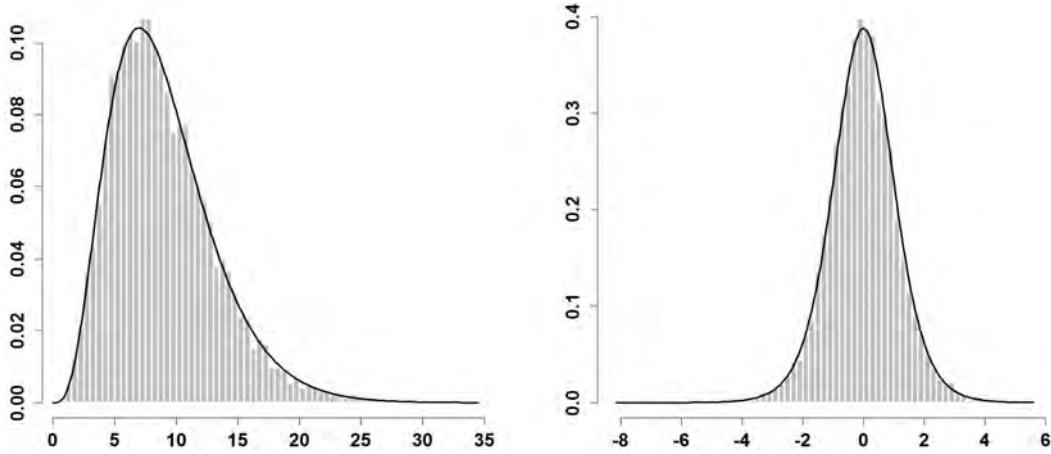


图 7.12: 对性质 7.9 和性质 7.10 的模拟试验: 令 $n = 10, \mu = 0, \sigma^2 = 1$, 经过 10^4 轮反复抽样, 得到 $(n-1)S^2/\sigma^2$ 和 $\sqrt{n}(\bar{X} - \mu)/S$ 的直方图。实线分别是 χ_{n-1}^2 和 t_{n-1} 分布的密度函数曲线。

例 7.13. 设样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ 的均值和方差分别为 \bar{X} 和 S^2 , 若从总体再抽取一个样本点 X_{n+1} , 问下面的统计量服从什么分布?

$$Y = \frac{X_{n+1} - \bar{X}}{S} \sqrt{\frac{n}{n+1}}$$

解. 由 $X_{n+1} - \bar{X} \sim N(0, (n+1)\sigma^2/n)$ 和性质 7.9 得 $Y \sim t_{n-1}$ 。

性质 7.11. 已知来自两个独立总体的样本 $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} N(\mu_X, \sigma_X^2)$ 和 $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2)$ 的样本均值和样本方差分别为 $\bar{X}, S_X^2, \bar{Y}, S_Y^2$, 则

$$\begin{aligned} \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} &\sim F_{m-1, n-1} \\ \frac{[\bar{X} - \bar{Y} - (\mu_X - \mu_Y)] \sqrt{\frac{m+n-2}{\sigma_X^2/m + \sigma_Y^2/n}}}{\sqrt{(m-1)S_X^2/\sigma_X^2 + (n-1)S_Y^2/\sigma_Y^2}} &\sim t_{m+n-2} \end{aligned}$$

证明. 因为这两个总体是独立的, 于是

$$\begin{aligned}\bar{X} - \bar{Y} &\sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right) \\ \frac{(m-1)S_X^2}{\sigma_X^2} + \frac{(n-1)S_Y^2}{\sigma_Y^2} &\sim \chi_{m+n-2}^2\end{aligned}$$

由性质 7.9 以及 F 分布、 t 分布的定义, 易证。 \square

※例 7.14. 设简单随机样本 X_1, X_2, \dots, X_n 来自总体 $X \sim F(x)$, 求第 j 个次序统计量 $X_{(j)}$ 的分布函数。

解. 由分布函数的定义, $X_{(j)}$ 的分布函数为 $F_{X_{(j)}}(x) = P(X_{(j)} \leq x)$, 进而

$$\begin{aligned}F_{X_{(j)}}(x) &= P\{X_1, X_2, \dots, X_n \text{ 中至少有 } j \text{ 个满足 } " \leq x "\} \\ &= \sum_{k=j}^n C_n^k [P(X \leq x)]^k [1 - P(X \leq x)]^{n-k} \\ &= \sum_{k=j}^n C_n^k [F(x)]^k [1 - F(x)]^{n-k}\end{aligned}$$

特别地, $X_{(1)}$ 的分布函数是 $F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n$, $X_{(n)}$ 的分布函数是 $F_{X_{(n)}}(x) = [F(x)]^n$ 。譬如, 如果总体是均匀分布 $U(0, 1)$, 则极值统计量 $X_{(1)}$ 和 $X_{(n)}$ 的密度函数分别是

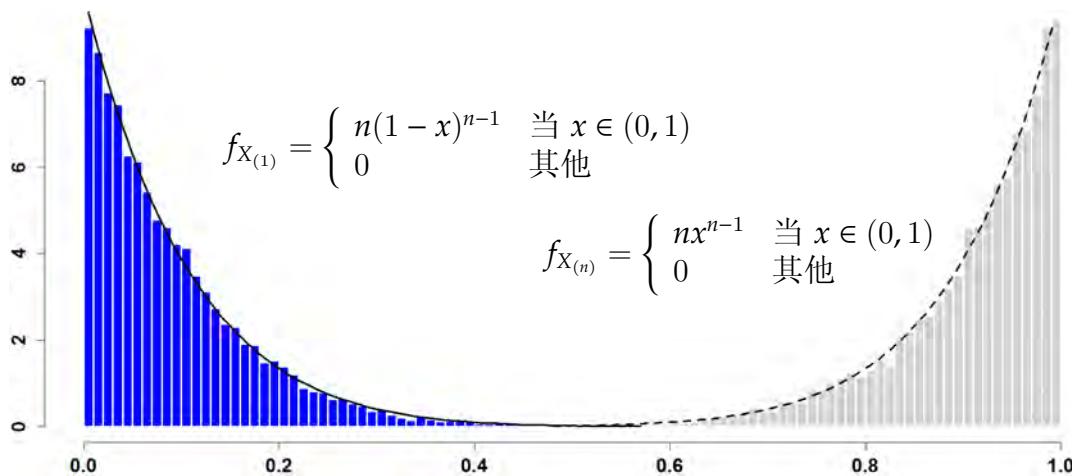


图 7.13: 令总体是均匀分布 $U(0, 1)$, 则极值统计量 $X_{(1)}$ 和 $X_{(n)}$ 的密度函数分别是图中的实线和虚线。经过 10^4 轮反复抽取容量为 10 的样本, 分别得到极小值和极大值的直方图。

定义 7.13. 设简单随机样本 $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top$ 来自二元正态总体 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 其中 $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ 未知。定义统计量如下,

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{j=1}^n X_j & \bar{Y} &= \frac{1}{n} \sum_{j=1}^n Y_j \\ S_X^2 &= \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2 & S_Y^2 &= \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})^2 \\ C_{XY} &= \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) & R_{XY} &= \frac{C_{XY}}{S_X S_Y}\end{aligned}$$

我们把 R_{XY} 称为样本相关系数, 把对称矩阵 $\begin{pmatrix} S_X^2 & C_{XY} \\ C_{XY} & S_Y^2 \end{pmatrix}$ 称为样本方差-协方差矩阵, 简称样本协方差矩阵。

定理 7.6. 定义 7.13 中的统计量具有以下性质:

- 随机向量 $(\bar{X}, \bar{Y})^\top$ 与 $(X_1 - \bar{X}, \dots, X_n - \bar{X}, Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})^\top$ 相互独立。进而, $(\bar{X}, \bar{Y})^\top$ 与 $(S_X^2, C_{XY}, S_Y^2)^\top$ 相互独立。
- 随机向量 $(\bar{X}, \bar{Y})^\top$ 服从二元正态分布如下,

$$(\bar{X}, \bar{Y})^\top \sim N\left(\mu_X, \mu_Y, \frac{\sigma_X^2}{n}, \frac{\sigma_Y^2}{n}, \rho\right) \quad (7.21)$$

证明. 仿照定理 3.16 的证明: 设随机向量 $(\bar{X}, \bar{Y}, X_1 - \bar{X}, \dots, X_n - \bar{X}, Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})^\top$ 的特征函数是 $\varphi(u, v, s_1, \dots, s_n, t_1, \dots, t_n)$, 则

$$\begin{aligned}\varphi(u, v, s_1, \dots, s_n, t_1, \dots, t_n) &= E \exp \left\{ iu\bar{X} + iv\bar{Y} + \sum_{k=1}^n is_k(X_k - \bar{X}) + \sum_{k=1}^n it_k(Y_k - \bar{Y}) \right\} \\ &= E \exp \left\{ \sum_{k=1}^n iX_k \left(\frac{u}{n} + s_k - \bar{s} \right) + iY_k \left(\frac{v}{n} + t_k - \bar{t} \right) \right\} \\ \text{其中, } \bar{s} &= \frac{s_1 + \dots + s_n}{n}, \text{ 且 } \bar{t} = \frac{t_1 + \dots + t_n}{n}\end{aligned}$$

根据例 3.7 所示二元正态分布的特征函数, 对 $\varphi(u, v, s_1, \dots, s_n, t_1, \dots, t_n)$ 进行整理得到下面的分解,

$$\begin{aligned}\varphi(u, v, s_1, \dots, s_n, t_1, \dots, t_n) &= \exp \left\{ iu\mu_X + iv\mu_Y - \frac{\sigma_X^2 u^2 + 2\rho\sigma_X\sigma_Y uv + \sigma_Y^2 v^2}{2n} \right\} \\ &\cdot \exp \left\{ -\frac{1}{2}\sigma_X^2 \sum_{k=1}^n (s_k - \bar{s})^2 - \rho\sigma_X\sigma_Y \sum_{k=1}^n (s_k - \bar{s})(t_k - \bar{t}) - \frac{1}{2}\sigma_Y^2 \sum_{k=1}^n (t_k - \bar{t})^2 \right\}\end{aligned}$$

$$= \varphi(u, v)\varphi(s_1, \dots, s_n, t_1, \dots, t_n)$$

利用定理 3.15 可证得独立性，利用 $(\bar{X}, \bar{Y})^\top$ 的特征函数 $\varphi(u, v)$ 可得其分布为二元正态分布 (7.21)。 \square

7.2.2 重抽样和自助法

设样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F(x)$ 的经验分布函数是 $\hat{F}_n(x)$, 如果 n 足够地大, 由 Glivenko 定理 7.1 有 $F(x) \approx \hat{F}_n(x)$ 。于是, 统计量 $T = T(X_1, \dots, X_n)$ 的抽样分布可以由 $T_* = T(X_*^{(1)}, \dots, X_*^{(n)})$ 来近似, 其中 $X_*^{(1)}, \dots, X_*^{(n)} \stackrel{\text{iid}}{\sim} \hat{F}_n(x)$ 。

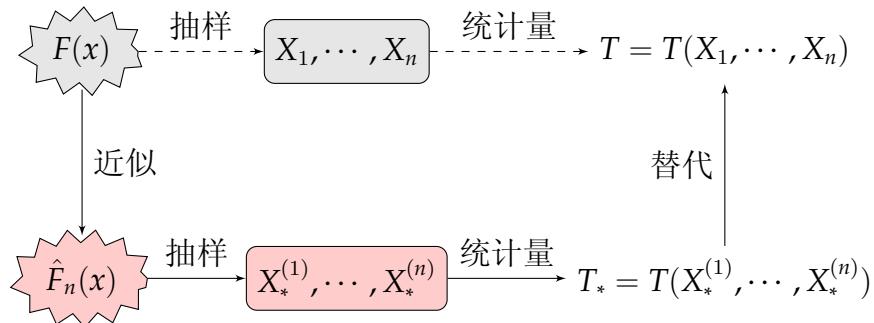


图 7.14: 偷梁换柱: 把从总体 $F(x)$ 抽样转化为从经验分布 $\hat{F}_n(x)$ 抽样。

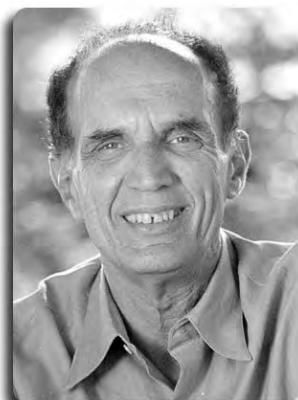
如何从总体 $\hat{F}_n(x)$ 抽取简单随机样本 $X_*^{(1)}, \dots, X_*^{(n)}$ 呢？从 $\hat{F}_n(x)$ 抽取 n 个随机数等同于从样本 X_1, \dots, X_n 有放回地抽取一个容量为 n 的样本 [157]——这便是自助法 (bootstrap method) 的理论基础。

练习 7.11. 观察到 $n = 20$ 样本如下，考虑随机试验：有放回地抽取 n 个样本求其均值。将这个随机试验独立重复 $m = 10^3$ 次，考察均值的直方图。

0.7938, -1.3182, -0.0186, 0.8658, 0.1547, 1.3976, -0.3154, -0.0692, -1.7258, 1.2550, 2.0547, 0.1922, 0.5905, -1.2801, 1.7199, 0.2308, 1.3225, 0.3340, 0.0219, 0.9353

自助法是美国著名统计学家 Bradley Efron (1938-) 于 1979 年提出的一种基于重抽样 (resampling) 的模拟方法 [30, 41, 42]，是很多统计推断问题的有效工具，如统计量 $T = T(X_1, \dots, X_n)$ 的抽样分布近似为 $T_*^{(1)}, T_*^{(2)}, \dots, T_*^{(m)}$ 的经验分布。

算法 7.3 (自助法). 令 $\hat{F}_n(x)$ 是简单随机样本 X_1, X_2, \dots, X_n 的经验分布函数, 统计量 $T = T(X_1, X_2, \dots, X_n)$ 的样本可用下面的方法近似求得。



- 有放回地从样本 X_1, X_2, \dots, X_n 中抽取 $X_*^{(1)}, X_*^{(2)}, \dots, X_*^{(n)}$ 。
 - 计算 $T_* = T(X_*^{(1)}, X_*^{(2)}, \dots, X_*^{(n)})$ 。
 - 重复上述两个步骤 m 遍得到样本 $T_*^{(1)}, T_*^{(2)}, \dots, T_*^{(m)}$ 。

例 7.15. 从样本 X_1, \dots, X_n 有放回地抽取一个容量为 n 的样本，其经验分布函数与 $\hat{F}_n(x)$ 的关系如何？利用自助法得到样本均值 \bar{X} 的分布情况如何？我们通过下面的模拟试验来了解。

- 产生 n 个 Laplace(0, 1) 分布的随机数 x_1, \dots, x_n ，其经验分布函数 $\hat{F}_n(x)$ 的曲线是下图中红色的粗实线。
- 有放回地从 x_1, \dots, x_n 抽取 n 个样本，绘出其经验分布函数。重复此过程 $m = 200$ 次，发现大多数的阶梯曲线都接近于 $\hat{F}_n(x)$ ，散落在 $\hat{F}_n(x)$ 的周围。

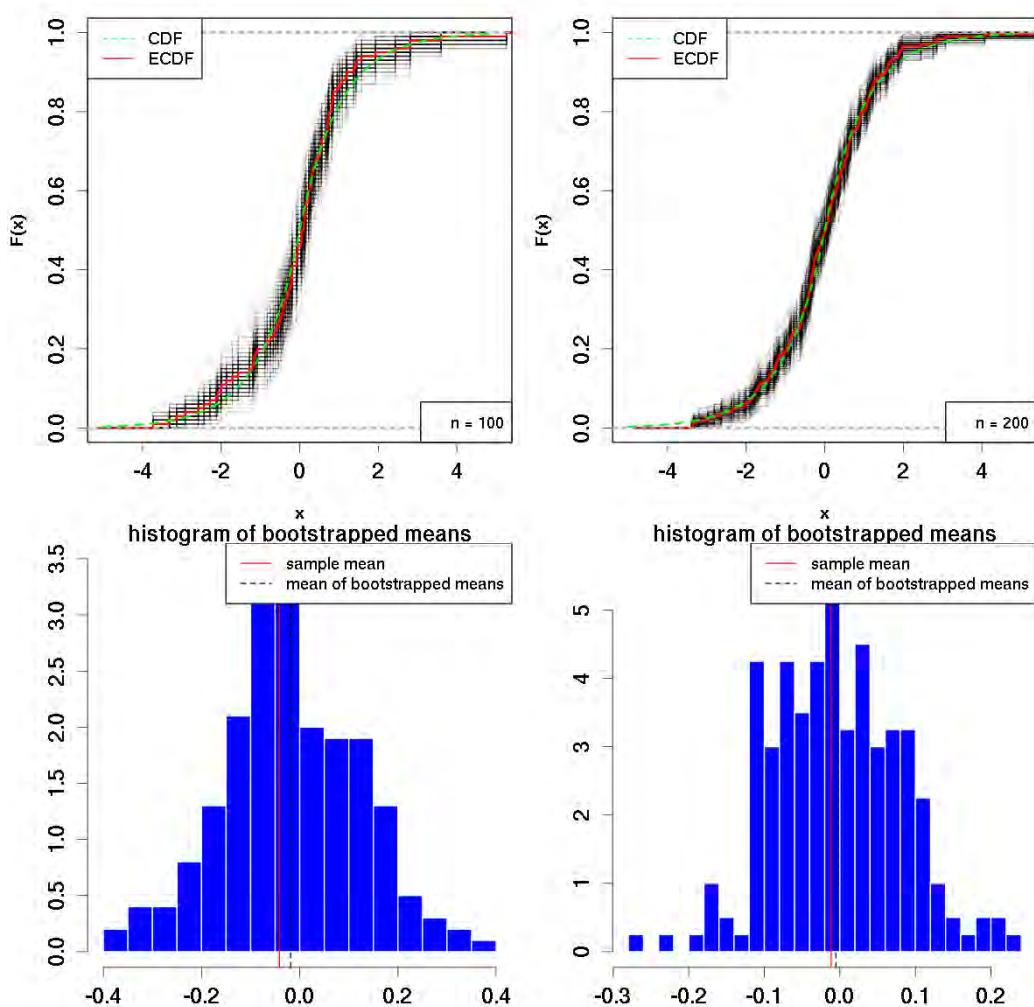


图 7.15：有放回地从观测值 x_1, \dots, x_n 中随机抽取 n 个样本，其经验分布函数通常与 x_1, \dots, x_n 的经验分布函数 $\hat{F}_n(x)$ 很接近，尤其当 n 很大的时候。

通过自助法得到样本均值 \bar{X} 这一统计量的 m 个样本 $\bar{X}_*^{(1)}, \dots, \bar{X}_*^{(m)}$ ，其直方图如图 7.15 所示。不难发现 $\bar{X}_*^{(1)}, \dots, \bar{X}_*^{(m)}$ 的均值与 \bar{X} 非常之接近。

例 7.16. 如果样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1)$, 其中 $n = 101$ 。请利用自助法（见[算法 7.3](#)）分别求样本中位数 $T = M(X_1, X_2, \dots, X_n)$ 和样本均值 $T = \bar{X}$ 的经验分布。

解. 由[例 7.14](#) 可得样本中位数 $X_{(d)}$ 的分布函数 $F_{X_{(d)}}(x)$, 其中 $d = (n + 1)/2$ 。进而根据[定理 5.23](#) 将 $F_{X_{(d)}}(x)$ 简化为

$$F_{X_{(d)}}(x) \approx \Phi\left(\frac{n - n\Phi(x)}{\sqrt{n\Phi(x)(1 - \Phi(x))}}\right) - \Phi\left(\frac{d - 1 - n\Phi(x)}{\sqrt{n\Phi(x)(1 - \Phi(x))}}\right)$$

样本中位数 $X_{(d)}$ 和样本均值 $\bar{X} \sim N(0, 1/n)$ 这两个统计量的分布函数在下图中用虚线绘出。令 $m = 10^3$, 利用[算法 7.3](#), 得到统计量 T_* 的样本 $T_*^{(1)}, T_*^{(2)}, \dots, T_*^{(m)}$ (直方图见下图左列), 进而得到对应的经验分布函数 (见下图右列中的实线)。

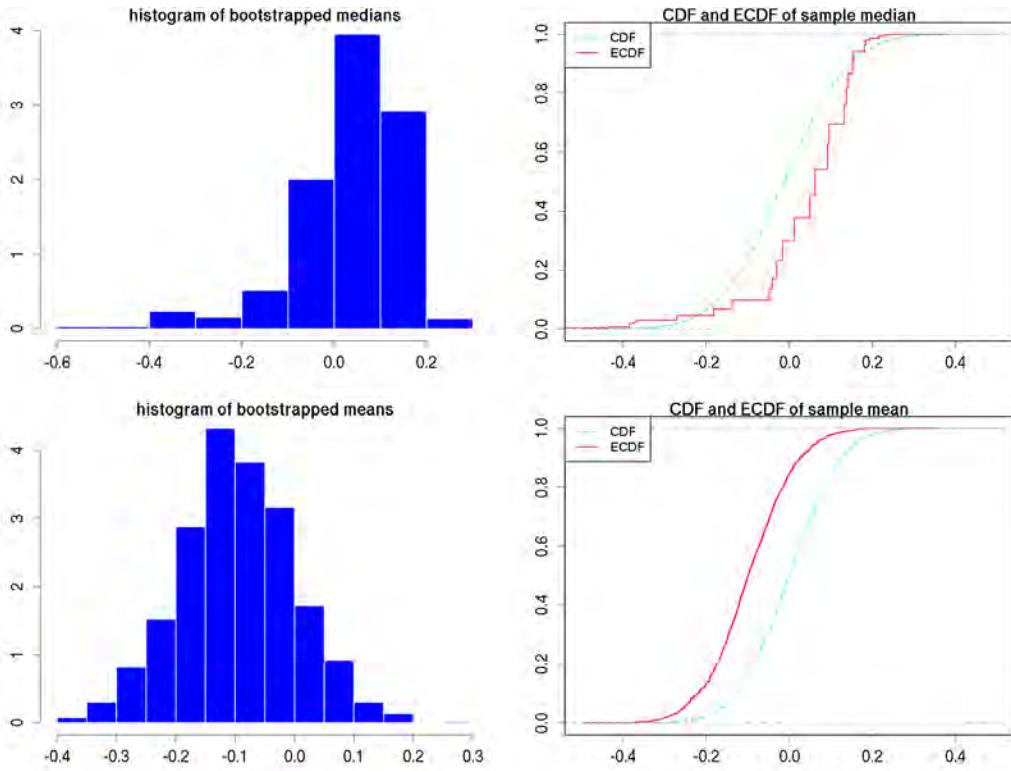


图 7.16: 利用自助法对样本 X_1, X_2, \dots, X_n 进行多次有放回的抽样, 分别得到样本中位数 $X_{(d)}$ 和样本均值 \bar{X} 的直方图和对应的经验分布函数曲线 (右列中的实线)。

通过试验我们发现, 样本 X_1, X_2, \dots, X_n 越真实地反映出总体的分布情况, 用自助法得到的统计量的经验分布与该统计量的真实分布就越接近。

例 7.17. 如果样本容量 n 足够地大, 统计量 $T = T(X_1, X_2, \dots, X_n)$ 的方差 $V_F(T)$ 近

似地为 $V_{\hat{F}_n}(T)$, 例如样本均值 \bar{X} 的方差为

$$\begin{aligned} V_F(\bar{X}) &= \frac{1}{n} \left\{ \int_{\mathbb{R}} x^2 dF(x) - \left[\int_{\mathbb{R}} x dF(x) \right]^2 \right\} \\ &\approx \frac{1}{n} \left\{ \int_{\mathbb{R}} x^2 d\hat{F}_n(x) - \left[\int_{\mathbb{R}} x d\hat{F}_n(x) \right]^2 \right\} \end{aligned}$$

如果 $V_{\hat{F}_n}(T) = \sigma^2$ 难于计算, 可用自助法近似求得。根据强大数律, 当 $m \rightarrow \infty$ 时有

$$\hat{\sigma}_{\text{boot}}^2 = \frac{1}{m} \sum_{k=1}^m \left(T_*^{(k)} - \frac{1}{m} \sum_{j=1}^m T_*^{(j)} \right)^2 \xrightarrow{a.s.} \sigma^2$$

算法 7.4. 如何由 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的 m 个观察样本再构造 k 个新样本?

- 由算法 7.1 分别独立产生 X_1, \dots, X_n 的随机数 k 个。
- 在 \mathbf{X} 的 m 个样本中有放回地随机抽取一个, 对其每个分量, 相应地在第一步产生的结果中寻找离它最近的随机数, 所得即一个新样本。同时, 相应地在第一步的结果中删掉刚选中的随机数。
- 独立重复第二步直至产生出 k 个新样本。

7.2.3 统计量的充分性

由样本构造的统计量与样本相比，多多少少丢失掉一些总体的信息。例如，后者包含未知参数 θ 的更多信息。

但在某些情况下，统计量包含了与样本同样多有关 θ 的信息。为定义如此好性质的统计量，1920 年，英国天才统计学家 Fisher 提出了“充分统计量”(sufficient statistic) 这一重要的概念，并于 1922 年给出了一个充要条件来判定任一给定的统计量是否是充分的，即 Fisher 因子分解定理*。

定义 7.14 (充分性). 已知 $T = T(X_1, X_2, \dots, X_n)$ 为一个统计量，如果 $\theta \in \Theta$ 样本的条件分布 $F_\theta(x_1, \dots, x_n | T = t)$ 与 θ 无关，则称 T (对未知参数 θ 而言) 是一个充分统计量。

充分统计量必定包含了未知参数的所有信息，以它做条件才会使得条件分布与未知参数无关。具体说来，总体 X 为离散型或连续型随机变量时，条件概率 $P_\theta\{X_1 = x_1, \dots, X_n = x_n | T = t\}$ 或条件密度函数 $f_\theta(x_1, \dots, x_n | T = t)$ 与 θ 无关。

充分统计量还有一个美妙且实用的好处：哪怕原始数据丢失了，凭借 $T = t$ 也能通过条件分布 $F(x_1, \dots, x_n | T = t)$ 来“恢复”原始数据。从这个角度充分统计量是原始数据的一个化简，或“无参数信息损失”的数据压缩。

例 7.18. 设样本 $X_1, X_2 \stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda)$ ，下面验证 $X_1 + X_2$ 对参数 λ 而言是一个充分统计量，但 $X_1 + 2X_2$ 不是。

$$\begin{aligned} P\{X_1 = x_1, X_2 = x_2 | X_1 + X_2 = t\} &= \frac{P\{X_1 = x_1, X_2 = t - x_1\}}{P\{X_1 + X_2 = t\}} \\ &= \begin{cases} C_t^{x_1}/2^t & \text{若 } x_1 + x_2 = t \\ 0 & \text{否则} \end{cases} \end{aligned}$$

上式的计算用到了练习 4.17。下面说明 $X_1 + 2X_2$ 不是充分统计量。

$$\begin{aligned} P\{X_1 = 0, X_2 = 1 | X_1 + 2X_2 = 2\} &= \frac{P\{X_1 = 0, X_2 = 1\}}{P\{X_1 + 2X_2 = 2\}} \\ &= \frac{e^{-\lambda}(\lambda e^{-\lambda})}{P\{X_1 = 0, X_2 = 1\} + P\{X_1 = 2, X_2 = 0\}} \\ &= \frac{\lambda e^{-2\lambda}}{\lambda e^{-2\lambda} + (\lambda^2/2)e^{-2\lambda}} = \frac{2}{\lambda + 2} \end{aligned}$$

*该结果于 1935 年被 J. Neyman 重新发现，所以有的文献中也称之为 Neyman 因子分解定理。该定理于 1949 年被两位美国数学家 Paul Richard Halmos (1916-2006) 和 Leonard Jimmie Savage (1917-1971) 严格证明，其中 Savage 是主观贝叶斯主义者。



例 7.19. 令样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 则 $T = \sum_{j=1}^n X_j$ 对参数 p 而言是一个充分统计量, 事实上

$$\begin{aligned} P\left\{X_1 = x_1, \dots, X_n = x_n \mid \sum_{j=1}^n X_j = t\right\} &= \frac{P\left\{X_1 = x_1, \dots, X_n = x_n, \sum_{j=1}^n X_j = t\right\}}{P\left\{\sum_{j=1}^n X_j = t\right\}} \\ &= \begin{cases} \frac{p^{\sum_{j=1}^n x_j} (1-p)^{n-\sum_{j=1}^n x_j}}{C_n^t p^t (1-p)^{n-t}} = \frac{1}{C_n^t} & \text{若 } \sum_{j=1}^n x_j = t \\ 0 & \text{否则} \end{cases} \end{aligned}$$

通过充分性的**定义 7.14** 和上面两个离散型的例子可以看出: $P_\theta\{\mathbf{X} = \mathbf{x}\}$ 可以分解为不含 θ 的有关 \mathbf{x} 的某函数与 $P_\theta\{T(\mathbf{X}) = T(\mathbf{x})\}$ 的乘积。1925 年 Fisher 提供了一个判定充分统计量的有效方法可以避开繁琐的条件概率计算, 这就是著名的 Fisher 因子分解定理。

定理 7.7 (Fisher 因子分解定理, 1925). 设样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的概率密度函数为 $f_\theta(\mathbf{x})$ 或概率函数为 $f_\theta(\mathbf{x}) = P_\theta\{\mathbf{X} = \mathbf{x}\}$, 其中 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$, 统计量 $T(\mathbf{X})$ 对未知参数 θ 而言是充分的当且仅当存在分解

$$f_\theta(\mathbf{x}) = h(\mathbf{x})g_\theta[T(\mathbf{x})]$$

其中, 非负 (可测) 函数 $h(\mathbf{x})$ 不依赖于 θ , 非负 (可测) 函数 $g_\theta[T(\mathbf{x})]$ 是关于 θ 和 $T(\mathbf{x})$ 的函数。

※证明. 严格的证明需用到测度论的知识, 感兴趣的读者可参阅陈希孺的《高等数理统计学》[168] 第一章的附录。这里仅考虑总体是离散型的。

□ 往证 “ \Rightarrow ”: 令 $T(\mathbf{X})$ 是充分统计量, 则 $P_\theta\{\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t\}$ 与参数 θ 无关。当 $T(\mathbf{x}) = t$ 时,

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}) &= P_\theta\{\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t\} \\ &= P_\theta\{\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t\} P_\theta\{T(\mathbf{X}) = t\} \end{aligned}$$

- 对那些满足 $\forall \theta \in \Theta, P_\theta(\mathbf{X} = \mathbf{x}) = 0$ 的 \mathbf{x} , 定义 $h(\mathbf{x}) = 0$ 。
- 对那些满足 $\exists \theta$ 使得 $P_\theta(\mathbf{X} = \mathbf{x}) > 0$ 的 \mathbf{x} , 定义 $h(\mathbf{x}) = P_\theta\{\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t\}$, 并定义 $g_\theta[T(\mathbf{x})] = P_\theta\{T(\mathbf{X}) = T(\mathbf{x}) = t\}$ 。

□ 下面往证“ \Leftarrow ”：对任意固定的 t_0 有

$$\begin{aligned} P_\theta\{T(\mathbf{X}) = t_0\} &= \sum_{\{\mathbf{x}: T(\mathbf{x})=t_0\}} P_\theta\{\mathbf{X} = \mathbf{x}\} \\ &= \sum_{\{\mathbf{x}: T(\mathbf{x})=t_0\}} h(\mathbf{x}) g_\theta[T(\mathbf{x})] \\ &= g_\theta(t_0) \sum_{\{\mathbf{x}: T(\mathbf{x})=t_0\}} h(\mathbf{x}) \end{aligned}$$

若 $P_\theta\{T(\mathbf{X}) = t_0\} = 0$, 结果是平凡的。设 $P_\theta\{T(\mathbf{X}) = t_0\} > 0$: 若 $T(\mathbf{x}) \neq t_0$, 则 $P_\theta\{\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t_0\} = 0$; 若 $T(\mathbf{x}) = t_0$, 则

$$\begin{aligned} P_\theta\{\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t_0\} &= \frac{P_\theta\{\mathbf{X} = \mathbf{x}\}}{P_\theta\{T(\mathbf{X}) = t_0\}} \\ &= \frac{h(\mathbf{x}) g_\theta(t_0)}{g_\theta(t_0) \sum_{\{\mathbf{x}: T(\mathbf{x})=t_0\}} h(\mathbf{x})} \\ &= \frac{h(\mathbf{x})}{\sum_{\{\mathbf{x}: T(\mathbf{x})=t_0\}} h(\mathbf{x})} \end{aligned}$$

不管怎样, $P_\theta\{\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t_0\}$ 都不依赖于 θ , 得证。 □

例 7.20. 已知样本 $X_1, X_2 \cdots, X_n \stackrel{\text{iid}}{\sim} U[0, \theta]$, 极值 $X_{(n)} = \max(X_1, X_2 \cdots, X_n)$ 对未知参数 θ 而言是充分的。直观上, $X_{(n)}$ 是样本中最接近 θ 的, 很自然它最能反映出 θ 。理论上, 样本的联合密度函数为

$$\begin{aligned} f_\theta(x_1, x_2, \cdots, x_n) &= \begin{cases} \theta^{-n} & \text{当 } 0 \leq x_1, x_2, \cdots, x_n \leq \theta \\ 0 & \text{其他} \end{cases} \\ &= J(x_{(1)})[\theta^{-n} J(\theta - x_{(n)})] \end{aligned}$$

其中, $x_{(1)} = \min(x_1, x_2, \cdots, x_n), x_{(n)} = \max(x_1, x_2, \cdots, x_n)$ 且 $J(\cdot)$ 是式 (2.10) 定义的非负判定函数。由因子分解定理 7.7, 证得 $X_{(n)}$ 对 θ 而言是充分的。

练习 7.12. 已知简单随机样本 $X_1, X_2 \cdots, X_n$ 来自离散均匀分布总体 $U\{1, 2, \cdots, m\}$, 其中 m 未知, 则 $X_{(n)}$ 对 m 而言是充分的。提示:

$$P(X_1 = x_1, \cdots, X_n = x_n) = J(x_{(1)} - 1)[m^{-n} J(m - x_{(n)})]$$

例 7.21. 令样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 其中参数 μ, σ^2 都是未知的, $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的概率密度函数为

$$\begin{aligned} f_{\theta}(x_1, x_2, \dots, x_n) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left\{ -\frac{\sum_{j=1}^n (x_j - \mu)^2}{2\sigma^2} \right\} \\ &= \exp \left\{ \frac{\mu \sum_{j=1}^n x_j}{\sigma^2} - \frac{\sum_{j=1}^n x_j^2}{2\sigma^2} - \frac{n}{2} \left[\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2) \right] \right\} \\ &= \exp \left\{ \frac{n\mu\bar{x}}{\sigma^2} - \frac{n\bar{x}^2 + (n-1)s^2}{2\sigma^2} - \frac{n}{2} \left[\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2) \right] \right\} \end{aligned}$$

于是统计量 $(\bar{X}, A_2)^\top$ 与 $(\bar{X}, S^2)^\top$ 对 $\theta = (\mu, \sigma^2)^\top$ 而言都是充分的。

练习 7.13. 令样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 试证明:

□ 若 σ^2 已知, 统计量 \bar{X} 对未知参数 μ 而言是充分的。

提示: 因为 $\sum_{j=1}^n (x_j - \mu)^2 = \sum_{j=1}^n x_j^2 - 2n\mu\bar{x} + n\mu$, 所以 $\prod_{j=1}^n \phi(x_j | \mu, \sigma^2)$ 满足 Fisher 因子分解定理 7.7 的条件。

□ 若 μ 已知, $V = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ 对未知参数 σ^2 而言是充分的。

练习 7.14. 令样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 试证明:

□ 若 σ^2 未知, \bar{X} 对未知参数 μ 而言不是充分的。

提示: 由 $\bar{X} \sim N(\mu, \sigma^2/n)$, 得到 $\mathbf{X}|\bar{X} = \bar{x}$ 的条件密度函数仍含有 μ, σ^2 , 按照定义 7.14, \bar{X} 对未知参数 μ 而言不是充分的。

□ 若 μ 未知, S^2 对未知参数 σ^2 而言不是充分的。

定义 7.15 (指数族). k -参数指数族 (k -parameter exponential family) $\{f_{\theta}(\mathbf{x}) : \theta \in \Theta \subseteq \mathbb{R}^k, \mathbf{x} \in \mathbb{R}^d\}$ 中每个密度函数或概率函数 $f_{\theta}(\mathbf{x})$ 都具有如下形式:

$$\begin{aligned} f_{\theta}(\mathbf{x}) &= h(\mathbf{x})\eta(\theta) \exp \left\{ \sum_{j=1}^k \lambda_j(\theta)T_j(\mathbf{x}) \right\} \\ &= h(\mathbf{x}) \exp \left\{ \sum_{j=1}^k \lambda_j(\theta)T_j(\mathbf{x}) + \beta(\theta) \right\} \end{aligned}$$

其中, 对任意的 $j = 1, 2, \dots, k$, 函数 $\eta(\theta) > 0, \beta(\theta) = \ln \eta(\theta)$ 和 $\lambda_j(\theta)$ 都是 Θ 上的实值函数, $h(\mathbf{x}) \geq 0$ 和 $T_j(\mathbf{x})$ 都是 \mathbb{R}^d 上的实值函数。

例 7.22. 二项分布 $B(m, p)$ 、Poisson 分布 $\text{Poisson}(\lambda)$ 、正态分布 $N(\mu, \sigma^2)$ ，还有**例 7.21** 中样本的概率密度函数 $f_\theta(x)$ 都属于指数族。

$$\begin{aligned} f(x) &= C_m^x \exp \left\{ x \ln \frac{p}{1-p} + m \ln(1-p) \right\}, \text{ 其中 } p \in (0, 1), x \in \{0, \dots, m\} \\ f(x) &= \frac{1}{x!} \exp \{x \ln \lambda - \lambda\}, \text{ 其中 } \lambda > 0 \text{ 且 } x \in \{0, 1, 2, \dots\} \\ \phi(x|\mu, \sigma^2) &= \exp \left\{ \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2} \left[\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2) \right] \right\}, \text{ 其中 } \mu \in \mathbb{R}, \sigma^2 > 0 \\ &= \exp \left\{ \lambda_1 T_1(x) + \lambda_2 T_2(x) + \left[\frac{\lambda_1^2}{4\lambda_2} - \frac{1}{2} \ln \left(-\frac{\pi}{\lambda_2} \right) \right] \right\} \\ \text{其中, } \lambda_1 &= \frac{\mu}{\sigma^2}, \lambda_2 = -\frac{1}{2\sigma^2}, T_1(x) = x, T_2(x) = x^2 \end{aligned}$$

练习 7.15. 假设参数都是未知的，验证 Gamma 分布、Beta 分布、多项分布、多元正态分布也都属于指数族。

人们为什么对指数族感兴趣呢？因为这类分布有一些好处，例如，其简单随机样本 X_1, X_2, \dots, X_n 的密度函数在形式上非常之简单，即

$$f_\theta(x_1, x_2, \dots, x_n) = \left[\prod_{i=1}^n h(x_i) \right] \exp \left\{ \sum_{j=1}^k \lambda_j(\boldsymbol{\theta}) \sum_{i=1}^n T_j(x_i) + n\beta(\boldsymbol{\theta}) \right\} \quad (7.22)$$

另外，指数族的充分统计量的构造非常之方便。基于 (7.22)，由 Fisher 因子分解定理 7.7 不难得到下面的性质。

性质 7.12. 如果总体分布属于 k -参数指数族，设 X_1, X_2, \dots, X_n 是来自该分布的简单随机样本，则下面的统计量是充分统计量。

$$T(X_1, X_2, \dots, X_n) = \left[\sum_{i=1}^n T_1(X_i), \dots, \sum_{i=1}^n T_k(X_i) \right]^\top$$

例如，对于正态总体， $(X_1 + \dots + X_n, X_1^2 + \dots + X_n^2)^\top$ 是充分统计量，进而样本均值和样本方差也是。

7.3 习题

7.1. 设 \bar{X} 是样本 X_1, \dots, X_n 的均值, 试证明:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \mathbf{X}(I - \mathbf{1}\mathbf{1}^\top)\mathbf{X}, \text{ 其中 } \mathbf{X} = (X_1, \dots, X_n)^\top, \mathbf{1} = (1, \dots, 1)^\top$$

7.2. 作为描述分布的手段, 试比较直方图和经验累积分布函数 (ECDF) 的优缺点。

7.3. 设 \bar{X} 是样本 X_1, \dots, X_n 的均值, 试证明: 当 $c = \bar{X}$ 时, $f(c) = \sum_{i=1}^n (X_i - c)^2$ 的值达到最小。

7.4. 设样本 X_1, X_2, \dots, X_n 与样本 Y_1, Y_2, \dots, Y_n 之间有关系 $Y_j = (X_j - a)/b, j = 1, 2, \dots, n$, 其中 $b \neq 0$ 和 a 都是常数, 求样本平均值 \bar{Y} 与 \bar{X} , 以及样本方差 S_Y^2 与 S_X^2 之间的关系。

7.5. 设简单随机样本 X_1, X_2, \dots, X_n 来自总体 $X \sim F(x)$ 。若 X 的二阶矩存在, \bar{X} 为样本均值, 试证明: $X_i - \bar{X}$ 与 $X_j - \bar{X}$ 的相关系数 $\rho = -(n-1)^{-1}$, 其中 $i, j = 1, 2, \dots, n$ 且 $i \neq j$ 。

7.6. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 。(1) 求样本均值 \bar{X} 的分布列以及 $E(\bar{X})$ 和 $V(\bar{X})$; (2) 若 S^2 为样本方差, 求 $E(S^2)$; (3) 若样本值有 m 个 1, 其余的为 0, 求其经验分布函数。

7.7. 设样本 $X_1, X_2, \dots, X_{10} \stackrel{\text{iid}}{\sim} N(\mu, 4^2)$, 令 S^2 为样本方差。若已知 $P\{S^2 > a\} = 0.1$, 求常数 a (已知 $\chi_9^2(0.9) \approx 14.684$)。

☆ 7.8. 已知样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(\lambda)$, 试求极值 $X_{(1)}$ 的均值与方差。

☆ 7.9. 设 \bar{X}_1 和 \bar{X}_2 分别是取自正态总体 $N(\mu, \sigma^2)$ 的容量为 n 的两个简单随机样本 $X_{11}, X_{12}, \dots, X_{1n}$ 和 $X_{21}, X_{22}, \dots, X_{2n}$ 的均值, 试确定 n 使得 $P(|\bar{X}_1 - \bar{X}_2| > \sigma) = 0.01$ 。

☆ 7.10. 设样本 $X_1, X_2 \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, 求概率 $P\{(X_1 + X_2)^2 / (X_1 - X_2)^2 < 4\}$ 。

7.11. 设 X_1, X_2, \dots, X_9 是来自正态总体的简单随机样本, 令 $Y_1 = (X_1 + X_2 + \dots + X_6)/6, Y_2 = (X_7 + X_8 + X_9)/3$ 且 $S^2 = \frac{1}{2} \sum_{k=7}^9 (X_k - Y_2)^2$, 试证明: $\sqrt{2}(Y_1 - Y_2)/S \sim t_2$ 。

7.12. 已知样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, 2^2)$ 均值为 \bar{X} , 要使 $E(\bar{X} - \mu)^2 \leq 0.1$ 成立, 则样本量 n 不小于多少?

- 7.13. 已知样本 $X_1, X_2 \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, (1) 试证明: $(X_1 + X_2)^2$ 与 $(X_1 - X_2)^2$ 相互独立; (2) 试求统计量 $Y = (X_1 + X_2)^2 / (X_1 - X_2)^2$ 的分布。
- 7.14. 设简单随机样本 X_1, X_2, \dots, X_n 和 Y_1, Y_2, \dots, Y_n 为分别来自总体 $X, Y \stackrel{\text{iid}}{\sim} N(0, 1)$, 问统计量 $W = (X_1 + X_2 + \dots + X_n) / \sqrt{Y_1^2 + Y_2^2 + \dots + Y_n^2}$ 服从什么分布?
- 7.15. 已知样本 $X_1, X_2, X_3, X_4 \stackrel{\text{iid}}{\sim} N(0, 2^2)$ 且 $Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2 \sim \chi_2^2$, 求 a, b 的值。
- 7.16. 已知样本 $X_1, X_2, \dots, X_5 \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ 且 $Y = a(X_1 + X_2) / \sqrt{X_3^2 + X_4^2 + X_5^2}$ 服从 t 分布, 问 a 为多少?
- 7.17. 求总体 $N(20, 3)$ 的容量分别为 10、15 的两个样本的均值差的绝对值大于 0.3 的概率。
- 7.18. 设样本 $X_1, X_2, \dots, X_8 \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, 求下列统计量的分布: $Y_1 = (X_1 + X_2)^2 / (X_4 - X_3)^2$, $Y_2 = [(X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2] / [3(X_7 + X_8)^2]$ 和 $Y_3 = \sqrt{2/3}(X_1 + X_2 + X_3) / |X_4 - X_5|$ 。
- 7.19. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 样本均值 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$, 令 $Y_j = X_j + a\bar{X}$, 其中 a 为常数。求 Y_j 的分布。
- ☆ 7.20. 简单随机样本 X_1, \dots, X_n 来自密度函数为 $f(x) = \begin{cases} 2x & \text{当 } x \in [0, 1] \\ 0 & \text{其他} \end{cases}$ 的总体, 试求次序统计量 $X_{(1)}, X_{(k)}, X_{(n)}$ 的密度函数 ($1 < k < n$)。
- 7.21. 不管简单随机样本 X_1, \dots, X_n 来自总体 $\text{Expon}(\lambda)$ 还是 $\text{Poisson}(\lambda)$, 试证明: $T = \sum_{j=1}^n X_j$ 对未知参数 λ 而言是充分统计量。
- ☆ 7.22. 设简单随机样本 $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top$ 来自二元正态总体 $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 其中参数 $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ 都未知, 试给出一个充分统计量。

第八章

参数估计理论

松下问童子，言师采药去。只在此山中，云深不知处。

贾岛《寻隐者不遇》

理统计学的基本问题之一就是根据样本所提供的信息，推断总体的分布或其数数字特征。其中“最简单”的情况就是总体分布的类型已知，只是某些参数未知，这种情况下的统计推断称为参数统计推断。例如，已知总体 X 服从正态分布 $N(\mu, \sigma^2)$ ，其中方差 σ^2 已知，而均值 μ 未知，人们可以利用样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的均值估计出 μ 的取值。我们把担当估计任务的统计量称作估计量。

参数 θ 的估计量常记作 $\hat{\theta}(\mathbf{X})$ 或 $\hat{\theta}$ ，有时为了突出样本量 n ，也记作 $\hat{\theta}_n(\mathbf{X})$ 或 $\hat{\theta}_n$ 。得到样本值 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 后，经过计算所得的数值（或向量） $T(\mathbf{x})$ 称作 θ 的估计值，也常记作 $\hat{\theta}(\mathbf{x})$ 或 $\hat{\theta}_n(\mathbf{x})$ ，在不引起歧义的前提下简记作 $\hat{\theta}$ 或 $\hat{\theta}_n$ 。有的时候需要估计 θ 的某个实值函数 $g(\theta)$ 的值，在统计中 $g(\theta)$ 也称作参数，其估计值记作 $\widehat{g(\theta)}$ 。下文中对参数 θ 的估计方法都适用于估计 $g(\theta)$ ，不再赘述。

为刻画样本里所含未知参数信息的多少，我们引入 Fisher 信息量这一非常美妙却仍有待继续挖掘的概念。它将出现在一些重要的场合，如 Jeffreys 先验分布、Cramér-Rao 不等式、信息几何等。无论在频率派还是贝叶斯学派那里，Fisher 信息量都是具有生命力的。

在频率派看来，参数都是固定值，不管它是已知的还是未知的。而贝叶斯学派则认为未知参数是随机变量，有先验分布和后验分布。根据这一观念上的差别可以区分这两个学派，以及它们主张的经典统计方法和贝叶斯方法。



频率派有两类传统的方法来估计未知参数 μ : 点估计 (point estimation) 和区间估计 (interval estimation)。

- 点估计要求 μ 的估计量 $\hat{\mu}(\mathbf{X})$ 具备一些“好品质”，如相合性、无偏性、有效性等。点估计的方法主要包括 K. Pearson 的矩方法和 R. A. Fisher 的最大似然法。二者各有千秋，在某些条件下最大似然法要略胜一筹（具体讨论见 §8.1.5）。
- 区间估计是给出以某个概率，譬如 $1 - \alpha$ ，覆盖住 μ 的区间表示 $[\underline{\mu}(\mathbf{X}), \bar{\mu}(\mathbf{X})]$ ，其中统计量 $\underline{\mu}(\mathbf{X}) < \bar{\mu}(\mathbf{X})$ ， α 是个很小的正数。换句话说，在大量可重复试验中，随机区间 $[\underline{\mu}(\mathbf{X}), \bar{\mu}(\mathbf{X})]$ 覆盖住 μ 的机会是 $1 - \alpha$ 。

参数估计的结果一般不唯一，如何挑选其中的优者呢？对于区间估计，我们可以约定这样的评判标准：在给定 α 的前提下，以概率 $1 - \alpha$ 覆盖住未知参数的随机区间，其长度越短意味着猜得越准。

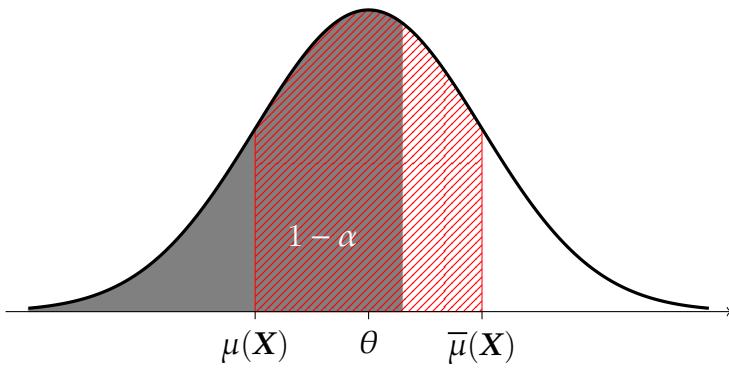


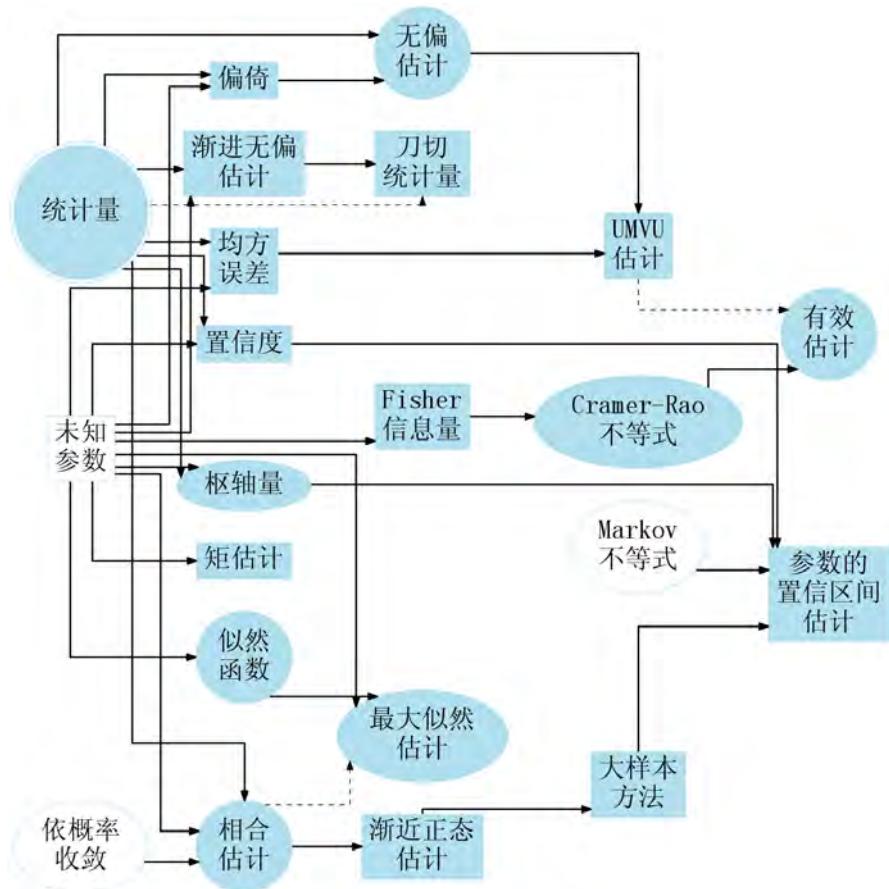
图 8.1: 能以概率 $1 - \alpha$ 覆盖住 θ 的随机区间 $[\underline{\mu}(\mathbf{X}), \bar{\mu}(\mathbf{X})]$ 不唯一，越短越好。例如，图中两种阴影部分的面积都是 $1 - \alpha$ ，所对应的区间长度差异却很大。

评价点估计的优劣要棘手一些。因为点估计要么猜中要么没猜中，而猜中的概率 $P\{\hat{\theta}(\mathbf{X}) = \theta\}$ 在一般情况下为零，所以我们转而研究点估计量的其他概率性质，如期望 $E[\hat{\theta}(\mathbf{X})]$ 是否命中 θ ，以及方差 $V[\hat{\theta}(\mathbf{X})]$ 有多大，等等。

有效估计 (efficient estimator) 是一类性质优良的估计，与它息息相关的是著名的 Cramér-Rao 不等式和一个有效性的判定定理 8.6。针对有偏估计，刀切法有助于修正偏倚，§8.1.3 对它做了简介。

本章只关注频率派的参数估计方法。作为补充，§8.2.4 将介绍 Fisher 的信任区间 (fiducial interval) 估计，它有别于传统的置信区间方法和贝叶斯方法，一直备受争议。有趣的是，在信任区间估计中，Fisher 也把未知参数视作随机变量，虽然 Fisher 本人自始至终是强烈反对贝叶斯学派的。更有趣的是，Fisher 信任推断得到的结果通过贝叶斯方法也能得到。

第八章的主要内容及其关系



8.1 点估计及其优良性

点 估计的目标就是构造统计量 $T = T(X_1, X_2, \dots, X_n)$ 使得用它对参数 θ 的估计时在某些标准下是“好的”，譬如用偏倚（bias，或称系统误差）和均方误差（mean squared error, MSE）来评价统计量 T 。

$$\text{BIAS}(\theta, T) = E_\theta(T) - \theta \quad (8.1)$$

$$\begin{aligned} \text{MSE}(\theta, T) &= E_\theta(T - \theta)^2 \\ &= E_\theta [T - E_\theta(T) + E_\theta(T) - \theta]^2 \\ &= E_\theta [T - E_\theta(T)]^2 + [E_\theta(T) - \theta]^2 \\ &= V_\theta(T) + [\text{BIAS}(\theta, T)]^2 \end{aligned} \quad (8.2)$$

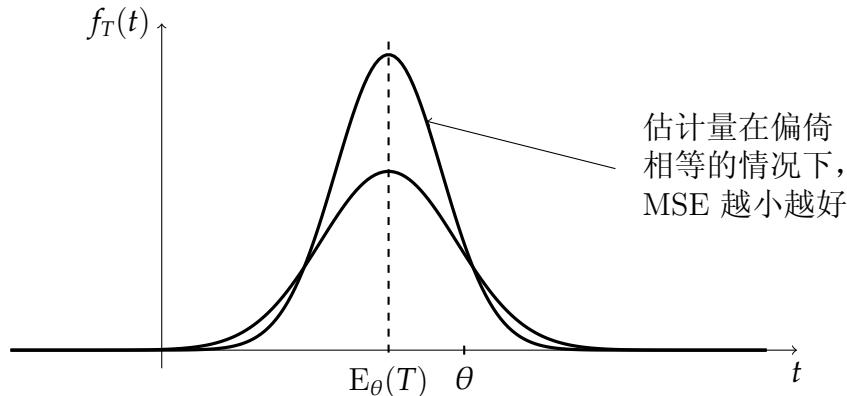


图 8.2: 偏倚和均方误差都是估计量 $T = T(X_1, X_2, \dots, X_n)$ 固有的特征，不依赖于样本的具体观测结果。当偏倚为零时，MSE 越小意味着估计的精度越高。

本节内容

第一和第二小节具体给出衡量点估计优劣的几个标准，即点估计的优良性，如相合性 (consistency)、渐近正态性、无偏性 (unbiasness)、有效性 (efficiency) 等，并继而研究了这些标准的性质和它们之间的关系。特别地，我们利用 Fisher 信息量和相关系数的性质来证明了 Cramér-Rao 不等式，利用刀切法对有偏估计量进行改造以降低偏倚。第三小节着重介绍了两个最常见的点估计方法*——矩方法和最大似然法，并且比较了它们的优劣。有关点估计的更多的内容请参阅 Lehmann 和 Casella 的经典之作《点估计理论》[103]。

关键知识

- (1) 相合性；(2) 无偏性；(3) 有效性；(4) Fisher 信息量；(5) Cramér-Rao 不等式；(6) 刀切法；(7) 矩方法；(8) 最大似然法。

*在第 14 章还将介绍另一常见的点估计方法——期望-最大化算法。

8.1.1 Fisher 信息量与信息矩阵

未知参数 θ 的信息隐藏于样本之中, 为了刻画样本包含未知参数信息的多少, Fisher 提出了信息量的概念 [48]。

定义 8.1 (Fisher 信息量). 设随机向量 $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ 的概率/密度函数为 $p_\theta(\mathbf{x})$, 未知参数 $\theta \in \Theta$ 的 Fisher 信息量 $\mathcal{I}(\theta)$ 定义为

$$\mathcal{I}(\theta) = E_\theta \left[\frac{\partial \ln p_\theta(\mathbf{X})}{\partial \theta} \right]^2 = \begin{cases} \int_{\mathcal{X}} \left[\frac{\partial \ln p_\theta(\mathbf{x})}{\partial \theta} \right]^2 p_\theta(\mathbf{x}) d\mathbf{x} & \text{连续型} \\ \sum_{j=1}^{\infty} \left[\frac{\partial \ln p_\theta(\mathbf{x}_j)}{\partial \theta} \right]^2 p_\theta(\mathbf{x}_j) & \text{离散型} \end{cases}$$

参考第 216 页的式 (2.98), Fisher 信息量还有一个等价定义:

$$\mathcal{I}(\theta) = -E_\theta \left[\frac{\partial^2 \ln p_\theta(\mathbf{X})}{\partial \theta^2} \right] = \begin{cases} - \int_{\mathcal{X}} \frac{\partial^2 \ln p_\theta(\mathbf{x})}{\partial \theta^2} p_\theta(\mathbf{x}) d\mathbf{x} & \text{连续型} \\ - \sum_{j=1}^{\infty} \frac{\partial^2 \ln p_\theta(\mathbf{x}_j)}{\partial \theta^2} p_\theta(\mathbf{x}_j) & \text{离散型} \end{cases}$$

例 8.1. 令总体为 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 其中参数 p 未知, 则

$$\begin{aligned} \mathcal{I}(p) &= -E_p \left\{ \frac{\partial^2 \ln[p^X(1-p)^{1-X}]}{\partial p^2} \right\} \\ &= E_p \left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2} \right] \\ &= \frac{1}{p(1-p)} \end{aligned}$$

当 $p = 1/2$ 时, Fisher 信息量达到最小, 此时熵是最大的。

练习 8.1. 若总体为 $Y \sim B(n, p)$, 其中参数 n 已知, 未知参数 p 的 Fisher 信息量为 $\mathcal{I}(p) = n/[p(1-p)]$ 。当 $p = 1/2$ 时, Fisher 信息量达到最小。

定义 8.2 (Fisher 信息矩阵). 设随机向量 $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ 的概率/密度函数为 $p_\theta(\mathbf{x})$, 未知向量参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top \in \Theta$ 的 Fisher 信息矩阵定义为一个 k 阶对阵矩阵 $\mathcal{I}(\boldsymbol{\theta}) = -E_\theta \{\nabla_{\boldsymbol{\theta}}^2 \ln p_\theta(\mathbf{X})\}$ (参见第 772 页的**定义 E.11**), 即 $\mathcal{I}(\boldsymbol{\theta})$ 的第 (i, j) 元素

定义为

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}} \left\{ \frac{\partial^2 \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \theta_i \partial \theta_j} \right\} = \begin{cases} - \int_{\mathcal{X}} \frac{\partial^2 \ln p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \theta_i \partial \theta_j} p_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{x} & \text{连续型} \\ - \sum_{j=1}^{\infty} \frac{\partial^2 \ln p_{\boldsymbol{\theta}}(\mathbf{x}_j)}{\partial \theta_i \partial \theta_j} p_{\boldsymbol{\theta}}(\mathbf{x}_j) & \text{离散型} \end{cases} \quad (8.3)$$

Fisher 信息矩阵亦可定义为 $\mathcal{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\mathbf{Y}\mathbf{Y}^\top\}$, 其中 $\mathbf{Y} = \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\mathbf{X})$, 即 $\mathcal{I}(\boldsymbol{\theta})$ 的第 (i, j) 元素定义为

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left\{ \frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \theta_i} \times \frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \theta_j} \right\}$$

Fisher 信息矩阵 $\mathcal{I}(\boldsymbol{\theta})$ 是一个 k 阶半正定对称矩阵, 在 k 维参数空间上定义了一个黎曼度量, 被称为 Fisher 信息度量, 它把统计学与微分几何学联系了起来从而发展成为一个交叉学科——信息几何学 (information geometry) [5], 通过几何不变量来研究统计不变量。

例 8.2. 已知总体 $X \sim N(\mu, \sigma^2)$, 分以下三种情况讨论参数的 Fisher 信息量。

□ 若参数 μ 未知, σ^2 已知, 则方差越小, μ 的 Fisher 信息量越大。这是因为

$$\mathcal{I}(\mu) = -E_{\mu} \left\{ \frac{\partial^2 \ln \phi(X|\mu, \sigma^2)}{\partial \mu^2} \right\} = E_{\mu} \left(\frac{1}{\sigma^2} \right) = \frac{1}{\sigma^2}$$

□ 若 μ 已知, σ^2 未知, 均值并不影响方差的 Fisher 信息量。

$$\mathcal{I}(\sigma^2) = -E_{\sigma^2} \left\{ -\frac{(X-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4} \right\} = \frac{1}{2\sigma^4}$$

□ 若 $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ 未知, Fisher 信息矩阵为

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

例 8.3. 令简单随机样本 X_1, \dots, X_n 来自概率/密度函数为 $p_{\boldsymbol{\theta}}(x)$ 的总体, 则样本的

Fisher 信息量为 $I_n(\theta) = nI(\theta)$, 这是因为

$$\begin{aligned} I_n(\theta) &= E_\theta \left[\frac{\partial \sum_{j=1}^n \ln p_\theta(X_j)}{\partial \theta} \right]^2 \\ &= E_\theta \left[\sum_{j=1}^n \frac{\partial \ln p_\theta(X_j)}{\partial \theta} \right]^2 \\ &= \sum_{j=1}^n E_\theta \left[\frac{\partial \ln p_\theta(X_j)}{\partial \theta} \right]^2 \\ &= nI(\theta) \end{aligned}$$

 **例 8.3** 的结果与直观认识是一致的: 容量为 n 的样本所含未知参数 θ 的 Fisher 信息量是单个样本点所含 θ 的 Fisher 信息量的 n 倍。样本量愈大, 样本所含 θ 的 Fisher 信息量就愈大。Fisher 信息量 $I(\theta)$ 是一个有关未知参数 θ 的函数, 通过**定理 8.5** (Cramér-Rao 不等式) 读者将会了解到 Fisher 信息量可以用来描述未知参数点估计的方差的下界。

※例 8.4. 已知随机变量 $X \sim p_\theta(x)$ 和一一映射 $\eta = g(\theta)$, 其中 $\theta \in \Theta \subseteq \mathbb{R}$ 。由链式法则有

$$\begin{aligned} \frac{\partial \ln p_\theta(x)}{\partial \theta} &= \frac{\partial \ln p_\theta(x)}{\partial \eta} \cdot \frac{d\eta}{d\theta} \\ \text{进而, } I(\theta) &= E \left[\frac{\partial \ln p_\theta(X)}{\partial \theta} \right]^2 \\ &= E \left[\frac{\partial \ln p_\theta(X)}{\partial \eta} \right]^2 \left(\frac{d\eta}{d\theta} \right)^2 \\ &= I(\eta) \left(\frac{d\eta}{d\theta} \right)^2 \end{aligned}$$

可得 Fisher 信息量 $I(\theta)$ 与 $I(\eta)$ 具有关系

$$\sqrt{I(\theta)} = \left| \frac{d\eta}{d\theta} \right| \sqrt{I(\eta)}$$

※例 8.5. 已知随机向量 $\mathbf{X} \sim p_\theta(\mathbf{x})$ 和一一映射 $\boldsymbol{\eta} = g(\boldsymbol{\theta})$, 其中参数 $\boldsymbol{\eta}$ 与 $\boldsymbol{\theta}$ 的维数相同。由**附录 E** 所述的链式法则 (E.3) 有

$$\frac{\partial \ln p_\theta(\mathbf{x})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \boldsymbol{\eta}^\top}{\partial \boldsymbol{\theta}} \right) \left[\frac{\partial \ln p_\theta(\mathbf{x})}{\partial \boldsymbol{\eta}} \right]$$

上式中 $d\boldsymbol{\eta}^\top/d\boldsymbol{\theta}$ 见第 771 页的定义 E.10。进而，信息矩阵 $\mathcal{I}(\boldsymbol{\theta})$ 与 $\mathcal{I}(\boldsymbol{\eta})$ 具有关系

$$\sqrt{\det \mathcal{I}(\boldsymbol{\theta})} = \left| \det \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \right) \right| \sqrt{\det \mathcal{I}(\boldsymbol{\eta})}$$

其中 $\det(A)$ 表示矩阵 A 的行列式。这是因为，

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= E \left\{ \left[\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \boldsymbol{\theta}} \right] \left[\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \boldsymbol{\theta}} \right]^\top \right\} \\ &= E \left\{ \frac{\partial \boldsymbol{\eta}^\top}{\partial \boldsymbol{\theta}} \left[\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \boldsymbol{\eta}} \right] \left[\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \boldsymbol{\eta}} \right]^\top \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \right\} \\ &= \frac{\partial \boldsymbol{\eta}^\top}{\partial \boldsymbol{\theta}} E \left\{ \left[\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \boldsymbol{\eta}} \right] \left[\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \boldsymbol{\eta}} \right]^\top \right\} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \\ &= \frac{\partial \boldsymbol{\eta}^\top}{\partial \boldsymbol{\theta}} \mathcal{I}(\boldsymbol{\eta}) \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \end{aligned}$$

在贝叶斯统计学中，Fisher 信息量和信息矩阵常用于定义未知参数的 Jeffreys 先验分布（见第 655 页的定义 12.14），其理论依据需要用到例 8.4 和例 8.5 的结果。

8.1.2 相合性与渐近正态性

结果 (7.19) 说明只要样本量足够地大, 可以以任意的精度用 k 阶样本矩来近似总体的 k 阶矩。为描述这样的大样本性质, 人们提出下述概念。

定义 8.3 (相合性). 当 $n \rightarrow \infty$ 时如果 $T_n = T(X_1, X_2, \dots, X_n) \xrightarrow{a.s.} \theta$, 称 T_n 是参数 θ 的强相合估计 (strong consistent estimator)。当 $n \rightarrow \infty$ 时如果 $T_n = T(X_1, X_2, \dots, X_n) \xrightarrow{P} \theta$, 称 T_n 是 θ 的弱相合估计或相合估计。

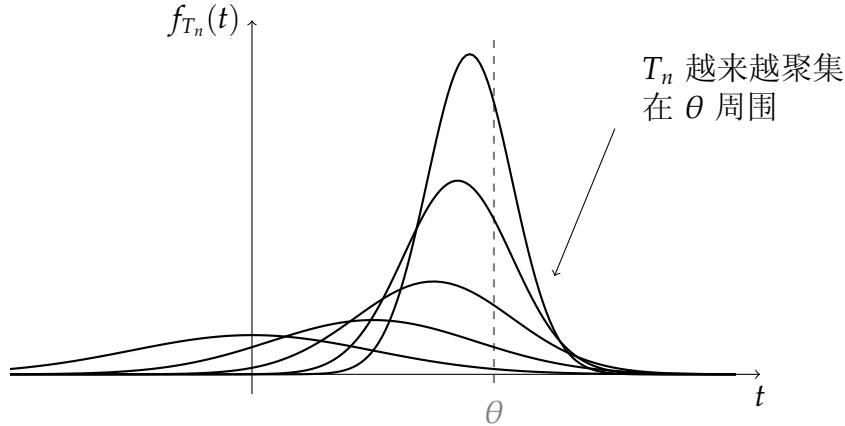


图 8.3: 随着样本容量 n 的增加, 相合估计量 T_n 的分布越来越聚集在参数 θ 的周围。只要 n 足够地大, T_n 的抽样就能非常地接近 θ 。

作为大数律的一个应用, 相合性并未描述收敛速度, 但如果一个估计量不具备相合性, 样本量再大对改善估计的精度也无济于事。所以, 相合性是对点估计的最低要求。

例 8.6. 相合估计不一定是唯一的。例如, 令样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} p|1\rangle + (1-p)|0\rangle$, 则

$$\begin{aligned} T_n &= \frac{1}{n} \sum_{j=1}^n X_j \xrightarrow{P} p, \text{ 并且} \\ T_n &= \frac{1}{n+2} \left(\sum_{j=1}^n X_j + 1 \right) \xrightarrow{P} p \end{aligned}$$

更一般地, $T'_n = T_n + c_n \xrightarrow{P} p$, 其中 $c_n \rightarrow 0$ 。

练习 8.2. 已知连续函数 $g(x)$ 和相合估计 $X_n \xrightarrow{P} \theta$, 则 $g(X_n)$ 是 $g(\theta)$ 的相合估计, 即 $g(X_n) \xrightarrow{P} g(\theta)$ 。提示: 利用第 361 页的定理 5.10。

练习 8.3. 已知相合估计 $T_n \xrightarrow{P} \alpha$ 和 $S_n \xrightarrow{P} \beta$, 则 $T_n + S_n$ 和 $T_n S_n$ 分别是 $\alpha + \beta$ 和 $\alpha\beta$ 的相合估计, 即 $T_n + S_n \xrightarrow{P} \alpha + \beta$, $T_n S_n \xrightarrow{P} \alpha\beta$; 若 $\beta \neq 0$, 还有 $T_n / S_n \xrightarrow{P} \alpha/\beta$ 。提示: 利用第 352 页的 Slutsky 定理可证。

例 8.7. 令样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 试证明: 样本方差 S^2 是总体方差 σ^2 的相合估计。

证明. 由 $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ 可知, $V(S^2) = \frac{2}{n-1}\sigma^4$ 且 $E(S^2) = \sigma^2$ 。根据 Chebyshev 不等式, $\forall \epsilon > 0$, 当 $n \rightarrow \infty$ 时,

$$P\{|S^2 - \sigma^2| \geq \epsilon\} \leq \frac{V(S^2)}{\epsilon^2} = \frac{2\sigma^4}{(n-1)\epsilon^2} \rightarrow 0 \quad \square$$

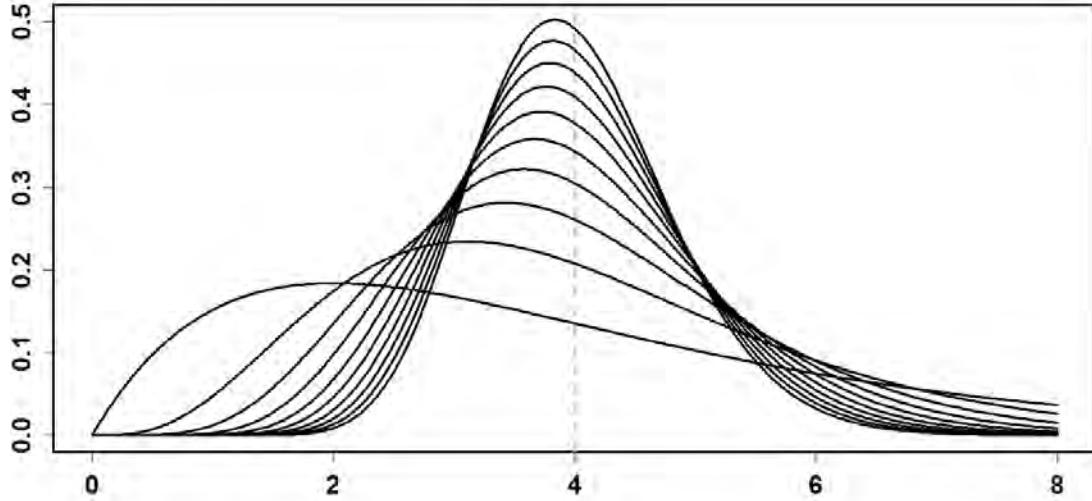


图 8.4: 例 8.7 中, 若 $\sigma^2 = 4$, 当样本容量 $n = 5, 10, 15, \dots, 50$ 时样本方差 S^2 的密度函数曲线随着 n 的增加越来越“高瘦”, 越来越凝聚在 σ^2 的周围。

定理 8.1. 令 $\{T_n = T(X_1, X_2, \dots, X_n)\}_{n=1}^\infty$ 是一个统计量的序列, 满足 $\lim_{n \rightarrow \infty} E(T_n) = \theta$ 且 $\lim_{n \rightarrow \infty} V(T_n) = 0$, 则 T_n 是 θ 的相合估计。

证明. 由 Chebyshev 不等式, 当 $n \rightarrow \infty$ 时有

$$\begin{aligned} P\{|T_n - \theta| \geq \epsilon\} &\leq \frac{E(T_n - ET_n + ET_n - \theta)^2}{\epsilon^2} \\ &= \frac{V(T_n) + (ET_n - \theta)^2}{\epsilon^2} \rightarrow 0 \end{aligned} \quad \square$$

定理 8.2. 已知总体 X 的均值 μ 和方差 σ^2 都存在, 设 \bar{X}, S^2, B_2 分别是来自总体 X 的简单随机样本 X_1, X_2, \dots, X_n 的样本均值、样本方差和样本二阶中心矩, 则 \bar{X} 是 μ 的相合估计, 且 S^2, B_2 都是 σ^2 的相合估计。

证明. 由 Chebyshev 弱大数律知 $\bar{X} \xrightarrow{P} \mu$ (即 \bar{X} 是 μ 的相合估计) 且

$$\frac{1}{n} \sum_{j=1}^n X_j^2 \xrightarrow{P} E(X^2)$$

而 $B_2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{1}{n} \sum_{j=1}^n X_j^2 - (\bar{X})^2$, 由性质 5.1 可证

$$\begin{aligned} B_2 &\xrightarrow{P} E(X^2) - [E(X)]^2 = \sigma^2 \\ \text{并且, } S^2 &= \frac{n}{n-1} B_2 \xrightarrow{P} \sigma^2 \end{aligned}$$
□

例 8.8 (陈希孺, 1994). 未知参数 θ 的相合估计并非总存在, 陈希孺院士曾构造过一个相合估计不存在的例子。令简单随机样本 X_1, X_2, \dots, X_n 来自总体 $X \sim P_\theta(X = 1)\langle 1 \rangle + (1 - P_\theta(X = 1))\langle 0 \rangle$, 其中 $0 \leq \theta \leq 1$ 并且

$$P_\theta(X = 1) = \begin{cases} \theta & \text{若 } \theta \text{ 是有理数} \\ 1 - \theta & \text{若 } \theta \text{ 是无理数} \end{cases}$$

经过很复杂的论证, 例 8.8 中未知参数 θ 的相合估计不存在。目前尚未发现相合估计存在性的充分必要条件, 下面的漂亮结果仅仅是一个充分条件。

~**定理 8.3.** 如果对任意的 $\theta \in \Theta, \epsilon > 0$, (8.4) 成立, 则未知参数 θ 的强相合估计存在。

$$\inf\{d(F_\theta, F_\eta) : \eta \in \Theta \text{ 满足 } \|\theta - \eta\| > \epsilon\} > 0 \quad (8.4)$$

~**证明.** 详见陈希孺的《高等数理统计学》[168] 第四章第一节。条件 (8.4) 的直观含义是“相隔较远的参数值所对应的分布函数不能太接近”。

估计量 T_n 的抽样分布通常很难求得, 但其极限分布有时却具有比较简单的形式。例如, 当样本容量足够大时, 式 (7.19) 描述了 k 阶样本矩 A_k 近似地服从一个正态分布, 即

$$\frac{A_k - m_k}{\sqrt{(m_{2k} - m_k^2)/n}} \xrightarrow{L} N(0, 1)$$

定义 8.4 (渐近正态性). 未知参数 $\mu_n(\theta)$ 的估计量 $T_n = T(X_1, X_2, \dots, X_n)$ 称为具有渐近正态性 (asymptotic normality) 或是渐近正态的 (asymptotically normal), 如果存在一个只依赖于 n, θ 的常量 $\sigma_n(\theta) > 0$ 或者统计量 $V_\theta(T_n)$ 使得

$$\begin{aligned} \frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} &\xrightarrow{L} N(0, 1) \\ \text{或者, } \frac{T_n - \mu_n(\theta)}{\sqrt{V_\theta(T_n)}} &\xrightarrow{L} N(0, 1) \end{aligned}$$

当 n 很大时, T_n 近似地服从一个正态分布, 简记作 $T_n \sim AN(\mu_n(\theta), \sigma_n^2(\theta))$ 。显然, $V_\theta(T_n)$ 越小表明 $\mu_n(\theta)$ 的估计量 $T_n = T(X_1, X_2, \dots, X_n)$ 越精确。

例 8.9. 如果总体的均值 μ 和方差 $\sigma^2 > 0$ 都存在, 由 Lindeberg-Lévy 中心极限定理 5.17, 样本均值 \bar{X} 具有渐近正态性 $\bar{X} \sim \text{AN}(\mu, \sigma^2/n)$, 即

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{L} N(0, 1), \text{ 其中 } n \text{ 为样本容量}$$

例如, 样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} U[0, \theta]$, 总体均值 $\mu = \theta/2$ 和总体方差 $\sigma^2 = \theta^2/12$ 都存在, 则

$$\frac{\sqrt{12n}(\bar{X} - \theta/2)}{\theta} \xrightarrow{L} N(0, 1)$$

再例如, 根据例 8.7 的结论, $S/\sigma \xrightarrow{P} 1$, 由 Slutsky 定理 5.3,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{S/\sigma} \xrightarrow{L} N(0, 1)$$

定理 8.4. 若总体的期望 μ 和方差 σ^2 都存在, 并且函数 $g: \mathbb{R} \rightarrow \mathbb{R}$ 在 μ 处可导, 则

$$\sqrt{n}(g(\bar{X}) - g(\mu)) \xrightarrow{L} N(0, \tau^2), \text{ 其中 } \tau^2 = \sigma^2[g'(\mu)]^2$$

证明. 由 $\sqrt{n}(\bar{X} - \mu) \xrightarrow{L} Y$, 其中 $Y \sim N(0, \sigma^2)$, 以及定理 5.4 可证。 □

 在例 8.9 的条件下, 样本均值 \bar{X} 是渐近正态的。如果定理 8.4 中 $g'(\mu) \neq 0$, 则 $g(\bar{X})$ 也是渐近正态的, 这是因为

$$\frac{\sqrt{n}(g(\bar{X}) - g(\mu))}{\sigma g'(\mu)} \xrightarrow{L} N(0, 1)$$

如果定理 8.4 中 $g'(\mu) \neq 0$ 并且导函数 g' 连续, 由定理 5.10, 则 $g'(\bar{X}) \xrightarrow{L} g'(\mu)$, 进而有 $Sg'(\bar{X})/[\sigma g'(\mu)] \xrightarrow{L} 1$ 。由 Slutsky 定理 5.3,

$$\frac{\sqrt{n}(g(\bar{X}) - g(\mu))}{Sg'(\bar{X})} = \frac{\sqrt{n}(g(\bar{X}) - g(\mu))/[\sigma g'(\mu)]}{Sg'(\bar{X})/[\sigma g'(\mu)]} \xrightarrow{L} N(0, 1)$$

8.1.3 无偏性和有效性

定义 8.5 (无偏性). 设 θ 是总体分布中的未知参数, 若统计量 T 满足 $E_\theta(T) = \theta$, 即 $BIAS(\theta, T) = 0$, 则称 T 是参数 θ 的无偏估计 (unbiased estimator), 否则称 T 是 θ 的有偏估计 (biased estimator)。

显然, 若按照 θ 的无偏估计 T 的分布有多个取值 t_1, t_2, \dots, t_m , 则当 m 很大时, $\frac{1}{m}(t_1 + t_2 + \dots + t_m)$ 可以很接近 θ 。

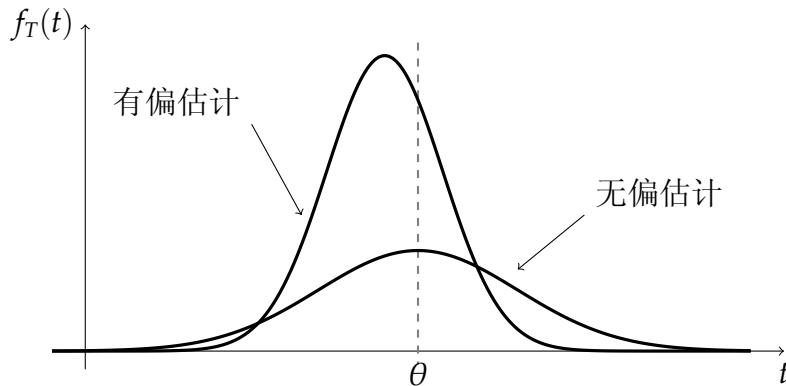


图 8.5: 不像要求估计量必须具备相合性那样, 无偏性是一个锦上添花的事情, 有自然好, 没有也无所谓。实践中, 有的时候宁愿要小方差的有偏估计, 也不要大方差的无偏估计。

不像要求估计量必须具备相合性那样, 无偏性是一个锦上添花的事情, 有自然好, 没有也无所谓。无偏估计并不唯一, 方差小的那个更受欢迎。实践中, 有的时候宁愿要小方差的有偏估计, 也不要大方差的无偏估计。

例 8.10. 简单样本 X_1, \dots, X_n 来自总体 $U[0, \theta]$, 其中 θ 是未知参数。令 $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$, 则 $2\bar{X}$ 是 θ 的无偏估计, 这是因为根据性质 4.16 有

$$\begin{aligned} E_\theta(2\bar{X}) &= 2E_\theta(\bar{X}) = \theta \\ \text{另外, } V_\theta(2\bar{X}) &= \frac{4}{n}V_\theta(\bar{X}) = \frac{\theta^2}{3n} \end{aligned}$$

还有一个估计 θ 的方法看起来更合理: 样本中的最大值 $X_{(n)} = \max(X_1, \dots, X_n)$ 。由例 7.20, 我们知道 $X_{(n)}$ 对 θ 而言是充分统计量。 $X_{(n)}$ 的分布函数为

$$P(X_{(n)} \leq x) = P(X_1 \leq x) \cdots P(X_n \leq x) = [P(X_1 \leq x)]^n = \left(\frac{x}{\theta}\right)^n$$

其中, $0 \leq x \leq \theta$ 。进而, 密度函数为

$$f_{X_{(n)}}(x) = \frac{d}{dx}P(X_{(n)} \leq x) = \frac{n}{\theta^n}x^{n-1}$$

$X_{(n)}$ 不是 θ 的无偏估计，这是因为

$$\begin{aligned} E_\theta X_{(n)} &= \int_0^\theta \frac{n}{\theta^n} x^n dx = \frac{n}{n+1} \theta \\ \text{然而, } V_\theta X_{(n)} &= \int_0^\theta \left(x - \frac{n}{n+1} \theta \right)^2 \frac{n}{\theta^n} x^{n-1} dx = \frac{n\theta^2}{(n+2)(n+1)^2} < V_\theta(2\bar{X}) \end{aligned}$$

我们对 $X_{(n)}$ 稍加改造， $\frac{n+1}{n}X_{(n)}$ 便是 θ 的无偏估计，并且其均方误差为

$$V_\theta \left(\frac{n+1}{n} X_{(n)} \right) = \frac{\theta^2}{n(n+2)} < V_\theta(2\bar{X})$$

通过此例，我们看到无偏估计不唯一，方差成为评判无偏估计优劣的标准。

例 8.11. 如果总体 X 的期望和方差存在，由性质 7.7 知，样本均值 \bar{X} 和样本方差 S^2 分别是对总体期望与方差的无偏估计。另外，如果 $E(X^k) = m_k$ 存在，则 k 阶样本矩 A_k 是 m_k 的无偏估计。

例 8.12. 无偏估计有时并不存在。例如，总体 $X \sim B(n, \theta)$ ，其中参数 $0 < \theta < 1$ 未知。设 $T = T(X)$ 是参数 $g(\theta) = \theta^{-1}$ 的无偏估计，则

$$\sum_{k=0}^n T(k) C_n^k \theta^k (1-\theta)^{n-k} = \frac{1}{\theta}, \text{ 其中 } 0 < \theta < 1$$

上式是不可能的，因为左边是有关 θ 的多项式，而右边不是。

定义 8.6 (渐近无偏性). 若统计量 $T_n = T(X_1, X_2, \dots, X_n)$ 满足 $\lim_{n \rightarrow \infty} E(T_n) = \theta$ ，则称之为 θ 的渐近无偏估计 (asymptotically unbiased estimator)。例 8.10 中的 $X_{(n)}$ 就是 θ 的渐近无偏估计。

例 8.13. 样本二阶中心矩

$$B_2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{n-1}{n} S^2$$

是总体方差的渐近无偏估计，而非无偏估计。

 参数 θ 的无偏估计 T 并不意味着每次估计都是精确的，它只保证基于不同的样本用 T 在对 θ 进行大量重复的估计时 $T - \theta$ 或正或负相互抵消，偏倚 $BIAS(\theta, T) = E_\theta(T) - \theta = 0$ 意味着 T 的均值是 θ 。另外，当 T 是 θ 的无偏估计时，均方误差 (8.2) 简化为 $MSE(\theta, T) = V_\theta(T)$ 。于是，比较 θ 的两个无偏估计的优劣即比较它们的方差大小，理论上比较容易处理，因此无偏性成为点估计的常见标准之一。无偏性与相合性没有逻辑上的关系。

定义 8.7 (UMVU 估计). 对于任意的 $\theta \in \Theta$, 如果 θ 的无偏估计 T_* 满足

$$V_\theta(T_*) = E_\theta(T_* - \theta)^2 \leq V_\theta(T) = E_\theta(T - \theta)^2 \quad (8.5)$$

其中 T 是 θ 的任一无偏估计, 则称 T_* 为 θ 的一致最小方差无偏估计 (uniformly minimum-variance unbiased estimator, UMVUE) 或简称 UMVU 估计。这里, 所谓的“一致”就是指对参数空间 Θ 内的所有 θ 来说, T_* 总是“最优的”。

UMVU 估计存在的情形并不多见, 利用**定义 8.7** 验证给定的统计量是 UMVU 估计绝非易事。很自然地人们对下述问题感兴趣, 因为如果得到肯定回答, 达到下界者一定是 UMVU 估计。

参数 θ 的所有无偏估计的方差是否存在非平凡的下界? 印度统计学家、Fisher 的学生 Calyampudi Radhakrishna Rao (1920-) 与瑞典统计学家 H. Cramér 分别于 1945 年和 1946 年独立对上述问题做出了肯定的回答, 并给出了著名的 Cramér-Rao 不等式, 其中所描述的下界被称为 Cramér-Rao 下界或简称 CR 界。



定理 8.5 (Cramér-Rao 不等式). 令简单随机样本 X_1, \dots, X_n 来自密度函数为 $f_\theta(x)$ 的总体 X , 统计量 $T = T(X_1, \dots, X_n)$ 满足 $V_\theta(T) < \infty$ 且

$$\frac{d}{d\theta} E_\theta(T) = \int_{\mathbb{R}^n} \cdots \int \frac{\partial}{\partial \theta} \left[T(x_1, \dots, x_n) \prod_{j=1}^n f_\theta(x_j) \right] dx_1 \cdots dx_n$$

记 $\psi(\theta) = E_\theta(T)$, 则 $V_\theta(T)$ 满足下面的不等式。

$$V_\theta(T) \geq \frac{[\psi'(\theta)]^2}{n I(\theta)} \quad (8.6)$$

对离散型的总体, 该结论也是同样的。特别地, 如果 T 是未知参数 θ 的无偏估计, 则 $V_\theta(T)$ 满足下面的不等式。

$$V_\theta(T) \geq \frac{1}{n I(\theta)} \quad (8.7)$$

证明. 令 $Z = \sum_{j=1}^n \partial \ln f_\theta(X_j) / \partial \theta$, 由相关系数的**性质 2.41** 易知,

$$\rho^2(T, Z) = \left[\frac{E_\theta(TZ) - E_\theta(T)E_\theta(Z)}{\sqrt{V_\theta(T)V_\theta(Z)}} \right]^2 \leq 1 \quad (8.8)$$

由式 (2.98) 的证明可知,

$$\begin{aligned} E_\theta(Z) &= 0, \text{ 并且 } V_\theta(Z) = nE\left[\frac{\partial \ln f_\theta(X)}{\partial \theta}\right]^2 = nI(\theta) \\ \psi'(\theta) &= \int_{\mathbb{R}^n} \cdots \int T(x_1, \dots, x_n) \left[\sum_{j=1}^n \frac{1}{f_\theta(x_j)} \frac{\partial f_\theta(x_j)}{\partial \theta} \right] \prod_{j=1}^n f_\theta(x_j) dx_1 \cdots dx_n \\ &= \int_{\mathbb{R}^n} \cdots \int T(x_1, \dots, x_n) \left[\sum_{j=1}^n \frac{\partial \ln f_\theta(x_j)}{\partial \theta} \right] \prod_{j=1}^n f_\theta(x_j) dx_1 \cdots dx_n \\ &= E_\theta(TZ) \end{aligned}$$

把这些结果代入式 (8.8) 即可证得结果 (8.6)。 \square

定义 8.8 (有效性). 如果 θ 的无偏估计 T_* 的方差 $V_\theta(T_*)$ 达到了式 (8.7) 所示的 Cramér-Rao 下界, 则称 T_* 为 θ 的有效估计 (efficient estimator), 它是无偏估计的“极品”。显然, 有效估计一定是 UMVU 估计, 反之则不然 (因为方差最小并不意味着它能达到 Cramér-Rao 下界)。

例 8.14. 已知简单随机样本 X_1, \dots, X_n 来自总体 $X \sim B(m, p)$, 其中 m 已知而 p 未知, 则 \bar{X}/m 是 p 的有效估计量。事实上, $V_p(\bar{X}/m) = \frac{pq}{mn}$, 其中 $q = 1 - p$ 。令 $p_k = P(X = k)$, 则参数 p 的 Fisher 信息量为

$$\begin{aligned} I(p) &= \sum_{k=0}^m \left[\frac{d}{dp} \ln(C_m^k p^k q^{m-k}) \right]^2 p_k \\ &= \sum_{k=0}^m \left(\frac{k}{p} - \frac{m-k}{1-p} \right)^2 p_k \\ &= \sum_{k=0}^m \left(\frac{k - mp}{pq} \right)^2 p_k = \frac{mpq}{p^2 q^2} = \frac{m}{pq} \end{aligned}$$

经过验证, $V_p(\bar{X}/m) = [nI(p)]^{-1}$, 因此 \bar{X}/m 确是 p 的有效估计量。

例 8.15. 设样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda)$, 其中参数 λ 未知, 则 \bar{X} 是 λ 的有效估计量。这是因为 $V_\lambda(\bar{X}) = \lambda/n$, 并且 $\partial \ln f_\lambda(x)/\partial \lambda = \partial(x \ln \lambda - \lambda - \ln x!)/\partial \lambda = (x - \lambda)/\lambda$ 。于是, 参数 λ 的 Fisher 信息量为

$$I(\lambda) = E_\lambda \left[\frac{\partial \ln f_\lambda(X)}{\partial \lambda} \right]^2 = \frac{E_\lambda(X - \lambda)^2}{\lambda^2} = \frac{1}{\lambda}$$

经过验证, $V_\lambda(\bar{X}) = [nI(\lambda)]^{-1}$, 因此 \bar{X} 确是 λ 的有效估计量。

练习 8.4. 已知样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 其中 μ 已知而 σ^2 未知, 则 S^2 不是 σ^2 的有效估计量。提示: $V(S^2) = \frac{2}{n-1}\sigma^4$, 由例 8.2 知 $I(\sigma^2) = \frac{1}{2}\sigma^{-4}$ 。 $V(S^2)$ 未达到 Cramér-Rao 下界 $\frac{2}{n}\sigma^4$ 。



从 Cramér-Rao 不等式的证明过程中还能得出什么结果? 假设 T 是 θ 的有效估计量, 从式 (8.8) 可知 $P\{Z = cT + d\} = 1$, 即几乎必然有 $Z = cT + d$, 其中 c, d 为常数。因为 T 是无偏的且 $E_\theta(Z) = 0$, 所以

$$\begin{aligned} E_\theta(Z) &= cE_\theta(T) + d = 0 \Rightarrow d = -c\theta \\ &\Rightarrow P\{Z = c(T - \theta)\} = 1 \end{aligned}$$

进而可知 T 对未知参数 θ 而言是充分的, 这是因为几乎必然有

$$\begin{aligned} Z &= \frac{\partial \ln \prod_{j=1}^n f_\theta(X_j)}{\partial \theta} = c(T - \theta) \Rightarrow \prod_{j=1}^n f_\theta(X_j) = h(X_1, \dots, X_n)g_\theta(T) \\ &\Rightarrow \frac{\partial \ln g_\theta(T)}{\partial \theta} = c(T - \theta), \text{ 其中 } g_\theta(T) > 0 \end{aligned}$$

定理 8.6 (有效性的判定). 条件与定理 8.5 相同, 未知参数 θ 的无偏估计 $T = T(\mathbf{X})$ 是有效的当且仅当

① 统计量 T 是充分的, 即样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的密度函数

$$\prod_{j=1}^n f_\theta(x_j) = h(\mathbf{x})g_\theta[T(\mathbf{x})], \text{ 其中 } \mathbf{x} = (x_1, x_2, \dots, x_n)^\top$$

② 函数 $g_\theta(t)$ 对于 $g_\theta(t) > 0$ 几乎必然满足方程

$$\frac{\partial \ln g_\theta(t)}{\partial \theta} = c(t - \theta), \text{ 其中 } c \text{ 与 } t \text{ 无关}$$

证明. “ \Rightarrow ” 已证, 现在往证 “ \Leftarrow ”, 即往证 $V_\theta(T) = [nI(\theta)]^{-1}$:

$$\begin{aligned} T \text{ 是 } \theta \text{ 的无偏估计} &\Rightarrow E_\theta T = \int_{\mathbb{R}^n} T(\mathbf{x})h(\mathbf{x})g_\theta[T(\mathbf{x})]d\mathbf{x} = \theta \\ &\Rightarrow \int_{\mathbb{R}^n} T(\mathbf{x})h(\mathbf{x})\frac{\partial g_\theta[T(\mathbf{x})]}{\partial \theta}d\mathbf{x} = 1 \\ h(\mathbf{x})g_\theta[T(\mathbf{x})] \text{ 是密度函数} &\Rightarrow \int_{\mathbb{R}^n} h(\mathbf{x})\frac{\partial g_\theta[T(\mathbf{x})]}{\partial \theta}d\mathbf{x} = 0 \end{aligned}$$

综合上述两个结果,

$$\begin{aligned}
 & \int_{\mathbb{R}^n} T(\mathbf{x}) h(\mathbf{x}) \frac{\partial g_\theta[T(\mathbf{x})]}{\partial \theta} d\mathbf{x} - \theta \int_{\mathbb{R}^n} h(\mathbf{x}) \frac{\partial g_\theta[T(\mathbf{x})]}{\partial \theta} d\mathbf{x} = 1 \\
 \Rightarrow & \int_{\mathbb{R}^n} [T(\mathbf{x}) - \theta] h(\mathbf{x}) \frac{\partial g_\theta[T(\mathbf{x})]}{\partial \theta} d\mathbf{x} = 1 \\
 \Rightarrow & \int_{\mathbb{R}^n} [T(\mathbf{x}) - \theta] h(\mathbf{x}) \frac{\partial \ln g_\theta[T(\mathbf{x})]}{\partial \theta} g_\theta[T(\mathbf{x})] d\mathbf{x} = 1 \\
 \Rightarrow & c \int_{\mathbb{R}^n} [T(\mathbf{x}) - \theta]^2 h(\mathbf{x}) g_\theta[T(\mathbf{x})] d\mathbf{x} = 1 \\
 \Rightarrow & c V_\theta(T) = 1
 \end{aligned}$$

由于 T 是充分统计量, 于是样本的 Fisher 信息量为

$$\begin{aligned}
 n \mathcal{I}(\theta) &= \mathcal{I}_n(\theta) \\
 &= E_\theta \left\{ \frac{\partial \ln f_\theta(\mathbf{X})}{\partial \theta} \right\}^2 \\
 &= E_\theta \left\{ \frac{\partial \ln g_\theta[T(\mathbf{X})]}{\partial \theta} \right\}^2 \\
 &= c^2 E_\theta (T - \theta)^2 \\
 &= c^2 V_\theta(T)
 \end{aligned}$$

与上式联立可得 $V_\theta(T) = 1/[n \mathcal{I}(\theta)]$, 达到了 Cramér-Rao 下界。 \square

※例 8.16. 接着练习 8.4, 试证明: σ^2 的无偏估计 $V = \frac{1}{n} \sum_{j=1}^n (X_j - \mu)^2$ 是有效的, 满足

$$V_{\sigma^2}(V) = \frac{2\sigma^4}{n}$$

证明. 显然 V 是 σ^2 的无偏估计。由 $nV/\sigma \sim \chi_n^2$ 及式 (4.16) 可得 V 的密度函数 $g_{\sigma^2}(v)$ 如下,

$$g_{\sigma^2}(v) = \frac{n^{n/2} v^{n/2-1}}{(2\sigma^2)^{n/2} \Gamma(n/2)} \exp \left\{ -\frac{nv}{2\sigma^2} \right\}$$

由练习 7.13 的结果, V 对 σ^2 而言还是充分的。另外,

$$\frac{\partial \ln g_{\sigma^2}(v)}{\partial \sigma^2} = \frac{n}{2\sigma^2} (v - \sigma^2), \text{ 其中 } \frac{n}{2\sigma^2} \text{ 与 } v \text{ 无关}$$

综上所述, 定理 8.6 的必要条件成立, 于是 V 是有效的。根据例 8.2 的结果, 不

难求得估计量 V 的方差。

$$V_{\sigma^2}(V) = \frac{1}{nI(\sigma^2)} = \frac{2}{n}\sigma^4$$

□

从练习 8.4 可知，在例 8.16 的条件下，样本方差 S^2 差一点儿就成为总体方差 σ^2 的有效估计量，与例 8.16 中的统计量 V 比较一下便知 S^2 差在没利用 μ 这一已知信息。

8.1.4 刀切法

刀切法 (jackknife method) [62, 144] 是一种普遍的方法, 能由给定的统计量 $T_n = T(X_1, X_2, \dots, X_n)$ 构造出具有更小偏倚的统计量。这种偏倚修正的方法最初由英国统计学家 Maurice Henry Quenouille (1924-1973) 于 1949、1956 年提出 [126, 127], 后由美国统计学家 John Wilder Tukey (1915-2000) 于 1958 年定名 [152]。

刀切法也是一种重采样技术, 它是自助法的线性近似。上世纪七十年代, Tukey 将刀切法发展为计算标准误差和置信区间 [157] 的非参数方法, 还应用于方差估计 [159] 等。下面我们粗略地介绍如何从给定的统计量构造刀切统计量, 并给出实例。

定义 8.9. 约定将 n 维向量 $\mathbf{X} = (X_1, \dots, X_j, \dots, X_n)^\top$ 中第 j 个分量 “切掉” 后所得的 $(n - 1)$ 维向量记作 $\mathbf{X}_{-j} = (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_n)^\top$, 并称之为 \mathbf{X} 的弃一 (leave-one-out) 表示。



定义 8.10 (刀切估计量). 如下构造的统计量称为 θ 的刀切估计量:

$$T_{\text{jack}} = nT(\mathbf{X}) - \frac{n-1}{n} \sum_{j=1}^n T(\mathbf{X}_{-j}) \quad (8.9)$$

下面说明式 (8.9) 所定义的刀切估计量的偏倚比 $T(\mathbf{X})$ 的有所改善。不妨设 $T(\mathbf{X})$ 的偏倚 $E_\theta[T(\mathbf{X})] - \theta$ 可用如下 n^{-1} 的幂级数展开:

$$E_\theta[T(\mathbf{X})] - \theta = \frac{c_1}{n} + \frac{c_2}{n^2} + \frac{c_3}{n^3} + \dots$$

随机向量 \mathbf{X}_{-j} 的维数是 $n - 1$, 构造估计量 $T(\mathbf{X}_{-j})$ 的方式与 $T(\mathbf{X})$ 的相同, 按照上式可将 $T(\mathbf{X}_{-j})$ 的偏倚表示为

$$E_\theta[T(\mathbf{X}_{-j})] - \theta = \frac{c_1}{n-1} + \frac{c_2}{(n-1)^2} + \frac{c_3}{(n-1)^3} + \dots$$

由 θ 的估计量 $T(\mathbf{X})$ 和 $T(\mathbf{X}_{-j})$ 的偏倚, 下面求估计量 T_{jack} 的偏倚。

$$\begin{aligned} E_\theta(T_{\text{jack}}) - \theta &= n\{E_\theta[T(\mathbf{X})] - \theta\} - \frac{n-1}{n} \sum_{j=1}^n \{E_\theta[T(\mathbf{X}_{-j})] - \theta\} \\ &= c_1 + \frac{c_2}{n} + \frac{c_3}{n^2} + \dots - \left\{ c_1 + \frac{c_2}{n-1} + \frac{c_3}{(n-1)^2} + \dots \right\} \\ &\sim O(1/n^2) \end{aligned}$$

当 $n \rightarrow \infty$ 时, 比起偏倚 $E_\theta[T(\mathbf{X})] - \theta \sim O(1/n)$, $E_\theta(T_{\text{jack}}) - \theta$ 收敛于 0 的速度更快一些。因此, T_{jack} 的偏倚比 $T(\mathbf{X})$ 的有所改善。

※例 8.17. 假设总体方差 σ^2 存在且未知。样本二阶中心矩 $T(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2$ 是 σ^2 的有偏估计, 按照式 (8.9) 构造刀切估计量。

$$T(\mathbf{X}_{-j}) = \frac{1}{n-1} \sum_{\substack{k=1 \\ k \neq j}}^n \left(X_k - \frac{n\bar{X} - X_j}{n-1} \right)^2, \text{ 代入到式 (8.9) 中}$$

$$\begin{aligned} \text{于是, } T_{\text{jack}} &= \sum_{j=1}^n (X_j - \bar{X})^2 + \frac{n}{n-1} S^2 - \frac{1}{n} \sum_{k,j=1}^n \left(X_k - \frac{n\bar{X} - X_j}{n-1} \right)^2 \\ &= (n-1)S^2 - \frac{1}{n} \sum_{k,j=1}^n \left(X_k - \bar{X} + \bar{X} - \frac{n\bar{X} - X_j}{n-1} \right)^2 \\ &= \frac{n}{n-1} S^2 - \frac{1}{n-1} S^2 = S^2, \text{ 为总体方差的无偏估计} \end{aligned}$$

8.1.5 点估计之矩方法和最大似然法

在参数的点估计方法中，矩方法和最大似然法分别受到 K. Pearson 和 R. A. Fisher 的推崇而引发了这两位统计学大师之间有关哪种方法更好的长期争论。Fisher 认为矩方法“没有理论上的合法性”，譬如矩方法对 Cauchy 分布的参数估计束手无策，而且矩估计的方差又大于最大似然估计的（渐近）方差。然而 K. Pearson 认为矩方法适用范围更广，譬如最大似然法就基本不适用于非参数统计学。历史的评价是这两种方法各有所长，下面依次介绍它们。

性质 8.1. 如果总体 X 的 k 阶矩 $m_k = E(X^k)$ 存在，则样本 j 阶矩 $A_j = \frac{1}{n} \sum_{i=1}^n X_i^j, j = 1, 2, \dots, k$ 是对 m_j 的（强）相合的无偏估计。

证明. 无偏性是显然的，（强）相合性直接由**性质 7.7** 可得。 \square

该性质启发了矩方法：若未知参数能通过总体矩表示出来，则将总体矩替换为相应的样本矩后，所得到的就是未知参数的估计量。

定义 8.11 (矩估计). 令 A_j 是样本 j 阶矩， $j = 1, \dots, k$ 。如果总体分布中的未知参数 θ 能表示成有限个总体矩 m_1, \dots, m_k 的函数 $\theta = h(m_1, \dots, m_k)$ ，其中 h 为 Borel 可测函数，这样就能保证 $h(A_1, \dots, A_k)$ 是一个统计量，称为 θ 的矩估计，记作 $\hat{\theta}$ 。

定理 8.7. 在**定义 8.11** 中，如果 h 是连续函数，则矩估计 $\hat{\theta} = h(A_1, \dots, A_k)$ 是 θ 的强相合估计。如果函数 h 对各个变量的一阶偏导数存在，则矩估计是渐近正态的。

证明. 强相合性，即 $\hat{\theta} = h(A_1, \dots, A_k) \xrightarrow{a.s.} \theta$ 由第 361 页的**定理 5.10** 可证。渐近正态性见 [29]，本书不作要求。 \square

 **定理 8.7** 是矩方法的理论依据。矩方法最大的优点是计算简单，只需把 $\theta = h(m_1, \dots, m_k)$ 中的各阶矩 m_1, \dots, m_k “偷梁换柱”为相应的样本矩即可。矩方法的另外一个优点是在一般情况下矩估计是强相合的。矩方法的缺点是当参数无法通过矩用 Borel 可测函数表示出来的时候就彻底无法使用，如 Cauchy 分布中的参数。矩方法是“统计学之父”K. Pearson 提出并大力推广的点估计经典方法，由于 K. Pearson 笃信大样本分析，他主推矩方法就不足为怪了。

例 8.18. 若总体方差 $\sigma^2 = m_2 - m_1^2$ 存在，从简单随机样本 X_1, X_2, \dots, X_n 可得到 σ^2 的矩估计 $\hat{\sigma}^2 = A_2 - A_1^2$ ，它是对 σ^2 的相合的、渐近正态的、渐近无偏的估计。根据练习 7.2 的结果， $\hat{\sigma}^2$ 可进一步表示为

$$\hat{\sigma}^2 = A_2 - A_1^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{n-1}{n} S^2 = B_2$$

因为 $nB_2/\sigma^2 \sim \chi_{n-1}^2$, 所以 $E(\hat{\sigma}^2) = (1 - 1/n)\sigma^2, V(\hat{\sigma}^2) = 2(n-1)\sigma^4/n^2$ 。当 n 很大时, 近似地有

$$\hat{\sigma}^2 \sim N\left(\left(1 - \frac{1}{n}\right)\sigma^2, \frac{2(n-1)}{n^2}\sigma^4\right)$$

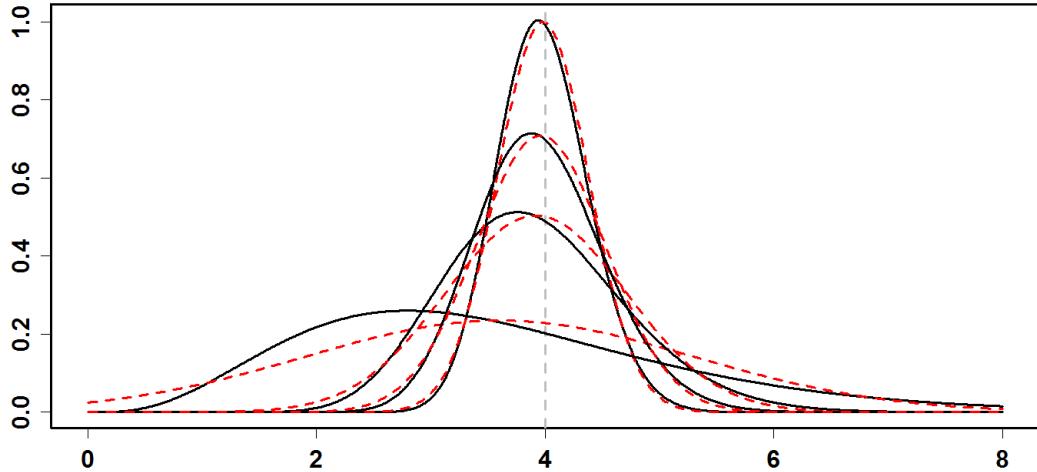


图 8.6: 若简单随机样本来自正态总体 $N(\mu, \sigma^2)$, 其中 $\sigma^2 = 4$ 。图中实线依次是样本容量 $n = 10, 50, 100, 200$ 时样本二阶中心矩 B_2 的密度函数曲线。随着 n 的增加, 曲线越来越“高瘦”, 越来越接近 $N((1 - \frac{1}{n})\sigma^2, \frac{2(n-1)}{n^2}\sigma^4)$ (虚线)。

例 8.19. 已知样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} B(m, p)$, 其中参数 m, p 都未知。由 $E(X) = mp, E(X^2) = V(X) + [E(X)]^2 = mp(1-p) + m^2p^2$, 解下方程组

$$\begin{cases} A_1 = mp, \text{ 其中 } A_1 \text{ 是样本一阶矩} \\ A_2 = mp(1-p) + m^2p^2, \text{ 其中 } A_2 \text{ 是样本二阶矩} \end{cases}$$

得到 m 和 p 的矩估计

$$\hat{m} = \frac{A_1^2}{A_1 + A_1^2 - A_2}$$

$$\hat{p} = \frac{A_1}{\hat{m}}$$

例 8.20. 设总体 X 的概率密度为 $f(x) = \begin{cases} \frac{1}{\theta_2} \exp\{-(x - \theta_1)/\theta_2\} & \text{当 } x > \theta_1 \\ 0 & \text{其他} \end{cases}$ 其中 $\theta_2 > 0$, 即 $X - \theta_1 \sim \text{Expon}(1/\theta_2)$ 。已知 X_1, X_2, \dots, X_n 是来自此总体的简单随机样本, 若参数 θ_1, θ_2 都未知, 求它们的矩估计。

解. 由指数分布 (见第 296 页的定义 4.17) 的期望和方差不难得到 $E(X) = \theta_1 +$

$\theta_2, E(X^2) = 2\theta_2^2 + 2\theta_1\theta_2 + \theta_1^2$, 列方程求得未知参数的矩估计。

$$\begin{cases} A_1 = \theta_1 + \theta_2 \\ A_2 = 2\theta_2^2 + 2\theta_1\theta_2 + \theta_1^2 \end{cases} \Rightarrow \begin{cases} \hat{\theta}_1 = \bar{X} - \sqrt{B_2} \\ \hat{\theta}_2 = \sqrt{B_2} \end{cases}$$

练习 8.5. 已知样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} U[\theta_1, \theta_2]$, 其中参数 θ_1, θ_2 都未知, 求它们的矩估计。答案: $\hat{\theta}_1 = \bar{X} - \sqrt{3B_2}, \hat{\theta}_2 = \bar{X} + \sqrt{3B_2}$ 。

例 8.21. 已知样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \text{Poisson}(\lambda)$, 其中参数 λ 未知。由 $m_1 = \lambda, m_2 = \lambda + \lambda^2$, 我们利用矩方法可得 \bar{X} 和 $\sum_{i=1}^n (X_i - \bar{X})^2/n$ 都是 λ 的矩估计。这两个统计量有着不同的量纲, 一般情况下是不同的, 选哪个作为参数 λ 的矩估计呢? 我们规定矩估计如果能通过低阶矩解决, 就不要通过高阶的。此例中, λ 的矩估计是 $\hat{\lambda} = \bar{X}$ 。

最大似然法是参数点估计理论的另一个经典方法, 最早由德国数学家 C. F. Gauss 于 1821 年提出*, 后被英国统计学家 R. A. Fisher 于 1912 年在论文《关于拟合频率曲线的一个绝对准则》中重新提及, 接着 Fisher 于 1922 年在他的一篇重要论文†《理论统计学的数学基础》中明确提出该方法(见 [48] 的第六节《估计问题的形式解》), Fisher 乐此不疲地以最大似然法挑战 K. Pearson 的权威, 不惜得罪 K. Pearson 并与之交恶, 通常文献中也把最大似然法归功于 Fisher。

Fisher 在《统计方法和科学推断》中批评 K. Pearson, “他的数学和科学工作中可怕的弱点归因于他缺乏自我批评的能力, 以及不愿承认自己有向他人学习的可能, 甚至在他知之甚少的生物学上也是如此。因此他的数学, 虽然总是充满活力, 但通常是笨拙的, 而且经常有误导。在他颇为沉迷的争论中, 他不断表现出自己缺乏公道。” Fisher 甚至嘲笑 K. Pearson 盛产垃圾论文, 既自明不凡又稀奇古怪, 基本上没啥价值。

定义 8.12 (似然函数). 设随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的密度函数为 $f_\theta(\mathbf{x})$, 其中 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$, 与密度函数差一个常数因子的任意函数 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ 被称为似然函数 (likelihood function), 即

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \propto f_\theta(\mathbf{x})$$

记作 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ 是为了突出似然函数是关于参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top \in \Theta$ 的函数。对数似然函数 (log-likelihood function) 定义为

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$$

*最大似然法的思想是简单而无懈可击的——观察到的现象多是以大概率发生的。历史上最大似然法还曾被其他很多数学家研究过, 如 J. L. Lagrange、Daniel Bernoulli (1700-1782)、L. Euler、P. S. Laplace 等等。

†另外 Fisher 还在此文中提出了充分统计量和 Fisher 信息量等关键概念。这篇经典论文 1955 年由英国皇家学会重印, 并作为 Fisher 的代表作收录于《统计学中的重大突破》第一卷 [96]。

例 8.22. 已知简单随机样本 X_1, X_2, \dots, X_n 来自于总体 $X \sim f_\theta(x)$, 则对数似然函数为

$$\ell(\boldsymbol{\theta}; x_1, x_2, \dots, x_n) = \sum_{j=1}^n \ln f_\theta(x_j)$$

定义 8.13. 给定 (对数) 似然函数, 未知 (向量) 参数 $\boldsymbol{\theta} \in \Theta$ 的最大似然估计 (maximum likelihood estimate, MLE) 定义为下述寻找最大值点的最优化问题。

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \ell(\boldsymbol{\theta}; \mathbf{x}) \quad (8.10)$$

当 Θ 为开集时极值可能达不到, 为了讨论的方便也常用 Θ 的闭包 Θ_1 来替换式 (8.10) 中的 Θ 。若在 Θ_1 的内点集 Θ_0 上, $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ 对 $\boldsymbol{\theta}$ 的各分量的一阶偏导数存在且 $\hat{\boldsymbol{\theta}} \in \Theta_0$, 则 $\hat{\boldsymbol{\theta}}$ 可通过求解似然方程组 $\partial \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) / \partial \theta_j = 0$ 或者 (对数) 似然方程组

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_j} = 0, \text{ 其中 } j = 1, 2, \dots, k$$

得到 (在解不唯一的时候, 需要判定哪个是最大值点)。然而, 似然方程组的解只是最大似然估计的“备选答案”, 有时需要讨论似然函数是否在 Θ_1 的边界上取得最大值, 式 (8.10) 的最优化问题可能很复杂。

例 8.23. 接着第 486 页的例 7.20, 对数似然函数是

$$\ell(\theta) = \ln f_\theta(x_1, x_2, \dots, x_n) = -n \ln \theta$$

然而, $d\ell(\theta)/d\theta = 0$ 给出 θ 的最大似然估计是 $\hat{\theta} = \infty$, 这显然是不对的。函数 $f_\theta(x_1, x_2, \dots, x_n)$ 在 $x_{(n)}$ 处达到最大, 因此 θ 的最大似然估计是 $\hat{\theta} = X_{(n)}$ 。

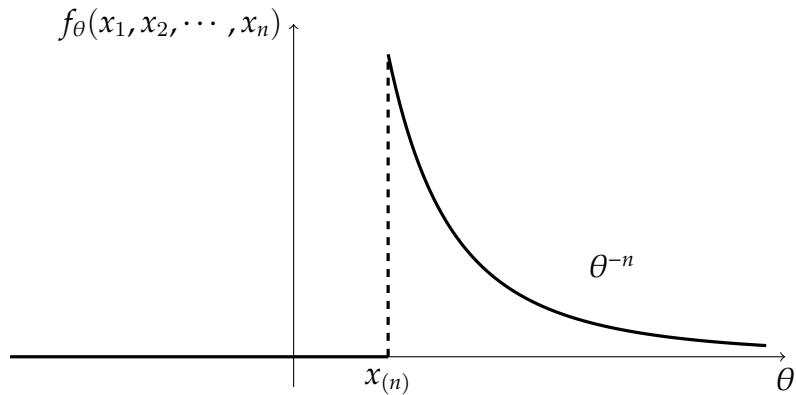


图 8.7: 密度函数 $f_\theta(x_1, x_2, \dots, x_n)$ 当 $\theta < x_{(n)}$ 时为零, 在 $x_{(n)}$ 处取得最大值。

例 8.24. 已知样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 其中 $\sigma^2 > 0$ 且 $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ 未知。由对数似然函数 $\ell(\boldsymbol{\theta}; x_1, x_2, \dots, x_n) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2 - \frac{n}{2} \ln(2\pi)$, 求解下面的似然方程组:

$$\begin{cases} \frac{1}{\sigma^2} \sum_{j=1}^n (x_j - \mu) = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^n (x_j - \mu)^2 = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 \end{cases}$$

因为对于任意的 $\mu \neq \bar{X}$, 皆有 $\sum_{j=1}^n (X_j - \mu)^2 > \sum_{j=1}^n (X_j - \bar{X})^2$ (见第 7 章课后习题), 所以

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \bar{x})^2 \right\} \geq \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} \exp \left\{ -\frac{1}{2\hat{\sigma}^2} \sum_{j=1}^n (x_j - \mu)^2 \right\}$$

而上式左端在 $\sigma^2 = \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$ 取得最大值。经过上面的验证, 所得的 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 确是 μ 和 σ^2 的最大似然估计。

练习 8.6. 在**例 8.24** 的条件下, 求未知参数的矩估计, 看是否与最大似然估计相同。
答案: 相同。

例 8.25. 令样本 $\mathbf{X} = (X_1, X_2, X_3, X_4)^\top \sim \text{Multin}(n; \frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4} - \frac{1}{4}\theta, \frac{1}{4} - \frac{1}{4}\theta, \frac{1}{4}\theta)$, 已知样本值 $\mathbf{x} = (x_1, x_2, x_3, x_4)^\top = (125, 18, 20, 34)^\top$, 求参数 θ 的最大似然估计。

解. 由题意, 随机向量 \mathbf{X} 的密度函数为

$$\frac{(x_1 + x_2 + x_3 + x_4)!}{x_1! x_2! x_3! x_4!} \left(\frac{1}{2} + \frac{1}{4}\theta \right)^{x_1} \left(\frac{1}{4} - \frac{1}{4}\theta \right)^{x_2} \left(\frac{1}{4} - \frac{1}{4}\theta \right)^{x_3} \left(\frac{1}{4}\theta \right)^{x_4}$$

定义似然函数 $\mathcal{L}(\theta; \mathbf{x}) = (2 + \theta)^{x_1} (1 - \theta)^{x_2 + x_3} \theta^{x_4}$, 则对数似然函数

$$\ell(\theta; \mathbf{x}) = x_1 \ln(2 + \theta) + (x_2 + x_3) \ln(1 - \theta) + x_4 \ln \theta$$

此函数在 $(0, 1)$ 上是单峰函数, 通过 $d\ell(\theta; \mathbf{x})/d\theta = 0$ 得到 θ 的最大似然估计 $\hat{\theta} \approx 0.6268215$ 。

性质 8.2. 若统计量 $T(\mathbf{X})$ 对 $\boldsymbol{\theta}$ 而言是充分的, 并且 $\boldsymbol{\theta}$ 的最大似然估计 $\hat{\boldsymbol{\theta}}$ 通过对数似然方程组解得, 那么它一定是 $T(\mathbf{X})$ 的函数。

证明. 因为 $T(\mathbf{X})$ 是充分统计量, 所以似然函数 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = h(\mathbf{x})g_{\boldsymbol{\theta}}[T(\mathbf{x})]$ 。最大似然估

计 $\hat{\boldsymbol{\theta}}$ 是 $T(\mathbf{X})$ 的函数, 这是因为

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_j} = 0 \Rightarrow \frac{\partial \ln g_{\boldsymbol{\theta}}[T(\mathbf{x})]}{\partial \theta_j} = 0, \text{ 其中 } j = 1, 2, \dots, k \quad \square$$

定理 8.8. 如果参数空间 $\Theta_0 \subseteq \mathbb{R}^k$ 为开凸集, (对数) 似然函数 $\ell(\boldsymbol{\theta}; \mathbf{x})$ 在 Θ_0 上对 $\boldsymbol{\theta}$ 存在一阶和二阶偏导数, 并且 $\forall \boldsymbol{\theta} \in \Theta_0$ 皆有 $-\nabla_{\boldsymbol{\theta}}^2 \ell(\boldsymbol{\theta}; \mathbf{x})$ 为正定矩阵, 则对数似然方程组的解 (若存在) 即为 $\boldsymbol{\theta}$ 的最大似然估计。

※证明. 见附录 E 中的定理 E.12 及其证明。 \square

定理 8.9. 已知参数空间 Θ 为一个开凸集, 设样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^{\top}$ 的密度函数为

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = h(\mathbf{x}) \exp \left\{ \sum_{j=1}^k \theta_j T_j(\mathbf{x}) + g(\boldsymbol{\theta}) \right\}, \text{ 其中 } \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$$

□ 随机向量 $\mathbf{T} = (T_1(\mathbf{X}), \dots, T_k(\mathbf{X}))^{\top}$ 的协方差矩阵为

$$\text{Cov}(\mathbf{T}, \mathbf{T}) = -\nabla_{\boldsymbol{\theta}}^2 g(\boldsymbol{\theta})$$

□ 若 $\text{Cov}(\mathbf{T}, \mathbf{T})$ 正定, 则对数似然方程组的解 (若存在) 即为 $\boldsymbol{\theta}$ 的最大似然估计。

※证明. 第一个命题的证明参见陈希孺的《高等数理统计学》[168] 的定理 1.1。在第二个命题的条件之下有 $\text{Cov}(\mathbf{T}, \mathbf{T}) = -\nabla_{\boldsymbol{\theta}}^2 g(\boldsymbol{\theta})$, 于是 $-\nabla_{\boldsymbol{\theta}}^2 \ell(\boldsymbol{\theta}; \mathbf{x})$ 为正定矩阵, 由定理 8.8 即可证得定理 8.9。 \square

例 8.26. 设简单随机样本 $(X_1, Y_1)^{\top}, \dots, (X_n, Y_n)^{\top}$ 来自二元正态总体 $N(0, 0, \sigma^2, \sigma^2, \rho)$, 其中 $|\rho| < 1$ 和 $0 < \sigma^2 < \infty$ 未知。试求参数 ρ, σ^2 的最大似然估计。

解. 在开凸集 $(0, 1) \times (0, \infty)$ 上, 似然函数为

$$\mathcal{L} = \left(2\pi\sigma^2 \sqrt{1-\rho^2} \right)^{-n} \exp \left\{ -\frac{\sum_{j=1}^n (x_j^2 + y_j^2 - 2\rho x_j y_j)}{2\sigma^2(1-\rho^2)} \right\}$$

引入新参数 $\theta_1 = -[2\sigma^2(1-\rho^2)]^{-1}$ 和 $\theta_2 = \rho[\sigma^2(1-\rho^2)]^{-1}$, 则上式简化为 θ_1, θ_2 和 $T_1 = \sum_{j=1}^n (x_j^2 + y_j^2), T_2 = \sum_{j=1}^n x_j y_j$ 的函数, 即

$$\mathcal{L} = \exp\{\theta_1 T_1 + \theta_2 T_2 + g(\theta_1, \theta_2)\}$$

$$\text{其中, } g(\theta_1, \theta_2) = \ln[(2\pi)^{-n} (4\theta_1^2 - \theta_2^2)^{-n/2}]$$

经验证满足定理 8.9 的条件, 对参数的最大似然估计可由对数似然方程组解得。将对数似然方程组和参数的逆变换联立可得参数 ρ, σ^2 的最大似然估计。

$$\begin{cases} -\frac{4n\theta_1}{4\theta_1^2 - \theta_2^2} = T_1 \\ \frac{2n\theta_2}{4\theta_1^2 - \theta_2^2} = T_2 \end{cases} + \begin{cases} \rho = -\frac{\theta_2}{\theta_1} \\ \sigma^2 = -\frac{2\theta_1}{4\theta_1^2 - \theta_2^2} \end{cases} \Rightarrow \begin{cases} \hat{\sigma}^2 = \frac{1}{2n} \left(\sum_{j=1}^n X_j^2 + \sum_{j=1}^n Y_j^2 \right) \\ \hat{\rho} = \frac{2 \sum_{j=1}^n X_j Y_j}{\sum_{j=1}^n X_j^2 + \sum_{j=1}^n Y_j^2} \end{cases}$$

例 8.27. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} U[\theta, 0]$, 其中未知参数 $\theta \in \Theta = (-\infty, 0)$, 试求 θ 的最大似然估计。

解. 似然函数 $\mathcal{L}(\theta; x_1, \dots, x_n) = \begin{cases} (-\theta)^{-n} & \text{当 } \theta \leq x_1, \dots, x_n \leq 0 \\ 0 & \text{其他} \end{cases}$

当 θ 取 $x_{(1)} = \min_{1 \leq j \leq n} x_j$ 时, \mathcal{L} 达到最大, 于是 θ 的最大似然估计为 $\hat{\theta} = \min_{1 \leq j \leq n} X_j$ 。

例 8.28. 设样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} U[\theta - 1/2, \theta + 1/2]$, 其中 θ 是未知参数, 试求 θ 的最大似然估计。

解. 似然函数 $\mathcal{L}(\theta; x_1, \dots, x_n) = \begin{cases} 1 & \text{当 } \theta - 1/2 \leq x_1, \dots, x_n \leq \theta + 1/2 \\ 0 & \text{其他} \end{cases}$

记 $x_{(1)} = \min(x_1, x_2, \dots, x_n), x_{(n)} = \max(x_1, x_2, \dots, x_n)$, 则

$$\theta - 1/2 \leq x_{(1)} \leq x_{(n)} \leq \theta + 1/2 \Leftrightarrow x_{(n)} - 1/2 \leq \theta \leq x_{(1)} + 1/2$$

于是满足 $x_{(n)} - 1/2 \leq T(X_1, X_2, \dots, X_n) \leq x_{(1)} + 1/2$ 的每个统计量 $T(X_1, X_2, \dots, X_n)$ 都是 θ 的最大似然估计, 如 $x_{(n)} - 1/2 + \alpha[1 + x_{(1)} - x_{(n)}]$, 其中 $0 < \alpha < 1$ 。

例 8.29. 求第 513 页的例 8.20 中未知参数 θ_1, θ_2 的最大似然估计。

解. 似然函数为

$$\mathcal{L}(\theta_1, \theta_2; x_1, \dots, x_n) = \theta_2^{-n} \exp \left\{ -\frac{1}{\theta_2} \sum_{k=1}^n (x_k - \theta_1) \right\}$$

其中 $x_k \geq \theta_1, k = 1, 2, \dots, n$ 。从而得到似然方程组

$$\begin{cases} \frac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_1} = \frac{n}{\theta_2} = 0 \\ \frac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{\theta_2} + \frac{1}{\theta_2^2} \sum_{k=1}^n (x_k - \theta_1) = 0 \end{cases}$$

由第二式可得 $\hat{\theta}_2 = \bar{X} - \hat{\theta}_1$, 但无论 θ_1 取何值都不能使第一式成立。为了使 $\mathcal{L}(\theta_1, \theta_2; x_1, \dots, x_n)$ 达到最大就要选 $\hat{\theta}_1 = \min_{1 \leq k \leq n} X_k$ 。

最大似然估计通常很难计算, 其相合性的证明也相当复杂 (若总体是指数族的, 在一般条件下最大似然估计是相合的)。1946 年, H. Cramér 在《统计学数学方法》[29] 中首次证明了在一定条件之下最大似然估计的弱相合性和渐近正态性。

定理 8.10 (Cramér, 1946). 在某些正则条件^{*}下, 最大似然估计 $\hat{\theta}_n$ 是相合的, 并且满足渐近正态性, 即

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \mathcal{I}^{-1}(\theta)), \text{ 其中 } \mathcal{I}(\theta) \text{ 是 } \theta \text{ 的 Fisher 信息量}$$

这个结果可以推广到高维未知参数 $\boldsymbol{\theta}$ 的估计, 即

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta})), \text{ 其中 } \mathcal{I}(\boldsymbol{\theta}) \text{ 是 } \boldsymbol{\theta} \text{ 的 Fisher 信息矩阵}$$

满足 Cramér 定理 8.10 条件的最大似然估计 $\hat{\theta}_n$ 是渐近无偏的并且也是渐近有效的, 即当样本量 $n \rightarrow \infty$ 时, $V(\hat{\theta}_n)$ 趋近 Cramér-Rao 下界, 而一般情况下矩估计的方差不是渐近有效的, 从这个角度比较最大似然估计比矩估计略胜一筹。有关最大似然估计强相合性的工作是由 Wald 于 1949 年做出的, 本书不作介绍。

例 8.30. 接着例 8.26, 分别求未知参数 σ^2, ρ 的最大似然估计的均方误差。

解. 统计量 $\hat{\sigma}^2, \hat{\rho}$ 都是渐近无偏的。令 $\boldsymbol{\theta} = (\sigma^2, \rho)^T$, 计算未知参数 $\boldsymbol{\theta}$ 的 Fisher 信息矩阵及其逆矩阵如下,

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= \frac{1}{1 - \rho^2} \begin{pmatrix} \frac{1 - \rho^2}{\sigma^4} & -\frac{\rho}{\sigma^2} \\ -\frac{\rho}{\sigma^2} & \frac{1 + \rho^2}{1 - \rho^2} \end{pmatrix} \\ \mathcal{I}^{-1}(\boldsymbol{\theta}) &= \begin{pmatrix} (1 + \rho^2)\sigma^4 & \rho(1 - \rho^2)\sigma^2 \\ \rho(1 - \rho^2)\sigma^2 & (1 - \rho^2)^2 \end{pmatrix} \end{aligned}$$

利用定理 8.10, 当样本容量 n 很大时, $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ 渐近服从二元正态分布, 其

*详见 P. J. Bickel 和 K. A. Doksum 的《数理统计: 基本思想和选题》[12] 第五章《渐近近似》或 Cramér 的《统计学数学方法》。

分量也都是渐近正态分布：

$$\begin{aligned}\sqrt{n}(\hat{\sigma}^2 - \sigma^2) &\xrightarrow{L} N(0, (1 + \rho^2)\sigma^4) \\ \sqrt{n}(\hat{\rho} - \rho) &\xrightarrow{L} N(0, (1 - \rho^2)^2) \\ \text{进而, } \text{MSE}(\sigma^2, \hat{\sigma}^2) &= \frac{(1 + \rho^2)\sigma^4}{n} \\ \text{MSE}(\rho, \hat{\rho}) &= \frac{(1 - \rho^2)^2}{n}\end{aligned}$$

显然, 样本容量越大均方误差越小。同样样本容量之下, 相关系数越接近 ± 1 , $\hat{\rho}$ 的均方误差越小; 对 $\hat{\sigma}^2$ 来说恰恰相反, 相关系数越接近 0, $\hat{\sigma}^2$ 的均方误差越小。因为 $0 \leq \rho^2 \leq 1$, 上述均方误差总有上下界如下:

$$\begin{aligned}\frac{1}{n}\sigma^4 &\leq \text{MSE}(\sigma^2, \hat{\sigma}^2) \leq \frac{2}{n}\sigma^4 \\ 0 &\leq \text{MSE}(\rho, \hat{\rho}) \leq \frac{1}{n}\end{aligned}$$

假设对数似然函数 $\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \sum_{j=1}^n \ln f(x_j | \boldsymbol{\theta})$ 的极大值点^{*} $\hat{\boldsymbol{\theta}} \in \Theta_1$ 是 $\boldsymbol{\theta}$ 的最大似然估计, 考虑 $\ell(\boldsymbol{\theta}; \mathbf{x})$ 在 $\hat{\boldsymbol{\theta}}$ 处的 Taylor 级数展开。

$$\begin{aligned}\ell(\boldsymbol{\theta}; \mathbf{x}) &\approx \ell(\hat{\boldsymbol{\theta}}; \mathbf{x}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \left. \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}} \right|_{\hat{\boldsymbol{\theta}}} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \left. \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}^2} \right|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &= \ell(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \left. \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}^2} \right|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\end{aligned}$$

定义 8.14. 我们称下面的正定矩阵 $\hat{\mathcal{I}}(\mathbf{x})$ 为观测的 Fisher 信息矩阵,

$$\hat{\mathcal{I}}(\mathbf{x}) = - \left. \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}^2} \right|_{\hat{\boldsymbol{\theta}}}$$

性质 8.3. 观测的 Fisher 信息矩阵 $\hat{\mathcal{I}}(\mathbf{x})$ 的 (i, j) 元素为

$$\hat{\mathcal{I}}_{ij}(\mathbf{x}) = - \sum_{k=1}^n \left. \frac{\partial^2 \ln f(x_k | \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

定理 8.11. 令 $\hat{\mathcal{I}}(\mathbf{x})$ 为观测的 Fisher 信息矩阵, 条件如上所述, 则

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \approx \phi(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}, [\hat{\mathcal{I}}(\mathbf{x})]^{-1})$$

^{*}如果解析地求 $\ell(\boldsymbol{\theta}; \mathbf{x})$ 的极大值点比较困难, 可以利用 Newton-Raphson 算法 (见第 773 页的算法 E.2) 求得极大值点的数值解。

如果样本容量 n 足够地大，亦有

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \approx \phi(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}, [\mathcal{I}(\hat{\boldsymbol{\theta}})]^{-1})$$

证明. 参数 $\boldsymbol{\theta}$ 的信息矩阵的 (i, j) 元素为

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = -n E_{\boldsymbol{\theta}} \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(X_1 | \boldsymbol{\theta}) \right\}$$

如果样本容量 n 足够地大，则 $\mathcal{I}_{ij}(\hat{\boldsymbol{\theta}}) \approx \hat{\mathcal{I}}_{ij}(\mathbf{x})$ 。 \square

按照频率派的观点，未知参数 $\boldsymbol{\theta}$ 是某一固定值，其最大似然估计 $\hat{\boldsymbol{\theta}}$ 是样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ 定义的随机向量，渐近地有

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, [\mathcal{I}(\boldsymbol{\theta})]^{-1})$$

8.2 Neyman 置信区间估计

频率派坚信未知参数是固定值，只是估计者尚不知道它而已。所以，频率派的置信区间估计就是利用样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 构造两个统计量 $\bar{\theta}(\mathbf{X})$ 和 $\underline{\theta}(\mathbf{X})$ 满足 $\underline{\theta}(\mathbf{X}) \leq \bar{\theta}(\mathbf{X})$ ，如果区间 $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ 以一个大的概率（譬如，95%）覆盖住未知参数 θ ，则称它为 θ 的一个区间估计。因为 $\underline{\theta}(\mathbf{X})$ 和 $\bar{\theta}(\mathbf{X})$ 是由样本 \mathbf{X} 构造的随机变量，区间 $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ 覆盖住未知参数 θ 的概率 $P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta \leq \bar{\theta}(\mathbf{X})\}$ 和样本具体取什么值没有半点关系。换句话说，没有观察数据，频率派也能够谈论 $P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta \leq \bar{\theta}(\mathbf{X})\}$ 。在观察到样本值 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 之后，区间 $[\underline{\theta}(\mathbf{x}), \bar{\theta}(\mathbf{x})]$ 要么覆盖住 θ ，要么未覆盖住 θ ，没有任何随机性可言。频率派区间估计的价值体现在大量重复的试验观察中，区间 $[\underline{\theta}(\mathbf{x}), \bar{\theta}(\mathbf{x})]$ 经常能够覆盖住未知参数*。

频率派统计学家跟“上帝”玩一个游戏：上帝选 θ ，然后根据分布 F_θ 产生样本 \mathbf{X} ，统计学家由此构造区间 $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ ，若该区间覆盖住 θ 则统计学家赢，否则统计学家输。统计学家该如何构造这个区间才能在大量独立重复的游戏中以一个大概率赢呢？

基于上述频率派区间估计的想法，波兰统计学家 Jerzy Neyman (1894-1981) 于 1934-1937 年间提出了置信区间 (confidence interval) 的理论 [114]。当年，Neyman 的这项研究工作受到了一些质疑和误解，其著名论文《基于经典概率论的统计估计理论纲要》[115] 颇费周折终于在 1937 年得以发表†。如今，Neyman 的置信区间估计理论已成为经典统计学的一部分，其基本思想就是用样本构造一个随机区间的上下限，使得该区间覆盖未知参数的概率不小于给定的正数 $1 - \alpha$ ，其中 $0 < \alpha < 1$ 。值得注意的是，如果对区间长度不作要求，置信区间一般不唯一。



定义 8.15 (置信度). 如果 $\forall \theta \in \Theta$ 皆有

$$P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta \leq \bar{\theta}(\mathbf{X})\} \geq 1 - \alpha, \text{ 其中常数 } \alpha \in (0, 1)$$

则称区间估计 $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ 具有置信水平或置信度 (confidence level) $1 - \alpha$ ，或称 $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ 是 θ 的置信度为 $1 - \alpha$ 的置信区间 (confidence interval)。通常 α 是一个接近 0 的正实数，如 $\alpha = 0.05, 0.01$ 等。

* 贝叶斯学派把未知参数 θ 视为随机变量，所以贝叶斯区间估计只谈论 θ 落于固定区间 $[\underline{\theta}(\mathbf{x}), \bar{\theta}(\mathbf{x})]$ 的概率。在贝叶斯学派看来，没有观察数据就没法讨论区间估计。

† 1930 年，Fisher 曾提出过未知参数的信任区间估计。Neyman 的置信区间理论多少受其影响，当时有很多统计学家将二者混淆。因为 Neyman 的置信区间理论和之前提出的假设检验理论是相通的，它们逐渐赢得了频率派多数的统计学家的认可而成为经典统计学的重要组成部分。而 Fisher 的信任区间理论由于种种原因却没能流行起来，详情见 §8.2.4。

显然, 对于任何 $\beta > \alpha$ 皆有 $P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta \leq \bar{\theta}(\mathbf{X})\} \geq 1 - \beta$, 即 $1 - \beta$ 也是置信度。于是, 就有了下面的概念。

定义 8.16 (置信系数). 对于未知参数 θ 的区间估计 $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$, 其置信度中最大者被称为置信系数 (confidence coefficient), 即

$$\inf_{\theta \in \Theta} P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta \leq \bar{\theta}(\mathbf{X})\}$$

某些具体问题所关心的置信区间是半开半闭的, 如电子元件的寿命问题只关心寿命的下限, 置信区间形如 $[\underline{\theta}(\mathbf{X}), \infty)$ 。

定义 8.17 (置信上限与置信下限). 如果 $\forall \theta \in \Theta$ 皆有

$$P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta\} \geq 1 - \alpha, \text{ 其中常数 } \alpha \in (0, 1)$$

$$P_\theta\{\theta \leq \bar{\theta}(\mathbf{X})\} \geq 1 - \beta, \text{ 其中常数 } \beta \in (0, 1)$$

则称 $\underline{\theta}(\mathbf{X})$ 是 θ 的置信度为 $1 - \alpha$ 的置信下限, 称 $\bar{\theta}(\mathbf{X})$ 是置信度为 $1 - \beta$ 的置信上限。

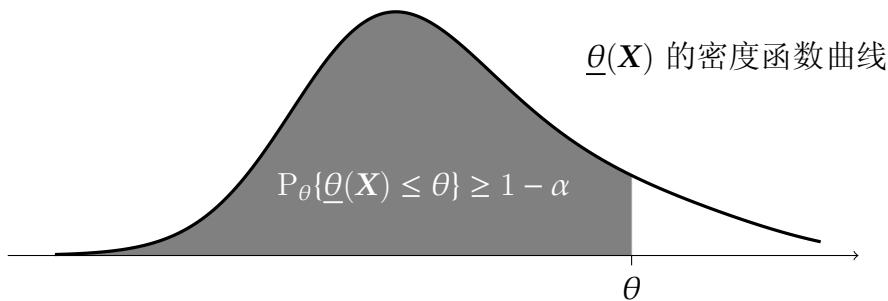


图 8.8: 未知参数 θ 的置信度为 $1 - \alpha$ 的置信下限 $\underline{\theta}(\mathbf{X})$ 满足 $P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta\} \geq 1 - \alpha$, 即图中阴影部分的面积大于 $1 - \alpha$ 。

本节内容

Neyman 的置信区间理论, 它与第 9 章即将介绍的 Neyman-Pearson 假设检验理论有着密切联系, 通过随机试验 (见图 8.13) 读者可以了解置信区间的概率意义。第一小节是基于 Markov 不等式的区间估计, 该方法是普适的, 但较粗糙。枢轴量法和大样本方法也是较常见的区间估计方法, 第二小节对它们进行了介绍。第三小节是对 Fisher 信任区间估计的简介。

关键知识

- (1) 置信区间; (2) 基于 Markov 不等式的区间估计; (3) 枢轴量法; (4) 大样本时, 置信区间的近似估计; (5) 信任区间。

8.2.1 基于 Markov 不等式的区间估计

利用 Markov 不等式 (2.72) 和参数 θ 的点估计 $\hat{\theta} = \hat{\theta}(\mathbf{X})$ 可以给出未知参数 θ 的置信区间估计, 不要求 $\hat{\theta}$ 是无偏的, 仅要求 $V_\theta(\hat{\theta}) < \infty$ 。该方法的优点是适用范围广泛, 缺点是所给出的置信区间较粗, 因为 Markov 不等式不擅长计算概率。

性质 8.4. 令 $\hat{\theta} = \hat{\theta}(\mathbf{X})$ 是未知参数 θ 的点估计, 则

$$P_\theta\{(\hat{\theta} - \theta)^2 \leq \epsilon^2 E_\theta(\hat{\theta} - \theta)^2\} \geq 1 - \frac{1}{\epsilon^2} \quad (8.11)$$

证明. 令 $Y = (\hat{\theta} - \theta)^2$ 且 $k = \epsilon^2 E(Y)$, 代入到 $P(Y \leq k) \geq 1 - E(Y)/k$ 中得证。 \square

算法 8.1. 有两种方法可以从 (8.11) 得到 θ 的置信度为 $1 - 1/\epsilon^2$ 的置信区间:

- 式 (8.11) 中的 $E_\theta(\hat{\theta} - \theta)^2$ 即 $\hat{\theta}$ 的均方误差 $MSE(\theta, \hat{\theta})$, 如果需要, 对其做一个适当的放大使其不再含有 θ 。或者,
- 直接求解式 (8.11) 中有关 θ 的不等式 $(\hat{\theta} - \theta)^2 \leq \epsilon^2 E_\theta(\hat{\theta} - \theta)^2$ 。

基于 Markov 不等式的区间估计方法是普适的, 在很弱的条件下就可以得到解。但有些时候, 该方法给出的结果略显粗糙。

例 8.31. 已知样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, 其中参数 μ 未知, $\sigma^2 > 0$ 已知。未知参数 μ 的最大似然估计是 $\bar{X} \sim N(\mu, \sigma^2/n)$ 。利用式 (8.11),

$$P\left\{\bar{X} - \frac{\sigma}{\sqrt{n}\alpha} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}\alpha}\right\} \geq 1 - \alpha$$

即未知参数 μ 的置信度为 $1 - \alpha$ 的置信区间是 $[\bar{X} - \sigma/\sqrt{n}\alpha, \bar{X} + \sigma/\sqrt{n}\alpha]$ 。请读者将本例的结果与性质 8.5 的进行比较, 看看哪个更优。

例 8.32. 已知样本 $X_1, \dots, X_n \stackrel{iid}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 参数 p 未知。显然, 未知参数 p 的无偏估计量 \bar{X} 使得 $MSE(p, \bar{X}) = V_p(\bar{X})$ 。

(1) 下面, 适当地扩大 $V_p(\bar{X})$ 使之不含 p 。

$$V_p(\bar{X}) = \frac{p(1-p)}{n} \leq \frac{1}{4n}$$

未知参数 p 的取值范围是 $[0, 1]$, 由式 (8.11) 得到参数 p 的置信度为 $1 - 1/\epsilon^2$ 的置信区间如下。

$$[0, 1] \cap \left[\bar{X} - \frac{\epsilon}{2\sqrt{n}}, \bar{X} + \frac{\epsilon}{2\sqrt{n}}\right]$$

(2) 将不等式 $(\hat{\theta} - \theta)^2 \leq \epsilon^2 E_\theta(\hat{\theta} - \theta)^2$ 整理为有关 θ 的不等式, 即

$$(\bar{X} - p)^2 \leq \frac{\epsilon^2 p(1-p)}{n} \Leftrightarrow \left(1 + \frac{\epsilon^2}{n}\right)p^2 - \left(2\bar{X} + \frac{\epsilon^2}{n}\right)p + \bar{X}^2 \leq 0$$

上式右端关于 p 的二次方程总存在两个不同的非负实根, 不妨设为 $p_1(\bar{X}) < p_2(\bar{X})$, 则 $P\{p_1(\bar{X}) \leq p \leq p_2(\bar{X})\} \geq 1 - 1/\epsilon^2$ 。区间 $[p_1(\bar{X}), p_2(\bar{X})]$ 是对 p 更精细的区间估计, 但解的具体形式比较复杂。

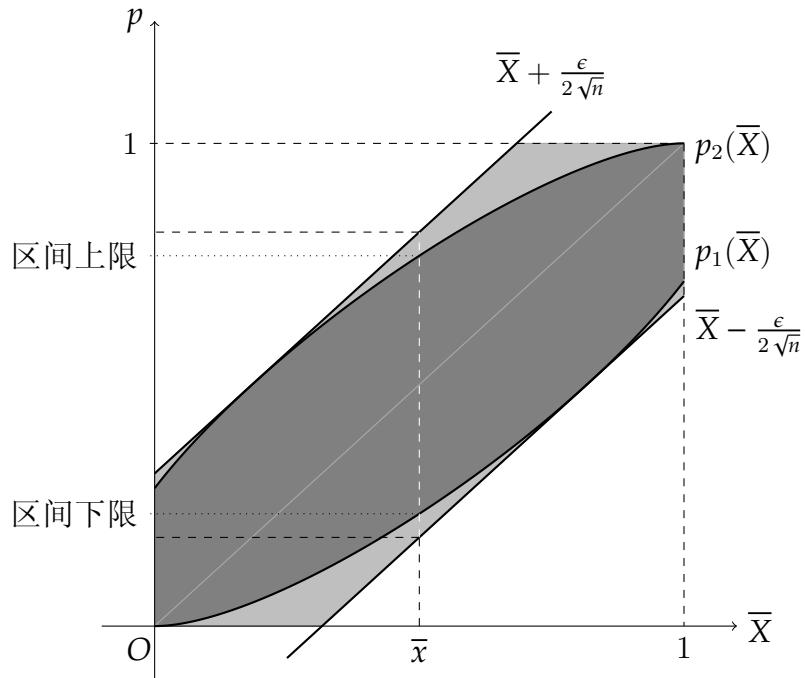


图 8.9: 在例 8.32 中, 利用 Markov 不等式给出的 p 的置信度为 $1 - 1/\epsilon^2$ 的置信区间估计, 显然第二种方法比第一种方法的效果要好一些。

8.2.2 枢轴量法

定义 8.18 (枢轴量). 如果简单随机样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 和未知参数 θ 的函数 $Y = h_\theta(\mathbf{X})$ 的分布与 θ 无关, 则称 $h_\theta(\mathbf{X})$ 为枢轴量 (pivot)。

算法 8.2. 枢轴量法的关键就是选择合适的枢轴量 $Y = h_\theta(\mathbf{X})$, 表达式中含有未知参数 θ , 但分布与 θ 无关。接着, 置信区间或置信限构造如下。

- 首先, 找到实数 $c_1 < c_2$ 使得

$$P\{c_1 \leq h_\theta(\mathbf{X}) \leq c_2\} \geq 1 - \alpha \quad (8.12)$$

一般地, 在式 (8.12) 中, 取 $c_1 = q_{\alpha/2}, c_2 = q_{1-\alpha/2}$, 其中 $q_{1-\alpha/2}$ 是 Y 的 $(1-\alpha/2)$ -分位数。特别地, 当枢轴量 $Y = h_\theta(\mathbf{X})$ 的密度函数关于 0 对称, 由式 (2.27),

$$c_1 = q_{\alpha/2} = -q_{1-\alpha/2}$$

- 然后, 解不等式 $c_1 \leq h_\theta(\mathbf{X}) \leq c_2$ 得到 $\underline{\theta}(\mathbf{X}) \leq \theta \leq \bar{\theta}(\mathbf{X})$ 即是 θ 的置信度为 $1 - \alpha$ 的置信区间。

性质 8.5. 已知样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 下面分别在不同的情况之下, 利用枢轴量法对参数 μ 和 σ^2 进行区间估计。

- ① 参数 μ 未知, 但参数 σ^2 已知。考虑枢轴量 $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$, 其密度函数关于 $z = 0$ 对称。令 $c_1 = -z_{1-\alpha/2}, c_2 = z_{1-\alpha/2}$, 它们可满足式 (8.12), 其中 $z_{1-\alpha/2}$ 是 $N(0, 1)$ 分布的 $(1 - \alpha/2)$ -分位数。于是, 得到 μ 的置信度为 $1 - \alpha$ 的置信区间

$$\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (8.13)$$

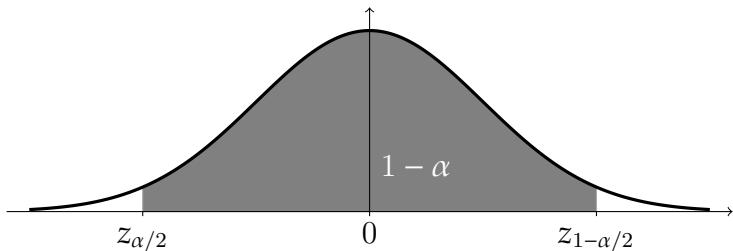


图 8.10: 枢轴量 $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$, 式 (8.12) 中 $c_1 = z_{\alpha/2} = -z_{1-\alpha/2}, c_2 = z_{1-\alpha/2}$ 。

② 参数 μ 已知, 但参数 σ^2 未知。考虑枢轴量 $Y = \sum_{j=1}^n (X_j - \mu)^2 / \sigma^2 \sim \chi_n^2$, 其密度函数非对称。令 $c_1 = \chi_{n,\alpha/2}^2, c_2 = \chi_{n,1-\alpha/2}^2$, 它们可满足式 (8.12), 其中 $\chi_{n,1-\alpha/2}^2$ 是 χ_n^2 分布的 $(1 - \alpha/2)$ -分位数。于是, 得到 σ^2 的置信度为 $1 - \alpha$ 的置信区间

$$\frac{\sum_{j=1}^n (X_j - \mu)^2}{\chi_{n,1-\alpha/2}^2} \leq \sigma^2 \leq \frac{\sum_{j=1}^n (X_j - \mu)^2}{\chi_{n,\alpha/2}^2} \quad (8.14)$$

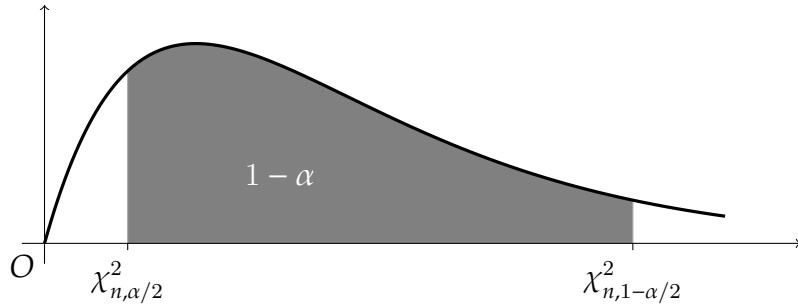


图 8.11: 枢轴量 $Y = \sum_{j=1}^n (X_j - \mu)^2 / \sigma^2 \sim \chi_n^2$, 式 (8.12) 中 $c_1 = \chi_{n,\alpha/2}^2, c_2 = \chi_{n,1-\alpha/2}^2$ 。

③ 参数 μ, σ^2 都未知。考虑枢轴量 $T = \sqrt{n}(\bar{X} - \mu) / S \sim t_{n-1}$, 其密度函数关于 $t = 0$ 对称。令 $c_1 = -t_{n-1,1-\alpha/2}, c_2 = t_{n-1,1-\alpha/2}$, 它们可满足式 (8.12), 其中 $t_{n-1,1-\alpha/2}$ 是 t_{n-1} 分布的 $(1 - \alpha/2)$ -分位数。于是, 得到 μ 的置信度为 $1 - \alpha$ 的置信区间

$$\bar{X} - t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}} \quad (8.15)$$

在这个区间估计中, 未知参数 σ^2 没有出现, 被称为冗余参数 (nuisance parameter)。

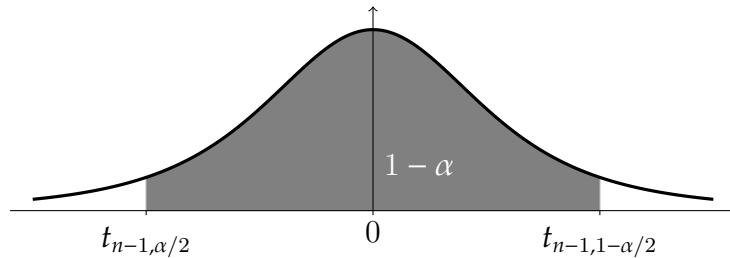


图 8.12: 枢轴量 $T = \sqrt{n}(\bar{X} - \mu) / S \sim t_{n-1}$, 式 (8.12) 中 $c_1 = -t_{n-1,1-\alpha/2}, c_2 = t_{n-1,1-\alpha/2}$ 。

考虑枢轴量 $(n-1)S^2 / \sigma^2 \sim \chi_{n-1}^2$, 不难得到 σ^2 的置信度为 $1 - \alpha$ 的置信区间 (μ

是冗余参数)

$$\frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} \quad (8.16)$$

 参数 θ 的置信度为 $1 - \alpha$ 的置信区间 $[\underline{\mu}(\mathbf{x}), \bar{\mu}(\mathbf{x})]$ 并不是指这个区间以（至少） $1 - \alpha$ 的概率覆盖住 θ 。事实上，区间 $[\underline{\mu}(\mathbf{x}), \bar{\mu}(\mathbf{x})]$ 是否覆盖住 θ 并无随机性。必须通过独立的重复试验——不断地从总体中抽取容量为 n 的简单随机样本，利用覆盖率赋予置信度以概率含义。所以，置信度 $1 - \alpha$ 的意义在于随机区间 $[\underline{\mu}(\mathbf{X}), \bar{\mu}(\mathbf{X})]$ 覆盖住未知参数 θ 的概率，而与具体观察到的样本值 \mathbf{x} 并无多大关系。置信度 $1 - \alpha$ 就像 $[\underline{\mu}(\mathbf{x}), \bar{\mu}(\mathbf{x})]$ 的出身证明，频率派拿只在假想试验中存在而实际尚未观察到的数据为当前的估计结果 $[\underline{\mu}(\mathbf{x}), \bar{\mu}(\mathbf{x})]$ “撑腰助威”的这一作法常被贝叶斯学派诟病。

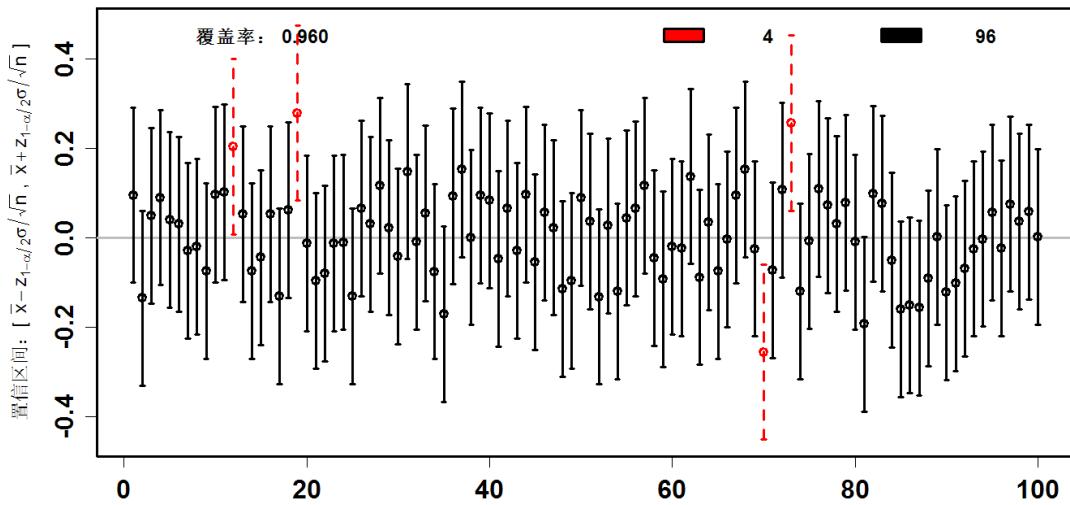


图 8.13: 置信度为 $1 - \alpha$ 的置信区间的频率解释：假设总体为 $N(0, 1)$ ，其中方差已知，均值 μ 未知。利用性质 8.5 的第一种情况，求得 μ 的置信度为 95% 的置信区间 $[\underline{\mu}(\mathbf{x}), \bar{\mu}(\mathbf{x})]$ 。在 100 次独立重复的随机试验中，发现有 96 次覆盖住 $\mu = 0$ （图中竖直实线），4 次未覆盖住 $\mu = 0$ （图中竖直虚线）。

例 8.33. 设食品厂生产的某袋装食品的重量服从正态分布 $N(\mu, \sigma^2)$ ，参数 μ, σ^2 都未知。现随机抽取 20 袋食品测得重量 0.1126, 0.0860, 0.0954, 0.0861, 0.0907, 0.0971, 0.1038, 0.1012, 0.1043, 0.1022, 0.0952, 0.1072, 0.0909, 0.1176, 0.1069, 0.1032, 0.1058, 0.1043, 0.1125, 0.0952 千克，分别求 μ, σ^2 的置信度为 95% 的置信区间。

解. 分别利用性质 8.5 中第二和第四种情况求未知参数置信度为 95% 的置信区间。

$$\mu \in [0.0968, 0.1050]，\text{并且} \sigma^2 \in [4.3936 \times 10^{-5}, 1.6206 \times 10^{-4}]$$

例 8.34. 已知样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(\beta)$, 其中参数 β 未知, 求 β 的置信度为 $1 - \alpha$ 的置信区间和置信下限。

解. 令 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$, 由第 473 页的**例 7.2** 中的结果, 考虑 $2n\beta\bar{X} \sim \chi^2_{2n}$ 是枢轴量。于是, 未知参数 β 的置信度为 $1 - \alpha$ 的置信区间是

$$\frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}} \leq \beta \leq \frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}$$

另外, β 的置信度为 $1 - \alpha$ 的置信下限是 $\chi^2_{2n,1-\alpha}/(2n\bar{X})$ 。

在参数 θ 所有可能的置信区间中, 人们最希望得到的是那个长度最短的区间 $[\underline{\theta}, \bar{\theta}]$, 因为它对参数的估计最精确。但有时候为了顾及形式上的简单, 如**性质 8.5** 中第三、四种情况, 并不奢求最短的区间。

例 8.35. 考虑**性质 8.5** 中的第一种情况, 下面验证它就是最短的置信区间。令 $a < b$, 其中 b 是 a 的函数, 它们满足

$$G(a) = P \left\{ a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b \right\} = \int_a^b \phi(t) dt = 1 - \alpha$$

显然, $dG/da = 0$ 。另外, 为了使得区间 $[\bar{X} - b\sigma/\sqrt{n}, \bar{X} - a\sigma/\sqrt{n}]$ 的长度 $L(a) = (b - a)\sigma/\sqrt{n}$ 取得最小, 令 $dL/da = 0$ 得到

$$\begin{cases} \frac{dG}{da} = \phi(b) \frac{db}{da} - \phi(a) = 0 \\ \frac{dL}{da} = \frac{\sigma}{\sqrt{n}} \left(\frac{db}{da} - 1 \right) = 0 \end{cases} \Rightarrow \phi(a) = \phi(b) \Rightarrow a = b \text{ 或 } a = -b$$

$a = b$ 之解无意义, 所以必有 $a = -b$, 进而 $b = z_{1-\alpha/2}, a = -z_{1-\alpha/2}$ 。

例 8.36. 接着**例 8.35**, 以置信度 $1 - \alpha$ 估计参数 μ 的置信区间, 令区间长度 $L = 2z_{1-\alpha/2}\sigma/\sqrt{n}$ 不超过 d , 样本量必须满足

$$n \geq 4z_{1-\alpha/2}^2 \frac{\sigma^2}{d^2}$$

练习 8.7. 仿照**例 8.35**, 请读者验证**性质 8.5** 中第二种情况所给出的也是最短的置信区间。事实上, 当枢轴量 $Y = h_\theta(\mathbf{X})$ 的密度函数关于 0 对称时, **算法 8.2** 给出的置信区间是最短的。

※例 8.37. 将**例 8.31** 和**性质 8.5** 第一种情况的结果进行比较, 因为**性质 8.5** 取得了置

信度为 $1 - \alpha$ 的最短的置信区间，不难得到下面的不等式。

$$z_{1-\alpha/2} \leq \frac{1}{\sqrt{\alpha}}, \text{ 其中 } 0 < \alpha < 1$$

见下图，任何在实线 $z_{1-\alpha/2}$ 之上的函数 $h(\alpha)$ ，例如 $h(\alpha) = 0.2 - \text{sign}(\alpha - 1) \sqrt{-1.6 \ln(\alpha(2 - \alpha))}$ (白线)，都可以用来构造例 8.31 中未知参数 μ 的置信度为 $1 - \alpha$ 的置信区间

$$\bar{X} - h(\alpha) \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + h(\alpha) \frac{\sigma}{\sqrt{n}}$$

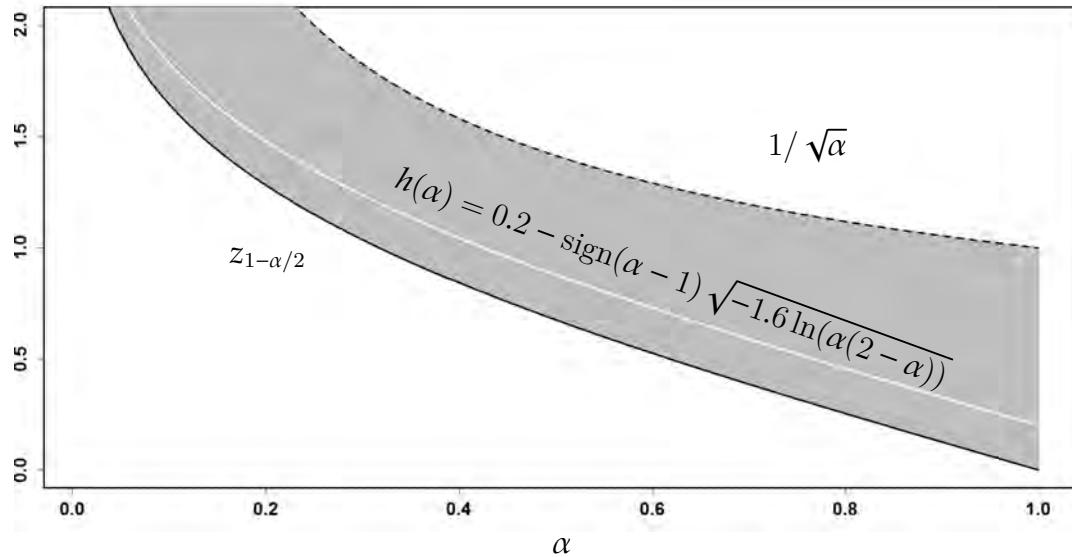


图 8.14: 实线是 $z_{1-\alpha/2}$ ，即标准正态分布 $Z \sim N(0, 1)$ 的 $(1 - \alpha/2)$ -分位数曲线。虚线是 $1/\sqrt{\alpha}$ ，白线是 $h(\alpha)$ ，它们都在曲线 $z_{1-\alpha/2}$ 之上。

8.2.3 大样本区间估计

前面所介绍的置信区间估计都是小样本方法。在大样本的情况下，可以利用由样本和参数所构造的随机变量的极限分布来求得未知参数的近似置信区间。或者，利用 Cramér 定理 8.10 所保证的最大似然估计的渐近正态性来求得近似置信区间。

性质 8.6. 设简单随机样本 X_1, X_2, \dots, X_n 来自总体 X ，已知总体具有有限期望和方差 $E(X) = \mu, V(X) = \sigma^2 > 0$ ，它们都是未知的。当样本容量足够地大，近似地有 μ 的置信度为 $1 - \alpha$ 的置信区间

$$\bar{X} - z_{1-\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

证明. 由 Lindeberg-Lévy 中心极限定理 5.17 知，

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{L} N(0, 1), \text{ 进而 } \frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{L} N(0, 1) \quad \square$$

性质 8.7. 令 $\hat{\theta}_n$ 是未知参数 θ 的最大似然估计，假设总体分布满足 Cramér 定理 8.10 的条件，当样本容量很大时近似地有

$$P\left\{-z_{1-\alpha/2} \leq \sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \leq z_{1-\alpha/2}\right\} = 1 - \alpha$$

求解花括号中的不等式，便得到 θ 的置信度为 $1 - \alpha$ 的置信区间。

例 8.38. 已知简单随机样本 X_1, \dots, X_n 来自正态总体 $N(\mu, \sigma^2)$ ，若样本容量 n 足够大，下面分三种情况讨论未知参数的区间估计。

- 若 μ 未知， σ^2 已知。由第 496 页的例 8.2， $I(\mu) = 1/\sigma^2$ 。当 n 很大时，由性质 8.7 得到 μ 的置信度为 $1 - \alpha$ 的置信区间如式 (8.13) 所示。
- 若 σ^2 未知， μ 已知，则 $I(\sigma^2) = 1/(2\sigma^4)$ 。 σ^2 的置信度为 $1 - \alpha$ 的置信区间为

$$\frac{\sum_{j=1}^n (X_j - \mu)^2}{n + \sqrt{2n}z_{1-\alpha/2}} \leq \sigma^2 \leq \frac{\sum_{j=1}^n (X_j - \mu)^2}{n - \sqrt{2n}z_{1-\alpha/2}}$$

比较这个结果和式 (8.14)，二者之间相差无几（见图 8.15）。另外，我们顺手得到了 χ^2 分布和标准正态分布的两类分位数之间的近似关系：

$$\begin{aligned} \chi_{n,1-\alpha/2}^2 &\approx n + \sqrt{2n}z_{1-\alpha/2} \\ \chi_{n,\alpha/2}^2 &\approx n - \sqrt{2n}z_{1-\alpha/2} \end{aligned}$$

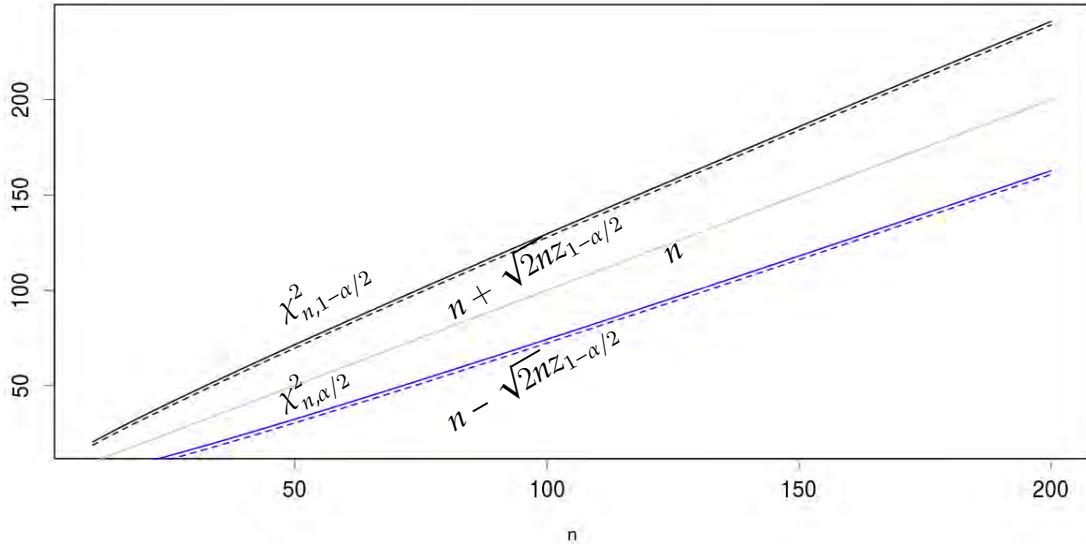


图 8.15: 令 $\alpha = 0.05$ 。图中实线是 $\chi^2_{n,1-\alpha/2}$ 和 $\chi^2_{n,\alpha/2}$, 虚线是 $n \pm \sqrt{2n}z_{1-\alpha/2}$ (为方便显示, 把散点连成线)。

- 若 $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ 未知, 例 8.2 给出 $\boldsymbol{\theta}$ 的 Fisher 信息矩阵。于是, σ^2 的置信度为 $1 - \alpha$ 的置信区间为

$$\frac{\sum_{j=1}^n (X_j - \bar{X})^2}{n + \sqrt{2n}z_{1-\alpha/2}} \leq \sigma^2 \leq \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{n - \sqrt{2n}z_{1-\alpha/2}}$$

请读者仿照图 8.15 比较上述结果和式 (8.16)。类似地, μ 的置信度为 $1 - \alpha$ 的置信区间为

$$\bar{X} - \sqrt{\frac{\sigma^2}{n}}z_{1-\alpha/2} \leq \mu \leq \bar{X} + \sqrt{\frac{\sigma^2}{n}}z_{1-\alpha/2}$$

用样本方差 S^2 替换 σ^2 , 便得到

$$\bar{X} - \frac{S}{\sqrt{n}}z_{1-\alpha/2} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}}z_{1-\alpha/2}$$

该结果与式 (8.15) 非常相似, 因为 n 很大时 $t_{n-1,1-\alpha/2} \approx z_{1-\alpha/2}$ (见第 357 页的例 5.10)。

例 8.39. 设样本 $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 和样本 $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} q\langle 1 \rangle + (1-q)\langle 0 \rangle$ 来自两个独立总体, 其中参数 p, q 未知, 当 m, n 都很大时, 求 $p - q$ 的置信度为 $1 - \alpha$ 的置信区间。

解. 当 m, n 都很大时, 漐近地有

$$\begin{aligned}\bar{X} &\sim N\left(p, \frac{\bar{X}(1-\bar{X})}{m}\right) \\ \bar{Y} &\sim N\left(q, \frac{\bar{Y}(1-\bar{Y})}{n}\right)\end{aligned}$$

于是,

$$\frac{\bar{X} - \bar{Y} - (p - q)}{\sqrt{\bar{X}(1-\bar{X})/m + \bar{Y}(1-\bar{Y})/n}} \xrightarrow{L} N(0, 1)$$

进而可得 $p - q$ 的置信度为 $1 - \alpha$ 的置信区间

$$\bar{X} - \bar{Y} \mp z_{1-\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{m} + \frac{\bar{Y}(1-\bar{Y})}{n}}$$

例 8.40 (Fisher-Behrens 问题). 设样本 $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} N(\mu_X, \sigma_X^2)$ 和样本 $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2)$ 来自两个独立总体, 若 $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ 都未知, 在大样本的情况下, 给出 $\mu_X - \mu_Y$ 的置信度为 $1 - \alpha$ 的置信区间。

解. 此问题没有适当的小样本解。根据性质 7.11,

$$\frac{[\bar{X} - \bar{Y} - (\mu_X - \mu_Y)] \sqrt{\frac{m+n-2}{\sigma_X^2/m + \sigma_Y^2/n}}}{\sqrt{(m-1)S_X^2/\sigma_X^2 + (n-1)S_Y^2/\sigma_Y^2}} \sim t_{m+n-2}$$

在大样本情况下, 即 m, n 充分大时, 有 $S_X^2/\sigma_X^2 \xrightarrow{L} 1$ 和 $S_Y^2/\sigma_Y^2 \xrightarrow{L} 1$, 于是漐近地有

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/m + S_Y^2/n}} \sim N(0, 1)$$

利用正态逼近得到 $\mu_X - \mu_Y$ 的置信度为 $1 - \alpha$ 的置信区间

$$\bar{X} - \bar{Y} \mp z_{1-\alpha/2} \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}$$

例 8.41. 接着第 486 页的例 7.20, 求未知参数 θ 的置信度为 $1 - \alpha$ 的置信区间。

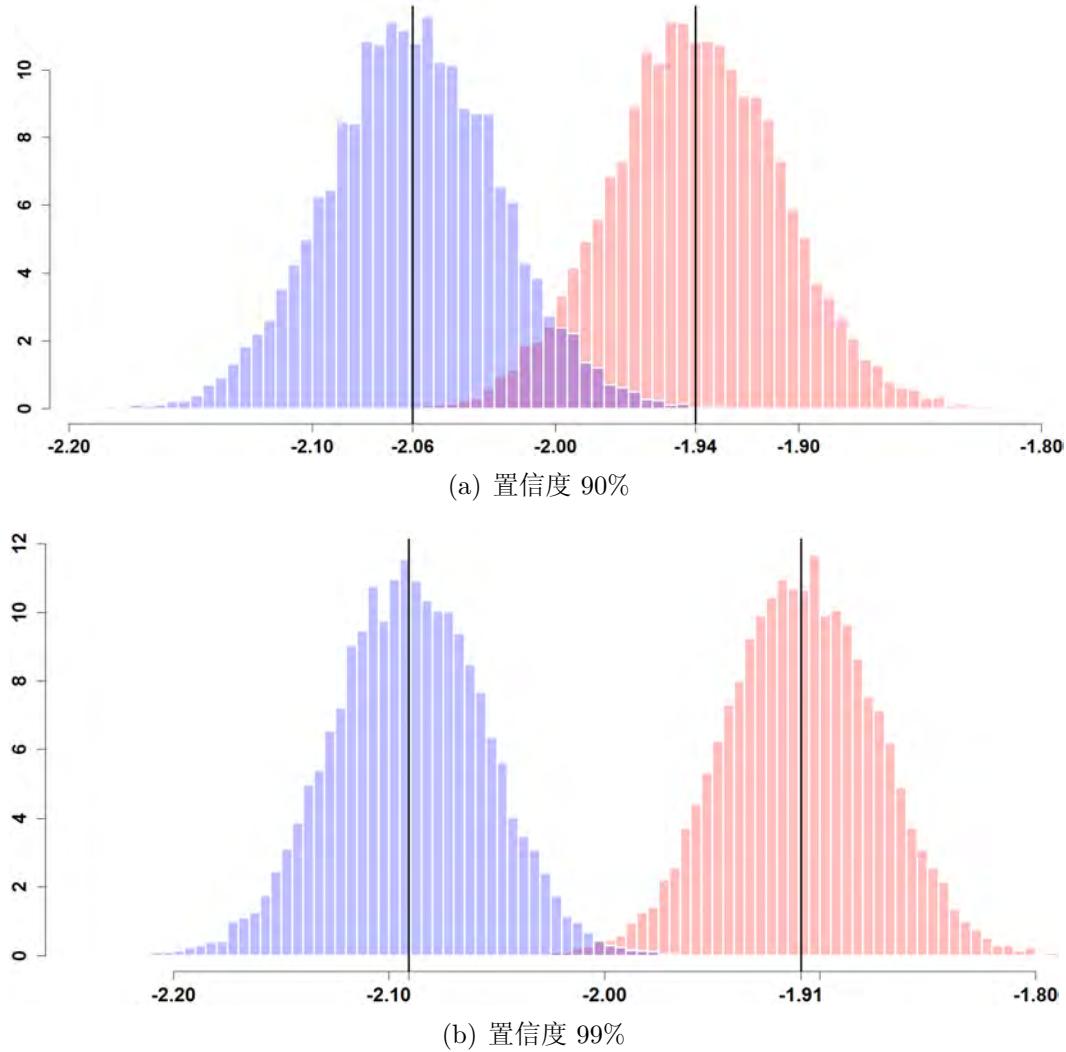


图 8.16: 在例 8.40 中, 设两个独立总体为 $X \sim N(0, 1)$, $Y \sim N(2, 0.25)$ 。利用大样本方法, 分别求出 $\mu_X - \mu_Y$ 的置信度为 90% 和 99% 的置信区间。在 10^4 次独立的随机试验中, 置信区间的上下限的分布如直方图所示。

解. 由例 8.23 知, $X_{(n)}$ 是 θ 的最大似然估计。由第 477 页的例 7.14 可得 $X_{(n)}$ 的分布函数为

$$F_{X_{(n)}}(x) = \begin{cases} (x/\theta)^n & \text{当 } 0 \leq x \leq \theta \\ 0 & \text{当 } x > \theta \end{cases}$$

于是, $P\left\{\sqrt[n]{\alpha}\theta \leq X_{(n)} \leq \theta\right\} = F_{X_{(n)}}(\theta) - F_{X_{(n)}}(\sqrt[n]{\alpha}\theta) = 1 - \alpha$

进而, $P\left\{X_{(n)} \leq \theta \leq \frac{1}{\sqrt[n]{\alpha}}X_{(n)}\right\} = 1 - \alpha$

即, $[X_{(n)}, X_{(n)}/\sqrt[n]{\alpha}]$ 是 θ 的置信度为 $1 - \alpha$ 的置信区间。

8.2.4 Fisher 的信任估计

1930 年, Fisher 提出从观测数据中获取参数分布的信任推断 (fiducial inference) 方法, 或多或少地影响了 Neyman 的置信区间理论 [116]。

定义 8.19. 把未知参数 θ 视作随机变量, 从枢轴量及其分布得到参数 θ 的分布 F , Fisher 把它称为未知参数 θ 的信任分布。Fisher 把满足条件 $F(\theta_2) - F(\theta_1) = 1 - \alpha$ 的区间 $[\theta_1, \theta_2]$ 作为 θ 的区间估计, 称作 θ 的信任区间 (一般要使得 $\theta_2 - \theta_1$ 最小), 把 $1 - \alpha$ 称作信任系数。

Fisher 对使用 Bayes 公式持非常谨慎的态度, 他是强烈抵触贝叶斯学派的, 尤其反对使用无信息先验。所以, 有别于贝叶斯学派通过参数的先验分布和观测数据得到参数的后验分布, 信任分布无先验与后验之说。遗憾的是 Fisher 并未给出信任推断的一般定义与一般方法, 只是处理了几个具体的例子, Fisher 也意识到信任推断的局限性, 该方法终因缺少系统理论的支持而未被广泛接受。

Fisher 把要作区间估计的参数看成随机变量不同于贝叶斯学派对未知参数的理解, 也不同于频率派把未知参数视为未知的固定常数。信任区间估计在 Fisher 的诸多成就中是颇受争议的, 与 Fisher 估计理论一贯坚持的似然方法也格格不入。在 Fisher 身后, 许多追随者企图发展 Fisher 的信任推断也未能取得实质成效, 所以该理论的应用并不广泛。美国数学家、贝叶斯学派统计学家 Leonard Jimmie Savage (1917-1971) 曾调侃, “Fisher 的信任论点是不想打破贝叶斯鸡蛋却想做出贝叶斯煎蛋的一个大胆的但不成功的尝试。”

美国统计学家 B. Efron (1978) 给信任推断盖棺定论, “多数的, 虽然不是全部的, 当代统计学家认为它或者是某种形式的客观贝叶斯主义, 或者就是一个简单的错误”。尽管如此, 它依旧能反映出一个伟大统计学家的思想的挣扎——在频率派和贝叶斯学派之间, 一方面有太多的来自哲学层面的思想冲突, 一方面它们的理论体系又相互地借鉴和促进。我们把 Fisher 的信任区间估计看作是两个学派之间一次不成功的调和。

例 8.42. 考虑样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 其中 μ 未知而 σ^2 已知。从事实 $Y = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$ 不难得到参数 μ 的信任分布

$$\mu = \bar{X} - \frac{\sigma}{\sqrt{n}} Y \sim N(\bar{X}, \sigma^2/n)$$

进而得到信任区间 $\bar{X} \mp z_{1-\alpha/2}\sigma / \sqrt{n}$, 与置信区间估计取得了相同的结果, 见第 526 页的性质 8.5 中的第一种情况。若 μ, σ^2 都未知, μ 的信任区间与置信区间的结论也是相同的。

例 8.43. 考虑例 8.40 (Fisher-Behrens 问题) 的信任区间估计。显然,

$$\xi = \frac{\bar{X} - \mu_X}{S_X / \sqrt{m}} \sim t_{m-1} \text{ 与 } \eta = \frac{\bar{Y} - \mu_Y}{S_Y / \sqrt{n}} \sim t_{n-1} \text{ 相互独立}$$

从 $\mu_X - \mu_Y - (\bar{X} - \bar{Y}) = \frac{S_X}{\sqrt{m}}\xi - \frac{S_Y}{\sqrt{n}}\eta$ 的信任分布可求得 δ 使得

$$P \left\{ \left| \frac{S_X}{\sqrt{m}}\xi - \frac{S_Y}{\sqrt{n}}\eta \right| \leq \delta \right\} = 1 - \alpha$$

于是, 便得到信任系数为 $1 - \alpha$ 的信任区间 $[\bar{X} - \bar{Y} - \delta, \bar{X} - \bar{Y} + \delta]$ 。

Fisher-Behrens 问题是说明信任区间估计有别于置信区间估计的一个典型案例, Neyman 曾就 $\alpha = 0.05, m = 12, n = 6, \sigma_X/\sigma_Y = 0.1, 1, 10$ 等情况考察过信任区间所对应的置信系数, 与 95% 相差很小。

8.3 习题

- 8.1. 设 X_1, X_2, \dots, X_n 是来自总体 X 的简单随机样本, 设 $E(X) = \mu$ 且 $V(X) = \sigma^2$ 。
 (1) 确定常数 c 使 $c \sum_{j=1}^{n-1} (X_{j+1} - X_j)^2$ 为 σ^2 的无偏估计。(2) 确定常数 c 使 $\bar{X}^2 - cS^2$ 为 μ^2 的无偏估计。
- 8.2. 设 $\hat{\theta}$ 是参数 θ 的无偏估计并且 $V(\hat{\theta}) > 0$, 试证明: $\hat{\theta}^2$ 不是 θ^2 的无偏估计。
- 8.3. 设 X_1, X_2 是总体 X 的一个简单随机样本, 已知 $E(X) = \mu, V(X) = \sigma^2$ 。试问:
 $\hat{\mu}_1 = \frac{1}{2}(X_1 + X_2)$ 和 $\hat{\mu}_2 = a_1 X_1 + a_2 X_2$ (其中 $a_1, a_2 > 0$ 满足 $a_1 + a_2 = 1$) 是否是 μ 的无偏估计量, 哪个方差更小?
- ☆ 8.4. 设样本 $X_1, \dots, X_n \stackrel{iid}{\sim} f_\theta(x) = \begin{cases} \exp(\theta - x) & \text{当 } x \geq \theta, \text{ 其中 } \theta \text{ 未知} \\ 0 & \text{当 } x < \theta \end{cases}$
 试证明: $\hat{\theta}_1 = \bar{X} - 1$ 和 $\hat{\theta}_2 = X_{(1)} - 1/n$ 都是 θ 的无偏估计, 并且 $V(\hat{\theta}_2) \leq V(\hat{\theta}_1)$ 。
- 8.5. 设样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, 其中参数 μ 已知, σ^2 未知。试证明: $\hat{\sigma} = \sqrt{\pi/2} \sum_{j=1}^n |X_j - \mu|/n$ 是 σ 的无偏估计。
- 8.6. 求 $X \sim \text{Expon}(\beta)$ 未知参数 β 的 Fisher 信息量 $I(\beta)$ 。
- 8.7. 设总体密度函数 $f_\theta(x) = \begin{cases} (\theta + 1)x^\theta & \text{当 } 0 < x < 1, \text{ 其中 } \theta > 1 \text{ 未知} \\ 0 & \text{其他} \end{cases}$
 若样本 $X_1, \dots, X_n \stackrel{iid}{\sim} f_\theta(x)$, 求参数 θ 的矩估计和最大似然估计。
- 8.8. 一个盒子里装有黑球和白球, 有放回地抽取 n 次共得 k 次白球, 求盒子里黑球数和白球数之比 θ 的最大似然估计。
- 8.9. 设样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \theta^2 \langle 1 \rangle + 2\theta(1-\theta) \langle 2 \rangle + (1-\theta)^2 \langle 3 \rangle$, 其中参数 $0 < \theta < 1$ 未知。试求: 参数 θ 的矩估计和最大似然估计。
- ☆ 8.10. 设简单随机样本来自对数正态分布 (见定义 4.12) 的总体 $X \sim \log N(\mu, \sigma^2)$, 即 $X_1, \dots, X_n \stackrel{iid}{\sim} \log N(\mu, \sigma^2)$, 其中参数 $\mu \in \mathbb{R}, \sigma > 0$ 皆未知, 试求参数 $\theta_1 = E(X)$ 和 $\theta_2 = V(X)$ 的最大似然估计。
- 8.11. 设样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} U[0, \theta]$, 其中未知参数 $\theta \in \Theta = (0, +\infty)$, 求 θ 的最大似然估计。
- ☆ 8.12. 设简单随机样本 $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top$ 来自二元正态总体 $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, 试求未知参数 $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ 的最大似然估计 $\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\rho}$ 及其均方误差。说明 $\hat{\mu}_X, \hat{\mu}_Y$ 与 $\hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\rho}$ 相互独立。

- 8.13. 设某电子元件的寿命服从均值为 μ 、方差为 σ^2 的分布，从总体中分别抽取容量为 n_1, n_2 的两个独立样本，样本均值分别是 \bar{X} 和 \bar{Y} ，确定常数 p 使 $T = p\bar{X} + (1-p)\bar{Y}$ 的方差达到最小。
- ☆ 8.14. 有 n 台测量光速的仪器，其中第 j 台的测量值 $X \sim N(\theta, \sigma_j^2)$, $j = 1, 2, \dots, n$ 。用这些仪器独立地对光速 θ 各测量一次，得到样本 X_1, X_2, \dots, X_n 。问 k_1, k_2, \dots, k_n 取何值时能使 $\hat{\theta} = \sum_{j=1}^n k_j X_j$ 是 θ 的无偏估计并且 $V(\hat{\theta})$ 最小？
- 8.15. 设样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ ，其中参数 μ, σ^2 未知，求 $\ln \sigma^2$ 的置信度为 $1 - \alpha$ 的置信区间。
- 8.16. 设样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ ，其中参数 $\mu, \sigma^2 > 0$ 未知，令 L 表示 μ 的置信度为 $1 - \alpha$ 的最短置信区间的长度，求 $E(L^2)$ 。
- ☆ 8.17. 设样本 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} U[0, \theta]$ ，证明：对于任意给定的 $1 - \alpha$ ，其中 $0 < \alpha < 1$ ，存在常数 c_n 使 $[\max(X_1, \dots, X_n), c_n \max(X_1, \dots, X_n)]$ 为 θ 的一个置信度为 $1 - \alpha$ 的置信区间。
- 8.18. 接着例 8.40，给出 σ_Y^2/σ_X^2 的置信度为 $1 - \alpha$ 的置信区间。
- ☆ 8.19. 设来自两个独立总体的样本 $X_1, \dots, X_m \stackrel{iid}{\sim} \text{Expon}(\lambda_1)$ 和 $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Expon}(\lambda_2)$ ，试求 λ_2/λ_1 的置信度为 $1 - \alpha$ 的置信区间。

第九章

假设检验

竹外桃花三两枝，春江水暖鸭先知。蒌蒿满地芦芽短，正是河豚欲上时。

苏轼《惠崇春江晚景》

关于总体分布的假设称为统计假设。统计假设与数学里“假设函数 $f(x)$ 光滑”、“假设 \mathcal{T} 为 X 上的一个拓扑”等假设不同，二者的区别在于，是否拒绝一个统计假设要依赖于观察数据，并通过某些合理的检验手段才能下结论。

假设检验 (hypothesis testing) 正是这样的手段，它是统计推断的一个重要组成部分。假设检验的目的就是在已知样本的基础上，对一个统计假设 H_0 进行判断以决定是否拒绝它。我们常把 H_0 称为零假设 (null hypothesis) 或原假设，把 H_0 的对立命题 H_1 称为备择假设 (alternative hypothesis)。

Fisher 认为“与任何试验关联的假设，我们都可以称之为零假设，应该指出的是，零假设从不被证明或公认，而有可能在试验中被否定。任何试验可以说仅为给事实一次反驳零假设的机会而存在。”

例 9.1. 工厂生产一批零件，其长度（单位：毫米）服从分布 $N(\mu, 10^{-2})$ ，其中参数 μ 未知， $\mu_0 = 100$ 为合格零件的长度。随机抽取 15 个零件测得其长度分别为：100.095, 100.101, 100.248, 100.156, 99.946, 100.243, 100.041, 100.145, 100.054, 100.113, 100.055, 100.080, 99.895, 100.135, 100.056。零假设是这批零件的长度合格，即 $H_0 : \mu = \mu_0$ 。备择假设是 $H_1 : \mu \neq \mu_0$ 。本章后续正文将给出**例 9.1** 假设检验的细节。

定义 9.1. 从不同的角度，统计假设可分为以下几种类型。

□ 从是否已知总体分布类型的角度，统计假设分为参数假设和非参数假设两类。

– 参数假设：总体分布类型已知且仅涉及未知参数的统计假设，如**例 9.1**。已知简单随机样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 来自总体 $X \sim F_\theta(x)$ ，其中 $\theta \in \Theta$

是未知参数（可以是向量）， Θ 是参数空间。令 $\Theta_0 \subseteq \Theta$ 且 $\Theta_1 = \Theta - \Theta_0$ ，通常把零假设和备择假设记作

$$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$$

这里符号“ \leftrightarrow ”表示的是“对比”(versus)的意思。例如， $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$ 或 $H_0 : \theta \leq \theta_0 \leftrightarrow H_1 : \theta > \theta_0$ 等等。

- 非参数假设：总体分布类型未知时，仅涉及总体分布类型的统计假设，譬如 H_0 ：总体分布 $F(x) \in$ 正态分布族。

□ 从能否能确定总体分布的角度，又可分为简单假设和复合假设两类。

- 简单假设 (simple hypothesis)：能让我们明确写出总体分布的统计假设，如例 9.1 中的零假设 $H_0 : \mu = \mu_0$ ，若它成立，总体则服从 $N(\mu_0, 10^{-2})$ 这一确定的分布。
- 复合假设 (composite hypothesis)：不是简单假设的统计假设，如非参数假设 H_0 ：两个样本来自同一总体。再如，对例 9.1 中的总体也可以做这样的零假设， $H_0 : |\mu - \mu_0| \leq 0.1$ ，这就是一个复合假设。另外，非参数假设“ H_0 ：总体分布 $F(x) \in$ 正态分布族”也是复合假设。

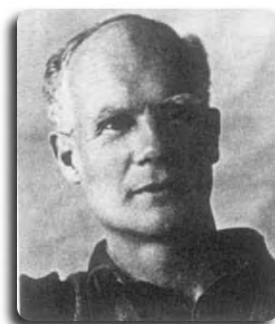
对零假设 $H_0 : \mu = \mu_0$ 只有两种行为可选择：拒绝或者不拒绝。“拒绝 H_0 ”意味着观测数据（即样本值）不支持零假设，“不拒绝 H_0 ”意味着观测数据不足以否定零假设。同样地，对备择假设 $H_1 : \mu \neq \mu_0$ 也只有拒绝或者不拒绝这两个选择。

对零假设之所以不用“接受或不接受”，其原因是，拒绝一个命题只需一个反例，而接受一个命题仅仅有一个佐证的例子是远远不够的——换言之，基于观察数据，拒绝零假设容易接受它难。

但是，在很多情况下为了表述的方便，只要不引起误解，我们也不严谨地用“接受”零假设来表示“不拒绝”零假设。由于零假设与备择假设是互为逆命题的，所以拒绝零假设和接受备择假设是一回事。

1926 年，E. Pearson (照片见右) 曾写信请教 Gosset 有关正态均值的检验问题，Gosset 在回信中指出，“即便发现某样本的机会非常之小，如 0.00001，检验就其本身来说并未证明该样本不是从假设的总体中随机采得。检验做的是，说明如果有某个备择假设（譬如样本来自另一总体，或样本不是随机的）将以一个更适度的概率，如 0.05，解释该样本的存在，你将更倾向于认为原假设不是真的。”

换句话说，拒绝一个统计假设的正当理由是它的对立假设能以更大的概率解释观察到的样本。令 D 是观察数据，当 H_0 和 H_1 都是简单假



设时, Gosset 认为拒绝 H_0 的条件应该是

$$P(D|H_0) < P(D|H_1)$$

在人们默认的理念中, 总是认为合理的假设应该使得观察到的事件以较大的概率发生。或者说, 哪个假设能更好地解释数据的由来, 哪个假设就更容易被接受。这个想法在贝叶斯学派那儿是再清楚不过的了, 因为

$$P(H_j|D) = \frac{P(D|H_j)P(H_j)}{P(D)}, \text{ 其中 } j = 0, 1$$

在得到观察数据之前, 零假设和备择假设满足 $P(H_0) = P(H_1) = 1/2$, 所以 $P(D|H_0) < P(D|H_1)$ 即意味着 $P(H_0|D) < P(H_1|D)$, 即数据更支持 H_1 。

例 9.2. 一个常喝牛奶加茶的妇女称, 她能区分出牛奶还是茶被先倒入杯子。对她进行了 10 次试验, 结果她都说对了。

令参数 θ 表示该妇女在每次试验中答对的概率, 参数空间 $\Theta = \{0.5, 0.9\}$ 。则零假设 $H_0 : \theta = 0.5$ 表示她每次是随机猜测的, 备择假设 $H_1 : \theta = 0.9$ 表示她很可能有此“特异功能”。

显然, $P(10 \text{ 次都答对了}|H_0) = 2^{-10}$, 这个概率值非常之小。但它不能作为拒绝 H_0 的理由, 合理的理由是备择假设使得以较大的概率观察到结果。

定义 9.2. 若样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} f_\theta(x)$, 似然函数 $\mathcal{L}(\theta; \mathbf{x}) = \prod_{j=1}^n f_\theta(x_j)$ 在 θ_1 和 θ_0 两点的函数值之比被称为似然比 (likelihood ratio), 记作 $\lambda(\mathbf{x}; \theta_0, \theta_1)$, 即

$$\begin{aligned}\lambda(\mathbf{x}; \theta_0, \theta_1) &= \frac{\mathcal{L}(\theta_1; \mathbf{x})}{\mathcal{L}(\theta_0; \mathbf{x})} \\ &= \frac{\prod_{j=1}^n f_{\theta_1}(x_j)}{\prod_{j=1}^n f_{\theta_0}(x_j)}\end{aligned}$$

在上下文信息不引起误解的情况下, 似然比有时也简记作 $\lambda(\mathbf{x})$ 。对于像**例 9.2** 这样的两个简单假设 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$ 的检验问题, 似然比 $\lambda(\mathbf{x}; \theta_0, \theta_1) > 1$ 意味着 $H_1 : \theta = \theta_1$ 更有可能成立。

练习 9.1. 计算**例 9.2** 中的似然比, 其中 $\theta_0 = 0.5, \theta_1 = 0.9$ 。

例 9.3. 已知样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 其中 σ^2 已知, μ 未知, 考虑双侧检验^{*} $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$ 。与**例 9.2** 不同的是, 此处备择假设是一个复合假设, 它

^{*}对于检验问题 $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$, 如果存在凸集 (见**定义 F.1**) C 使得 $\Theta_1 \subseteq C$ 且 $C \cap \Theta_0 = \emptyset$, 则称该问题是单侧检验 (one-sided testing problem), 否则就称双侧检验 (two-sided testing problem)。例如, $H_0 : \theta \leq \theta_0 \leftrightarrow H_1 : \theta > \theta_0$ 是单侧检验, 而 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$ 是双侧检验。

不能确定总体分布，进而无法计算 $P(X_1 = x_1, \dots, X_n = x_n | H_1)$ ，也不能计算似然比。我们如何实现 Gosset 的想法呢？

- 假定零假设 H_0 成立，即 $\mu = \mu_0$ ，则样本均值 \bar{X} 是参数 μ 的最大似然估计，且 $\bar{X} \sim N(\mu_0, \sigma^2/n)$ 。显然，样本均值以大概率落于区间 $A = [\mu_0 - 3\sigma/\sqrt{n}, \mu_0 + 3\sigma/\sqrt{n}]$ 之内。若给定样本容量 n ，这个区间是固定的。在大量重复试验中，样本均值的观察结果 \bar{x} 多落于此区间内。
- 如果 $\bar{x} \notin A$ ，意味着小概率事件 $\bar{X} \notin A$ 发生了。在这种情况下，小概率事件的发生让我们倾向于认为零假设 H_0 不成立而拒绝它，因为总体设为 $N(\bar{x}, \sigma^2)$ 能更好地解释数据的由来，我们更愿意相信 $\mu = \bar{x} \neq \mu_0$ 。

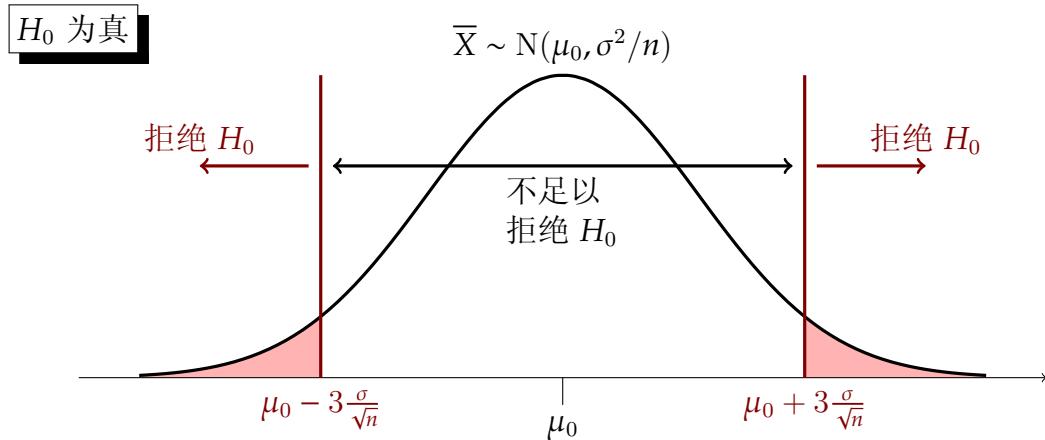


图 9.1: 如果零假设 $H_0 : \mu = \mu_0$ 成立，则样本均值 \bar{X} 落在区间 $A = [\mu_0 - 3\sigma/\sqrt{n}, \mu_0 + 3\sigma/\sqrt{n}]$ 之外的概率 $P(\bar{X} \notin A | H_0)$ 很小。如果这个小概率事件发生，与 $H_0 : \mu = \mu_0$ 相比， $N(\bar{x}, \sigma^2)$ 能更好地解释观察结果，于是我们拒绝 H_0 而接受 $H_1 : \mu = \bar{x} \neq \mu_0$ 。

E. Pearson 受到 Gosset 的启发，他写信给 Neyman 提议当备择假设之下样本的最大似然远大于原假设之下样本的最大似然时，可用似然比标准来拒绝原假设。E. Pearson 认为似然比为假设检验提供了一般框架，在与 Neyman 的通信中二人开始了在假设检验理论方面的合作。

定义 9.3. 1926-1928 年，Neyman 与 E. Pearson 的合作研究中指出，对假设 $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$ 的检验有可能犯以下两种类型的错误。

第一类错误: 当零假设 H_0 真（即 $\theta \in \Theta_0$ ）时，拒绝了 H_0 或者接受了 H_1 。第一类错误也称为拒真错误。

第二类错误: 当零假设 H_0 假（即 $\theta \in \Theta_1$ ）时，接受了 H_0 或者拒绝了 H_1 。第二类错误也称为取伪错误。

表 9.1: 假设检验的两类错误。

行为	真实情况	
	H_0 为真	H_0 为假
拒绝 H_0	第一类错误	正确
接受 H_0	正确	第二类错误

定义 9.4. 第一类错误的概率 $\alpha = P\{\text{拒绝 } H_0 | \theta \in \Theta_0\}$ 也称为拒真概率。不犯第二类错误的概率 β 称为检验对备择假设的功效或势 (power), 即

$$\beta = P\{\text{拒绝 } H_0 | \theta \in \Theta_1\}$$

显然, 第二类错误的概率 (也称为取伪概率) 是

$$\text{取伪概率} = 1 - P\{\text{拒绝 } H_0 | \theta \in \Theta_1\} = 1 - \beta$$

在检验的拒真概率不超过一个给定的接近零的正实数 α 的时候, β 越大表明检验拒伪的能力越强。人们当然希望一个检验的拒真概率和取伪概率都足够地小, 然而当样本量一定时, 我们将论证同时无限制地减少这两类错误的概率是不可能的。

例 9.4. 哪类错误更应引起注意呢? 这依赖于零假设和具体的应用对象。譬如,

- “ H_0 : 某人有癌症 $\leftrightarrow H_1$: 某人没有癌症” 对医院来说, 还有什么比挽救生命更重要的? 所以拒真错误比取伪错误更让人难以容忍。
- “ H_0 : 产品合格 $\leftrightarrow H_1$: 产品不合格” 对生产者来说, 取伪错误过大将带来产品质量的下降, 拒真错误过大将导致生产成本的增加。而对消费者来说, 减小取伪错误比减小拒真错误更重要。

通常情况下, 人们把力图否定的命题约定为零假设 (如例 9.1), 假设检验就像是在用数据“抬杠”, 总想对零假设说“不”。或者, 把更关注的统计假设当作零假设 (如 H_0 : 某人有癌症)。在这样的约定之下, 拒真错误往往显得比较重要。

1928-1930 年, E. Pearson 和 Neyman 联名发表了几篇论文, 提出了备择假设、两类错误等基本概念, 研究了两样本问题和多样本问题的似然比检验。

因为 Neyman 是频率派的忠实支持者, 所以他终生反对贝叶斯学派。Neyman 觉得似然比与贝叶斯方法 (回顾第 83 页的例 1.64) 有牵连, 因此他并不满足于似然比标准。Neyman 力图寻找更底层的原则和更坚实的基础, 为此他提出了所谓的“Neyman-Pearson 原则”。

定义 9.5 (Neyman-Pearson 原则). 1928 年, Neyman 和 E. Pearson 提出一个评价检验优劣的原则, 称为 Neyman-Pearson 原则: 在控制住拒真概率的前提下, 使取伪概率尽可能地小或不犯取伪错误的概率 $\beta = P\{\text{拒绝 } H_0 | \theta \in \Theta_1\}$ 尽可能地大。

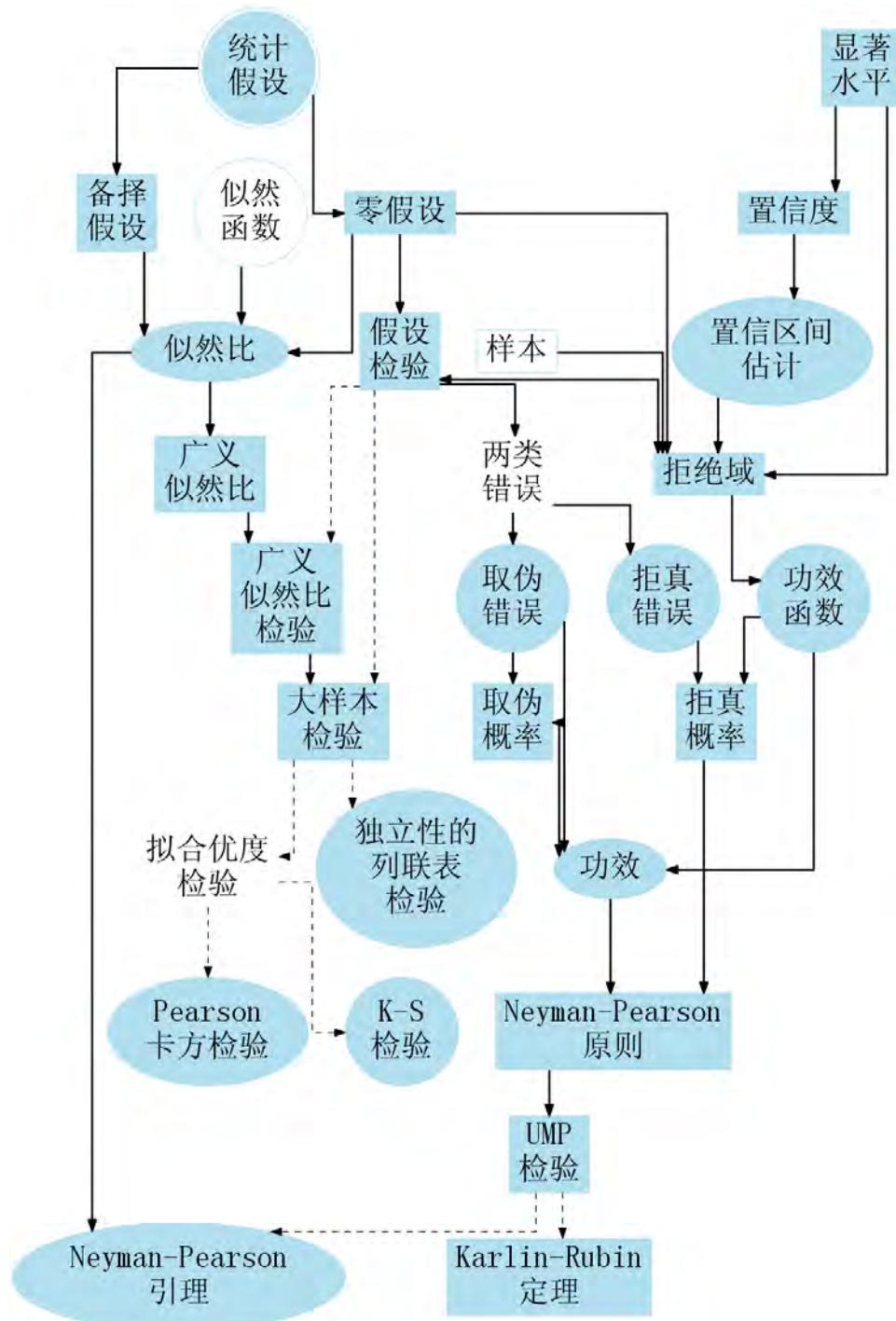
- 在该原则之下, 1930 年 Neyman 构造出了当原假设和备择假设都是简单假设之时的最优检验, 即著名的 Neyman-Pearson 引理。
- 1933 年, Neyman 和 E. Pearson 发表了著名论文《关于统计假设的最有效检验问题》, 奠定了 Neyman-Pearson 假设检验理论的基础, 这篇论文被收录在《统计学中的重大突破》第一卷 [96] 中。

Neyman 的统计哲学是把统计学视作决策行为的指导原则, 他的假设检验理论与置信区间理论同出一辙, 都是保证在长期的经验中不经常地犯错, 而不是针对某次具体的决策。Neyman 的这一观点遭到了 Fisher 的强烈反对, Fisher 认为 Neyman 的理论只适用于可重复的情形, 而作为科学推断的一般方法却是不适宜的。

对 Neyman-Pearson 假设检验理论的反对之声还有一些来自技术层面的, 譬如在未获得观察数据之前所有的检验步骤都已确定, 因为统计实践往往要根据观察数据来选择模型, Neyman-Pearson 检验程序难免显得过于死板。另外, 在 Neyman-Pearson 的框架里, 最优的检验要么不存在, 要么难以求得, 很多问题不得不局限在指数族里讨论。

即便如此, Neyman-Pearson 假设检验理论仍旧是经典统计学的重要组成部分, 其后续发展在统计实践中一直扮演着重要角色。

第九章的主要内容及其关系



9.1 Neyman-Pearson 假设检验理论

在 Neyman 和 E. Pearson 的论文《关于统计假设的最有效检验问题》的导言部分，作者提到 Bertrand 曾悲观地认为，依据样本的某个特征的概率之大小来评判假设的真伪是得不到可靠的结果的，然而 Borel 相信只要特征选得好，该方法还是可行的。

Neyman 和 E. Pearson 的观点是，检验的目的并不在于了解每个具体的假设为真或为假，而是寻找在长期实践中很少犯错的检验规则。譬如这样的检验规则 R_{H_0} ：对于假设 H_0 ，计算观察数据的某特征 x ，如果 $x > x_0$ 就拒绝 H_0 ，否则就接受 H_0 ，其中 x_0 是一个阈值。Neyman 和 E. Pearson 希望规则 R_{H_0} 在大量重复使用中犯错误的概率很小。

这一检验规则虽不能告诉人们某次具体的检验是否得出正确的结论，但如果我们将证明在反复实践中按此规则行事将很少犯错，那么每次具体的检验就会让人觉得“八九不离十”。打个比方，一个高超的赌博策略虽不能保证每次都赢，但一直赌下去必定是赢多输少。

Neyman 和 E. Pearson 对假设检验的认知深受频率派思想的影响。事实上，Neyman 的置信区间理论延续了他的统计哲学，请读者回顾置信度的频率解释，再来体会 Neyman 不纠缠检验规则“一时的成败”，而是通过大量反复的试验对检验规则给出一个综合评价。

定义 9.6. Neyman-Pearson 假设检验理论的基本想法是设定一个小的概率阈值 α ，譬如 $\alpha = 0.05$ 或 0.01 ，假定零假设 $H_0 : \theta \in \Theta_0$ 成立，如果观察到样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的概率不超过 α ，即如果 $P\{\mathbf{X}|\theta \in \Theta_0\} \leq \alpha$ ，则拒绝零假设 H_0 （见例 9.2 和例 9.3）。我们称这个概率阈值 α 为显著水平 (significance level) 或水平，并称在水平 α 拒绝 H_0 。

在实践中，为了避免概率计算上的麻烦，这个决策问题常常转化为判断样本 \mathbf{X} 是否落于样本空间的某子区域 R ，称之为该假设检验的拒绝域 (rejection region)，当样本 \mathbf{X} 落于 R 中时拒绝零假设 H_0 。拒绝域 R 的补集 $A = R^c$ 称为接受域，当样本 \mathbf{X} 落于 A 中时接受零假设 H_0 。不管拒绝还是接受，都是针对零假设 $H_0 : \theta \in \Theta_0$ 而言的。

定义 9.7. 假设检验的目标就是构造样本空间的子区域 R ，我们把 R 的指示函数 $\delta(\mathbf{x}) = I_R(\mathbf{x})$ 称作检验函数 (test function)，构造拒绝域 R 与构造检验函数 $\delta(\mathbf{x})$ 是一回事。

例 9.5. 以 $H_0 : \theta \leq \theta_0$ 为例，如果样本均值 \bar{X} 是 θ 的点估计，则 \bar{X} 越小 H_0 成立的可能就越大。不妨设一个临界值 c ，把样本空间划分为 $R = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} > c\}$ 和 $R^c = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} \leq c\}$ 两部分：当样本 \mathbf{X} 落于 R 中时，就拒绝 H_0 ，否则就接受 H_0 。

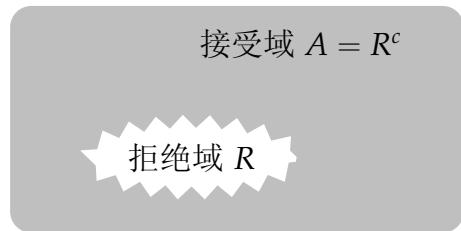


图 9.2: 拒绝域和接受域都是针对零假设 H_0 而言的, 当样本落于拒绝域时就拒绝 H_0 。而当样本落于接受域时, 该样本无法拒绝 H_0 。

给定拒绝域 R 后, 凭借样本 \mathbf{X} 是否落于 R 中来拒绝或接受 H_0 便带来了随机性——在假设检验中, 变的是样本, 不变的是拒绝域。如何构造出合理的拒绝域 R 呢?

本节内容

衡量假设检验优劣的标准是拒真概率和取伪概率, 为了统一地刻画它们, 第一小节描述了功效和功效函数等概念。在样本量一定的情况下, 通过几个实例说明了这样的事实: 降低一类错误的概率必然增大另一类错误的概率。按照 Neyman-Pearson 原则, 在拒真概率被控制住的前提下检验的取伪概率越小越好, 于是一致最大功效 (UMP) 检验就成为梦寐以求的检验。第二小节介绍的 Neyman-Pearson 引理在零假设和备择假设都是简单假设的情况下给出了 UMP 检验。如果分布族对某统计量具有单调似然比, Karlin-Rubin 定理为某类复合检验提供了 UMP 检验。

关键知识

- (1) 两类检验错误; (2) 功效函数; (3) UMP 检验; (4) 单调似然比; (5) Neyman-Pearson 引理; (6) Karlin-Rubin 定理; (7) 正态总体下对参数的假设检验。

9.1.1 功效函数与两类错误的概率

拒真概率与取伪概率有怎样的关系？要搞清楚这个问题需要功效函数这一工具，它源于 Neyman 对 t 检验第二类错误概率的研究。功效函数把两类错误的概率统一地表示出来，形式上很方便。以后讨论拒真概率和取伪概率都可以通过功效函数来完成。

定义 9.8. 样本 \mathbf{X} 落于拒绝域 R 概率 $P_\theta\{\mathbf{X} \in R\} = P_\theta\{I_R(\mathbf{X}) = 1\}$ 即拒绝 H_0 的概率 $P_\theta\{\text{拒绝 } H_0\}$ ，显然它是定义于参数空间 Θ 上关于 θ 的函数，称为功效函数或势函数 (power function)，记作 $\beta_\delta(\theta)$ 。

性质 9.1. 由**定义 9.8**，显然，功效函数 $\beta_\delta(\theta)$ 满足：

$$\begin{aligned}\beta_\delta(\theta) &= E_\theta I_R(\mathbf{X}) \\ &= P_\theta\{I_R(\mathbf{X}) = 1\} \\ &= P_\theta\{\text{拒绝 } H_0\} \\ &= \begin{cases} \text{拒真概率} & \text{如果 } \theta \in \Theta_0 \\ 1 - \text{取伪概率} & \text{如果 } \theta \in \Theta_1 \end{cases}\end{aligned}$$

定义 9.9. 对于检验函数 $\delta(\mathbf{x})$ ，如果 $\forall \theta \in \Theta_0$ 皆有 $\beta_\delta(\theta) \leq \alpha$ ，则称 δ 是一个水平 α 检验，它犯拒真错误的概率不超过 α 。后文中，所有水平 α 检验构成的集合记作 Δ_α 。

如果拒绝域 R 由 $T(\mathbf{x}) > c$ 给出，其中 $T(\mathbf{X})$ 是一个统计量（称为检验统计量，test statistic）， c 为一待定常数称为临界值 (critical value)，可根据检验统计量 T 的分布构造出拒绝域，只要保证拒真概率的上确界 $\alpha(c)$ 不超过给定的水平 α ，即

$$\begin{aligned}\alpha(c) &= \sup_{\theta \in \Theta_0} P_\theta\{\text{拒绝 } H_0\} \\ &= \sup_{\theta \in \Theta_0} P_\theta\{T(\mathbf{X}) > c\} \leq \alpha\end{aligned}$$

算法 9.1. 假设检验的一般过程是：

- 首先把整个参数空间 Θ 划分为 Θ_0 和 Θ_1 ，列出零假设 $H_0 : \theta \in \Theta_0$ 和备择假设 $H_1 : \theta \in \Theta_1$ 。给出显著水平 $0 < \alpha < 1$ ，譬如 $\alpha = 0.01$ 或 0.05 等。零假设一般为欲否定的命题。
- 定义拒绝域为 $R = \{\mathbf{x} \in \mathbb{R}^n : T(\mathbf{x}) \geq c\}$ ，其中 T 为某一检验统计量，临界值 c 待定。若数据落在边界 $\partial R = \{\mathbf{x} \in \mathbb{R}^n : T(\mathbf{x}) = c\}$ 上，“拒绝”还是“接受” H_0 本来就是模棱两可的事情，所以有时也将拒绝域定义为 $R = \{\mathbf{x} \in \mathbb{R}^n : T(\mathbf{x}) > c\}$ ，并不影响统计推断。

由检验统计量的分布 $T(\mathbf{X}) \sim G_\theta(t)$, 其中 $G_\theta(t)$ 是一个含未知参数 θ 的分布, 得到拒绝 H_0 的概率 $P_\theta\{T(\mathbf{X}) > c\}$ 的表达式。为使得拒真错误不超过 α , 由 $\alpha(c) = \sup_{\theta \in \Theta_0} P_\theta\{T(\mathbf{X}) > c\} = \alpha$ 解出临界值 c 。

当 $\mathbf{X} \in R$ 时, 在水平 α 拒绝零假设 H_0 ; 否则, 接受 H_0 。

例 9.6. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 其中参数 σ^2 已知, μ 未知, 设 μ 的参数空间为 $\Theta = \{\mu_0, \mu_1\}$, 其中 $\mu_0 < \mu_1$ 。在水平 α 对简单假设 $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu = \mu_1$ 进行检验并求该检验的取伪概率。

解. 显然, 样本均值 \bar{X} 越大, 越倾向于否定 H_0 , 拒绝域定义为 $R = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} > c\}$, 其中 c 是待定的常数。定义功效函数如下并得到拒真概率的上确界 $\alpha(c)$,

$$\beta_\delta(\mu) = P_\mu\{\bar{X} > c\} = 1 - \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

$$\alpha(c) = \sup_{\mu=\mu_0} \beta_\delta(\mu) = \beta_\delta(\mu_0) = 1 - \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right)$$

由 $\alpha(c) = \alpha$ 可得 $c = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$, 即拒绝 H_0 的条件是

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}, \text{ 或者等价地 } \bar{x} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

进而得到该检验的取伪概率为

$$P_{\mu_1}\{\bar{X} \leq c\} = 1 - \beta_\delta(\mu_1) = \Phi\left[z_{1-\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}\right]$$

明显地, 拒真概率 $\leq \alpha$ 。当 $\alpha \rightarrow 0$ 时, 取伪概率 $\rightarrow 1$ 。当样本量一定时, 拒真概率和取伪概率就像跷跷板的两端, 不能同时被降低。

例 9.7. 在**例 9.6** 中, 已知 $\sigma^2 = 1$, 观察数据如下: 1.39, 0.39, 1.35, 0.92, -0.61, -0.87, -0.84, -1.59, -0.87, 0.06, -0.06, -0.11, -0.16, -1.08, -0.20, -0.75, 1.63, -1.21, -0.64, 0.26。在水平 $\alpha = 0.05$ 对 $H_0 : \mu = 0 \leftrightarrow H_1 : \mu = 1/2$ 进行检验。

解. 求得样本均值 $\bar{x} = -0.1495$, 并且

$$\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 0 + z_{0.95}\frac{1}{\sqrt{20}} \approx 0.3678005$$

根据**例 9.6** 的结果, 在水平 $\alpha = 0.05$ 数据不足以否定零假设 $H_0 : \mu = 0$ 。另外,

该检验的取伪概率为

$$\Phi\left[z_{1-\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}\right] = \Phi\left(z_{0.95} - \frac{\sqrt{20}}{2}\right) \approx 0.2771884$$

下图给出了拒绝域、拒真概率、取伪概率、备择假设的功效等概念的直观图示。无论观察数据如何，只要样本容量 n 和检验水平 α 给定，该检验问题的拒绝域都固定是 $(0.3678005, \infty)$ ，进而取伪概率也总是 0.2771884。

图 9.3: 例 9.7 中检验的拒绝域、拒真概率、取伪概率、备择假设的功效等概念的直观图示。不难看出，当样本量一定时，拒真概率趋向于 0 将导致取伪概率趋向于 1，反之亦然。也就是说，在样本量一定时，两类错误的概率无法同时被减小。

例 9.8. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ ，其中参数 σ^2 已知， $\mu \in \mathbb{R}$ 未知。在水平 α 对假设 $H_0 : \mu \leq \mu_0 \leftrightarrow H_1 : \mu > \mu_0$ 进行检验。

解. 与例 9.6 类似，样本均值 \bar{X} 越大，越倾向于否定 H_0 。定义功效函数如下并得到拒真概率的上确界 $\alpha(c)$ ，

$$\begin{aligned}\beta_\delta(\mu) &= P_\mu\{\bar{X} > c\} = 1 - \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right) \\ \alpha(c) &= \sup_{\mu \leq \mu_0} \beta_\delta(\mu) = \beta_\delta(\mu_0) = 1 - \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right)\end{aligned}$$

由 $\alpha(c) = \alpha$ 可得 $c = \mu_0 + z_{1-\alpha}\sigma / \sqrt{n}$ 。拒绝 H_0 的条件与例 9.6 的相同，并且该检验的功效函数为

$$\beta_\delta(\mu) = \Phi\left[\frac{\sqrt{n}(\mu - \mu_0)}{\sigma} - z_{1-\alpha}\right]$$

显然有，

$$\lim_{\mu \rightarrow \mu_0} \beta_\delta(\mu) = \alpha$$

$$\lim_{\mu \rightarrow \infty} \beta_\delta(\mu) = 1$$

它的含义也很明显： μ 较之 μ_0 越大，拒绝 H_0 的概率就越大。

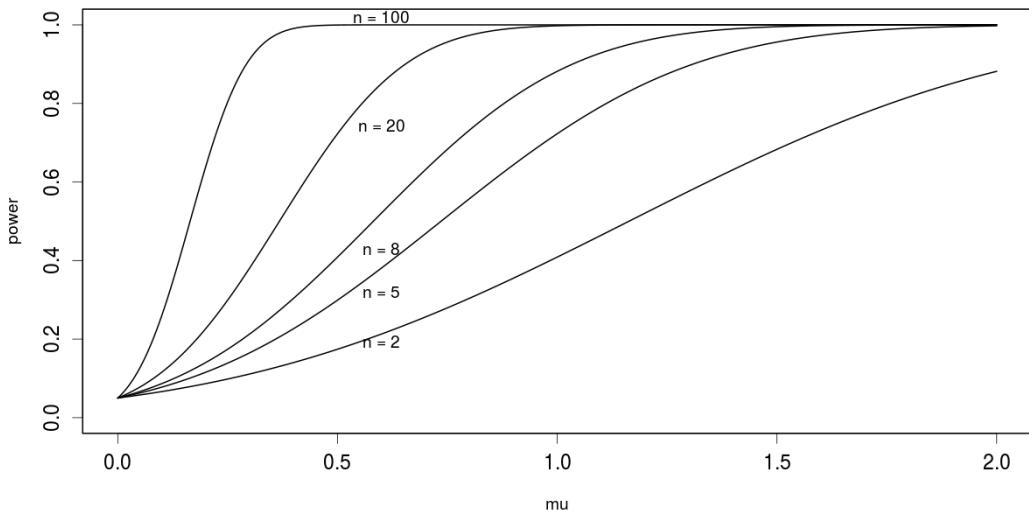


图 9.4: 在例 9.8 中，不妨令 $\mu_0 = 0, \sigma = 1$ ，在水平 $\alpha = 0.05$ ，不同的 n 所对应的功效函数如图所示。显然，样本容量 n 越大，功效函数 $\beta_\delta(\mu)$ 随着 μ 的增大趋向于 1 的速度就越快。

在面对同一数据、同一假设检验问题时，由于显著水平选取的不同，可能导致不同的结论。为了避免标准不一引起的不便，在实践中人们更多地使用 p -值 (p -value) 或检验的显著概率 (significance probability) 来报告假设检验的结果。

定义 9.10 (p -值). 基于样本观测值 $\mathbf{x} = (x_1, \dots, x_n)^\top$ ，定义 p -值为零假设 $H_0 : \theta \in \Theta_0$ 成立时，检验统计量 T 取值不小于 $T(\mathbf{x})$ 的最大概率。即，

$$p\text{-值} = \alpha[T(\mathbf{x})] = \sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) \geq T(\mathbf{x})\}$$

如果 p -值小于给定的显著水平 α ，则拒绝零假设 H_0 。

例 9.9. 在例 9.6 和例 9.8 中, p -值都为

$$p\text{-值} = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right), \text{ 其中 } \bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

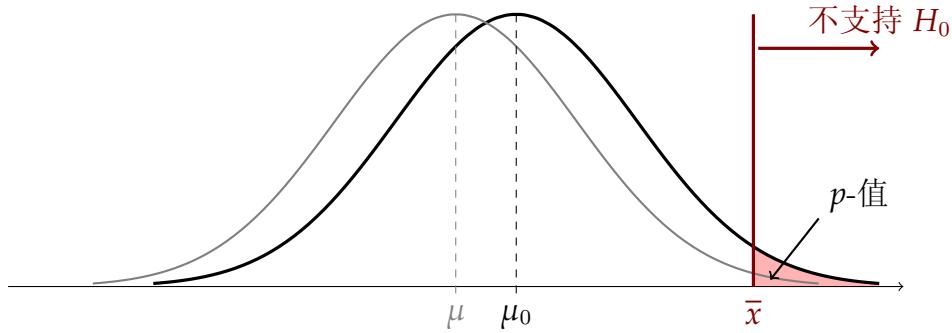


图 9.5: 在例 9.6 和例 9.8 中, 在零假设 H_0 为真的条件下, p -值的直观解释是观察到 $\{\bar{X} \geq \bar{x}\}$ 的概率, 即图中阴影部分的面积。显然, $\{\bar{X} = \bar{x}\}$ 以及更极端情形 $\{\bar{X} > \bar{x}\}$ 的概率越小, 表示 \bar{x} 越远离 μ_0 , 我们越倾向于拒绝 H_0 。

接着例 9.7 的试验, 计算得 p -值 = 0.7481197, 它远大于通常选定的显著水平 $\alpha = 0.05$ 或 $\alpha = 0.01$, 所以该观察数据无法拒绝零假设 H_0 。

9.1.2 Neyman-Pearson 引理与似然比检验

根据 Neyman-Pearson 原则, 当检验 δ 的拒真概率被控制在不超过 α 的前提之下, 该检验对备择假设 $H_1 : \theta \in \Theta_1$ 的功效越大越好。换句话说, 对于水平 α 检验 $\delta_1, \delta_2 \in \Delta_\alpha$, 如果 $\forall \theta \in \Theta_1$ 皆有 $\beta_{\delta_1}(\theta) \geq \beta_{\delta_2}(\theta)$, 则称检验 δ_1 优于 δ_2 , 记作 $\delta_1 \geq \delta_2$ 。显然, \geq 是定义在 Δ_α 上的偏序关系。

定义 9.11. 若水平 α 检验 $\delta^* \in \Delta_\alpha$ 优于任意的 $\delta \in \Delta_\alpha$, 即 $\forall \theta \in \Theta_1$ 皆有

$$\beta_{\delta^*}(\theta) \geq \beta_\delta(\theta), \text{ 其中 } \delta \in \Delta_\alpha$$

则称 δ^* 是显著水平 α 的一致最大功效 (uniformly most powerful, UMP) 检验, 或一致最优检验, 简称为水平 α -UMP 检验。

1933 年, J. Neyman 和 E. S. Pearson 在零假设和备择假设都是简单假设的情况下给出了如何构造 UMP 检验 [117], 即 Neyman-Pearson 引理, 该结果常被称为“数理统计学的基本引理”。

引理 9.1 (Neyman-Pearson, 1933). 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} f_\theta(x)$, 对于 $k \geq 0$, 定义简单假设 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$ 的检验函数 δ_k 为

$$\delta_k(\mathbf{x}) = \begin{cases} 1 & \text{若似然比 } \lambda(\mathbf{x}; \theta_0, \theta_1) = \frac{\mathcal{L}(\theta_1; \mathbf{x})}{\mathcal{L}(\theta_0; \mathbf{x})} \geq k \\ 0 & \text{否则} \end{cases}$$

若 $E_{\theta_0}[\delta_k(\mathbf{X})] = \alpha$, 则 δ_k 是水平 α -UMP 检验。

证明. 令 $\mathcal{L}(\theta_0; \mathbf{x})$ 和 $\mathcal{L}(\theta_1; \mathbf{x})$ 分别是 H_0 和 H_1 成立时的似然函数值。对于任意的检验函数 $\delta \in \Delta_\alpha$, 构造函数 $g(\mathbf{x})$ 如下。

$$\begin{aligned} g(\mathbf{x}) &= [\delta_k(\mathbf{x}) - \delta(\mathbf{x})] [\mathcal{L}(\theta_1; \mathbf{x}) - k\mathcal{L}(\theta_0; \mathbf{x})] \\ &= \begin{cases} [1 - \delta(\mathbf{x})] [\mathcal{L}(\theta_1; \mathbf{x}) - k\mathcal{L}(\theta_0; \mathbf{x})] & \text{当 } \lambda(\mathbf{x}; \theta_0, \theta_1) \geq k \\ [-\delta(\mathbf{x})] [\mathcal{L}(\theta_1; \mathbf{x}) - k\mathcal{L}(\theta_0; \mathbf{x})] & \text{当 } \lambda(\mathbf{x}; \theta_0, \theta_1) < k \end{cases} \end{aligned}$$

显然, $g(\mathbf{x}) \geq 0$, 于是

$$\int_{\mathbb{R}^n} g(\mathbf{x}) d\mathbf{x} = E_{\theta_1} [\delta_k(\mathbf{X})] - E_{\theta_1} [\delta(\mathbf{X})] - k \{E_{\theta_0} [\delta_k(\mathbf{X})] - E_{\theta_0} [\delta(\mathbf{X})]\} \geq 0$$

对于水平 α 检验 δ , 由 $E_{\theta_0} [\delta(\mathbf{X})] \leq E_{\theta_0} [\delta_k(\mathbf{X})] = \alpha$ 易得 $E_{\theta_1} [\delta(\mathbf{X})] \leq E_{\theta_1} [\delta_k(\mathbf{X})]$, 即 $\beta_\delta(\theta_1) \leq \beta_{\delta_k}(\theta_1)$, 加之 $\Theta_1 = \{\theta_1\}$, 所以 $\delta_k \geq \delta$ 。 \square

由引理 9.1 可知, 对于给定的显著水平 α , 检验函数 $\delta_k(\mathbf{x}) = 1$, 即拒绝 $H_0 : \theta = \theta_0$ 的条件是似然比大于某常数 k , 即

$$\lambda(\mathbf{x}; \theta_0, \theta_1) = \frac{\mathcal{L}(\theta_1; \mathbf{x})}{\mathcal{L}(\theta_0; \mathbf{x})} \geq k, \text{ 其中 } k \text{ 由 } E_{\theta_0}[\delta_k(\mathbf{X})] = \alpha \text{ 待定} \quad (9.1)$$

显然, 似然比越大越倾向于拒绝零假设 H_0 , 我们把 (9.1) 这样的检验称为似然比检验 (likelihood-ratio test)。Neyman-Pearson 引理 9.1 保证了两个简单假设之间的似然比检验是 UMP 检验, 然而一般很难直接由 $\lambda(\mathbf{x}) \geq k$ 解出拒绝域, 所幸的是有时通过函数 $\lambda(\mathbf{x})$ 关于由 \mathbf{x} 构造的某个量的单调性可以大大简化求解拒绝域的过程, 见下例。

例 9.10. 在显著水平 α , 给出例 9.6 的似然比检验。

解. 备择假设 $H_1 : \mu = \mu_1$ 和零假设 $H_0 : \mu = \mu_0$ 成立时的似然比是

$$\begin{aligned} \lambda(\mathbf{x}) &= \frac{\exp\left\{-\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu_1)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu_0)^2\right\}} \\ &= \exp\left\{\sum_{j=1}^n x_j \left(\frac{\mu_1}{\sigma^2} - \frac{\mu_0}{\sigma^2}\right) + n\left(\frac{\mu_0^2}{2\sigma^2} - \frac{\mu_1^2}{2\sigma^2}\right)\right\} \end{aligned}$$

显然 $\lambda(\mathbf{x})$ 是关于 $\sum_{j=1}^n x_j$ 的增函数。定义检验函数为

$$\delta_k(\mathbf{x}) = \begin{cases} 1 & \text{若 } \lambda(\mathbf{x}) \geq k \\ 0 & \text{若 } \lambda(\mathbf{x}) < k \end{cases}$$

函数 $\lambda(\mathbf{x}) \geq k$ 当且仅当 $\sum_{j=1}^n x_j \geq k_1$, 于是检验函数简化为

$$\delta(\mathbf{x}) = \begin{cases} 1 & \text{若 } \sum_{j=1}^n x_j \geq k_1 \\ 0 & \text{若 } \sum_{j=1}^n x_j < k_1 \end{cases}$$

上式中的 k_1 是由条件 $E_{\mu_0}[\delta(\mathbf{X})] = \alpha$ 确定的, 即

$$\begin{aligned} \alpha &= P_{\mu_0} \left\{ \sum_{j=1}^n X_j \geq k_1 \right\} \\ &= P \left\{ \frac{\sum_{j=1}^n X_j - n\mu_0}{\sigma \sqrt{n}} \geq \frac{k_1 - n\mu_0}{\sigma \sqrt{n}} \right\} \\ &= 1 - \Phi \left(\frac{k_1 - n\mu_0}{\sigma \sqrt{n}} \right) \end{aligned}$$

解之得 $k_1 = z_{1-\alpha} \sigma \sqrt{n} + n\mu_0$, 拒绝零假设 $H_0 : \mu = \mu_0$ 的条件是 $\sum_{j=1}^n x_j \geq z_{1-\alpha} \sigma \sqrt{n} + n\mu_0$, 即 $\bar{x} \geq \mu_0 + z_{1-\alpha} \sigma / \sqrt{n}$, 与例 9.6 的结果相同。由 Neyman-Pearson 引理 9.1, 检验 $\delta(x)$ 是水平 α -UMP 检验。

例 9.11. 在显著水平 α , 基于单个样本点 X 对总体分布的简单假设 $H_0 : X \sim N(0, 1) \leftrightarrow H_1 : X \sim \text{Cauchy}(0, 1)$ 进行似然比检验。

解. 备择假设 H_1 和零假设 H_0 成立时的似然比是

$$\begin{aligned}\frac{f_1(x)}{f_0(x)} &= \frac{1/(\pi + \pi x^2)}{1/\sqrt{2\pi} \exp\{-x^2/2\}} \\ &= \sqrt{\frac{2}{\pi}} \frac{\exp\{x^2/2\}}{1+x^2}\end{aligned}$$

由 Neyman-Pearson 引理 9.1, 该问题的 UMP 检验具有形式

$$\delta(x) = \begin{cases} 1 & \text{如果 } \sqrt{\frac{2}{\pi}} \frac{\exp\{x^2/2\}}{1+x^2} \geq k \\ 0 & \text{其他} \end{cases}$$

其中 k 由 $E_0[\delta(X)] = \alpha$ 唯一确定, $E_0[\delta(X)]$ 表示零假设成立时 $\delta(X)$ 的期望, 直接计算很困难。容易发现当 $|x| > 1$ 时, $f_1(x)/f_0(x)$ 是关于 $|x|$ 的非减函数, 尝试着定义检验函数为

$$\delta(x) = \begin{cases} 1 & \text{如果 } |x| \geq k_1 \\ 0 & \text{如果 } |x| < k_1 \end{cases}$$

其中 k_1 由 $E_0[\delta(X)] = 2[1 - \Phi(k_1)] = \alpha$ 唯一确定, 解之得 $k_1 = z_{1-\alpha/2}$ 。因为通常 α 为接近零的正数, 所以能保证 $k_1 > 1$ 。当样本值 x 满足 $|x| \geq z_{1-\alpha/2}$ 时拒绝零假设 H_0 。这样的检验 δ 对备择假设的功效是

$$\begin{aligned}E_1[\delta(X)] &= 1 - \int_{-k_1}^{k_1} \frac{1}{\pi(1+x^2)} dx \\ &= 1 - \frac{2}{\pi} \arctan z_{1-\alpha/2}\end{aligned}$$

如果显著水平 $\alpha = 0.05$, 则第二类错误的概率为

$$1 - E_1[\delta(X)] = \frac{2}{\pi} \arctan z_{1-\alpha/2} \approx 0.6996524$$

显然, α 越小, 第二类错误的概率越大。

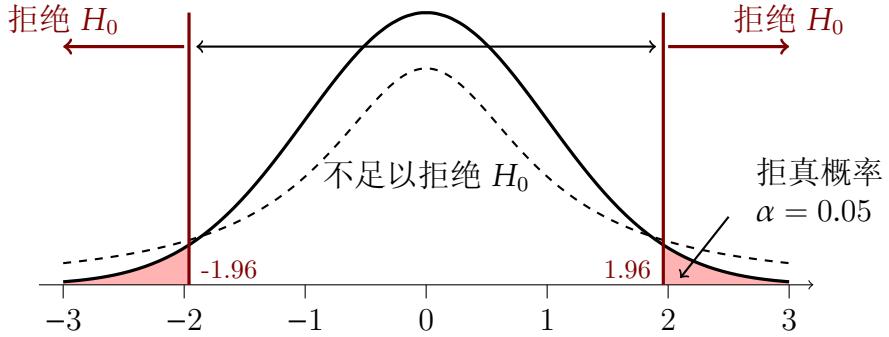


图 9.6: 实线是 $N(0, 1)$ 的密度曲线, 虚线是 $\text{Cauchy}(0, 1)$ 的密度曲线。例 9.11 中, 若样本 X 落于 H_0 的接受域 $[-1.96, 1.96]$ 上, X 更像是来自 $N(0, 1)$ 。

练习 9.2. 利用大量独立重复的试验, 以拒真频率和取伪频率来粗略考察例 9.11 中两类错误的概率。

定义 9.12. 分布族 $\{f_\theta(x) : \theta \in \Theta\}$ 称为对统计量 $T(\mathbf{X})$ 具有单调似然比 (monotone likelihood ratio, MLR), 如果对 $\theta_0 < \theta_1$, 密度函数 $f_{\theta_0} \neq f_{\theta_1}$ 且似然比 $\lambda(x) = f_{\theta_1}(x)/f_{\theta_0}(x)$ 是关于 $T(x)$ 的非减函数。

例 9.12. 令 $X \sim \text{Cauchy}(\theta, 1)$, 则当 $x \rightarrow \pm\infty$ 时,

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = \frac{(x - \theta_0)^2 + 1}{(x - \theta_1)^2 + 1} \rightarrow 1, \text{ 其中 } \theta_0 < \theta_1$$

于是 $\text{Cauchy}(\theta, 1)$ 没有单调似然比。

例 9.13. 单参数指数族 $\{f_\theta(x) = h(x)\eta(\theta) \exp[\lambda(\theta)T(x)] : \theta \in \Theta \subseteq \mathbb{R}$ 且实值函数 $\lambda(\theta)$ 关于 θ 非减 } 对统计量 $T(\mathbf{X})$ 具有单调似然比。

1956 年, 美国应用数学家、统计学家 Samuel Karlin (1924-2007) 和 Herman Rubin (1926-) 推广了 Neyman-Pearson 引理, 得到了一类复合假设的 UMP 检验。

定理 9.1 (Karlin-Rubin, 1956). 设样本 $\mathbf{X} \sim f_\theta(x)$, 其中未知参数 $\theta \in \Theta$, 如果 $\{f_\theta\}$ 对统计量 $T(\mathbf{X})$ 具有单调似然比, 则单侧检验问题 $H_0 : \theta \leq \theta_0 \leftrightarrow H_1 : \theta > \theta_0$ 的检验函数

$$\delta(x) = \begin{cases} 0 & \text{若 } T(x) \leq t_0 \\ 1 & \text{若 } T(x) > t_0 \end{cases}$$

具有非减的功效函数且是水平 $\alpha = P\{T > t_0\}$ 的 UMP 检验。

证明. 见 V. K. Rohatgi 的《概率论及数理统计导论》[137] 第九章第四节 (第 420-421 页) 或 G. Casella 和 R. L. Berger 的《统计推断》[24] 第八章第三节。 \square

推论 9.1. 令 $\theta_0 < \theta_1$, 例 9.13 中的单参数指数族存在对 $H_0 : \theta \leq \theta_0$ 或 $\theta \geq \theta_1 \leftrightarrow H_1 : \theta_0 < \theta < \theta_1$ 的 UMP 检验如下,

$$\delta(\mathbf{x}) = \begin{cases} 0 & \text{若 } T(\mathbf{x}) \leq c_1 \text{ 或 } T(\mathbf{x}) \geq c_2 \\ 1 & \text{若 } c_1 < T(\mathbf{x}) < c_2 \end{cases}$$

其中 c_1, c_2 由 $E_{\theta_0}\delta(\mathbf{X}) = E_{\theta_1}\delta(\mathbf{X}) = \alpha$ 解出, α 是给定的显著水平。

例 9.14. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, 1)$, 其中参数 μ 未知。在显著水平 α 给出复合假设 $H_0 : \mu \leq \mu_0$ 或 $\mu \geq \mu_1 \leftrightarrow H_1 : \mu_0 < \mu < \mu_1$ 的 UMP 检验。

解. 样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的密度函数 $f_\mu(\mathbf{x})$ 属于单参数指数族, 对统计量 $T(\mathbf{X}) = \sum_{j=1}^n X_j$ 具有单调似然比, 这是因为

$$f_\mu(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp \left\{ -\frac{1}{2} \sum_{j=1}^n x_j^2 \right\} \exp \left\{ -\frac{n\mu^2}{2} \right\} \exp \left\{ \mu \sum_{j=1}^n x_j \right\}$$

根据推论 9.1, 复合假设 $H_0 : \mu \leq \mu_0$ 或 $\mu \geq \mu_1 \leftrightarrow H_1 : \mu_0 < \mu < \mu_1$ 的 UMP 检验函数为

$$\delta(\mathbf{x}) = \begin{cases} 0 & \text{若 } \sum_{j=1}^n x_j \leq c_1 \text{ 或 } \sum_{j=1}^n x_j \geq c_2 \\ 1 & \text{若 } c_1 < \sum_{j=1}^n x_j < c_2 \end{cases}$$

其中 c_1, c_2 由 $E_{\mu_0}\delta(\mathbf{X}) = E_{\mu_1}\delta(\mathbf{X}) = \alpha$ 解出。例如, 由 $E_{\mu_0}\delta(\mathbf{X}) = \alpha$ 得到,

$$\begin{aligned} \alpha &= P_{\mu_0} \left\{ c_1 < \sum_{j=1}^n X_j < c_2 \right\} \\ &= P_{\mu_0} \left\{ \frac{c_1 - n\mu_0}{\sqrt{n}} < \frac{\sum_{j=1}^n X_j - n\mu_0}{\sqrt{n}} < \frac{c_2 - n\mu_0}{\sqrt{n}} \right\} \\ &= \Phi \left(\frac{c_2 - n\mu_0}{\sqrt{n}} \right) - \Phi \left(\frac{c_1 - n\mu_0}{\sqrt{n}} \right) \end{aligned}$$

同理, $\Phi[(c_2 - n\mu_1)/\sqrt{n}] - \Phi[(c_1 - n\mu_1)/\sqrt{n}] = \alpha$, 与上式联立解出 c_1, c_2 即可。 c_1, c_2 没有解析表达式, 但可以利用 Newton 法求它们的数值解。

UMP 检验虽好, 但可遇不可求, 对于大多数的检验问题并不存在。试想, 一个水平 α 检验 δ 可以在 Θ_1 中的某些参数上取得很大的功效, 同时在另一些参数上取得很小的功效也未尝。水平 α -UMP 检验 δ^* 要满足 $\forall \delta \in \Delta_\alpha$, 函数 $\beta_{\delta^*}(\theta) \geq \beta_\delta(\theta)$ 这一严酷的条件, 即在每个参数 $\theta \in \Theta_1$ 上 δ^* 都拔得头筹, UMP 检验的凤毛麟角就不难想象了。

为了制定出评估检验优劣的合理标准，人们提出了无偏检验、相似检验等概念，把检验限制在某个函数类中再精挑细选出 UMP 检验。即便如此兴师动众，找到受限的 UMP 检验也非易事。想深入了解功效函数和检验优良性等内容的读者可参阅 [101]。

9.1.3 广义似然比检验

受 Fisher 最大似然估计思想的影响, 1928 年, E. S. Pearson 和 J. Neyman 提出了广义似然比检验。(1) 该方法的适用范围较广, 不管样本容量是大是小, 它都具有一定的可行性, 但在一般情况下有计算上的困难。(2) 广义似然比检验不一定是 UMP 的, 但当样本量足够大时, 取伪概率也能控制得不错, 通常是渐近最优的。广义似然比检验在假设检验理论中的地位如同最大似然估计在点估计理论中的地位一样崇高, 它在正态分布样本上取得了漂亮的结果, 见例 9.16 至例 9.20。

定义 9.13 (广义似然比). 令 $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$ 为一个向量参数, 样本 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的似然函数为 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$, 其中 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 。考虑零假设 $H_0: \boldsymbol{\theta} \in \Theta_0$, 令 $\hat{\boldsymbol{\theta}}_0$ 和 $\hat{\boldsymbol{\theta}}$ 分别是参数限定在参数空间 Θ_0 和 Θ 上的最大似然估计。广义似然比 (generalized likelihood ratio, GLR) 定义为

$$\lambda(\mathbf{x}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x})}{\sup_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x})} = \frac{\mathcal{L}(\hat{\boldsymbol{\theta}}_0; \mathbf{x})}{\mathcal{L}(\hat{\boldsymbol{\theta}}; \mathbf{x})} \quad (9.2)$$

有的文献把广义似然比定义为 $1/\lambda(\mathbf{x})$, 采用哪种定义只是行文习惯不同而已, 后续的方法是类似的。本书之所以采用定义 (9.2), 是因为下面的性质。

性质 9.2. 式 (9.2) 定义的广义似然比 $\lambda(\mathbf{x})$ 满足 $0 \leq \lambda(\mathbf{x}) \leq 1$ 。

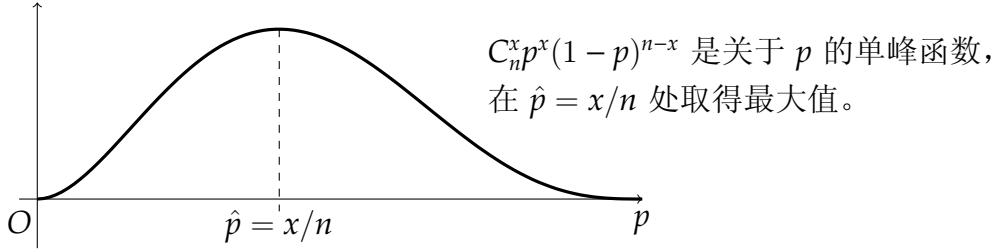
直观上, 如果 $H_0: \boldsymbol{\theta} \in \Theta_0$ 为真, 则广义似然比 (9.2) 必然接近 1。换句话说, 如果这个比值很小就应该否定 H_0 。我们把“拒绝 $H_0: \boldsymbol{\theta} \in \Theta_0$ 当且仅当 $\lambda(\mathbf{x}) < c$ ”这样的检验称为广义似然比检验或 GLR 检验。如何待定出临界值 $c \in (0, 1)$? 为了使检验的拒真概率不超过给定的显著水平 α , 临界值 c 可通过求解下述方程得到。

$$\sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}\{\lambda(\mathbf{X}) < c\} = \alpha \quad (9.3)$$

式 (9.3) 中左边即拒真错误的上确界, 其中 $P_{\boldsymbol{\theta}}\{\lambda(\mathbf{X}) < c\}$ 有时需要利用函数 $\lambda(\mathbf{x})$ 的单调性来求解——这是广义似然比检验的技巧所在。

例 9.15. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} p\langle 1 \rangle + (1-p)\langle 0 \rangle$, 其中参数 p 未知, 对假设 $H_0: p \leq p_0 \leftrightarrow H_1: p > p_0$ 进行水平为 α 的广义似然比检验。

解. 令 $X = X_1 + X_2 + \dots + X_n$, 则 $X \sim \text{B}(n, p)$ 。



$$\sup_{0 \leq p \leq 1} C_n^x p^x (1-p)^{n-x} = C_n^x (x/n)^x (1-x/n)^{n-x}$$

$$\sup_{p \leq p_0} C_n^x p^x (1-p)^{n-x} = \begin{cases} C_n^x p_0^x (1-p_0)^{n-x} & \text{如果 } p_0 < x/n \\ C_n^x (x/n)^x (1-x/n)^{n-x} & \text{如果 } p_0 \geq x/n \end{cases}$$

由式 (9.2) 计算广义似然比,

$$\begin{aligned} \lambda(x) &= \frac{\sup_{p \leq p_0} C_n^x p^x (1-p)^{n-x}}{\sup_{0 \leq p \leq 1} C_n^x p^x (1-p)^{n-x}} \\ &= \begin{cases} 1 & \text{如果 } x \leq np_0 \\ \frac{p_0^x (1-p_0)^{n-x}}{(x/n)^x (1-x/n)^{n-x}} & \text{如果 } x > np_0 \end{cases} \end{aligned}$$

广义似然比 $\lambda(x)$ 是一个关于 x 的减函数, 于是 $\lambda(x) < c \Leftrightarrow x > c'$ 。即如果 X 的观测值 $x > c'$, 广义似然比检验否定 H_0 。

$$\begin{aligned} \sup_{p \leq p_0} P_p\{X > c'\} &= P_{p_0}\{X > c'\} \\ &= 1 - \sum_{k=0}^{\lfloor c' \rfloor} C_n^k p_0^k (1-p_0)^{n-k} \end{aligned}$$

因为 X 是离散型随机变量, 可通过下面的方法求得临界值 c' : $P_{p_0}\{X > c'\} \leq \alpha$ 且 $P_{p_0}\{X > c' - 1\} > \alpha$ 。

练习 9.3. 在例 9.15 中, 对假设 $H_0: p = p_0 \leftrightarrow H_1: p \neq p_0$ 进行水平为 α 的广义似然比检验。提示: 模仿例 9.15 的解法。

下面是针对正态总体中未知参数的广义似然比假设检验, 也要用到类似例 9.15 的技巧, 即利用广义似然比函数的单调性来求解临界值。

例 9.16. 总体的分布为 $N(\mu, \sigma^2)$, 其中参数 $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ 未知, 在水平 α 对假设 $H_0: \mu = \mu_0 \leftrightarrow H_1: \mu \neq \mu_0$ 进行广义似然比检验。

解. 下面分别考虑 $\sup_{\theta \in \Theta_0} f_\theta(\mathbf{x})$ 和 $\sup_{\theta \in \Theta} f_\theta(\mathbf{x})$, 其中 Θ_0 是零假设成立时的参数空间, Θ 是整个参数空间, $\mathbf{x} = (x_1, \dots, x_n)^\top$ 。

□ 零假设成立时的参数空间为 $\Theta_0 = \{(\mu_0, \sigma^2)^\top : \sigma^2 > 0\}$, 似然函数的上确界为

$$\begin{aligned}\sup_{\theta \in \Theta_0} f_\theta(\mathbf{x}) &= \sup_{\sigma^2 > 0} \left[\left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{\sum_{j=1}^n (x_j - \mu_0)^2}{2\sigma^2} \right\} \right] \\ &= \left(\frac{1}{\sqrt{2\pi e \hat{\sigma}^2}} \right)^n\end{aligned}$$

其中, $\hat{\sigma}^2$ 是参数 σ^2 的最大似然估计值, 即

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^n (x_j - \mu_0)^2 \\ &= (\bar{x} - \mu_0)^2 + \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2\end{aligned}$$

□ 在整个参数空间 $\Theta = \{(\mu, \sigma^2)^\top : \mu \in \mathbb{R}, \sigma^2 > 0\}$ 上, 参数 $\theta = (\mu, \sigma^2)^\top$ 的最大似然估计值和似然函数的上确界如下。

$$\begin{aligned}\hat{\theta} &= \left(\frac{1}{n} \sum_{j=1}^n x_j, \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 \right)^\top = (a_1, b_2)^\top \\ \sup_{\theta \in \Theta} f_\theta(\mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi e b_2}} \right)^n\end{aligned}$$

于是, 广义似然比为

$$\lambda(\mathbf{x}) = \left(\frac{b_2}{\hat{\sigma}^2} \right)^{n/2} = \left[1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right]^{-n/2}$$

因为 $\lambda(\mathbf{x})$ 是有关 $n(\bar{x} - \mu_0)^2 / \sum_{j=1}^n (x_j - \bar{x})^2$ 的减函数, 如果 $\lambda(\mathbf{x}) < c$ 是广义似然比检验拒绝零假设 H_0 的条件, 则该条件等价于

$$\left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2}} \right| > c' \text{ 或者等价地, } \left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \right| > c''$$

其中 $s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$ 。由性质 7.10, 当 H_0 成立时,

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

故选取 $c'' = t_{n-1,1-\alpha/2}$, 即在水平 α 拒绝零假设 H_0 的条件是

$$\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} > t_{n-1,1-\alpha/2}$$

例 9.17. 接着例 9.16, 在水平 α 对 $H_0 : \mu \leq \mu_0 \leftrightarrow H_1 : \mu > \mu_0$ 进行广义似然比检验。

解. 似然函数 $f_\theta(\mathbf{x}) = \prod_{j=1}^n \phi(x_j | \mu, \sigma^2)$ 关于 μ 是单峰函数, 在整个参数空间 Θ 上, $f_\theta(\mathbf{x})$ 都在 $\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$ 处取得最大值。考虑下面两种情况。

- 如果 $\bar{x} \leq \mu_0$, 在子空间 $\Theta_0 = \{(\mu, \sigma^2)^\top : \mu \leq \mu_0, \sigma^2 > 0\}$ 上, 似然函数 $f_\theta(\mathbf{x})$ 也在 $\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$ 处取得最大值。
- 如果 $\bar{x} > \mu_0$, 在 Θ_0 上, $f_\theta(\mathbf{x})$ 在 $\hat{\mu} = \mu_0, \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \mu_0)^2$ 处取得最大值。

类似例 9.16, 求得广义似然比 $\lambda(\mathbf{x})$ 如下,

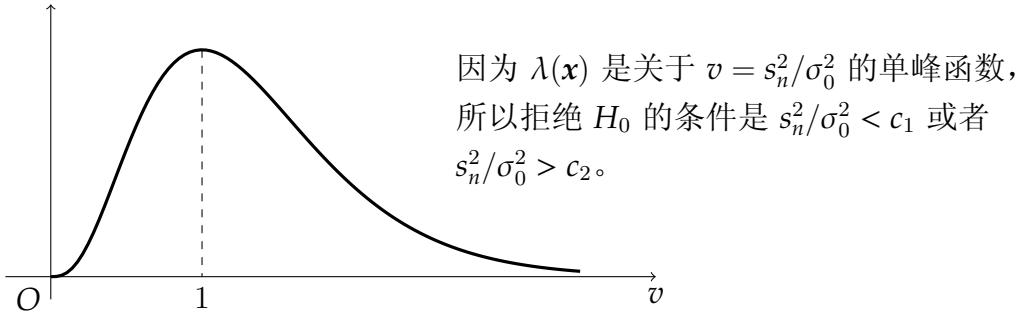
$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{当 } \bar{x} \leq \mu_0 \\ \left(1 + \frac{t^2}{n-1}\right)^{-n/2} & \text{当 } \bar{x} > \mu_0, \text{ 其中 } t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \end{cases}$$

显然, 广义似然比 $\lambda(\mathbf{x})$ 是关于 $t = \sqrt{n}(\bar{x} - \mu_0)/s$ 的减函数, 因为 $\sqrt{n}(\bar{X} - \mu_0)/S \sim t_{n-1}$, 所以在水平 α 拒绝 H_0 的条件是 $t > t_{n-1,1-\alpha}$ 。

例 9.18. 接着例 9.16, 在水平 α 对 $H_0 : \sigma^2 = \sigma_0^2 \leftrightarrow H_1 : \sigma^2 \neq \sigma_0^2$ 进行广义似然比检验。

解. 仿照例 9.16 求得广义似然比 $\lambda(\mathbf{x})$ 如下,

$$\begin{aligned} \lambda(\mathbf{x}) &= \frac{(2\pi\sigma_0^2)^{-n/2} \exp\{-\frac{1}{2\sigma_0^2} \sum_{j=1}^n (x_j - \bar{x})^2\}}{(2\pi s_n^2)^{-n/2} \exp\{-n/2\}} \\ &= \left(\frac{s_n^2}{\sigma_0^2} \exp\left\{1 - \frac{s_n^2}{\sigma_0^2}\right\} \right)^{n/2} \end{aligned}$$



等价地，拒绝 H_0 的条件是 $(n-1)s^2/\sigma_0^2 < c'_1$ 或者 $(n-1)s^2/\sigma_0^2 > c'_2$ 。由性质 7.9，当 H_0 成立时， $(n-1)S^2/\sigma_0^2 \sim \chi_{n-1}^2$ ，选择 $c'_1 = \chi_{n-1,\alpha/2}^2$ 和 $c'_2 = \chi_{n-1,1-\alpha/2}^2$ ，即在水平 α 拒绝 H_0 的条件是

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi_{n-1,\alpha/2}^2 \text{ 或者 } \frac{(n-1)s^2}{\sigma_0^2} > \chi_{n-1,1-\alpha/2}^2$$

※例 9.19. 设样本 $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} N(\mu_X, \sigma_X^2)$ 和 $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2)$ ，且两总体是独立的，其中参数 $\theta = (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2)^\top$ 未知，在水平 α 对假设 $H_0 : \sigma_X^2 = \sigma_Y^2 \leftrightarrow H_1 : \sigma_X^2 \neq \sigma_Y^2$ 进行广义似然比检验。

解. 整个未知参数空间是 $\Theta = \{(\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2)^\top : \mu_X, \mu_Y \in \mathbb{R}, \sigma_X^2, \sigma_Y^2 > 0\}$ ，两个样本 $(X_1, \dots, X_m, Y_1, \dots, Y_n)^\top$ 的联合密度函数为

$$f_\theta(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^{(m+n)/2} \sigma_X^m \sigma_Y^n} \exp \left\{ -\frac{\sum_{j=1}^m (x_j - \mu_X)^2}{2\sigma_X^2} - \frac{\sum_{j=1}^n (y_j - \mu_Y)^2}{2\sigma_Y^2} \right\}$$

□ 在参数空间 Θ 上， $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ 的最大似然估计值分别为

$$\begin{aligned} \hat{\mu}_X &= \bar{x} & \hat{\sigma}_X^2 &= \frac{1}{m} \sum_{j=1}^m (x_j - \bar{x})^2 \\ \hat{\mu}_Y &= \bar{y} & \hat{\sigma}_Y^2 &= \frac{1}{n} \sum_{j=1}^n (y_j - \bar{y})^2 \end{aligned}$$

□ 当 H_0 成立的时候，在参数空间 $\Theta_0 = \{(\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2)^\top : \mu_X, \mu_Y \in \mathbb{R}, \sigma_X^2 = \sigma_Y^2 = \sigma^2 > 0\}$ 上， μ_X, μ_Y, σ^2 的最大似然估计值分别为

$$\begin{aligned} \hat{\mu}_X &= \bar{x} & \hat{\mu}_Y &= \bar{y} \\ \hat{\sigma}^2 &= \frac{1}{m+n} \left[\sum_{j=1}^m (x_j - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2 \right] \end{aligned}$$

经过简单的计算，得到广义似然比如下。

$$\begin{aligned}\lambda(\mathbf{x}, \mathbf{y}) &= \sqrt{\frac{(m+n)^{m+n} \left[\sum_{j=1}^m (x_j - \bar{x})^2 \right]^m \left[\sum_{j=1}^n (y_j - \bar{y})^2 \right]^n}{m^m n^n \left[\sum_{j=1}^m (x_j - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2 \right]^{m+n}}} \\ &= \sqrt{\frac{(m+n)^{m+n}}{m^m n^n \left[1 + \frac{m-1}{n-1} f \right]^n \left[1 + \frac{n-1}{m-1} \cdot \frac{1}{f} \right]^m}} \\ \text{其中, } f &= \frac{\sum_{j=1}^m (x_j - \bar{x})^2 / (m-1)}{\sum_{j=1}^n (y_j - \bar{y})^2 / (n-1)}\end{aligned}$$

因为广义似然比 $\lambda(\mathbf{x}, \mathbf{y})$ 是关于 f 的单峰函数，所以 $\lambda(\mathbf{x}, \mathbf{y}) < c$ 等价于 $f < c_1$ 或 $f > c_2$ 。根据性质 7.11 的第一个结论，当 H_0 成立时，

$$F = \frac{\sum_{j=1}^m (X_j - \bar{X})^2 / (m-1)}{\sum_{j=1}^n (Y_j - \bar{Y})^2 / (n-1)} \sim F_{m-1, n-1}$$

选取 $c_1 = F_{m-1, n-1, \alpha/2}$ 和 $c_2 = F_{m-1, n-1, 1-\alpha/2}$ ，在水平 α 拒绝零假设 H_0 的条件是 $f < F_{m-1, n-1, \alpha/2}$ 或者 $f > F_{m-1, n-1, 1-\alpha/2}$ 。

例 9.20. 设简单随机样本 $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top$ 来自二元正态总体 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ，其中参数 $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)^\top$ 未知。众所周知，随机变量 X, Y 相互独立当且仅当 $\rho = 0$ 。令零假设是 X, Y 相互独立，在水平 α 对假设 $H_0 : \rho = 0 \leftrightarrow H_1 : \rho \neq 0$ 进行广义似然比检验。

解. 零假设 H_0 成立和一般情况下的最大似然函数如下，

$$\begin{aligned}f_{\hat{\boldsymbol{\theta}}_0}(\mathbf{x}, \mathbf{y}) &= \frac{1}{(2\pi\hat{\sigma}_X^2\hat{\sigma}_Y^2)^n} \exp\left(-\frac{n}{2}\right) \\ f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{y}) &= \frac{1}{(2\pi\hat{\sigma}_X^2\hat{\sigma}_Y^2)^n} \cdot \frac{1}{\sqrt{(1-\hat{\rho}^2)^n}} \exp\left(-\frac{n}{2}\right) \\ \text{其中, } \hat{\rho} &= \frac{1}{n\hat{\sigma}_X\hat{\sigma}_Y} \sum_{i=1}^n (x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y) \\ \hat{\mu}_X &= \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\sigma}_X^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_X)^2\end{aligned}$$

广义似然比函数 $\lambda(\mathbf{x}, \mathbf{y}) = \sqrt{(1-\hat{\rho}^2)^n}$ 关于 $\hat{\rho}^2$ 递减，于是拒绝 H_0 的条件 $\lambda(\mathbf{x}, \mathbf{y}) < c$ 等价于 $|\hat{\rho}| > c'$ 。

参考习题 8.12 的结果，当 n 足够大时，渐近地有

$$\frac{\sqrt{n}(\hat{\rho} - \rho)}{1 - \rho^2} \sim N(0, 1)$$

若 H_0 成立，则 $\sqrt{n}\hat{\rho} \sim N(0, 1)$ 。于是，拒绝 H_0 的条件是 $|\hat{\rho}| > z_{1-\alpha/2}/\sqrt{n}$ 。

9.1.4 假设检验与置信区间估计的关系

假设检验与置信区间估计有着密切的联系，它们共同反映出 Neyman 的统计思想。假设检验理论在先，置信区间估计理论在后，Neyman 的初衷是在二者之间建立联系，把假设检验的某些结果转化为区间估计的结果。考虑到本书介绍这两个理论的次序，我们把区间估计的某些结果转化为假设检验的结果。

若 θ 是未知参数，在显著水平 α ，构造简单假设 $H_0 : \theta = \theta_0$ 的接受域：如果 H_0 成立， θ 的置信度为 $1 - \alpha$ 的置信区间，就是由样本 \mathbf{X} 构造的、以不小于 $1 - \alpha$ 的概率覆盖住 θ 的随机区间 $[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$ ，即 $P_{\theta_0}[\underline{\theta}(\mathbf{X}) \leq \theta_0 \leq \bar{\theta}(\mathbf{X})] \geq 1 - \alpha$ 。仿照置信区间的构造方式，定义检验函数 $\delta(\mathbf{x})$ 如下，

$$\delta(\mathbf{x}) = \begin{cases} 0 & \text{若 } \theta_0 \in [\underline{\theta}(\mathbf{x}), \bar{\theta}(\mathbf{x})] \\ 1 & \text{否则} \end{cases} \quad (9.4)$$

区域 $A = \{\mathbf{x} \in \mathbb{R}^n : \theta_0 \in [\underline{\theta}(\mathbf{x}), \bar{\theta}(\mathbf{x})]\}$ 是水平 α 下对简单假设 $H_0 : \theta = \theta_0$ 的接受域，使得检验 $\delta(\mathbf{x})$ 犯拒真错误的概率不超过 α ，这是因为

$$\begin{aligned} P_{\theta_0}\{\delta(\mathbf{X}) = 1\} &= 1 - P_{\theta_0}[\underline{\theta}(\mathbf{X}) \leq \theta_0 \leq \bar{\theta}(\mathbf{X})] \\ &\leq 1 - (1 - \alpha) = \alpha \end{aligned}$$

于是，在显著水平 α ，零假设 $H_0 : \theta = \theta_0$ 的拒绝域是 $(-\infty, \underline{\theta}(\mathbf{x})) \cup (\bar{\theta}(\mathbf{x}), \infty)$ 。显著水平 α 与置信度 $1 - \alpha$ 的频率解释（见图 8.13）类似，即在大量重复的随机试验中，基于式 (9.4) 的检验犯拒真错误的频率接近 α ，例如图 8.13 所示的 100 次重复试验中，有 4 次拒绝 $H_0 : \mu = 0$ ，拒真错误的频率是 4%，接近 $\alpha = 5\%$ 。

如果 $\underline{\theta}(\mathbf{X})$ 是 θ 的置信度为 $1 - \alpha$ 的置信下限，即 $P_{\theta}\{\underline{\theta}(\mathbf{X}) \leq \theta\} = 1 - \alpha$ ，则在显著水平 α 拒绝 $H_0 : \theta \leq \theta_0$ 当且仅当 $\theta_0 < \underline{\theta}(\mathbf{x})$ ，拒真错误的概率不超过 α 。类似地，如果 $\bar{\theta}(\mathbf{X})$ 是 θ 的置信度为 $1 - \alpha$ 的置信上限，即 $P_{\theta}\{\theta \leq \bar{\theta}(\mathbf{X})\} = 1 - \alpha$ ，则在显著水平 α 拒绝 $H_0 : \theta \geq \theta_0$ 当且仅当 $\theta_0 > \bar{\theta}(\mathbf{x})$ 。

例 9.21. 设样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ ，样本均值和方差分别为 \bar{X} 和 S^2 （它们的观察值分别记为 \bar{x} 和 s^2 ）。将参数 σ^2 分为已知和未知两种情况，在显著水平 α 对未知参数 μ 进行假设检验。

□ 当 σ^2 已知时，考虑假设 $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$ 的检验问题：如果零假设 H_0 成立，由性质 8.5 中的第一种情况给出 μ_0 的置信度为 $1 - \alpha$ 的置信区间

$$\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

于是，拒绝 $H_0 : \mu = \mu_0$ 的条件如下，

$$\frac{|\bar{x} - \mu_0|}{\sigma / \sqrt{n}} > z_{1-\alpha/2}$$

像这种用到了正态分布分位数的双侧或单侧检验统称为 z 检验。

- 当 σ^2 未知时，考虑假设 $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$ 的检验问题：由性质 8.5 中的第二种情况给出拒绝 $H_0 : \mu = \mu_0$ 的条件如下，

$$\frac{|\bar{x} - \mu_0|}{s / \sqrt{n}} > t_{n-1,1-\alpha/2}$$

像这种用到了 t 分布分位数的双侧或单侧检验统称为 t 检验。

- 当 σ^2 已知时，考虑假设 $H_0 : \mu \leq \mu_0 \leftrightarrow H_1 : \mu > \mu_0$ 的检验问题：参数 μ 的置信度为 $1 - \alpha$ 的置信下限是 $\bar{X} - z_{1-\alpha}\sigma / \sqrt{n}$ ，拒绝 $H_0 : \mu \leq \mu_0$ 的条件是 $\mu_0 < \bar{x} - z_{1-\alpha}\sigma / \sqrt{n}$ ，即

$$\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_{1-\alpha}$$

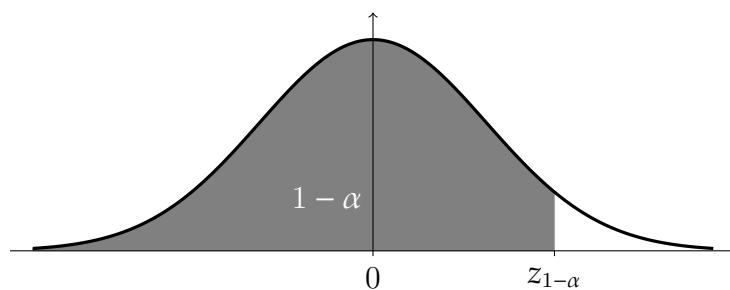


图 9.7: 因为 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ ，所以 $P_\mu \left\{ \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha} \right\} = 1 - \alpha$ 。

表 9.2: 设总体 $X \sim N(\mu, \sigma^2)$ ，表中给出了在水平 α 拒绝零假设 H_0 的条件。

$H_0 \leftrightarrow H_1$	σ^2 已知	σ^2 未知
$\mu = \mu_0 \leftrightarrow \mu \neq \mu_0$	$\frac{ \bar{x} - \mu_0 }{\sigma / \sqrt{n}} > z_{1-\alpha/2}$	$\frac{ \bar{x} - \mu_0 }{s / \sqrt{n}} > t_{n-1,1-\alpha/2}$
$\mu \leq \mu_0 \leftrightarrow \mu > \mu_0$	$\frac{ \bar{x} - \mu_0 }{\sigma / \sqrt{n}} > z_{1-\alpha}$	$\frac{ \bar{x} - \mu_0 }{s / \sqrt{n}} > t_{n-1,1-\alpha}$
$\mu \geq \mu_0 \leftrightarrow \mu < \mu_0$	$\frac{ \bar{x} - \mu_0 }{\sigma / \sqrt{n}} < z_\alpha$	$\frac{ \bar{x} - \mu_0 }{s / \sqrt{n}} < t_{n-1,\alpha}$

练习 9.4. 请读者仿照例 9.21, 验证表 9.2 中其他结果。

例 9.22. 工厂生产某种灯管, 假定灯管寿命 X 服从正态分布 $N(\mu, \sigma^2)$, 其中参数 μ, σ^2 都未知。随机抽取 25 个样本, 测得寿命值 (单位: 天) 如下: 285, 271, 328, 538, 81, 585, 308, 416, 228, 374, 187, 459, 216, 347, 664, 159, 304, 143, 412, 339, 84, 155, 31, 76, 137。在显著水平 $\alpha = 0.05$ 对假设 $H_0 : \mu \geq 300 \leftrightarrow H_0 < 300$ 进行检验。

解. 利用例 9.21 的结果, 发现当前观察数据无法拒绝零假设 H_0 。

例 9.23. 设样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, 其中 σ^2 未知。将参数 μ 分为已知和未知两种情况, 利用性质 4.26 和性质 7.9, 不难得到在显著水平 α 拒绝零假设 H_0 的条件如下。

$H_0 \leftrightarrow H_1$	μ 已知	μ 未知
$\sigma^2 = \sigma_0^2 \leftrightarrow \sigma^2 \neq \sigma_0^2$	$\begin{cases} \frac{\sum_{j=1}^n (x_j - \mu)^2}{\sigma_0^2} < \chi_{n,\alpha/2}^2 \\ \text{或} \\ \frac{\sum_{j=1}^n (x_j - \mu)^2}{\sigma_0^2} > \chi_{n,1-\alpha/2}^2 \end{cases}$	$\begin{cases} \frac{(n-1)s^2}{\sigma_0^2} < \chi_{n-1,\alpha/2}^2 \\ \text{或} \\ \frac{(n-1)s^2}{\sigma_0^2} > \chi_{n-1,1-\alpha/2}^2 \end{cases}$
$\sigma^2 \leq \sigma_0^2 \leftrightarrow \sigma^2 > \sigma_0^2$	$\frac{\sum_{j=1}^n (x_j - \mu)^2}{\sigma_0^2} > \chi_{n,1-\alpha}^2$	$\frac{(n-1)s^2}{\sigma_0^2} > \chi_{n-1,1-\alpha}^2$
$\sigma^2 \geq \sigma_0^2 \leftrightarrow \sigma^2 < \sigma_0^2$	$\frac{\sum_{j=1}^n (x_j - \mu)^2}{\sigma_0^2} < \chi_{n,\alpha}^2$	$\frac{(n-1)s^2}{\sigma_0^2} < \chi_{n-1,\alpha}^2$

练习 9.5. 接着例 9.22, 在显著水平 $\alpha = 0.05$ 对总体方差的假设 $H_0 : \sigma^2 \leq 100^2 \leftrightarrow H_1 : \sigma^2 > 100^2$ 进行检验。答案: 利用例 9.23 的结果, 发现观察数据拒绝零假设 H_0 。

例 9.24. 设样本 $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} N(\mu_X, \sigma_X^2)$ 和样本 $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2)$ 来自两个独立的总体。样本均值 \bar{X}, \bar{Y} 相互独立, 如果 $\mu_X = \mu_Y$, 则

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} \sim N(0, 1)$$

设两样本方差分别为 S_X^2 和 S_Y^2 , 两个样本的合并样本方差 (pooled sample variance) 定义为

$$S^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

将总体方差分为已知和未知 (但知道 $\sigma_X^2 = \sigma_Y^2$) 两种情况, 在显著水平 α 拒绝零假设 H_0 的条件如下 (右列结果根据的是性质 7.11)。

$H_0 \leftrightarrow H_1$	σ_X^2, σ_Y^2 已知	$\sigma_X^2 = \sigma_Y^2 = \sigma^2$ 未知
$\mu_X = \mu_Y \leftrightarrow \mu_X \neq \mu_Y$	$\frac{ \bar{x} - \bar{y} }{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} > z_{1-\alpha/2}$	$\frac{ \bar{x} - \bar{y} }{s\sqrt{1/m + 1/n}} > t_{m+n-2, 1-\alpha/2}$
$\mu_X \leq \mu_Y \leftrightarrow \mu_X > \mu_Y$	$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} > z_{1-\alpha}$	$\frac{\bar{x} - \bar{y}}{s\sqrt{1/m + 1/n}} > t_{m+n-2, 1-\alpha}$
$\mu_X \geq \mu_Y \leftrightarrow \mu_X < \mu_Y$	$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} < z_\alpha$	$\frac{\bar{x} - \bar{y}}{s\sqrt{1/m + 1/n}} < t_{m+n-2, \alpha}$

例 9.25. 设样本 $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top \stackrel{\text{iid}}{\sim} N(\boldsymbol{\mu}, \Sigma)$, 其中期望 $\boldsymbol{\mu} = (\mu_X, \mu_Y)^\top$ 且协方差矩阵 $\Sigma = [\sigma_X^2, \rho\sigma_X\sigma_Y; \rho\sigma_X\sigma_Y, \sigma_Y^2]$ 未知。

令 $D_j = X_j - Y_j, j = 1, 2, \dots, n$, 我们得到新的样本 D_1, D_2, \dots, D_n 。显然, $D_1, D_2, \dots, D_n \stackrel{\text{iid}}{\sim} N(\mu_X - \mu_Y, \sigma^2)$, 其中 $\sigma^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$ 。

样本均值为 $\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j$ 并且样本方差为 $S^2 = \frac{1}{n-1} \sum_{j=1}^n (D_j - \bar{D})^2$ 。利用例 9.21 的结果, 在显著水平 α 拒绝零假设 H_0 的条件如下。

$H_0 \leftrightarrow H_1$	拒绝零假设 H_0 的条件
$\mu_X - \mu_Y = d_0 \leftrightarrow \mu_X - \mu_Y \neq d_0$	$\frac{ \bar{d} - d_0 }{s/\sqrt{n}} > t_{n-1, 1-\alpha/2}$
$\mu_X - \mu_Y \leq d_0 \leftrightarrow \mu_X - \mu_Y > d_0$	$\frac{\bar{d} - d_0}{s/\sqrt{n}} > t_{n-1, 1-\alpha}$
$\mu_X - \mu_Y \geq d_0 \leftrightarrow \mu_X - \mu_Y < d_0$	$\frac{\bar{d} - d_0}{s/\sqrt{n}} < t_{n-1, \alpha}$

例 9.26. 考察 9 个人在新的饮食计划实施前后的体重 (千克), 以判定该饮食计划是否有助于减轻体重。假定前后的体重满足 $(X, Y)^\top \stackrel{\text{iid}}{\sim} N(\boldsymbol{\mu}, \Sigma)$, 在显著水平 $\alpha = 0.01$ 对假设 $H_0 : \mu_X - \mu_Y \leq 0 \leftrightarrow H_1 : \mu_X - \mu_Y > 0$ 进行检验。

	1	2	3	4	5	6	7	8	9
实施计划之前 X :	132	139	126	114	122	132	142	119	126
实施计划之后 Y :	124	141	118	116	114	132	145	123	121

解. 利用例 9.25 的结果对 $H_0 : \mu_X - \mu_Y \leq 0 \leftrightarrow H_1 : \mu_X - \mu_Y > 0$ 进行检验, 发现数据无法拒绝 H_0 , 即该饮食计划无助于减轻体重。

例 9.27. 已知来自两个独立总体的样本 $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} N(\mu_X, \sigma_X^2)$ 与 $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2)$, 样本方差分别为 S_X^2 和 S_Y^2 , 定义

$$F = \frac{\sum_{j=1}^m (X_j - \mu_X)^2 / m}{\sum_{j=1}^n (Y_j - \mu_Y)^2 / n}$$

在得到样本值后, 令 s_X^2, s_Y^2, f 分别是 S_X^2, S_Y^2, F 的取值。对于下面的假设检验, 在显著水平 α 拒绝零假设 H_0 的条件如下。

$H_0 \leftrightarrow H_1$	μ_X, μ_Y 已知	μ_X, μ_Y 未知
$\sigma_X^2 = \sigma_Y^2 \leftrightarrow \sigma_X^2 \neq \sigma_Y^2$	$\left\{ \begin{array}{l} f > F_{m,n,1-\alpha/2} \\ \text{或 } 1/f > F_{n,m,1-\alpha/2} \end{array} \right.$	$\left\{ \begin{array}{l} s_X^2/s_Y^2 > F_{m-1,n-1,1-\alpha/2} \\ \text{如果 } s_X^2 > s_Y^2 \text{ 或} \\ s_Y^2/s_X^2 > F_{n-1,m-1,1-\alpha/2} \\ \text{如果 } s_X^2 < s_Y^2 \end{array} \right.$
$\sigma_X^2 \leq \sigma_Y^2 \leftrightarrow \sigma_X^2 > \sigma_Y^2$	$f > F_{m,n,1-\alpha}$	$s_X^2/s_Y^2 > F_{m-1,n-1,1-\alpha}$
$\sigma_X^2 \geq \sigma_Y^2 \leftrightarrow \sigma_X^2 < \sigma_Y^2$	$1/f > F_{n,m,1-\alpha}$	$s_Y^2/s_X^2 > F_{n-1,m-1,1-\alpha}$

例 9.28. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(\beta)$, 其中参数 β 未知, 利用例 8.34, 在显著水平 α 对未知参数 β 进行假设检验。

$H_0 \leftrightarrow H_1$	拒绝零假设 H_0 的条件
$\beta = \beta_0 \leftrightarrow \beta \neq \beta_0$	$2\beta_0 n \bar{x} < \chi_{2n,\alpha/2}^2$ 或 $> \chi_{2n,1-\alpha/2}^2$
$\beta \leq \beta_0 \leftrightarrow \beta > \beta_0$	$2\beta_0 n \bar{x} > \chi_{2n,1-\alpha}^2$
$\beta \geq \beta_0 \leftrightarrow \beta < \beta_0$	$2\beta_0 n \bar{x} < \chi_{2n,\alpha}^2$

9.2 大样本检验

如果样本量允许趋向无穷，人们就可以凭借检验统计量的渐近分布（见性质 8.6 和性质 8.7）构造合理的检验，很多情况下也能带来形式上的简化。例如，广义似然比 $\lambda(\mathbf{x})$ 可以很复杂，当样本量足够地大时，在一定的条件之下 $-2 \ln \lambda(\mathbf{X})$ 渐近地服从 χ^2 分布，于是临界值 c 可近似求得。

定理 9.2. 令 m 是参数空间 Θ 与 Θ_0 中独立参数个数之差，则随着样本量趋向无穷，广义似然比具有渐近分布

$$\chi^2 = -2 \ln \lambda(\mathbf{X}) \sim \chi_m^2$$

例 9.29. 练习 9.3 的广义似然比是

$$\lambda(x) = \frac{p_0^x (1-p_0)^{n-x}}{(x/n)^x (1-x/n)^{n-x}}$$

由定理 9.2 知，

$$\chi^2 = -2 \ln \lambda(\mathbf{X}) = 2X \ln \frac{X}{np_0} + 2(n-X) \ln \frac{1-X/n}{1-p_0} \sim \chi_1^2$$

在显著水平 α , $H_0 : p = p_0$ 的拒绝域是 $R = [\chi_{1,\alpha}^2, \infty)$ 。即，当 $\chi^2 \in R$ 时，拒绝零假设 $H_0 : p = p_0$ 。

例 9.30 (共现的判定). 在语料库语言学 (corpus linguistics) 里，两个单词 w 和 w' 之间存在共现 (co-occurrence) 关系，指的是它们共同出现在上下文 (譬如，同一句子) 中并不是随机的，而是具有一定的统计显著性。

在包含 n 个句子的随机语料中，记 N_w 是包含单词 w 的句子个数， $N_{w,w'}$ 是同时包含 w 和 w' 的句子个数。由第 369 页的例 5.18， $N_{w,w'}$ 服从二项分布，不妨设 $N_{w,w'} \sim B(n, p)$ ，其中 $p = P(w, w')$ 是未知参数，其点估计为 $N_{w,w'}/n$ 。如果 w, w' 相互独立，则 $p = P(w)P(w')$ ，进而其点估计亦为 $N_w N_{w'}/n^2$ 。

现在考虑假设检验 $H_0 : p = p_0 \leftrightarrow p \neq p_0$ ，其中 $p_0 = n_w n_{w'}/n^2$ ，而 $n_w, n_{w'}$ 分别是 $N_w, N_{w'}$ 的观察结果。由例 9.29 的结果不难得到，

$$\chi^2 = 2N_{w,w'} \ln \frac{N_{w,w'}}{np_0} + 2(n - N_{w,w'}) \ln \frac{1 - N_{w,w'}/n}{1 - p_0} \sim \chi_1^2$$

对于给定的显著水平 α ，当 $\chi^2 > \chi_{1,\alpha}^2$ 时拒绝零假设，进而判定 w, w' 不是独立的，即二者有共现关系。

例 9.31. 接着例 9.16, 零假设 H_0 成立时, 独立参数只有 σ^2 一个, 按照定理 9.2,

$$-2 \ln \lambda(\mathbf{X}) = n \ln \left\{ 1 + \frac{n(\bar{X} - \mu_0)^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right\} \sim \chi_1^2$$

对于给定的置信水平 α , 零假设 H_0 的接受条件是

$$0 \leq -2 \ln \lambda(\mathbf{x}) = n \ln \left\{ 1 + \left[\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sqrt{n-1}s} \right]^2 \right\} \leq \chi_{1,1-\alpha}^2$$

令 $z = \sqrt{n}(\bar{x} - \mu_0)/s$, 上式等价为

$$z^2 \leq (n-1) \left(\exp \left\{ \frac{1}{n} \chi_{1,1-\alpha}^2 \right\} - 1 \right)$$

利用 $e^z = 1 + \sum_{k=1}^{\infty} z^k/k!$, 当 n 很大时, 上式近似为 $z^2 \leq \chi_{1,1-\alpha}^2$, 而例 9.16 的结论是 H_0 的接受条件是 $z^2 \leq t_{n-1,1-\alpha/2}^2$ 。

定理 9.3 (Wald 检验^{*}). 考虑假设检验 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$, 若 θ 的点估计 $\hat{\theta}_n$ 满足渐近正态性, 即

$$W_n = \frac{\hat{\theta}_n - \theta_0}{V(\hat{\theta}_n)} \xrightarrow{D} N(0, 1)$$

在显著水平 α , 若 $|W_n| > z_{\alpha/2}$ 则拒绝 H_0 。

例 9.32. 练习 9.3 在大样本的情况下, 利用二项分布的正态近似, 对 $H_0 : p = p_0 \leftrightarrow H_1 : p \neq p_0$ 的检验将变得简单, 因为 $n \rightarrow \infty$ 时渐近地有

$$\frac{X - np}{\sqrt{np(1-p)}} \sim N(0, 1)$$

当 $|x - np_0|/\sqrt{np_0(1-p_0)} > z_{1-\alpha/2}$ 时, 拒绝零假设 $H_0 : p = p_0$ 。这个判定条件还可以更精细一些, 利用性质 4.10,

$$\begin{aligned} P(X \leq m) &= \sum_{k=0}^m C_n^k p^k (1-p)^{n-k} \approx \Phi \left(\frac{m + \frac{1}{2} - np}{\sqrt{np(1-p)}} \right) \\ P(X \geq m) &= \sum_{k=m}^n C_n^k p^k (1-p)^{n-k} \approx 1 - \Phi \left(\frac{m - \frac{1}{2} - np}{\sqrt{np(1-p)}} \right) \end{aligned}$$

^{*}罗马尼亚裔美国统计学家 Abraham Wald (1902-1950) 是决策理论、统计序列分析的奠基人。

$H_0 \leftrightarrow H_1$	在显著水平 α 拒绝零假设 H_0 的条件
$p = p_0 \leftrightarrow p \neq p_0$	$\frac{x + \frac{1}{2} - np_0}{\sqrt{np_0(1-p_0)}} < z_{\alpha/2}$ 或 $\frac{x - \frac{1}{2} - np_0}{\sqrt{np_0(1-p_0)}} > z_{1-\alpha/2}$
$p \leq p_0 \leftrightarrow p > p_0$	$\frac{x - \frac{1}{2} - np_0}{\sqrt{np_0(1-p_0)}} > z_{1-\alpha}$
$p \geq p_0 \leftrightarrow p < p_0$	$\frac{x + \frac{1}{2} - np_0}{\sqrt{np_0(1-p_0)}} < z_\alpha$

例 9.33 (两比率的大样本检验). 接着第 532 页的例 8.39, 在显著水平 α 给出假设 $H_0 : p = q \leftrightarrow H_1 : p \neq q$ 的检验。

解. 根据例 8.39 的结果, 拒绝零假设 $H_0 : p = q$ 的条件是

$$\frac{|\bar{x} - \bar{y}|}{\sqrt{\bar{x}(1-\bar{x})/m + \bar{y}(1-\bar{y})/n}} > z_{1-\alpha/2}$$

例 9.34. 甲乙两厂生产同一种产品, 随机地从甲厂抽取 400 件发现 20 件次品, 从乙厂抽取 300 件发现 22 件次品, 在显著水平 $\alpha = 0.05$ 下, 甲乙两厂的次品率是否有显著差异?

解. 分别以 p_1, p_2 表示甲乙两厂的次品率, 在水平 $\alpha = 0.05$ 对假设 $H_0 : p_1 = p_2 \leftrightarrow p_1 \neq p_2$ 进行检验, 数据无法拒绝 H_0 , 即甲乙两厂的次品率无显著差异。

练习 9.6. 仿照例 9.33, 在置信水平 α 对例 8.39 中有关未知参数 $p - q$ 的假设进行检验, 拒绝零假设 H_0 的条件如下。

$H_0 \leftrightarrow H_1$	拒绝零假设 H_0 的条件
$p - q = d_0 \leftrightarrow p - q \neq d_0$	$\frac{ \bar{x} - \bar{y} - d_0 }{\sqrt{\bar{x}(1-\bar{x})/m + \bar{y}(1-\bar{y})/n}} > z_{1-\alpha/2}$
$p - q \leq d_0 \leftrightarrow p - q > d_0$	$\frac{\bar{x} - \bar{y} - d_0}{\sqrt{\bar{x}(1-\bar{x})/m + \bar{y}(1-\bar{y})/n}} > z_{1-\alpha}$
$p - q \geq d_0 \leftrightarrow p - q < d_0$	$\frac{\bar{x} - \bar{y} - d_0}{\sqrt{\bar{x}(1-\bar{x})/m + \bar{y}(1-\bar{y})/n}} < z_\alpha$

本节内容

为检验总体服从是某一给定的分布, 或属于某一分布族, 第一小节介绍了几个拟合优度检验: Pearson χ^2 检验、Kolmogorov 检验以及判定两总体具有相同的分布函数的 Smirnov 检验。用于检验独立性的列联表检验是 Pearson χ^2 检验的一个应用, 它是第二节所讨论的内容。

关键知识

(1) 掌握拟合优度的 Pearson χ^2 检验和 Kolmogorov 检验; (2) 了解 Smirnov 检验、列联表检验。

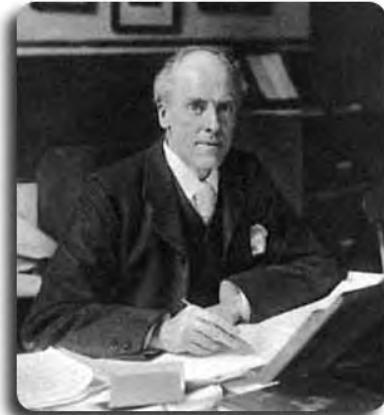
9.2.1 拟合优度检验

已知简单随机样本 X_1, X_2, \dots, X_n 来自总体 $X \sim F(x)$, 对假设 $H_0 : F = F_0 \leftrightarrow H_1 : F \neq F_0$ 进行的检验, 或者更一般地, 对假设 $H_0 : F \in \mathcal{F} \leftrightarrow H_1 : F \notin \mathcal{F}$ 进行的检验, 称为拟合优度检验 (goodness-of-fit test), 其中 F_0 是某一具体的分布 (不含未知参数), $\mathcal{F} = \{F_{\theta}(x) : \theta \in \Theta\}$ 是一个分布族。为方便起见, 备择假设常省略不说。例如, 零假设认为某骰子均匀, 即 $H_0 : F = \frac{1}{6}\langle 1 \rangle + \dots + \frac{1}{6}\langle 6 \rangle$ 。

大样本检验的第一个重要结果是统计学之父 K. Pearson (也是统计学家 E. Pearson 的父亲) 于 1900 年给出的下述引理, 它是统计学最重要的成果之一, 也是统计学从以描述为主的第一阶段进入以严格数学方法为基础的第二阶段的标志。

引理 9.2 (K. Pearson, 1900). 已知随机向量 $\mathbf{Y} = (Y_1, \dots, Y_k)^T$ 服从多项分布 $\text{Multin}(n; p_1, p_2, \dots, p_k)$, 其中 $\sum_{j=1}^k Y_j = n$, Y_j 被称为经验频次, np_j 被称为理论频次。定义 Pearson χ^2 统计量为

$$\chi^2(\mathbf{Y}) = \sum_{j=1}^k \frac{(Y_j - np_j)^2}{np_j} = \frac{1}{n} \sum_{j=1}^k \frac{Y_j^2}{p_j} - n \quad (9.5)$$



$\chi^2(\mathbf{Y})$ 刻画了经验频次 $\mathbf{Y} = (Y_1, \dots, Y_k)^T$ 与理论频次 $(np_1, \dots, np_k)^T$ 之间的差异。当 $n \rightarrow \infty$ 时, 渐近地有

$$\chi^2(\mathbf{Y}) = \sum_{j=1}^k \frac{(Y_j - np_j)^2}{np_j} \sim \chi^2_{k-1}$$

证明. 见陈希孺的《高等概率统计学》[168] 第六章第二节。 □

按照分点 $a_0 < \dots < a_k$ 把实数轴 \mathbb{R} 划分成 k 个两两不交的区间: $A_1 = (a_0, a_1], A_2 = (a_1, a_2], \dots, A_k = (a_{k-1}, a_k)$, 其中 $a_0 = -\infty, a_k = \infty$ 。

定理 9.4 (Pearson χ^2 检验). 若零假设 $H_0 : F = F_0$ 成立, 即总体的分布为 F_0 , 令 $p_j = P(X \in A_j) = F_0(a_j) - F_0(a_{j-1}) > 0, j = 1, 2, \dots, k$, 显然 $\sum_{j=1}^k p_j = 1$ 。定义随机变量 Y_j 为 X_1, \dots, X_n 落于区间 A_j 内的个数, 则

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)^T \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$$

由引理 9.2, 在显著水平 α , 当 Pearson χ^2 统计量的观察结果 $\chi^2(\mathbf{y}) > \chi^2_{k-1, 1-\alpha}$ 时拒绝零假设 $H_0 : F = F_0$; 当 $\chi^2(\mathbf{y}) \leq \chi^2_{k-1, 1-\alpha}$ 时接受零假设。

例 9.35. 某机器在周一至周五共发生了 $n = 30$ 次故障，每天的故障数依次为 4, 7, 10, 3, 6 次，在水平 $\alpha = 0.05$ 检验假设“故障率与周几有关”。

解. 设每天的故障数 $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)^\top \sim \text{Multin}(n; p_1, p_2, p_3, p_4, p_5)$ ，当零假设“ H_0 ：故障率与周几无关”成立时， $p_j = 1/5, j = 1, 2, \dots, 5$ ，即每天理论故障数为 $np_j = 6$ 。计算得

$$\chi^2(\mathbf{y}) = \sum_{j=1}^5 \frac{(y_j - np_j)^2}{np_j} > \chi^2_{4,1-\alpha}, \text{ 其中 } \mathbf{y} = (4, 7, 10, 3, 6)^\top$$

经过 Pearson χ^2 检验，数据在水平 $\alpha = 0.05$ 无法否认零假设。

若定理 9.4 中的总体改为参数总体 $F_\theta(x)$ ，K. Pearson 认为渐近性质 $\chi^2(\mathbf{Y}) \sim \chi^2_{k-1}$ 依然成立——这是 K. Pearson 的一个重大失误。1922 年，Fisher 发现了这个错误，并于 1924 年撰文《 χ^2 作为度量观测值与假设间的偏差的条件》给出了正确的结果，即下面的定理。

定理 9.5 (Fisher, 1924). 已知样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F_\theta(x)$ ，假设未知参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^\top$ 的最大似然估计存在，设为 $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_r)^\top$ 。设 $\hat{p}_j = P_{\hat{\boldsymbol{\theta}}}\{X \in A_j\} > 0, j = 1, \dots, k$ ，定义随机变量 Y_j 为样本 X_1, \dots, X_n 落于 A_j 内的个数（被称为经验频次）， $n\hat{p}_j$ 被称为理论频次。当 $n \rightarrow \infty$ 时，渐近地有

$$\chi^2(\mathbf{Y}) = \sum_{j=1}^k \frac{(Y_j - n\hat{p}_j)^2}{n\hat{p}_j} \sim \chi^2_{k-1-r} \quad (9.6)$$

对零假设 $H_0 : F \in \mathcal{F} = \{F_\theta(x)\}$ 的检验可转换为对 $H_0 : F = F_{\hat{\boldsymbol{\theta}}}(x)$ 的 Pearson χ^2 检验：当 $\chi^2(\mathbf{y}) > \chi^2_{k-1-r, 1-\alpha}$ 时，在水平 α 拒绝 H_0 。

起初 K. Pearson 并不承认自己的疏忽，在参数情况下自由度是否应该减小这个问题上曾与 Fisher 有过激烈的、不愉快的争论。但权威战胜不了真理，在很多事实面前 K. Pearson 终于不得不接受 Fisher 的意见。非不废是，瑕不掩瑜，历史为纪念 K. Pearson 原创发现引理 9.2 的学术功绩，习惯上把基于定理 9.5 的检验也称作拟合优度的 Pearson χ^2 检验。学术江湖的这点恩怨，都被真理化解了，可谓“千江有水千江月，万里无云万里天”。

例 9.36 ([137]). 三天之内全国发生了 306 起交通事故，按小时将三天等分为 $n = 72$ 个单位时间段，每小时事故数的观察结果见下表左边两列，例如单位时间段内发生 0 或 1 次事故的有 4 次，单位时间段内发生 2 次事故的有 10 次，……。问在水平 $\alpha = 0.05$ 之下，每小时的事故数 X 是否服从 Poisson 分布？

每小时事故数	观测值 y_j	拟合值 $n\hat{p}_j$
0 或 1	5	5.391880
2	10	9.275318
3	15	13.140034
4	12	13.961286
5	12	11.867093
6	6	8.405858
7	5	5.103556
8 或更多	7	4.854974

解. 在零假设 $H_0 : X \sim \text{Poisson}(\lambda)$ 成立的情况下, 参数 λ 的最大似然估计是 $\hat{\lambda} = \bar{X} = 306/72 = 4.25$ 。根据递归关系

$$\frac{P_{\hat{\lambda}}(X = j+1)}{P_{\hat{\lambda}}(X = j)} = \frac{\hat{\lambda}}{j+1}$$

以及初始值 $\hat{p}_0 = P_{\hat{\lambda}}(X = 0) = \exp\{-\hat{\lambda}\} = 0.0143$, 可以得到 $\hat{p}_j = P_{\hat{\lambda}}(X = j)$, 进而求得 $n\hat{p}_j$ (表中最右列), 其中 $j = 0, 1, 2, \dots$ 。

根据 $k - 1 - r = 8 - 1 - 1 = 6$ 以及式 (9.6), 在水平 $\alpha = 0.05$ 之下,

$$\sum_{j=1}^8 \frac{(y_j - n\hat{p}_j)^2}{n\hat{p}_j} = 2.263792 < \chi^2_{6,0.95} = 12.59159$$

于是, 在水平 $\alpha = 0.05$ 数据不能拒绝零假设 H_0 , 即每小时的事故数 X 服从 Poisson 分布。

练习 9.7. 仿照上例, 考虑第 273 页的例 4.18, 验证在水平 $\alpha = 0.01$, Rutherford-Geiger 实验数据无法拒绝零假设 $H_0 : X$ 服从 Poisson 分布。

Pearson χ^2 检验必须将样本分组, 多了一些任意性。当一维总体分布函数 $F(x)$ 连续时, 功效更大的检验是基于第 467 页的定理 7.2 的 Kolmogorov 检验, 下面介绍它。

令 $F_n^*(x)$ 是由简单随机样本 X_1, X_2, \dots, X_n 构造的经验分布函数, 根据 Glivenko 定理 7.1, 当 n 很大时统计量 $D_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)|$ 接近 0。如何计算 D_n 呢?

算法 9.2. 从图 7.6 可见 $|F_n^*(x) - F(x)|$ 的最大值点只可能出现在 $F_n^*(x)$ 的跳跃点 $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ 当中, 所以

$$D_n = \max\{D_n^+, D_n^-\}$$

其中, D_n^+, D_n^- 称为单侧 Kolmogorov 统计量, 定义如下

$$D_n^+ = \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F(X_{(j)}) \right\}$$

$$D_n^- = \max_{1 \leq j \leq n} \left\{ F(X_{(j)}) - \frac{j-1}{n} \right\}$$

如果零假设 $H_0 : F(x) = F_0(x)$ 成立, 一个合理的检验是当 $D_n(x) = \sup |F_n^*(x) - F_0(x)| > c$ 时拒绝零假设, 利用定理 7.2 可得 Kolmogorov 检验 (也称为单样本的 Kolmogorov-Smirnov 检验, 简称 K-S 检验):

- 在水平 $\alpha = 0.05$ 之下, 取 $c = K_{1-\alpha}/\sqrt{n} = K_{0.95}/\sqrt{n} = 1.358/\sqrt{n}$, 其中 $K_{1-\alpha}$ 是式 (7.14) 所定义的 Kolmogorov 分布 $K(z)$ 的 $(1 - \alpha)$ -分位数。
- 在水平 $\alpha = 0.01$ 之下, 取 $c = 1.628/\sqrt{n}$ 。

或者, 在总体分布的零假设 $H_0 : F(x) = F_0(x)$ 之下, 计算统计量 D_n 的观测值和检验的 p -值 (见定义 9.10)。如果 p -值小于预先给定的水平 α , 则拒绝零假设。通过 Kolmogorov 检验, 可以判定伪随机数是否来自某指定的分布。

例 9.37. 掷某骰子 60 次, 1 至 6 点的次数依次为 16, 7, 8, 8, 9, 12, 问该骰子是否均匀?

解. 令零假设 $H_0 : P\{X = x\} = 1/6$, 其中 $x = 1, 2, \dots, 6$, 即骰子均匀。

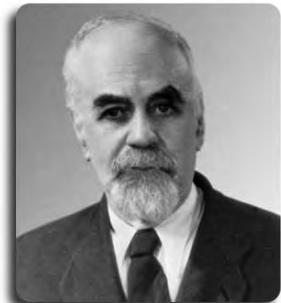
x	< 1	1	2	3	4	5	6	> 6
频次	0	16	7	8	8	9	12	0
$F_{60}^*(x)$	0	16/60	23/60	31/60	39/60	48/60	1	1
$F_0(x)$	0	1/6	2/6	3/6	4/6	5/6	1	1
$ F_{60}^*(x) - F_0(x) $	0	6/60	3/60	1/60	1/60	2/60	0	0

因为 $D_{60} = \sup |F_{60}^*(x) - F_0(x)| = 6/60 = 0.1 < 1.358/\sqrt{60} \approx 0.175$, 所以对于 K-S 检验, 在水平 $\alpha = 0.05$ 观察数据无法拒绝 H_0 , 即该骰子均匀。

然而 Pearson χ^2 检验在水平 $\alpha = 0.05$ 拒绝 H_0 , 因为 $\chi^2 = 284.1 > \chi_{5,0.95}^2 \approx 11.07$ 。两种不同的假设检验方法可以在相同的水平得出不同的结论!

1944 年, 苏联数学家、名著《高等数学教程》的作者 Vladimir Ivanovich Smirnov (1887-1974) 在定理 7.2 的基础上证明了统计量 D_n^+ 具有下面的极限性质 (D_n^- 也有相同的结果)。

$$\lim_{n \rightarrow \infty} P\{\sqrt{n}D_n^+ \leq z\} = \begin{cases} 1 - e^{-2z^2} & \text{若 } z > 0 \\ 0 & \text{若 } z \leq 0 \end{cases} \quad (9.7)$$



此外, Smirnov 还证明了下面的定理, 基于此定理可以检验两样本的总体 F_1 和 F_2 之间的关系, 例如 $F_1 = F_2$ 。

定理 9.6 (Smirnov, 1944). 设简单随机样本 X_{j1}, \dots, X_{jn_j} 来自具有一维连续分布函数 $F_j(x)$ 的总体, 其中 $j = 1, 2$, 记它们的经验分布函数为 $F_{1n_1}^*(x)$ 和 $F_{2n_2}^*(x)$ 。若 $F_1(x) = F_2(x)$, Smirnov 统计量 $D_{n_1, n_2} = \sup\{|F_{1n_1}^*(x) - F_{2n_2}^*(x)|\}$ 和单侧 Smirnov 统计量 $D_{n_1, n_2}^+ = \sup\{F_{1n_1}^*(x) - F_{2n_2}^*(x)\}$ 具有如下极限性质。

$$\lim_{\substack{n_1 \rightarrow \infty \\ n_2 \rightarrow \infty}} P \left\{ \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1, n_2} \leq z \right\} = K(z)$$

$$\lim_{\substack{n_1 \rightarrow \infty \\ n_2 \rightarrow \infty}} P \left\{ \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1, n_2}^+ \leq z \right\} = \begin{cases} 1 - \exp(-2z^2) & \text{若 } z > 0 \\ 0 & \text{若 } z \leq 0 \end{cases}$$

基于定理 9.6 可给出两个总体是否具有相同的连续分布函数的 Smirnov 检验(或两样本的 Kolmogorov-Smirnov 检验, 简称 K-S 检验): 当 $D_{n_1, n_2}(x_1, x_2) > K_{1-\alpha}/\sqrt{n}$ 时拒绝零假设 $H_0: F_1(x) = F_2(x)$, 其中 $n = n_1 n_2 / (n_1 + n_2)$ 。

例 9.38. 观察 A, B 两个牌子的电池的使用寿命(小时)如下, 牌子 A : 116, 76, 111, 99, 97, 103, 100, 116, 125, 72, 牌子 B : 121, 97, 108, 92, 93, 90, 78, 96, 97, 93。试问: 它们在使用寿命这一性能上是否相同?

解. 令零假设 H_0 : 两个牌子的电池的使用寿命服从相同的分布。先分别计算经验分布函数 $F_{10}^*(x)$ 和 $G_{10}^*(x)$, 见下图。

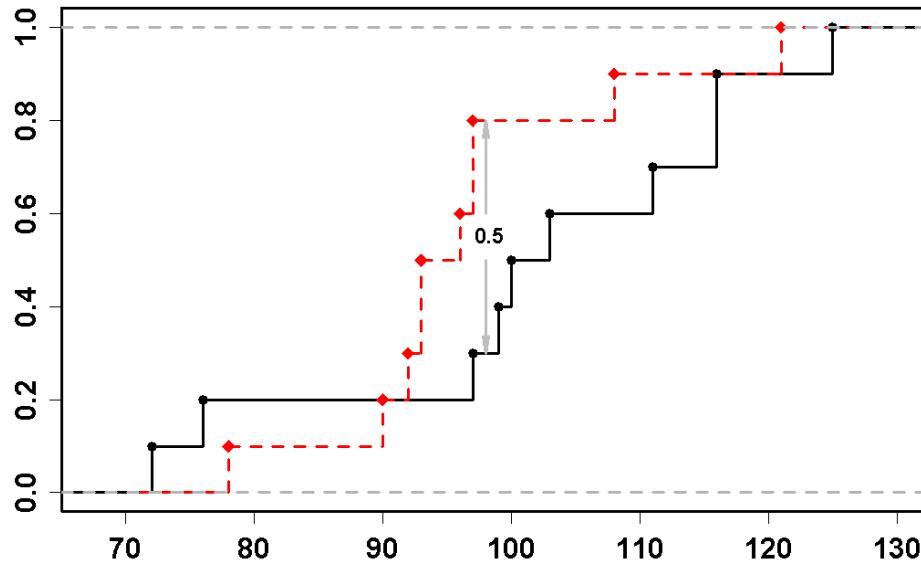


图 9.8: 例 9.38 中的两组数据对应的经验分布函数 $F_{10}^*(x)$ 和 $G_{10}^*(x)$, 经过计算得到 $D_{10,10} = \sup\{|F_{10}^*(x) - G_{10}^*(x)|\} = 0.5$ 。

利用 Smirnov 检验, 在水平 $\alpha = 0.05$ 有 $D_{10,10} = 0.5 < 1.358/\sqrt{5} \approx 0.607$, 于是数据无法拒绝零假设 H_0 , 即这两个牌子的电池的使用寿命服从相同的分布。

推论 9.2. 当 $n \rightarrow \infty$ 时, 统计量 D_n^+ 和 D_{n_1,n_2}^+ 具有下面的渐近性质。

$$\begin{aligned} 4n(D_n^+)^2 &\sim \chi_2^2 \\ \frac{4n_1 n_2}{n_1 + n_2} (D_{n_1,n_2}^+)^2 &\sim \chi_2^2 \end{aligned}$$

证明. 由结果 (9.7), $\lim_{n \rightarrow \infty} P\{4n(D_n^+)^2 \leq x\} = 1 - \exp(-x/2)$ 即是 χ_2^2 的分布函数。类似地, 由定理 9.6 可证得第二个结果。 \square

例 9.39. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} F(x)$, 其中 $F(x)$ 是一维连续分布函数。对于零假设 $H_0 : F(x) \geq F_0(x)$ 的单侧检验是当 $D_n^+(x) > c$ 时拒绝零假设。根据推论 9.2, 在水平 α 之下, 取临界值 c 如下。

$$c = \frac{1}{2} \sqrt{\frac{\chi_{2,1-\alpha}^2}{n}} = \begin{cases} 1.223873/\sqrt{n} & \text{当 } \alpha = 0.05 \\ 1.517427/\sqrt{n} & \text{当 } \alpha = 0.01 \end{cases} \quad (9.8)$$

例 9.40. 设两样本 $X_{11}, X_{12}, \dots, X_{1n_1} \stackrel{\text{iid}}{\sim} F_1(x)$ 和 $X_{21}, X_{22}, \dots, X_{2n_2} \stackrel{\text{iid}}{\sim} F_2(x)$, 其中 $F_1(x), F_2(x)$ 都是一维连续分布函数。对于零假设 $H_0 : F_1(x) \geq F_2(x)$ 的单侧检验是当 $D_{n_1,n_2}^+(x_1, x_2) > c$ 时拒绝零假设。根据推论 9.2, 在水平 α 之下, 按照式 (9.8) 取临界值 c , 其中 $n = n_1 n_2 / (n_1 + n_2)$ 。

9.2.2 独立性的列联表检验

当人们对某事物的两个不同属性 A, B (譬如 $A = \text{受教育程度}$, $B = \text{收入}$) 是否相互关联感兴趣时, 常把属性 A 分为 r 个等级 A_1, A_2, \dots, A_r , 把属性 B 分为 s 个等级 B_1, B_2, \dots, B_s , 这样共产生 rs 个组合子类。

定义 9.14 (列联表). 从总体中随机抽取 n 个样本点, 发现其中分到 (A_i, B_j) 子类的有 N_{ij} 个, 其中 $i = 1, \dots, r, j = 1, \dots, s$ 。如下构造的 $r \times s$ 数据表被称为二维列联表 (contingency table), 其中, $N_i = \sum_{j=1}^s N_{ij}$ 且 $N_j = \sum_{i=1}^r N_{ij}$ 。列联表分析是离散多元分析的研究内容之一, 利用它对 A, B 的独立性进行的假设检验称为列联表检验。

表 9.3: 列联表: N_{ij} 表示观察到 (A_i, B_j) 的个数。

A^B	B_1	\cdots	B_j	\cdots	B_s	和
A_1	N_{11}	\cdots	N_{1j}	\cdots	N_{1s}	$N_{1\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_i	N_{i1}	\cdots	N_{ij}	\cdots	N_{is}	$N_{i\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_r	N_{r1}	\cdots	N_{rj}	\cdots	N_{rs}	$N_{r\cdot}$
和	$N_{\cdot 1}$	\cdots	$N_{\cdot j}$	\cdots	$N_{\cdot s}$	n

定义随机向量 $(X, Y)^\top$ 满足 $P(X = i, Y = j) = p_{ij}$, 其中 p_{ij} 表示 $(X, Y)^\top$ 属于 (A_i, B_j) 子类的概率, $i = 1, 2, \dots, r$ 且 $j = 1, 2, \dots, s$ 。

对零假设 “ $H_0 : X, Y$ 相互独立” 的检验即验证存在非负常数 p_1, \dots, p_r 和 $p_{\cdot 1}, \dots, p_{\cdot s}$ 满足 $\sum_{i=1}^r p_i = \sum_{j=1}^s p_{\cdot j} = 1$ 并使得

$$P(X = i, Y = j) = p_i p_{\cdot j}$$

若 H_0 成立, 为确定 $(X, Y)^\top$ 的分布, 必须把未知参数 p_1, \dots, p_r 和 $p_{\cdot 1}, \dots, p_{\cdot s}$ 确定下来, 这其中只有 $r+s-2$ 个自由参数 (因为有两个约束条件)。这些参数的最大似然估计为 $\hat{p}_i = N_{i\cdot}/n, \hat{p}_{\cdot j} = N_{\cdot j}/n$, 进而得到经验频次 N_{ij} 和理论频次 $n\hat{p}_{ij} = N_{i\cdot}N_{\cdot j}/n$ 。

推论 9.3 (列联表检验). 由表 9.3 构造统计量 χ^2 如下, 利用定理 9.5 可得 $n \rightarrow \infty$ 时, 渐近地有 $\chi^2 \sim \chi_m^2$, 其中, $m = rs - 1 - (r + s - 2) = (r - 1)(s - 1)$ 。

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(N_{ij} - N_{i\cdot}N_{\cdot j}/n)^2}{N_{i\cdot}N_{\cdot j}/n} = \sum_{i=1}^r \sum_{j=1}^s \frac{(nN_{ij} - N_{i\cdot}N_{\cdot j})^2}{nN_{i\cdot}N_{\cdot j}} \sim \chi_m^2 \quad (9.9)$$

若 $\chi^2 > \chi_{m, 1-\alpha}^2$, 则在水平 α 下拒绝 $H_0 : X, Y$ 相互独立。

例 9.41. 接着考虑例 9.30 的单词共现问题, 为检验假设 “ $H_0 =$ 在句子中单词或短语 w 和 w' 相互独立”, 随机选取 n 个句子, 得到如下 2×2 列联表。

表 9.4: $N_{11} = N_{w,w'}$, $N_{12} = N_{w,\neg w'}$ 表示包含 w 而不含 w' 的句子的个数。类似地, $N_{21} = N_{\neg w,w'}$ 和 $N_{22} = N_{\neg w,\neg w'}$ 。令 $N_{k \cdot} = N_{k1} + N_{k2}$, $N_{\cdot k} = N_{1k} + N_{2k}$, 其中 $k = 1, 2$ 。

频次	出现 w'	不出现 w'	求和
出现 w	N_{11}	N_{12}	$N_{1 \cdot}$
不出现 w	N_{21}	N_{22}	$N_{2 \cdot}$
求和	$N_{\cdot 1}$	$N_{\cdot 2}$	$n = N_{11} + N_{12} + N_{21} + N_{22}$

由推论 9.3, 当 $n \rightarrow \infty$ 时渐近地有

$$\chi^2 = \frac{n(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1 \cdot}N_{2 \cdot}N_{\cdot 1}N_{\cdot 2}} \sim \chi_1^2 \quad (9.10)$$

根据列联表检验, 如果 $\chi^2 > \chi_{1,1-\alpha}^2$, 则观察数据在水平 α 拒绝 H_0 。

练习 9.8. 某药品有注射和口服两种给药方式, 对 $n = 1000$ 个病人进行给药方式和效果的考察, 结果如下。问效果与给药方式是否独立?

	有效	无效	总计
口服	226	278	504
注射	255	241	496
总计	481	519	1000

答案: 代入式 (9.10), 得到 $\chi^2 = 4.322482$ 。

9.3 习题

- 9.1. 接着第 554 页的例 9.10, 已知该 UMP 检验犯第一类、第二类错误的概率分别为 α, γ , 求样本容量 n 。
- 9.2. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, 1)$, 在水平 α 给出假设 $H_0 : \mu = 1 \leftrightarrow H_1 : \mu = 2$ 的 UMP 检验, 并求此检验犯第二类错误的概率。
- ☆ 9.3. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} U[0, \theta]$, 其中参数 $\theta > 0$ 未知。令 $X_{(n)} = \max_{0 \leq j \leq n} \{X_j\}$, 对假设 $H_0 : 3 \leq \theta \leq 4 \leftrightarrow H_1 : \theta < 3$ 或 $\theta > 4$ 取检验函数
- $$\delta(x) = \begin{cases} 0 & \text{若 } 2.9 \leq x_{(n)} \leq 4.2 \\ 1 & \text{若 } x_{(n)} < 2.9 \text{ 或 } x_{(n)} > 4.2 \end{cases}$$
- (1) 求此检验法的功效函数。(2) 问样本容量 n 至少取多大时, 可使犯第一类错误的概率不超过 0.1?
- 9.4. 在显著水平 α , 给出 $H_0 : (X, Y)^\top \sim N(0, 0, 1, 1, 0.6) \leftrightarrow H_1 : (X, Y)^\top \sim N(1, 1, 1, 1, 0.6)$ 的似然比检验。
- ☆ 9.5. 接着第 560 页的例 9.16, 在水平 α 对假设 $H_0 : \sigma^2 \leq \sigma_0^2 \leftrightarrow H_1 : \sigma^2 > \sigma_0^2$ 进行广义似然比检验。
- ☆ 9.6. 已知来自两个独立总体的样本 $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ 和 $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, 其中参数 μ_1, μ_2, σ^2 皆未知。在水平 α 对假设 $H_0 : \mu_1 = \mu_2 \leftrightarrow H_1 : \mu_1 \neq \mu_2$ 进行广义似然比检验。
- 9.7. 考虑第 539 页的例 9.1, 在显著水平 $\alpha = 0.05$ 下, 问这批零件的长度是否合格? ($z_{0.975} = 1.959964, z_{0.95} = 1.644854$)
- 9.8. 按照 Mendel 的遗传规律, 开粉红花的豌豆的子代可分为红花、粉红花和白花三类, 其比例为 1 : 2 : 1。为检验这一规律, 随机地考察了 100 株子代豌豆, 结果发现开红花 30 株, 开粉红花 48 株, 开白花 22 株。问 Mendel 遗传定律是否成立 (水平 $\alpha = 0.05$)?
- 9.9. 连续抛一枚硬币直至出现正面算完成一局, 令随机变量 X 表示每局的抛次。共完成 1000 局, 对应于抛次 k 的频次 n_k 如下, 问此硬币是否均匀 (水平 $\alpha = 0.05$)?

抛次 k	1	2	3	4	5	6	≥ 7
频次 n_k	505	257	122	63	21	15	17

9.10. 当 $k = 2$ 时, 证明引理 9.2。

第十章

回归分析与方差分析

若言琴上有琴声，放在匣中何不鸣？若言声在指头上，何不于君指上听？

苏轼《琴诗》

在经典数学、物理学的理论中，变量之间确定性的关系通常用函数来刻画，如圆的面积 S 与半径 r 有 $S = \pi r^2$ ，力 F 与加速度 a 有 $F = ma$ ，等等。这些确定的关系揭示了自然的本质规律，除此之外，变量之间还有一种非确定性的关系，即所谓的相关关系。笼统地说，相关关系就是在确定的函数关系上附加一个随机扰动。

定义 10.1 (相关关系). 输入变量 a_1, \dots, a_k 是可精确观测或者可精确控制的普通自变量，称为解释变量；因变量 X 是普通函数 $f(a_1, \dots, a_k)$ 和随机误差 $\epsilon \sim N(0, \sigma^2)$ 之和，在不同场合称为目标变量、响应变量。这里，函数 $f(a_1, \dots, a_k)$ 被称为回归函数。式 (10.1) 所描述的响应变量 X 与解释变量 a_1, \dots, a_k 之间的关系不是普通的函数关系，而是其推广（试想一下 $\sigma^2 = 0$ 的情形），被称为相关关系。

$$X = f(a_1, \dots, a_k) + \epsilon, \text{ 其中 } \epsilon \sim N(0, \sigma^2) \quad (10.1)$$

随机变量 $X \sim N(f(a_1, \dots, a_k), \sigma^2)$ 由两部分组成：确定的 $f(a_1, \dots, a_k)$ 和由随机因素引起的不确定的 ϵ 。该随机误差项的期望可以不失一般性地设为 0，这是因为我们总可以把 $\epsilon - E\epsilon \sim N(0, \sigma^2)$ 当作随机误差项，即

$$X = f(a_1, \dots, a_k) + E\epsilon + (\epsilon - E\epsilon)$$

因此，式 (10.1) 有时也简单地表示为

$$EX = f(a_1, \dots, a_k)$$

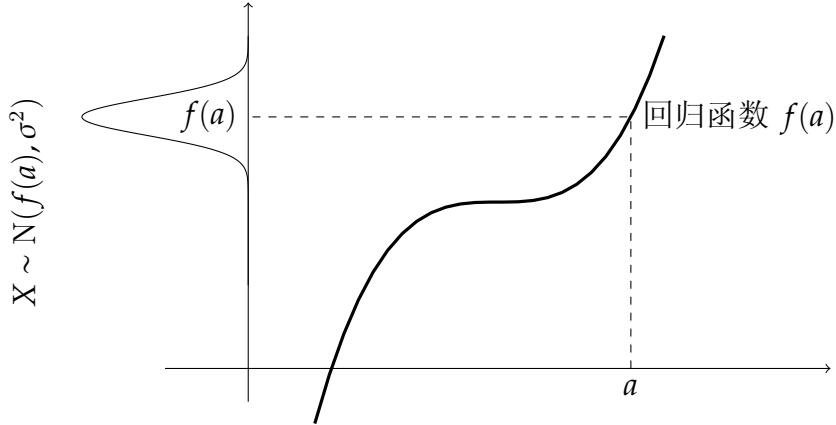


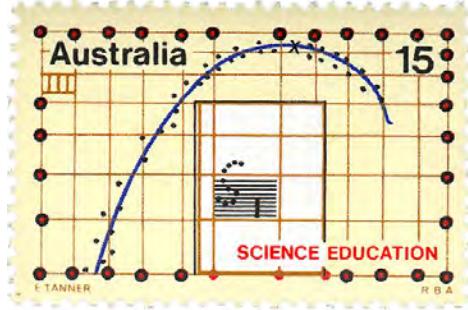
图 10.1: 响应变量 X 和解释变量 a 之间的相关关系: 当自变量 a 确定时, 相应的因变量的取值以 $N(f(a), \sigma^2)$ 的方式分布在 $f(a)$ 的周围。

统计学中的回归分析 (regression analysis) 研究的就是变量之间的这种非确定性的相关关系, 利用它们的数量表达式进行统计推断 [37, 158, 170], 其中包括寻找一个满意的 f 使得随机误差项的方差足够地小, 这样在实践中就能通过观察 a_1, \dots, a_n 来预测 X , 或通过控制 a_1, \dots, a_n 来控制 X 。譬如, 在烧砖生产中, 通过控制土质、烧结温度、烧制时间等来控制砖的硬度。

例 10.1. 人的体重和身高之间大致存在关系: 越高越重 (这里身高是自变量), 而且对于身高为 a 的人群, 体重 X 呈现出正态分布。

定义 10.2. 在自变量 a_1, \dots, a_k 取值为 a_{i1}, \dots, a_{ik} 时, 观测到样本 $X = x_i, i = 1, 2, \dots, n$, 记作 $(x_1|a_{11}, \dots, a_{1k}), \dots, (x_i|a_{i1}, \dots, a_{ik}), \dots, (x_n|a_{n1}, \dots, a_{nk})$, 或者

$$A = \begin{pmatrix} a_1 & \cdots & a_k \\ a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ik} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nk} \end{pmatrix} \mapsto x = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} \text{ 或者 } \begin{array}{|c|ccc|} \hline & a_1 & \cdots & a_k \\ \hline x_1 & a_{11} & \cdots & a_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ x_i & a_{i1} & \cdots & a_{ik} \\ \vdots & \vdots & \vdots & \vdots \\ x_n & a_{n1} & \cdots & a_{nk} \\ \hline \end{array}$$



矩阵 A 称为变量 a_1, \dots, a_k 的观测矩阵或数据矩阵。基于模型 (10.1), 我们得到下面的理论值 (在不同场合下也称作回归值、预测值、拟合值等), 它们是自变量

a_1, \dots, a_k 取某些值时的回归函数值 $f(a_1, \dots, a_k)$ 。

$$f_i = f(a_{i1}, \dots, a_{ik}), \text{ 其中 } i = 1, 2, \dots, n$$

有时也将理论值记作 \hat{x}_i , 把 $x_i - \hat{x}_i$ 称作残差 (residual), 把 $x - \hat{x}$ 称作残差向量, 其中 $x = (x_1, \dots, x_n)^\top$ 是实际观测值, $\hat{x} = (\hat{x}_1, \dots, \hat{x}_n)^\top$ 是理论值。

例 10.2. 软件 R 自带了 cars 数据, 散点图如下, 它记录了小汽车的速度和紧急刹车的滑行距离, 数据采集于二十世纪二十年代。大致来说, 汽车的速度越快, 紧急刹车后滑行得越远。

在实践中, 回归函数的类型经常是未知的。对此例, 我们可以采用 k 阶多项式来描述相关关系, 其中 $k = 1, 2, 3$ 如下图所示。有时, 甚至需要用更复杂的回归函数才能取得好的拟合效果。

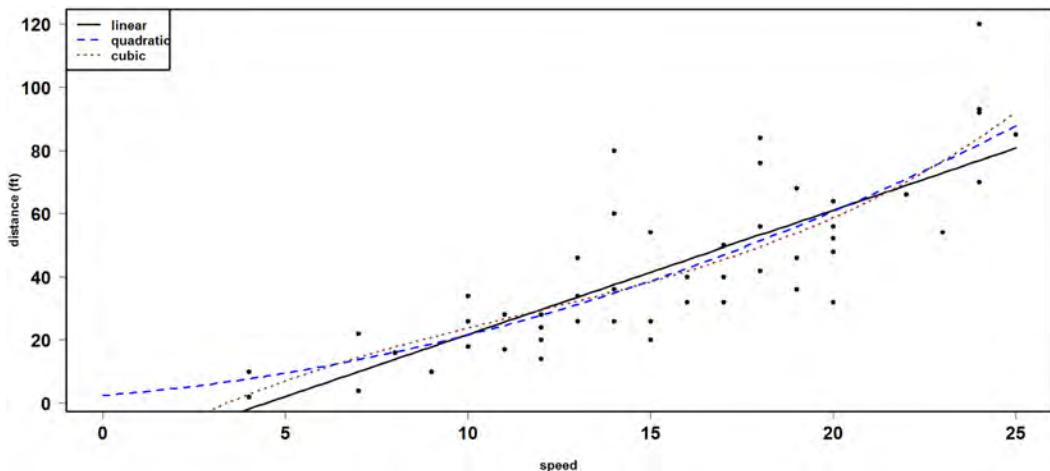


图 10.2: cars 数据的散点图: 横轴表示速度 a , 纵轴表示滑行距离 X , 分别用直线、二次曲线、三次曲线来描述 X 与 a 之间的相关关系。

例 10.3. 误差 (error) 和残差 (residual) 是两个容易混淆的概念, 我们通过下面的例子来区分它们。已知简单样本 X_1, \dots, X_n 来自总体 $N(\mu, \sigma^2)$, 则误差分别为

$$e_i = X_i - \mu, \text{ 其中 } i = 1, \dots, n$$

样本均值 $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ 是对总体均值 μ 的估计, 残差分别为

$$r_i = X_i - \bar{X}, \text{ 其中 } i = 1, \dots, n$$

误差和残差有着不同的统计性质，例如

$$\frac{1}{\sigma^2} \sum_{i=1}^n e_i^2 \sim \chi_n^2$$

$$\frac{1}{\sigma^2} \sum_{i=1}^n r_i^2 \sim \chi_{n-1}^2$$

简而言之，误差是观测值和真实值之间的差异，残差是观测值和理论值（或预测值）之间的差异。

定义 10.3. 最简单的回归函数是线性函数，即

$$X = \beta_0 + \beta_1 a_1 + \cdots + \beta_k a_k + \epsilon, \text{ 其中 } \epsilon \sim N(0, \sigma^2) \quad (10.2)$$

变量之间的关系可以很复杂，(10.2) 具有代表性吗？如果变量之间的关系不是线性的，某些情况下可以通过变换使之成为线性的。譬如，

$$\begin{aligned} \frac{1}{x} &= \beta_0 + \frac{\beta_1}{a} & \xrightarrow[a'=1/a]{x'=1/x} & x' = \beta_0 + \beta_1 a' \\ x &= \beta_1 \exp(\beta_0 a) & \xrightarrow[a'=\exp(\beta_0 a)]{x'=x} & x' = \beta_1 a' \\ x &= \beta_0 + \beta_1 a + \cdots + \beta_k a^k & \xrightarrow[a'_i=a^i, i=1,2,\dots,k]{x'=x} & x' = \beta_0 + \beta_1 a'_1 + \cdots + \beta_k a'_k \end{aligned}$$

在式 (10.2) 中，未知参数 $\beta_0, \beta_1, \dots, \beta_k$ 称为回归系数，我们要利用 X 的观察样本，通过参数估计的方法把 $\beta_0, \beta_1, \dots, \beta_k, \sigma^2$ 都估计出来（详见 §10.1.1）。不妨设这些估计值分别是 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k, \hat{\sigma}^2$ ，则理论值为

$$\hat{x}_i = \hat{\beta}_0 + \hat{\beta}_1 a_{i1} + \cdots + \hat{\beta}_k a_{ik}, \text{ 其中 } i = 1, 2, \dots, n \quad (10.3)$$

英国统计学家 George E. P. Box (1919-2013) 说，“设计简单而令人回味的模型的能力是伟大科学家的签名，过度细致和过度参数化常常是平庸的标志。”大量而广泛的应用已经证明线性模型是简单且有效的。

例 10.4. 美国社会学家 Otis Dudley Duncan (1921-2004) 调查了 1950 年美国 45 种职业的收入、受教育程度和社会威望。为考察社会威望 X 与收入 a_1 、受教育程度 a_2 之间的相关关系，观察 Duncan 数据的散点图。目标变量 X 与解释变量 a_1, a_2 之间的相关关系可以是 $X = \beta_0 + \beta_1 a_1 + \beta_2 a_2 + \epsilon$ ，其中 $\epsilon \sim N(0, \sigma^2)$ ，参数 $\beta_0, \beta_1, \beta_2, \sigma^2$ 未知。

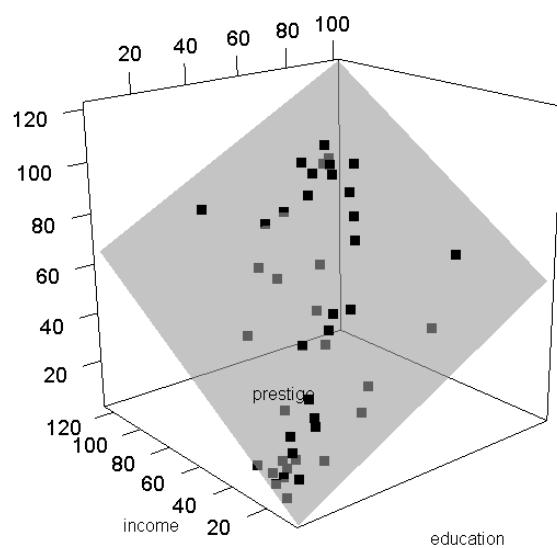
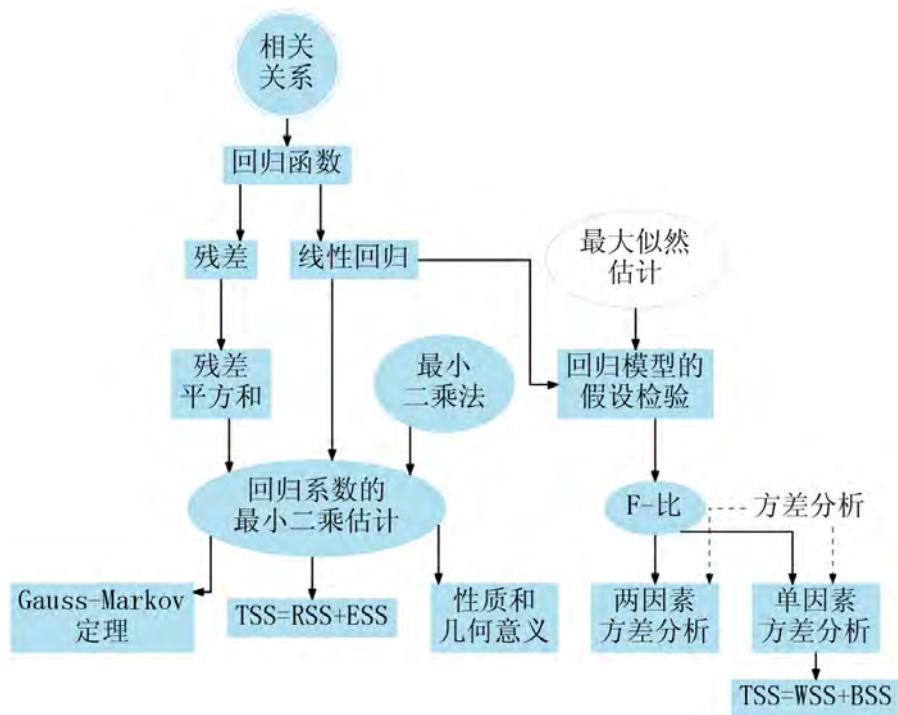


图 10.3: Duncan 数据的散点图: a_1 轴表示收入, a_2 轴表示受教育程度, 纵轴表示社会威望 X。观察数据散布在回归平面周围, 不难看出, 受教育程度和收入都偏低的人群的社会威望也偏低。

第十章的主要内容及其关系



10.1 线性回归模型

设 式 (10.2) 中自变量 a_1, \dots, a_k 的取值分别为 a_{i1}, \dots, a_{ik} 的时候观察到样本点 X_i , $i = 1, \dots, n$ 且 $n > k$, 于是便得到 n 个线性方程 $X_i = \beta_0 + \beta_1 a_{i1} + \dots + \beta_k a_{ik} + \epsilon_i$ 构成的方程组, 其中假定 $\epsilon_1, \dots, \epsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, σ^2 未知。

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} 1 & a_{11} & \cdots & a_{1k} \\ 1 & a_{21} & \cdots & a_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_{n1} & \cdots & a_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \text{其中 } \epsilon_1, \dots, \epsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad (10.4)$$

如果样本 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 使得式 (10.4) 成立, 则称它满足一个线性模型 (linear model)。方程组 (10.4) 称为 k 元线性回归模型, 我们将利用最小二乘法对未知参数 $\beta_0, \beta_1, \dots, \beta_k, \sigma^2$ 进行估计, 相应的估计值记作 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k, \hat{\sigma}^2$ 。

式 (10.4) 中, 条件 $\epsilon_1, \dots, \epsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ 是一个假设, 被称为方差齐性 (homogeneity of variance) 假设, 主要是为了简化模型便于计算。在使用这个假设之前, 应该对其进行假设检验。

例 10.5 (多项式回归模型). 若回归函数是某个一元 k 次多项式 $f(a) = \beta_0 + \beta_1 a + \dots + \beta_k a^k$, 当观察到解释变量 a 的 n 个取值 a_1, \dots, a_n 及其对应的响应值 x_1, \dots, x_n , 其中 $n > k$, 回归分析可抽象为一个 k 元线性回归模型。

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} 1 & a_1 & \cdots & a_1^k \\ 1 & a_2 & \cdots & a_2^k \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_n & \cdots & a_n^k \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (10.5)$$

定义 10.4. 我们定义常数 $a_0 = 1$, 令 A 是变量 a_0, a_1, \dots, a_n 的数据矩阵, 具体如下。

$$A = \begin{pmatrix} 1 & a_{11} & \cdots & a_{1k} \\ 1 & a_{21} & \cdots & a_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_{n1} & \cdots & a_{nk} \end{pmatrix} = \begin{pmatrix} a_{10} & a_{11} & \cdots & a_{1k} \\ a_{20} & a_{21} & \cdots & a_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n0} & a_{n1} & \cdots & a_{nk} \end{pmatrix}_{n \times (k+1)}$$

其中, $a_{10} = a_{20} = \dots = a_{n0} = 1$ 。为了记述和推导的方便, 我们把线性回归模型 (10.4) 整理为矩阵的形式。

$$\mathbf{X} = A\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{其中 } \boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top \sim N_n(\mathbf{0}, \sigma^2 I) \quad (10.6)$$

上式中，回归系数向量 $\beta = (\beta_0, \beta_1, \dots, \beta_k)^\top \in \mathbb{R}^{k+1}$ 未知待定。

练习 10.1. 随机向量 X (10.6) 的密度函数为

$$f(\mathbf{x}|\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{(\mathbf{x} - A\boldsymbol{\beta})^\top(\mathbf{x} - A\boldsymbol{\beta})}{2\sigma^2} \right\} \quad (10.7)$$

定义 10.5. 若**定义 10.4** 中误差 $\epsilon_1, \dots, \epsilon_n$ 的分布非正态，或者经过某变换后得到的线性模型 $E(\mathbf{X}) = g^{-1}(A\boldsymbol{\beta})$ 称为广义线性模型 (generalized linear model, GLM)，其中 g 是连接函数 (link function)，如 $g(z) = \ln z, 1/z, \ln \frac{z}{1-z}$ 等。广义线性模型的内容不在本书的范围之内，有关知识详见 [108]。

既然回归模型 (10.6) 是用来做预测或控制的，人们当然希望理论值 $\hat{\mathbf{x}}$ 和观测值 \mathbf{x} 越接近越好。为此，参数 $\boldsymbol{\beta}$ 的估计值 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)^\top$ 要使得下述残差平方和 (residual sum of squares, RSS) 达到最小。

$$\text{RSS} = \sum_{i=1}^n (x_i - \hat{x}_i)^2 = (\mathbf{x} - \hat{\mathbf{x}})^\top(\mathbf{x} - \hat{\mathbf{x}}) = \|\mathbf{x} - A\hat{\boldsymbol{\beta}}\|_2^2 \quad (10.8)$$

采用残差平方和比采用残差绝对值之和在理论推导和计算上更简捷些，所以习惯上用残差平方 (10.8) 和来构造最优化问题中的目标函数。在此标准之下，对未知参数 $\boldsymbol{\beta}$ 的点估计就归结为一个最优化的问题，它的求解方法被称为最小二乘法。

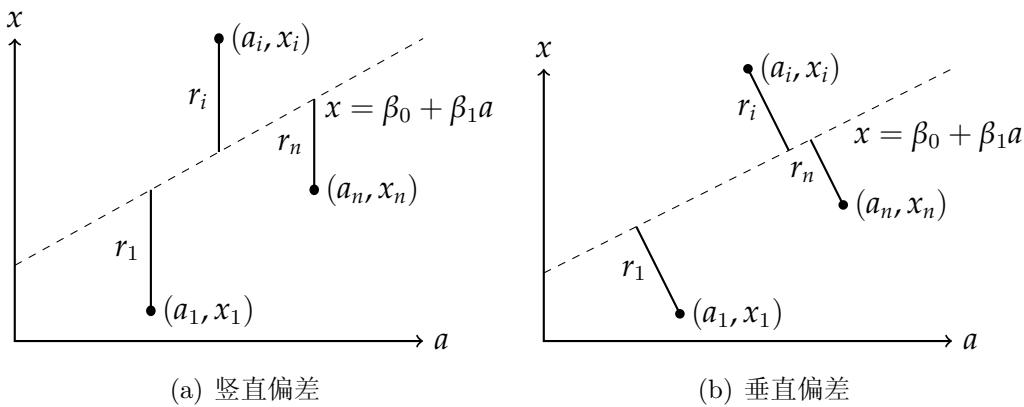


图 10.4: 响应变量 X 的观测值 x_i 与理论值 $\hat{x}_i = \beta_0 + \beta_1 a_i$ 之间的残差 r_i 可以有多种定义，式 (10.8) 的几何解释基于 (a) 的竖直偏差，它比 (b) 的垂直偏差易于计算。

本节内容

线性模型中未知参数的点估计方法——最小二乘法。

关键知识

- (1) 最小二乘估计；(2) 正则方程；(3) 参数最小二乘估计的分布；(4) 线性模型中回归系数的线性假设。

10.1.1 最小二乘估计

最小二乘法给出了一类最优化^{*}的标准，它的思想是朴素的。举个例子，对某物体长度的多次测量得到了观察结果 x_1, x_2, \dots, x_n ，测量误差分别为 $\epsilon_i = x_i - \theta, i = 1, 2, \dots, n$ ，其中 θ 为真实长度。对 θ 什么样的估计才是好的？一个简单而有效的评判标准是如下定义的误差平方和。

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (x_i - \theta)^2$$

显然，误差平方和越小越好。为了使其达到最小，未知参数 θ 的估计值 $\hat{\theta}$ 应是样本值的算术平均 \bar{x} ，即

$$\hat{\theta} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

最小二乘法深刻地影响了统计学的发展，曾是十九世纪的数学研究的热点之一，它的重要性“犹如微积分之于数学”[169]。德国数学家 Gauss 与法国数学家 Legendre 之间曾有过最小二乘法优先权之争，激烈程度堪比 Newton 和 Leibniz 的微积分优先权之战，真实情况可能是 Gauss 对最小二乘法的研究在先发表在后[†]。最小二乘法被视为 Gauss 应用数学解决实际问题的典型成就，见图 10.5。从最小二乘法优先权的公案，我们也可以感受到数学家对该方法之重视，以及它在数学史中地位之关键。



图 10.5: Gauss 数学应用奖的奖章背面为一条穿过圆和正方形的曲线，代表 Gauss 利用最小二乘法算出谷神星的轨道。2006 年，第一届 Gauss 奖授予了伊藤清。

^{*}最优化 (Optimization) 理论，有时也称作数学规划 (Mathematical Programming)，是应用数学的一个重要领域，它研究在约束条件之下目标函数的极值问题。

[†]为解出最小二乘估计，Gauss 还提出了线性方程组的“Gauss 消去法”。Legendre 没有研究最小二乘法的误差分析问题，这部分工作由 Gauss 于 1809 年完成，并对统计学产生了深远的影响。

最小二乘法虽然重要却并非完美，它的缺点是稳健性^{*}欠佳，这是因为平方函数增长较快，因此有人建议用比平方增长慢的函数来构造目标函数，譬如误差的绝对值之和，但计算上会稍复杂一些。

定义 10.6. 对于线性回归模型 (10.6)，如果 $\forall \beta \in \mathbb{R}^{k+1}$ 都有

$$\|\mathbf{x} - A\hat{\beta}\|_2^2 \leq \|\mathbf{x} - A\beta\|_2^2 \quad (10.9)$$

则称 $\hat{\beta}$ 是未知参数 β 的最小二乘估计 (least square estimate, LSE)。直观上，最小二乘估计使得理论值 $\hat{\mathbf{x}} = A\hat{\beta}$ 与观测值 \mathbf{x} 的欧氏距离最近。

性质 10.1. β 的最小二乘估计 (10.9) 也是最大似然估计。

证明. 由线性回归模型 (10.6)，随机向量 $\mathbf{X} \sim N_n(A\beta, \sigma^2 I)$ ，其密度函数为

$$\phi(\mathbf{x}|A\beta, \sigma^2 I) \propto \frac{1}{\sigma} \exp \left\{ -\frac{\|\mathbf{x} - A\beta\|_2^2}{2\sigma^2} \right\} \quad \square$$

首先，线性回归模型 (10.6) 中参数的最小二乘估计总是存在的（见定理 10.2）。该问题所要最小化的目标函数是一个关于 β 的函数 $L(\beta)$ ，即

$$\begin{aligned} L(\beta) &= \epsilon^\top \epsilon \\ &= (\mathbf{x} - A\beta)^\top (\mathbf{x} - A\beta) \\ &= \mathbf{x}^\top \mathbf{x} - \mathbf{x}^\top A\beta - \beta^\top A^\top \mathbf{x} + \beta^\top A^\top A\beta \end{aligned}$$

欲寻找合适的 β 使得上式最小，根据附录 E 中定理 E.11，不难从 $\partial L(\beta)/\partial \beta = 0$ 得到下面所谓的正则方程 (regular equation)：

$$A^\top A\beta = A^\top \mathbf{x}, \text{ 或者 } A^\top (\mathbf{x} - A\beta) = 0 \quad (10.10)$$

正则方程 (10.10) 的几何意义是：向量 $\epsilon = \mathbf{x} - A\beta$ 与 A 的每个列向量都正交。我们将在定理 10.2 的证明中深入探讨最小二乘估计的几何意义。

定理 10.1. 若矩阵 $A^\top A$ 非奇异（即存在逆矩阵），则线性回归模型 (10.6) 中未知参数 β 的最小二乘估计存在且唯一，就是

$$\hat{\beta} = (A^\top A)^{-1} A^\top \mathbf{x} \quad (10.11)$$

^{*}稳健性 (robustness) 是衡量统计方法优劣的标准之一，它考察的是方法是否容易受样本中异常值的影响。例如，中位数的稳健性优于均值。

当 $\mathbf{X} = \mathbf{x}$ 时, 我们从 (10.11) 得到 $\hat{\boldsymbol{\beta}} = (A^\top A)^{-1} A^\top \mathbf{x}$, 以及拟合值的向量

$$\hat{\mathbf{x}} = A\hat{\boldsymbol{\beta}} = A(A^\top A)^{-1} A^\top \mathbf{x}$$

推论 10.1. 在定理 10.1 的条件下,

$$(\mathbf{X} - A\boldsymbol{\beta})^\top (\mathbf{X} - A\boldsymbol{\beta}) = (\mathbf{X} - A\hat{\boldsymbol{\beta}})^\top (\mathbf{X} - A\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top A^\top A (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

证明. 因为 $A^\top (\mathbf{X} - A\hat{\boldsymbol{\beta}}) = 0$, 所以 $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top A^\top (\mathbf{X} - A\hat{\boldsymbol{\beta}}) = 0$, 进而

$$\begin{aligned} (\mathbf{X} - A\boldsymbol{\beta})^\top (\mathbf{X} - A\boldsymbol{\beta}) &= [(X - A\hat{\boldsymbol{\beta}}) + (A\hat{\boldsymbol{\beta}} - A\boldsymbol{\beta})]^\top [(X - A\hat{\boldsymbol{\beta}}) + (A\hat{\boldsymbol{\beta}} - A\boldsymbol{\beta})] \\ &= (\mathbf{X} - A\hat{\boldsymbol{\beta}})^\top (\mathbf{X} - A\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top A^\top A (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \end{aligned} \quad \square$$

矩阵 $A^+ = (A^\top A)^{-1} A^\top$ 也称作列满秩矩阵 A 的左逆 (left inverse), 这是因为 $A^+ A = I$ 。左逆是一类特殊的 Moore-Penrose 伪逆 (pseudo-inverse) 矩阵*, 显然

$$A^+ (A^+)^T = (A^\top A)^{-1}$$

练习 10.2. 对于线性回归模型 (10.6) 中未知参数 $\boldsymbol{\beta}$ 的最小二乘估计, 请验证

$$\text{RSS} = \mathbf{x}^\top (I - AA^+) \mathbf{x}, \text{ 其中 } A^+ = (A^\top A)^{-1} A^\top$$

提示: 显然, AA^+ 是对称阵。利用式 (10.11), $\text{RSS} = \|\mathbf{x} - A\hat{\boldsymbol{\beta}}\|_2^2 = \|\mathbf{x} - AA^+ \mathbf{x}\|_2^2 = \|(I - AA^+) \mathbf{x}\|_2^2 = \mathbf{x}^\top (I - AA^+) (I - AA^+) \mathbf{x} = \mathbf{x}^\top (I - AA^+ - AA^+ + AA^+ AA^+) \mathbf{x} = \mathbf{x}^\top (I - AA^+) \mathbf{x}$ 。

例 10.6. 考虑一元线性回归模型 $\mathbf{X} = A\boldsymbol{\beta} + \boldsymbol{\epsilon}$, 其中 $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, $A^\top = (\frac{1}{a_1} \cdots \frac{1}{a_n})$ 且 $A^\top A$ 可逆。给定 \mathbf{X} 的观测值 $\mathbf{x} = (x_1, \dots, x_n)^\top$, 求未知参数 $\boldsymbol{\beta}$ 的最小二乘估计 $\hat{\boldsymbol{\beta}}$ 。

解. 直接利用定理 10.1 来求解 $\boldsymbol{\beta}$ 的最小二乘估计 $\hat{\boldsymbol{\beta}}$ 如下,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (A^\top A)^{-1} (A^\top \mathbf{x}) \\ &= \begin{pmatrix} n & \sum_{i=1}^n a_i \\ \sum_{i=1}^n a_i & \sum_{i=1}^n a_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n a_i x_i \end{pmatrix} \\ &= \frac{1}{n \sum_{i=1}^n a_i^2 - (\sum_{i=1}^n a_i)^2} \begin{pmatrix} \sum_{i=1}^n a_i^2 \sum_{i=1}^n x_i - \sum_{i=1}^n a_i \sum_{i=1}^n a_i x_i \\ n \sum_{i=1}^n a_i x_i - \sum_{i=1}^n a_i \sum_{i=1}^n x_i \end{pmatrix} \end{aligned}$$

* 伪逆矩阵是对逆矩阵的推广, 分别由美国数学家 E. H. Moore (1862-1932) 和英国物理学家兼数学家 R. Penrose (1931-) 于 1920 和 1955 年提出。对于实矩阵 $A_{m \times n}$, 如果矩阵 $A_{n \times m}^+$ 满足以下三个条件, 则称为 A 的伪逆: (1) $AA^+A = A$, (2) $A^+AA^+ = A^+$, (3) AA^+ 和 A^+A 都是对称矩阵。伪逆总是存在且唯一的。特别地, 若方阵 A 可逆, 其伪逆就是 A 的逆。一般地, $A^+A = I$ 和 $AA^+ = I$ 并不成立。

也可以通过最小化目标函数 $L(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} = \sum_{i=1}^n (x_i - \beta_0 - \beta_1 a_i)^2$, 得到未知参数 β_0, β_1 的最小二乘估计如下。

$$\hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{a}, \text{ 其中 } \bar{a} = \frac{1}{n} \sum_{i=1}^n a_i \text{ 且 } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (a_i - \bar{a})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

例 10.7. 在第 591 页的**例 10.5** 中, 若 a_1, \dots, a_n 中至少有 $k+1$ 个两两不等, 试证明: 多项式回归模型 (10.5) 中未知参数 $\beta_0, \beta_1, \dots, \beta_k$ 的最小二乘估计存在。

证明. 若 a_1, \dots, a_n 中至少有 $k+1$ 个两两不等, 则如下定义的 Vandermonde 矩阵 A 是列满秩的 (即, 所有列向量线性无关)。

$$A = \begin{pmatrix} 1 & a_1 & \cdots & a_1^k \\ 1 & a_2 & \cdots & a_2^k \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_n & \cdots & a_n^k \end{pmatrix} \text{ 且 } \begin{pmatrix} 1 & a_1 & \cdots & a_1^k \\ 1 & a_2 & \cdots & a_2^k \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_{k+1} & \cdots & a_{k+1}^k \end{pmatrix} = \prod_{1 \leq i < j \leq k+1} (a_j - a_i)$$

由线性代数的知识, 矩阵 A 列满秩当且仅当 $A^\top A$ 可逆。由**定理 10.1**, 证得参数的最小二乘估计存在且唯一。特别地, 当 $n = k+1$ 时, $\beta_0, \beta_1, \dots, \beta_k$ 就是 Lagrange 插值多项式的系数。 \square

定理 10.1 给出的是“批量”估计的结果, 即变量 a_1, \dots, a_k 的观测矩阵 A 是已经给定了的。如果观测结果是一个一个地给 (这种情况在实践中很常见), 我们不能等到攒够一定规模的观测结果后去估计未知参数 $\boldsymbol{\beta}$, 而是希望给一个观测结果估计一次参数。笨的方法是每次按照式 (10.11) 来计算, 聪明的方法是下面增量学习 (incremental learning) 的算法, 它的基础是以下事实:

我们依次得到 $\mathbf{a} = (a_0, a_1, \dots, a_k)^\top$ 的观测值 $\mathbf{a}_1 = (a_{10}, a_{11}, \dots, a_{1k})^\top, \dots, \mathbf{a}_n = (a_{n0}, a_{n1}, \dots, a_{nk})^\top$, 其中 $a_{10} = \dots = a_{n0} = 1$ 。根据**定义 10.4**, 显然数据矩阵为 $A_n = (\mathbf{a}_1, \dots, \mathbf{a}_n)^\top$ 。由**定理 10.1** 我们有

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n &= (A_n^\top A_n)^{-1} A_n^\top \mathbf{x}, \text{ 利用事实 } A_n = (A_{n-1}, \mathbf{a}_n) \\ &= (A_{n-1}^\top A_{n-1} + \mathbf{a}_n \mathbf{a}_n^\top)^{-1} \left(\sum_{i=1}^{n-1} \mathbf{a}_i x_i + \mathbf{a}_n x_n \right), \text{ 利用第 770 页的定理 E.9} \\ &= \left(\Gamma_{n-1} - \frac{\Gamma_{n-1} \mathbf{a}_n \mathbf{a}_n^\top \Gamma_{n-1}}{1 + \mathbf{a}_n^\top \Gamma_{n-1} \mathbf{a}_n} \right) \left(\sum_{i=1}^{n-1} \mathbf{a}_i x_i + \mathbf{a}_n x_n \right), \text{ 其中 } \Gamma_{n-1} = (A_{n-1}^\top A_{n-1})^{-1} \\ &= \hat{\boldsymbol{\beta}}_{n-1} - \frac{\Gamma_{n-1} \mathbf{a}_n \mathbf{a}_n^\top \hat{\boldsymbol{\beta}}_{n-1}}{1 + \mathbf{a}_n^\top \Gamma_{n-1} \mathbf{a}_n} + \left(\Gamma_{n-1} - \frac{\Gamma_{n-1} \mathbf{a}_n \mathbf{a}_n^\top \Gamma_{n-1}}{1 + \mathbf{a}_n^\top \Gamma_{n-1} \mathbf{a}_n} \right) \mathbf{a}_n x_n \end{aligned}$$

上式给出了 $\hat{\beta}_n$ 与 $\hat{\beta}_{n-1}$ 之间的递归关系，只需 Γ_{n-1} 和 \mathbf{a}_n 的帮忙便可完成结果的更新。由此，我们得到下面的线性回归的增量学习算法。

算法 10.1. 初始化 $\hat{\beta}_0 = \mathbf{0} \in \mathbb{R}^{k+1}$ 且 $\Gamma_0 = I_{k+1}$ ，按照下述方法算得的 $\hat{\beta}_n$ 便是定理 10.1 的 $\hat{\beta}$ ，并且 $\Gamma_n = (A^\top A)^{-1}$ 。

$$\begin{aligned} C_i &= \frac{\Gamma_{i-1} \mathbf{a}_i \mathbf{a}_i^\top}{1 + \mathbf{a}_i^\top \Gamma_{i-1} \mathbf{a}_i} \\ \Gamma_i &= \Gamma_{i-1} - C_i \Gamma_{i-1} \\ \hat{\beta}_i &= \hat{\beta}_{i-1} - C_i \hat{\beta}_{i-1} + \Gamma_i \mathbf{a}_i x_i, \text{ 其中 } i = 1, \dots, n \end{aligned}$$

一般情况下， $n > k$ 。算法 10.1 和“批量”方法 (10.11) 的算法复杂度都是 $O(nk^2)$ ，但每一步所需要的存储空间只有 $O(k^2)$ ，比“批量”方法要小很多。

10.1.2 线性回归的若干性质

本小节所涉及的范数都是 2-范数，简记 $\|\mathbf{x}\|_2$ 为 $\|\mathbf{x}\|$ 。

\curvearrowleft 定理 10.2. 线性回归模型 (10.6) 参数 β 的最小二乘估计总是存在的，且 $\hat{\beta}$ 为 β 的最小二乘估计当且仅当 $\hat{\beta}$ 满足正则方程 (10.10)。

证明. 由矩阵 $A_{n \times (k+1)} = (\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_k)$ 的所有列向量张成的线性空间即为

$$\begin{aligned} \text{span}(\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_k) &= \{\beta_0 \mathbf{a}_0 + \beta_1 \mathbf{a}_1 + \dots + \beta_k \mathbf{a}_k : \beta = (\beta_0, \beta_1, \dots, \beta_k)^\top \in \mathbb{R}^{k+1}\} \\ &= \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} = A\beta, \text{ 其中 } \beta \in \mathbb{R}^{k+1}\} \end{aligned}$$

令 $A\tilde{\beta}$ 是向量 $\mathbf{x} \in \mathbb{R}^n$ 在线性空间 $\mathbb{S} = \text{span}(\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_k)$ 上的投影，记作 $\text{proj}_{\mathbb{S}}\mathbf{x}$ 。显然 $\|\mathbf{x} - A\tilde{\beta}\| \leq \|\mathbf{x} - A\beta\|$ ，即 $\tilde{\beta}$ 是 β 的最小二乘估计，存在性得证。

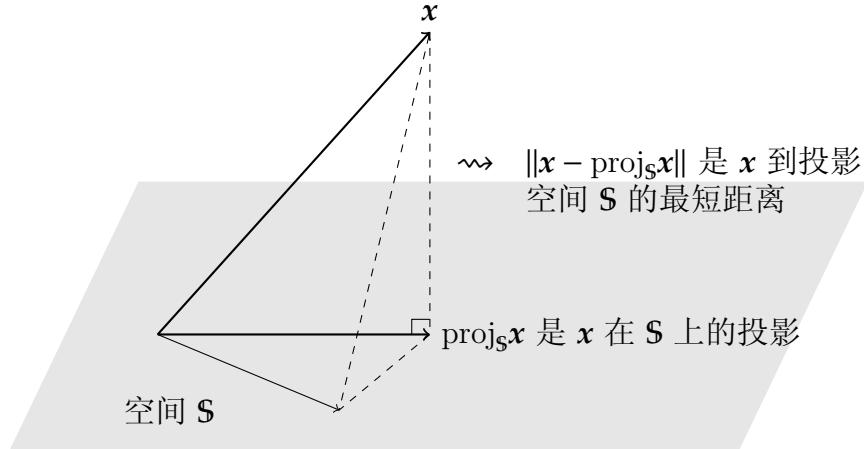


图 10.6: 向量 $\mathbf{x} \in \mathbb{R}^n$ 在线性空间 $\mathbb{S} \subseteq \mathbb{R}^n$ 里的最佳逼近是 \mathbf{x} 在 \mathbb{S} 上的投影向量 $\text{proj}_{\mathbb{S}}\mathbf{x}$ ，其几何直观源自直和分解 $\mathbf{x} = \text{proj}_{\mathbb{S}}\mathbf{x} + (\mathbf{x} - \text{proj}_{\mathbb{S}}\mathbf{x})$ ，其中向量 $(\mathbf{x} - \text{proj}_{\mathbb{S}}\mathbf{x}) \perp \mathbb{S}$ 。该分解满足勾股定理 $\|\mathbf{x}\|^2 = \|\text{proj}_{\mathbb{S}}\mathbf{x}\|^2 + \|\mathbf{x} - \text{proj}_{\mathbb{S}}\mathbf{x}\|^2$ 。

如果 $\hat{\beta}$ 是 β 的最小二乘估计，则必有 $A\hat{\beta} = \text{proj}_{\mathbb{S}}\mathbf{x}$ ，如若不然，将导致 $\|\mathbf{x} - A\hat{\beta}\| < \|\mathbf{x} - A\hat{\beta}\|$ ，矛盾！于是条件 $A\hat{\beta} = \text{proj}_{\mathbb{S}}\mathbf{x}$ 是 $\hat{\beta}$ 为 β 的最小二乘估计的充要条件，它等价于 $\epsilon = \mathbf{x} - A\hat{\beta} \perp \mathbb{S}$ ，即残差向量 $\epsilon = \mathbf{x} - A\hat{\beta}$ 垂直于 A 的每个列向量，也就是说 $A^\top(\mathbf{x} - A\hat{\beta}) = \mathbf{0}_{k+1}$ ，故 $\hat{\beta}$ 满足正则方程 (10.10)。□

 如果 A 是列满秩的，线性回归模型 (10.6) 参数的最小二乘估计是唯一的。定理 10.2 揭示了 $\hat{\mathbf{x}} = A\hat{\beta}$ 的几何意义，即 \mathbf{x} 在空间 \mathbb{S} 上的投影。如果 A 不是列满秩的， A 的列向量就是线性相关的，用这些列向量来线性表示 $\text{proj}_{\mathbb{S}}\mathbf{x}$ 就可能不唯一，即最小二乘估计可能不唯一。对模型 (10.6) 中的未知参数 β 若有其他额外的约束条件，如 $\beta^\top \beta \leq 2$ 等， β 的最小二乘估计可以通过 Lagrange 乘子法得到。

\nwarrow 定理 10.3. 对于线性回归模型 (10.6), 若 $A_{n \times (k+1)}$ 为列满秩 (即秩为 $k+1$), 令 $A^+ = (A^\top A)^{-1} A^\top$, 则回归系数的最小二乘估计 $\hat{\beta} = A^+ X$ 和 $\hat{\sigma}^2 = \|X - A\hat{\beta}\|^2/(n-k-1)$ 分别是 β 和 σ^2 的无偏估计。更细致的结果是

$$\begin{aligned}\hat{\beta} &\sim N_{k+1}(\beta, \sigma^2(A^\top A)^{-1}) \\ \frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} &\sim \chi_{n-k-1}^2\end{aligned}$$

※证明. 若矩阵 $A_{n \times (k+1)}$ 为列满秩, 则 $A^\top A$ 为正定矩阵, 存在逆矩阵。

□ 已知 X 服从正态分布, 根据第 335 页的定理 4.10, 经过线性变换后 $\hat{\beta} = A^+ X$ 依然服从正态分布, 只需计算其均值和协方差阵便可知该正态分布的细节。

$$\begin{aligned}E\hat{\beta} &= A^+(EX) = A^+ A \beta = \beta \\ \text{Cov}(\hat{\beta}, \hat{\beta}) &= \text{Cov}[A^+(A\beta + \epsilon), A^+(A\beta + \epsilon)] \\ &= \text{Cov}(A^+\epsilon, A^+\epsilon) \\ &= (A^\top A)^{-1} A^\top \text{Cov}(\epsilon, \epsilon) A (A^\top A)^{-1} \\ &= \sigma^2 (A^\top A)^{-1}\end{aligned}$$

□ 首先, 把 $\hat{\beta} = A^+ X$ 代入 $\|X - A\hat{\beta}\|^2$ 将之化简, 再计算其期望。

$$\begin{aligned}\|X - A\hat{\beta}\|^2 &= x^\top X - x^\top A A^+ X \\ &= x^\top (I - A A^+) X, \text{ 令 } B = I - A A^+ \\ &= x^\top B X, \text{ 根据附录 E 中定理 E.1 可得} \\ &= \text{tr}(B X x^\top) \\ E\|X - A\hat{\beta}\|^2 &= E[\text{tr}(B X x^\top)] \\ &= \text{tr}[B \cdot E(X x^\top)] \\ &= \text{tr}[B \cdot (\sigma^2 I + A \beta \beta^\top A^\top)] \\ &= \sigma^2 \text{tr}(B), \text{ 因为 } BA = O_{n \times (k+1)} \\ &= \sigma^2 \{\text{tr}(I) - \text{tr}(A A^+)\} \\ &= \sigma^2 \{n - \text{tr}(A^+ A)\} \\ &= \sigma^2(n - k - 1)\end{aligned}$$

于是, $\hat{\sigma}^2 = \|X - A\hat{\beta}\|^2/(n - k - 1)$ 是 σ 的无偏估计。

□ 欲证 $(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-k-1}^2$, 只需往证

$$\frac{\|\mathbf{X} - A\hat{\boldsymbol{\beta}}\|^2}{\sigma^2} = \frac{\mathbf{x}^\top (I - AA^\top) \frac{\mathbf{X}}{\sigma}}{\sigma} \sim \chi_{n-k-1}^2$$

矩阵 $B = I - AA^\top$ 是对称幂等矩阵, 根据定理 4.12, $\|\mathbf{X} - A\hat{\boldsymbol{\beta}}\|^2/\sigma^2$ 服从 χ^2 分布。再由其期望为 $n - k - 1$, 得证 $\|\mathbf{X} - A\hat{\boldsymbol{\beta}}\|^2/\sigma^2 \sim \chi_{n-k-1}^2$ 。 □

定理 10.4 (Gauss-Markov). 对于线性回归模型 (10.6), 若 $A_{n \times (k+1)}$ 列满秩, 则在 $\boldsymbol{\beta}$ 的所有形为 $C_{(k+1) \times n}\mathbf{X}$ 的线性无偏估计当中, 最小二乘估计 $\hat{\boldsymbol{\beta}} = A^\top \mathbf{X}$ 是方差最小的。

证明. 将矩阵 $C_{(k+1) \times n}$ 分解为 $C = A^\top + D_{(k+1) \times n}$, 则 $\tilde{\boldsymbol{\beta}} = CX$ 是 $\boldsymbol{\beta}$ 的无偏估计当且仅当 $DA = O$ 。这是因为

$$E(\tilde{\boldsymbol{\beta}}) = E(CX) = E[(A^\top + D)(A\boldsymbol{\beta} + \boldsymbol{\epsilon})] = (A^\top + D)A\boldsymbol{\beta} = (I + DA)\boldsymbol{\beta}$$

由定理 10.3 知 $V(\hat{\boldsymbol{\beta}}) = \sigma^2(A^\top A)^{-1}$, 下面往证 $V(\tilde{\boldsymbol{\beta}}) \geq V(\hat{\boldsymbol{\beta}})$ 。

$$V(\tilde{\boldsymbol{\beta}}) = V(CX) = CV(X)C^\top = \sigma^2 CC^\top = \sigma^2(A^\top A)^{-1} + \sigma^2 DD^\top \geq V(\hat{\boldsymbol{\beta}}) \quad \square$$

定义 10.7. 令 $\hat{\boldsymbol{\beta}}$ 是线性回归模型 (10.6) 的未知参数 $\boldsymbol{\beta}$ 的最小二乘估计。由观测数据 $\mathbf{x} = (x_1, \dots, x_n)^\top$ 及其均值 $\bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$, 以及理论值 $\hat{\mathbf{x}} = A\hat{\boldsymbol{\beta}} = (\hat{x}_1, \dots, \hat{x}_n)^\top$ 构造下面的三类平方和: 总平方和 (total sum of squares, TSS), 回归平方和 (explained sum of squares, ESS), 残差平方和 (residual sum of squares, RSS)。

$$\text{总平方和: } TSS = \sum_{j=1}^n (x_j - \bar{x})^2 = \|\mathbf{x} - \bar{x}\mathbf{1}\|^2, \text{ 其中 } \mathbf{1} = (\underbrace{1, \dots, 1}_n)^\top$$

$$\text{回归平方和: } ESS = \sum_{j=1}^n (\hat{x}_j - \bar{x})^2 = \|\hat{\mathbf{x}} - \bar{x}\mathbf{1}\|^2$$

$$\text{残差平方和: } RSS = \sum_{j=1}^n (x_j - \hat{x}_j)^2 = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$$

性质 10.2. 总平方和是回归平方和与残差平方和二者之和, 即

$$TSS = RSS + ESS$$

证明. 由定理 10.2 的证明, $\hat{\mathbf{x}} - \bar{x}\mathbf{1} \in \mathbb{S}$ 。因为 $(\mathbf{x} - \hat{\mathbf{x}}) \perp \mathbb{S}$, 所以 $\mathbf{x} - \hat{\mathbf{x}}$ 与 $\hat{\mathbf{x}} - \bar{x}\mathbf{1}$ 正交,

即二者内积为零。于是，

$$\begin{aligned}\text{TSS} &= \|\boldsymbol{x} - \hat{\boldsymbol{x}} + \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\mathbf{1}\|^2 \\ &= \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + \|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\mathbf{1}\|^2 + 2(\boldsymbol{x} - \hat{\boldsymbol{x}})^\top (\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\mathbf{1}) \\ &= \text{RSS} + \text{ESS}\end{aligned}$$

□

10.1.3 回归模型的假设检验

对于线性回归模型 (10.6) 中的未知参数 $\beta_0, \beta_1, \dots, \beta_k$, 除了对它们进行估计, 有时还会遇到如下一些假设检验的问题。

- 变量 X 与 a_1, a_2, \dots, a_k 之间是否存在线性关系? 若无线性关系, 则 $\beta_0 = \beta_1 = \dots = \beta_k = 0$, 欲通过样本说明这一事实需要对零假设 $H_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0$ 进行检验。
- 如果变量 X 与因素 a_1, a_2, \dots, a_k 之间的确存在线性关系, 是否每个因素都起作用呢? 若因素 a_i 对 X 的影响不是显著的, 则 $\beta_i = 0$, 欲通过样本说明这一事实需要对零假设 $H_0 : \beta_i = 0$ 进行检验。

定义 10.8. 包括上述两个假设检验问题, 凡是有关 $\beta_0, \beta_1, \dots, \beta_k$ 之间线性关系的假设都可以统一地表示为

$$H_0 : H\beta = \mathbf{0}, \text{ 其中 } H_{r \times (k+1)} \text{ 是一个秩为 } r \leq k+1 \text{ 的矩阵} \quad (10.12)$$

这类假设被称为线性假设 (linear hypothesis), 它陈述的是参数 $\beta_0, \beta_1, \dots, \beta_k$ 满足 r 个独立的线性约束。

定理 10.5. 考虑线性回归模型 (10.6), 线性假设为 $H_0 : H\beta = \mathbf{0}$, 其中 H 是一个秩为 $r \leq k+1$ 的 $r \times (k+1)$ 矩阵。令 $\hat{\beta}$ 是 β 的最大似然估计, 令 $\tilde{\beta}$ 是零假 H_0 成立时 β 的最大似然估计。构造统计量如下,

$$F = \frac{(X - A\hat{\beta})^\top (X - A\hat{\beta}) - (X - A\tilde{\beta})^\top (X - A\tilde{\beta})}{(X - A\hat{\beta})^\top (X - A\hat{\beta})} \quad (10.13)$$

则下面的结果成立,

$$\frac{n-k-1}{r} F \sim F_{r, n-k-1} \quad (10.14)$$

※证明. 详见 [137] 的第十二章。 □

算法 10.2. 统计量 F 如定理 10.5 所描述, 在给定的显著水平 α 之下, 似然比检验拒绝零假设 $H_0 : H\beta = \mathbf{0}$ 的条件是

$$\frac{n-k-1}{r} F > F_{r, n-k-1, 1-\alpha}$$

例 10.8. 考虑一元线性回归模型 $X_i = \beta_0 + \beta_1 a_i + \epsilon_i, i = 1, \dots, n$, 并假定 $\epsilon_1, \dots, \epsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ 。该模型写成式 (10.6) 的形式, 其中

$$\boldsymbol{\beta} = (\beta_0, \beta_1)^\top, \boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top, \text{ 并且 } A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ a_1 & a_2 & \cdots & a_n \end{pmatrix}^\top$$

零假设为 $H_0 : \beta_1 = 0$, 整理成线性假设的形式 $H_0 : H\boldsymbol{\beta} = \mathbf{0}$, 其中 $H = (0, 1)$, 于是 $r = 1, k = 2$ 。我们知道, 随机向量 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的密度函数为

$$f(\mathbf{x}; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \beta_0 - \beta_1 a_i)^2 \right\}$$

□ 利用最大似然估计的方法, 容易得到参数点估计的结果:

$$\begin{aligned} \hat{\beta}_0 &= \bar{X} - \hat{\beta}_1 \bar{a}, \text{ 其中 } \bar{a} = \frac{1}{n} \sum_{i=1}^n a_i \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (a_i - \bar{a})(X_i - \bar{X})}{\sum_{i=1}^n (a_i - \bar{a})^2}, \text{ 其中 } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\beta}_0 - \hat{\beta}_1 a_i)^2 \end{aligned}$$

□ 在零假设 $H_0 : \beta_1 = 0$ 成立时, 参数最大似然估计的结果是:

$$\begin{aligned} \tilde{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \\ \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{aligned}$$

根据式 (10.24), 构造统计量如下

$$\begin{aligned} F &= \frac{\sum_{i=1}^n (X_i - \bar{X} + \hat{\beta}_1 \bar{a} - \hat{\beta}_1 a_i)^2 - \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X} + \hat{\beta}_1 \bar{a} - \hat{\beta}_1 a_i)^2} \\ &= \frac{\hat{\beta}_1^2 \sum_{i=1}^n (a_i - \bar{a})^2}{\sum_{i=1}^n (X_i - \bar{X} + \hat{\beta}_1 \bar{a} - \hat{\beta}_1 a_i)^2} \end{aligned}$$

根据结果 (10.14) 有 $(n-2)F \sim F_{1, n-2}$ 。由 t 分布的定义知 $\sqrt{(n-2)F} \sim t_{n-2}$ 。于是, 若 $\sqrt{(n-2)F} \geq t_{n-2, 1-\alpha}$, 则在水平 α 拒绝 H_0 假设。

例 10.9. 还是例 10.8 中的线性回归模型。零假设换为 $H_0 : \beta_0 = 0$, 取 $H = (1, 0)$ 。零

假设 H_0 成立时，参数的最大似然估计为

$$\begin{aligned}\tilde{\beta}_1 &= \frac{\sum_{i=1}^n a_i X_i}{\sum_{i=1}^n a_i^2} = \hat{\beta}_1 + \frac{n \hat{\beta}_0 \bar{a}}{\sum_{i=1}^n a_i^2} \\ \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \tilde{\beta}_1 a_i)^2\end{aligned}$$

根据式 (10.24)，构造统计量如下

$$\begin{aligned}F &= \frac{\sum_{i=1}^n (X_i - \bar{X} + \hat{\beta}_1 \bar{a} - \hat{\beta}_1 a_i)^2 - \sum_{i=1}^n (X_i - \tilde{\beta}_1 a_i)^2}{\sum_{i=1}^n (X_i - \bar{X} + \hat{\beta}_1 \bar{a} - \hat{\beta}_1 a_i)^2} \\ &= \frac{\hat{\beta}_0^2 n \sum_{i=1}^n (a_i - \bar{a})^2 / \sum_{i=1}^n a_i^2}{\sum_{i=1}^n (X_i - \bar{X} + \hat{\beta}_1 \bar{a} - \hat{\beta}_1 a_i)^2}\end{aligned}$$

与上例类似，若 $\sqrt{(n-2)F} \geq t_{n-2, 1-\alpha}$ ，则在水平 α 拒绝 H_0 假设。

10.1.4 正交多项式回归

定义 10.9 (正交多项式). 令 $p_d(x)$ 表示 d 次多项式, 如果多项式 $p_0(x), p_1(x), \dots, p_k(x)$ 满足下面的条件, 则称它们是 x_1, \dots, x_m 上的正交多项式 (orthogonal polynomials)。

$$\begin{aligned} \sum_{j=1}^m p_d^2(x_j) &= c_d \neq 0, \text{ 其中 } d = 0, 1, \dots, k \text{ 且 } k < m \\ \sum_{j=1}^m p_d(x_j)p_{d'}(x_j) &= 0, \text{ 其中 } d \neq d' \end{aligned}$$

为方便起见, 我们把向量 $(p_d(x_1), \dots, p_d(x_m))^\top \in \mathbb{R}^m$ 记作 $p_d(\mathbf{x})$, 或者简记作 \mathbf{p}_d , 其中 $d = 0, 1, \dots, k, \mathbf{x} = (x_1, \dots, x_m)^\top$ 。上述条件简单地表示为

$$\begin{aligned} \|p_d(\mathbf{x})\|^2 &= c_d \neq 0, \text{ 其中 } d = 0, 1, \dots, k \text{ 且 } k < m \\ \langle p_d(\mathbf{x}), p_{d'}(\mathbf{x}) \rangle &= 0, \text{ 其中 } d \neq d' \end{aligned}$$

即, 多项式 p_0, p_1, \dots, p_k 把 \mathbf{x} 映射为正交向量 $p_0(\mathbf{x}), p_1(\mathbf{x}), \dots, p_k(\mathbf{x})$, 它们不必是单位长度的。我们把相应的单位正交向量记为 $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_k$, 即

$$\mathbf{u}_d = \frac{p_d(\mathbf{x})}{\|p_d(\mathbf{x})\|}, \text{ 其中 } d = 0, 1, \dots, k$$

定理 10.6 (G. E. Forsythe, 1957). 按照下面的方法构造的多项式是 x_1, \dots, x_m 上首系数为 1 的正交多项式 [87]。

$$\begin{aligned} p_0(\mathbf{x}) &= 1 \\ p_1(\mathbf{x}) &= x - \bar{x}, \text{ 其中 } \bar{x} = \frac{x_1 + \dots + x_m}{m} \\ p_{d+1}(\mathbf{x}) &= (x - \alpha_{d+1})p_d(\mathbf{x}) - \alpha'_d p_{d-1}(\mathbf{x}), \text{ 其中 } d = 1, \dots, k-1 \\ \alpha_{d+1} &= \frac{\langle \mathbf{x}, p_d^2(\mathbf{x}) \rangle}{\|p_d(\mathbf{x})\|^2} \text{ 且 } \alpha'_d = \frac{\|p_d(\mathbf{x})\|^2}{\|p_{d-1}(\mathbf{x})\|^2} \end{aligned}$$

证明. 我们只需验证对于任意的 $d \in \mathbb{N}$, 皆有

$$\langle p_{d+1}(\mathbf{x}), p_i(\mathbf{x}) \rangle = 0, \text{ 其中 } i = 0, 1, \dots, d$$

$$\text{或者, } \langle p_d(\mathbf{x}), \mathbf{x} * p_i(\mathbf{x}) \rangle - \alpha_{d+1} \langle p_d(\mathbf{x}), p_i(\mathbf{x}) \rangle - \alpha'_d \langle p_{d-1}(\mathbf{x}), p_i(\mathbf{x}) \rangle = 0 \quad (10.15)$$

显然, $d = 1$ 时上式是成立的。对 d 进行归纳验证,

□ 若 $i < d - 1$, 则 $x p_i(x)$ 可由 $p_0(x), \dots, p_{d-1}(x)$ 唯一线性表出, 由归纳假设, 式

(10.15) 左边三项皆为 0。

□ 若 $i = d - 1$, 则上式右边为

$$\begin{aligned}\langle p_d(\mathbf{x}), \mathbf{x} * p_{d-1}(\mathbf{x}) \rangle - \alpha'_d \|p_{d-1}(\mathbf{x})\|^2 &= \langle p_d(\mathbf{x}), \mathbf{x} * p_{d-1}(\mathbf{x}) \rangle - \|p_d(\mathbf{x})\|^2 \\ &= \langle p_d(\mathbf{x}), \mathbf{x} * p_{d-1}(\mathbf{x}) - p_d(\mathbf{x}) \rangle\end{aligned}$$

因为 $xp_{d-1}(x) - p_d(x)$ 的次数不超过 d , 所以式 (10.15) 左边一定为 0。

□ 若 $i = d$, 由 α_{d+1} 的定义, 式 (10.15) 左边为 $\langle \mathbf{x}, p_d^2(\mathbf{x}) \rangle - \alpha_{d+1} \|p_d(\mathbf{x})\|^2 = 0$ 。 □

推论 10.2. 由定理 10.6 的证明, 不难得到

$$\begin{aligned}\mathbf{p}_{d+1} &= \mathbf{x} * \mathbf{p}_d - \frac{\langle \mathbf{x}, \mathbf{p}_d^2 \rangle}{\|\mathbf{p}_d\|^2} \mathbf{p}_d - \frac{\|\mathbf{p}_d\|^2}{\|\mathbf{p}_{d-1}\|^2} \mathbf{p}_{d-1} \\ \|\mathbf{p}_{d+1}\|^2 &= \langle \mathbf{p}_{d+1}, \mathbf{x} * \mathbf{p}_d \rangle\end{aligned}\tag{10.16}$$

定义 10.10. 给定 $k + 1$ 个 x_1, \dots, x_m 上的正交多项式 p_0, \dots, p_k , 我们把下面的回归问题称为正交多项式回归 (orthogonal polynomial regression, OPR)。

$$y_j = \beta_0 p_0(x_j) + \dots + \beta_k p_k(x_j) + \epsilon_j, \text{ 其中 } \epsilon_1, \dots, \epsilon_m \stackrel{\text{iid}}{\sim} N(0, \sigma^2), j = 1, \dots, m$$

令 $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^\top$, 上式可简单地表示为

$$\begin{aligned}\mathbf{y} &= \beta_0 p_0(\mathbf{x}) + \beta_1 p_1(\mathbf{x}) + \dots + \beta_k p_k(\mathbf{x}) + \boldsymbol{\epsilon} \\ &= P\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ 其中 } P_{m \times (k+1)} = (\mathbf{1}, p_1(\mathbf{x}), \dots, p_k(\mathbf{x})), \boldsymbol{\epsilon} \sim N_m(\mathbf{0}, \sigma^2 I)\end{aligned}\tag{10.17}$$

有的时候, 我们把下面的回归问题也称为正交多项式回归。

$$\begin{aligned}\mathbf{y} &= \eta_0 + \eta_1 \mathbf{u}_1 + \dots + \eta_k \mathbf{u}_k + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N_m(\mathbf{0}, \sigma^2 I) \\ &= U\boldsymbol{\eta} + \boldsymbol{\epsilon}, \text{ 其中 } U_{m \times (k+1)} = (\mathbf{1}, \mathbf{u}_1, \dots, \mathbf{u}_k), \boldsymbol{\eta} = (\eta_0, \eta_1, \dots, \eta_k)^\top\end{aligned}\tag{10.18}$$

正交多项式回归具有多项式逼近和正交表示双重优势: 只要控制变量 \mathbf{x} 给好, 正交系 $p_0(\mathbf{x}), p_1(\mathbf{x}), \dots, p_k(\mathbf{x})$ 就是确定了的。由该正交系张成的空间的维数是 $k + 1$, $\mathbf{y} \in \mathbb{R}^m$ 在这个空间上的投影便是它的最佳近似。

定理 10.7. 正交多项式回归 (10.17) 中的回归系数 $\boldsymbol{\beta}$ 为

$$\beta_d = \frac{\langle \mathbf{y}, p_d(\mathbf{x}) \rangle}{\|p_d(\mathbf{x})\|^2}, \text{ 其中 } d = 0, 1, \dots, k\tag{10.19}$$

正交多项式回归 (10.18) 中的回归系数 η 为

$$\eta_d = \langle \mathbf{y}, \mathbf{u}_d \rangle = \|\mathbf{p}_d\| \beta_d, \text{ 其中 } d = 0, 1, \dots, k \quad (10.20)$$

证明. 首先, $P^\top P = \text{diag}(\|p_0(\mathbf{x})\|^2, \|p_1(\mathbf{x})\|^2, \dots, \|p_k(\mathbf{x})\|^2)$ 是正定矩阵, 所以

$$\begin{aligned} \boldsymbol{\beta} &= (P^\top P)^{-1} P^\top \mathbf{y} \\ &= \text{diag}(\|p_0(\mathbf{x})\|^{-2}, \dots, \|p_k(\mathbf{x})\|^{-2})(\langle \mathbf{y}, p_0(\mathbf{x}) \rangle, \dots, \langle \mathbf{y}, p_k(\mathbf{x}) \rangle)^\top \end{aligned}$$

结果 (10.20) 的证明是类似的, 留给读者补全。 \square

定理 10.7 的几何意义是很直观的: 向量 \mathbf{y} 在 \mathbf{p}_d 上的投影为

$$\text{proj}_{\mathbf{p}_d} \mathbf{y} = \frac{\langle \mathbf{y}, \mathbf{p}_d \rangle}{\|\mathbf{p}_d\|^2} \mathbf{p}_d = \langle \mathbf{y}, \mathbf{u}_d \rangle \mathbf{u}_d = \text{proj}_{\mathbf{u}_d} \mathbf{y}, \text{ 其中 } d = 0, 1, \dots, k$$

正交多项式回归 (10.17) 和 (10.18) 也可以表示为

$$\begin{aligned} \mathbf{y} &= \sum_{d=0}^k \text{proj}_{\mathbf{p}_d} \mathbf{y} + \boldsymbol{\epsilon} \\ &= \sum_{d=0}^k \text{proj}_{\mathbf{u}_d} \mathbf{y} + \boldsymbol{\epsilon} \end{aligned}$$

欲求正交多项式回归 (10.17) 的系数 $\boldsymbol{\beta}$, 不必通过**定理 10.6** 具体计算正交多项式。下面给出两个等价的算法, 绕过构造正交多项式, 直接计算系数 $\boldsymbol{\beta}$ 。

算法 10.3. 由递归式 (10.16) 和两个初始向量 $\mathbf{p}_0 = \mathbf{1}, \mathbf{p}_1 = \mathbf{x} - \bar{\mathbf{x}} = (x_1 - \bar{x}, \dots, x_m - \bar{x})^\top$, 可求得向量 $\mathbf{p}_2, \dots, \mathbf{p}_k$ 。进而,

- 由式 (10.19) 得到模型 (10.17) 的正交多项式回归系数 $\boldsymbol{\beta}$ 。
- 由式 (10.20) 得到模型 (10.18) 的正交多项式回归系数 η 。

算法 10.4. 模型 (10.18) 中, 正交多项式回归系数 η 可按下面的方法求得。

- 令 $\mathbf{z} \leftarrow \mathbf{x} - \bar{\mathbf{x}} = (x_1 - \bar{x}, \dots, x_m - \bar{x})^\top$ 。
- 构造 $Z_{m \times (k+1)} \leftarrow (\mathbf{1}, \mathbf{z}, \dots, \mathbf{z}^k)$, 设该 Vandermonde 矩阵的 QR 分解 (见第 766 页的**定理 E.2**) 如下。

$$Z_{m \times (k+1)} = Q_{m \times (k+1)} R_{(k+1) \times (k+1)}$$

其中, $Q = (\mathbf{q}_1, \dots, \mathbf{q}_{k+1})$ 是正交矩阵, R 是对角线元素为正数的上三角矩阵。

定义 $U_{m \times (k+1)} = (\mathbf{1}, \mathbf{q}_2, \dots, \mathbf{q}_{k+1})$, 则正交多项式回归系数为

$$\boldsymbol{\eta} = (U^\top U)^{-1} U^\top \mathbf{y}$$

证明. 我们只需证明在算法 10.4 中, $\mathbf{u}_d = \mathbf{q}_{d+1}$, 其中 $d = 0, 1, \dots, k$ 。事实上, 对于每个 $d = 0, 1, \dots, k$, 多项式 $(x - \bar{x})^d$ 存在唯一的线性表示

$$(x - \bar{x})^d = k_0 p_0(x) + \dots + k_{d-1} p_{d-1}(x) + p_d(x)$$

$$\text{进而, } z^d = (x - \bar{x})^d = \frac{k_0}{\|\mathbf{p}_0\|} \mathbf{u}_0 + \dots + \frac{k_{d-1}}{\|\mathbf{p}_{d-1}\|} \mathbf{u}_{d-1} + \frac{1}{\|\mathbf{p}_d\|} \mathbf{u}_d$$

于是, 矩阵 Z 可以表示为正交矩阵 $(\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_k)$ 与某个对角线元素为正数的上三角矩阵的乘积, 根据 QR 分解的唯一性, 我们有

$$(\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_k) = (\mathbf{q}_1, \dots, \mathbf{q}_{k+1}) \quad \square$$

现在, 我们考虑正交多项式回归 (10.17) 和多项式回归的关系。不妨设多项式回归模型如下,

$$y_j = \alpha_0 + \alpha_1 x_j + \dots + \alpha_k x_j^k + \epsilon_j, \text{ 其中 } j = 1, 2, \dots, m$$

或者, $\mathbf{y} = X\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, 其中 $X = (\mathbf{1}, \mathbf{x}, \dots, \mathbf{x}^k)$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_k)^\top$, $\boldsymbol{\epsilon} \sim N_m(\mathbf{0}, \sigma^2 I)$

值得注意的是, $\mathbf{1}, \mathbf{x}, \dots, \mathbf{x}^k$ 可能是线性相关的, 倘若如此, \mathbf{y} 在 $\text{span}(\mathbf{1}, \mathbf{x}, \dots, \mathbf{x}^k)$ 上的投影不能唯一地被 $\mathbf{1}, \mathbf{x}, \dots, \mathbf{x}^k$ 线性表出。而正交多项式回归则没有这个顾虑。不失一般性, 假设

$$p_d(x) = \sum_{i=0}^d f_{di} x^i, \text{ 其中 } d = 0, 1, \dots, k$$

将回归函数由正交多项式的表示变为多项式的表示, 即

$$\begin{aligned} \sum_{d=0}^k \beta_d p_d(x) &= \sum_{d=0}^k \beta_d \left(\sum_{i=0}^d f_{di} x^i \right) \\ &= \sum_{d=0}^k \left(\sum_{i=d}^k \beta_i f_{id} \right) x^d \end{aligned}$$

于是，我们得到多项式回归系数和正交多项式回归系数的关系如下。

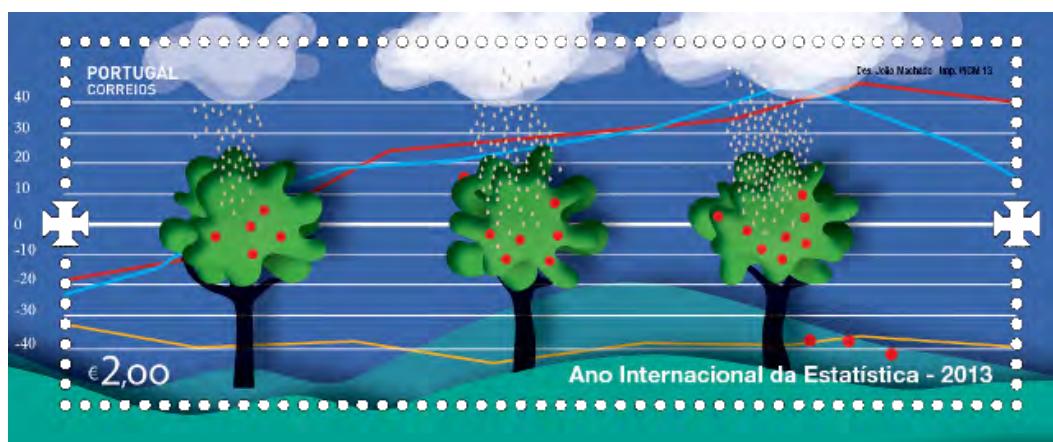
$$\alpha_d = \sum_{i=d}^k \beta_i f_{id}, \text{ 其中 } d = 0, 1, \dots, k$$

或者， $\boldsymbol{\alpha} = R\boldsymbol{\beta}$ ，其中 R 是一个上三角阵

10.2 方差分析模型

科 学研究、社会调查和生产实践中，经常需要通过试验来考察若干指定的因素（也称因子）对某一或某些指标的影响。影响指标的因素之间或制约或依存，它们的共同作用决定着指标值。譬如，

- 为研究家庭因素和学校因素对小学生成绩的影响，我们要考察同年龄段出自不同家庭环境和就学环境的小学生的学业情况。
- 几种药物对某疾病的疗效，如通过多种抗病毒药物联合使用来治疗艾滋病的鸡尾酒疗法。
- 考虑受教育的程度这单个因素对经济收入的影响，社会学家对受教育程度越高收入越少的现象感兴趣。
- 农业学家关心土壤、肥料、日照时间等因素对某农作物产量的影响，他们寻找最有利于该农作物生长的环境。
- 通过考察不同的饲料配方对家禽体重增长的效果，寻找优产、低耗、高质的饲养方法。如果几种饲料在增肥效果上有明显差异，要搞清楚这差异是饲料因素引起的还是试验误差造成的。



对每个因素，我们相应地设置几个水平。譬如，受教育的程度以最后毕业的学历可设为 7 个水平：无学历、小学、初中、高中、大学本科、硕士研究生、博士研究生。而不同的药物组合或者饲料配方就是不同的水平。再如，

例 10.10 ([137]). 把三个品牌的电池定为三个水平，以电池的寿命（单位：小时）为指标，下面是观察到的 $n = 15$ 个电池寿命的样本值，其中每个因素对应的样本数为 $n_1 = 5, n_2 = 4, n_3 = 6$ 。

	电池品牌		
	X_1	X_2	X_3
电	40	60	60
池	30	40	50
寿	50	55	70
命	50	65	65
	30		75
			40
均值	40	55	60

例 10.11 ([137]). 一般情况下, 教师和教学方法是影响学生成绩的两个重要因素。我们将教师和教学方法各自分为三个水平, 以学生的成绩为指标, 观察到三位教师分别使用三种不同教学方法导致学生取得如下的成绩。

教学 方法	教师		
	I	II	III
1	95	60	86
	85	90	77
	74	80	75
	74	70	70
2	90	89	83
	80	90	70
	92	91	75
	82	86	72
3	70	68	74
	80	73	86
	85	78	91
	85	93	89

本章将着重介绍全面试验的 Fisher 方差分析方法*, 包括单因素的和两因素的两种情况。顾名思义, 单因素方差分析 (one-way ANOVA, single-factor ANOVA) 在试验中仅考虑一个因素, 譬如例 10.10 中电池的品牌; 两因素方差分析 (two-way ANOVA, two-factor ANOVA) 在试验中考虑两个因素, 允许这两个因素存在相互作用, 譬如例 10.11 中的教师和教学方法。上述这些试验的目的无外乎以下几点:

- 通过数据分析找出显著影响指标的因素。

*Fisher 于 1918-1925 年间提出方差分析 (analysis of variance, ANOVA) 的理论。

- 对某个因素而言，哪个水平使得指标值最大或最小？
- 因为因素之间存在相互作用，所有因素以什么样的水平搭配才能使得指标最优？

如果我们把每个因素的所有水平都纳入考察范围，这样的试验被称作全面试验^{*}。对于更多因素的情况，为了避免组合爆炸导致的大工作量和高成本，我们不应简单地进行全面试验，而是要通过试验设计 (design of experiments) 的方法选取一些有代表性的水平组合来进行试验。试验设计是统计学的一个重要的分支，经历了方差分析、正交试验法、调优设计法三个发展阶段，它探究用何种方法经济地、科学地安排试验。

^{*} “可能影响结果的非可控原因永远是无穷多的。”

— R. A. Fisher

10.2.1 单因素方差分析

单因素方差分析的目的在于比较因素各水平上指标值的差别变化。我们把单因素分为 k 个水平，第 i 个水平上有 n_i 个观察样本 X_{i1}, \dots, X_{in_i} 。考虑下面的线性模型

$$X_{ij} = \mu_i + \epsilon_{ij}, \text{ 其中 } \epsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2), i = 1, 2, \dots, k \text{ 且 } j = 1, 2, \dots, n_i \quad (10.21)$$

显然，样本的密度函数为

$$f(\mathbf{x}; \boldsymbol{\mu}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right\} \quad (10.22)$$

参数最大似然估计的结果为

$$\begin{aligned} \hat{\mu}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} = \bar{X}_{i \cdot}, \text{ 其中 } i = 1, 2, \dots, k \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i \cdot})^2 \end{aligned}$$

假设共有 n 个观察样本，记作 $\mathbf{X} = (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}, \dots, X_{k1}, \dots, X_{kn_k})^\top$ ，其中 $\sum_{i=1}^k n_i = n$ 。线性模型 (10.21) 也可以简单地表示为

$$\mathbf{X} = A\boldsymbol{\mu} + \boldsymbol{\epsilon}, \text{ 其中 } \boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^\top$$

其中，误差向量 $\boldsymbol{\epsilon} = (\epsilon_{11}, \dots, \epsilon_{1n_1}, \epsilon_{21}, \dots, \epsilon_{2n_2}, \dots, \epsilon_{k1}, \dots, \epsilon_{kn_k})^\top$ ，矩阵 A 定义为

$$A_{n \times k} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_k} \end{pmatrix}^\top, \text{ 其中列向量 } \mathbf{1}_{n_1} = \overbrace{(1, 1, \dots, 1)}^{n_1 \uparrow}^\top$$

如果认为不同水平对指标的影响没有差异，零假设可设置为 $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ ，或者等价地用线性假设的形式表示为 $H_0 : H\boldsymbol{\mu} = \mathbf{0}$ ，其中矩阵 $H_{(k-1) \times k}$ 的秩为 $k-1$ ，具体为

$$H = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}$$

在零假设 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ 成立的情况下，设 $\mu_1 = \mu_2 = \dots = \mu_k = \mu$ ，则样本的密度函数 (10.22) 简化为

$$f(\mathbf{x}; \boldsymbol{\mu}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \mu)^2 \right\}$$

参数 μ, σ^2 的最大似然估计的结果为

$$\begin{aligned} \tilde{\mu} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij} = \bar{X} \\ \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 \end{aligned}$$

按照定理 10.5，我们构造统计量

$$F = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 - \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2} \quad (10.23)$$

■ 定义 10.11. 为了化简统计量 (10.23)，我们提出下面的概念。

□ 总偏差平方和 (total sum of squares, TSS):

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

□ 组内偏差平方和 (within sum of squares, WSS):

$$WSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$$

□ 组间偏差平方和 (between sum of squares, BSS):

$$BSS = \sum_{i=1}^k n_i (\bar{X}_{i\cdot} - \bar{X})^2$$

性质 10.3. 总偏差平方和可以分解为组内偏差平方和与组间偏差平方和之和，即

$$TSS = WSS + BSS$$

证明. 由于 $\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot}) = 0$, 所以 TSS 有如下的分解

$$\begin{aligned}\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot} + \bar{X}_{i\cdot} - \bar{X})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^k n_i (\bar{X}_{i\cdot} - \bar{X})^2\end{aligned}\quad \square$$

表 10.1: 各种偏差平方和的自由度。

来源	偏差平方和	自由度	平均偏差平方和
组间	$BSS = \sum_{i=1}^k n_i (\bar{X}_{i\cdot} - \bar{X})^2$	$k - 1$	$BSS/(k - 1)$
组内	$WSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$	$n - k$	$WSS/(n - k)$
总的	$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$	$n - 1$	$TSS/(n - 1)$

性质 10.4. 式 (10.23) 即为 $F = BSS/WSS$, 进一步由定理 10.5 可得

$$\frac{n-k}{k-1} F = \frac{BSS/(k-1)}{WSS/(n-k)} \sim F_{k-1, n-k} \quad (10.24)$$

上式中, $\frac{n-k}{k-1} F$ 被称为 F -比 (F -ratio), 它是组间与组内的平均偏差平方和之比。当 F -比 $\geq F_{k-1, n-k, 1-\alpha}$ 时, 在水平 α 下似然比检验拒绝零假设 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ 。

例 10.12. 接着例 10.10, 若零假设 H_0 是三个电池品牌之间没有差异。

$$\begin{aligned}\bar{x} &= \frac{200 + 220 + 360}{15} = 52, \quad \sum_{j=1}^5 (x_{1j} - \bar{x}_{1\cdot})^2 = 400 \\ &\quad \sum_{j=1}^4 (x_{2j} - \bar{x}_{2\cdot})^2 = 350, \quad \sum_{j=1}^6 (x_{3j} - \bar{x}_{3\cdot})^2 = 850\end{aligned}$$

不难求得组内偏差平方和为 $WSS = 400 + 350 + 850 = 1600$, 自由度为 12; 组间偏差平方和为 $BSS = 5(40 - 52)^2 + 4(55 - 52)^2 + 6(60 - 52)^2 = 1140$, 自由度为 2。

进而, 由 (10.24) 求得 F -比为 4.28。因为 F -比 $> F_{2, 12, 0.05} = 3.89$, 所以在水平 $\alpha = 0.05$ 拒绝零假设 $H_0: \mu_1 = \mu_2 = \mu_3$ 。

10.2.2 两因素方差分析

在实际应用中，人们经常会遇到一个指标被两个因素影响的情况，有时这两个因素之间还会有相互作用。譬如在例 10.11 中，教师和教学方法之间有着微妙的关系，有的教学方法有助于特定的教师改进教学效果而对其他教师适得其反。为了分析这两个因素谁对观察结果的影响大，以及两个因素之间有无相互作用，仿照例 10.11，不失一般性，我们把观察样本整理成下面的样子。

表 10.2: 观察数据按照两因素的不同水平分成若干个组。

因素 1 的水平	因素 2 的水平				均值
	1	2	...	b	
1	X_{111}	X_{121}	...	X_{1b1}	
:	:	:	:	:	:
1	X_{11m}	X_{12m}	...	X_{1bm}	$\bar{X}_{1..}$
:	:	:	:	:	:
a	X_{a11}	X_{a21}	...	X_{ab1}	
:	:	:	:	:	
a	X_{a1m}	X_{a2m}	...	X_{abm}	$\bar{X}_{a..}$
均值	$\bar{X}_{.1..}$	$\bar{X}_{.2..}$...	$\bar{X}_{.b..}$	\bar{X}

定义 10.12. 四种均值：样本均值、两因素第 ij 水平的均值、因素 1 第 i 水平的均值、因素 2 第 j 水平的均值分别定义如下。

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{s=1}^m X_{ijs} & \bar{X}_{ij.} &= \frac{1}{m} \sum_{s=1}^m X_{ijs} \\ \bar{X}_{i..} &= \frac{1}{mb} \sum_{j=1}^b \sum_{s=1}^m X_{ijs} & \bar{X}_{.j..} &= \frac{1}{ma} \sum_{i=1}^a \sum_{s=1}^m X_{ijs}\end{aligned}$$

定义 10.13. 四种偏差平方和：由因素 1 引起的偏差平方和、由因素 2 引起的偏差平方和、由相互作用引起的偏差平方和、由误差项引起的偏差平方和。

$$\begin{aligned}SS_1 &= bm \sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2 & SS_2 &= am \sum_{j=1}^b (\bar{X}_{.j..} - \bar{X})^2 \\ SSI &= m \sum_{i=1}^a \sum_{j=1}^b (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j..} + \bar{X})^2 & SSE &= \sum_{i=1}^a \sum_{j=1}^b \sum_{s=1}^m (\bar{X}_{ijs} - \bar{X}_{ij.})^2\end{aligned}$$

表 10.3: 各种偏差平方和的自由度、均方偏差和 F-比。

偏差来源	平方和	自由度	均方偏差	F-比
因素 1	SS ₁	$a - 1$	$MS_1 = SS_1/(a - 1)$	MS_1/MSE
因素 2	SS ₂	$b - 1$	$MS_2 = SS_2/(b - 1)$	MS_2/MSE
相互作用	SSI	$(a - 1)(b - 1)$	$MSI = SSI/[(a - 1)(b - 1)]$	MSI/MSE
误差项	SSE	$ab(m - 1)$	$MSE = SSE/[ab(m - 1)]$	

下面，我们考虑线性模型

$$X_{ijs} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijs}, \text{ 其中 } \epsilon_{ijs} \stackrel{iid}{\sim} N(0, \sigma^2) \quad (10.25)$$

并且 $i = 1, \dots, a, j = 1, \dots, b, s = 1, \dots, m$

$$\text{满足约束条件: } \sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = \sum_{i=1}^a \gamma_{ij} = \sum_{j=1}^b \gamma_{ij} = 0 \quad (10.26)$$

在模型 (10.25) 中, γ_{ij} 是 α_i 和 β_j 相互作用的结果。为什么会有约束 (10.26) 呢? 假设 $\mu, \alpha_i, \beta_j, \gamma_{ij}$ 不满足 (10.26), 可以构造 $\mu', \alpha'_i, \beta'_j, \gamma'_{ij}$ 如下, 使之满足 (10.26)。

$$\begin{aligned} \mu' &= \mu + \bar{\alpha} + \bar{\beta} + \bar{\gamma} \\ \alpha'_i &= \alpha_i - \bar{\alpha} + \bar{\gamma}_i - \bar{\gamma} \\ \beta'_j &= \beta_j - \bar{\beta} + \bar{\gamma}_{.j} - \bar{\gamma} \\ \gamma'_{ij} &= \gamma_{ij} - \bar{\gamma}_i - \bar{\gamma}_{.j} + \bar{\gamma} \end{aligned}$$

$$\text{其中, } \bar{\alpha} = \frac{1}{a} \sum_{i=1}^a \alpha_i, \bar{\beta} = \frac{1}{b} \sum_{j=1}^b \beta_j, \text{ 并且}$$

$$\bar{\gamma}_i = \frac{1}{b} \sum_{j=1}^b \gamma_{ij}, \bar{\gamma}_{.j} = \frac{1}{a} \sum_{i=1}^a \gamma_{ij}, \bar{\gamma} = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b \gamma_{ij}$$

练习 10.3. 请读者验证上述 $\mu', \alpha'_i, \beta'_j, \gamma'_{ij}$ 满足 $\mu' + \alpha'_i + \beta'_j + \gamma'_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ 和约束条件 (10.26)。提示:

$$\begin{aligned} \sum_{i=1}^a \alpha'_i &= \sum_{i=1}^a \alpha_i - a\bar{\alpha} + \sum_{i=1}^a \bar{\gamma}_i - a\bar{\gamma} = 0 \\ \sum_{i=1}^a \gamma'_{ij} &= \sum_{i=1}^a \gamma_{ij} - \sum_{i=1}^a \bar{\gamma}_i - a\bar{\gamma}_{.j} + a\bar{\gamma} = 0 \end{aligned}$$

性质 10.5. 我们经常考虑以下两种类型的零假设: 一种是针对某个因素; 另一种是针对两因素的相互作用。

① 对模型 (10.25), 我们考虑零假设 $H_\alpha : \alpha_1 = \dots = \alpha_a = 0$, 即因素 1 不影响指标。这是一个线性假设, 其中 $n = abm, k = ab, r = a - 1, n - k = ab(m - 1)$ 。参数的最大似然估计如下:

$$\begin{aligned}\hat{\mu} &= \tilde{\mu} = \bar{X}, \quad \hat{\alpha}_i = \bar{X}_{i..} - \bar{X}, \quad \hat{\beta}_j = \tilde{\beta}_j = \bar{X}_{.j} - \bar{X} \\ \hat{\gamma}_{ij} &= \tilde{\gamma}_{ij} = \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j} + \bar{X}\end{aligned}$$

构造统计量如下

$$\begin{aligned}F &= \frac{\sum_{i,j,s} (X_{ijs} - \bar{X}_{ij.} + \bar{X}_{i..} - \bar{X})^2 - \sum_{i,j,s} (X_{ijs} - \bar{X}_{ij.})^2}{\sum_{i,j,s} (X_{ijs} - \bar{X}_{ij.})^2} \\ &= \frac{bm \sum_i (\bar{X}_{i..} - \bar{X})^2}{\sum_{i,j,s} (X_{ijs} - \bar{X}_{ij.})^2} = \frac{\text{SS}_1}{\text{SSE}}\end{aligned}$$

由式 (10.24) 可得 F -比的分布如下,

$$\frac{ab(m-1)}{a-1} F = \frac{\text{MS}_1}{\text{MSE}} \sim F_{a-1, ab(m-1)}$$

② 对模型 (10.25), 我们考虑零假设 $H_\gamma : \gamma_{ij} = 0, \forall i, j$, 即两因素相互独立, 没有相互作用。这种情况下, $n = abm, k = ab, r = (a-1)(b-1), n - k = ab(m-1)$ 。参数的最大似然估计分别是:

$$\tilde{\mu} = \bar{X}, \quad \tilde{\alpha}_i = \bar{X}_{i..} - \bar{X}, \quad \tilde{\beta}_j = \bar{X}_{.j} - \bar{X}$$

构造统计量如下

$$\begin{aligned}F &= \frac{\sum_{i,j,s} (X_{ijs} - \bar{X}_{i..} - \bar{X}_{.j} + \bar{X})^2 - \sum_{i,j,s} (X_{ijs} - \bar{X}_{ij.})^2}{\sum_{i,j,s} (X_{ijs} - \bar{X}_{ij.})^2} \\ &= \frac{m \sum_{i,j,s} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j} + \bar{X})^2}{\sum_{i,j,s} (X_{ijs} - \bar{X}_{ij.})^2} = \frac{\text{SSI}}{\text{SSE}}\end{aligned}$$

由式 (10.24) 可得 F -比的分布如下,

$$\frac{ab(m-1)}{(a-1)(b-1)}F = \frac{\text{MSI}}{\text{MSE}} \sim F_{a-1, ab(m-1)}$$

例 10.13. 接着**例 10.11**, 求得几类均值如下。

	$\bar{X}_{ij.}$			$\bar{X}_{i..}$
82	75	77	78.0	
86	89	75	83.3	
80	78	85	81.0	
$\bar{X}_{.j}$	82.7	80.7	79.0	$\bar{X} = 80.8$

求得四种偏差平方和如下。

$$SS_1 = bm \sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2 = 3 \times 4 \times 14.13 = 169.56$$

$$SS_2 = am \sum_{j=1}^b (\bar{X}_{.j} - \bar{X})^2 = 82.32$$

$$SSI = 561.80$$

$$SSE = 1830.00$$

偏差来源	偏差平方和	自由度	均方偏差	F -比
方法	169.56	2	84.78	1.25
教师	82.32	2	41.16	0.61
相互作用	561.80	4	140.45	2.07
误差项	1830.00	27	67.78	

给定显著水平 $\alpha = 0.05$, 由 $F_{2,27,0.05} = 3.35, F_{4,27,0.05} = 2.73$, 我们不能拒绝三种方法是等效的, 也不能拒绝三位教师是等效的, 也不能拒绝两因素没有相互作用。

10.3 习题

- 10.1. 已知线性模型 $\begin{cases} X_1 = \beta_1 + \epsilon_1 \\ X_2 = 2\beta_1 - \beta_2 + \epsilon_2 \\ X_3 = \beta_1 + 2\beta_2 + \epsilon_3 \end{cases}$ 中 $\epsilon_1, \epsilon_2, \epsilon_3 \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ 。

试求：(1) 计算 β_1 和 β_2 的最小二乘估计 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ ；(2) 分别给出 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的分布并证明 $\hat{\beta}_1, \hat{\beta}_2$ 相互独立。

- 10.2. 在定理 10.3 的条件下，试证明：

$$\frac{\|X - A\hat{\beta}\|^2}{\sigma^2} \sim \chi_n^2$$

- 10.3. 设有一元线性回归模型 $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, \dots, n$, 其中 $\epsilon_i \sim N(0, \sigma^2)$, 且彼此独立, 试导出检验假设 $H_0 : \beta_0 = 0 \leftrightarrow H_1 : \beta_0 \neq 0$ 的检验统计量, 指出在原假设成立时该统计量的分布, 并对检验水平 $\alpha (0 < \alpha < 1)$ 给出此检验法的拒绝域。
- 10.4. 将抗生素注入人体会产生抗生素和血浆蛋白质结合的现象, 以致减少了药效。下表列出了 5 种常用的抗生素注入到牛的体内时, 抗生素与血浆蛋白质结合的百分比。试在水平 $\alpha = 0.05$ 下检验这些百分比的均值有无显著差异。

青霉素	四环素	链霉素	红霉素	氯霉素
29.6	27.3	5.8	21.6	29.2
24.3	32.6	6.2	17.4	32.8
28.5	30.8	11.0	18.3	25.0
32.0	34.8	8.3	19.0	24.2

- 10.5. 已知线性模型 $y_i = a + bx_i + \epsilon_i$, $i = 1, 2, \dots, n$, 其中 $\epsilon_1, \dots, \epsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ 。设 \hat{a} 和 \hat{b} 是回归系数 a 和 b 的最小二乘估计, 试求: (1) a 的 $1 - \alpha$ 置信区间; (2) b 的 $1 - \alpha$ 置信区间; (3) σ^2 的 $1 - \alpha$ 置信区间。

- 10.6. 下表列出在不同重量下 6 跟弹簧的长度:

重量 x (g)	5	10	15	20	25	30
长度 y (cm)	7.25	8.12	8.95	9.90	10.9	11.8

- (1) 求出回归方程; (2) 试在 $x = 16$ 时作出 Y 的 95% 的预测区间。

第十一章

时间序列分析

古今多少事，都付笑谈中。

杨慎《临江仙》

以随机过程为数学基础，时间序列分析是现代统计学的一个重要分支，研究的是以时间排序的数据中所包含的统计规律。时间序列分析和经典统计学是两个截然不同的工具，前者被视为离散指标的随机过程的统计学，关心的是序列数据内在的依赖关系；后者的研究对象是基于独立性假设的简单样本，没有把时间变量纳入研究框架。

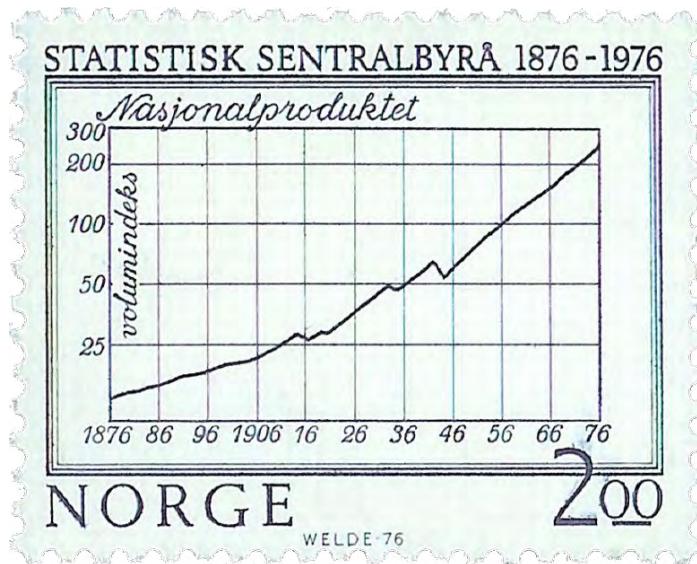


图 11.1: 挪威 1876-1976 百年间国内生产总值的变化，两次世界大战期间有所下降。

时间序列有时候指随机变量的序列 $\{X_1, X_2, \dots\}$ ，有时候指按时间顺序取得的具体观察值的序列 $\{x_1, x_2, \dots\}$ （即随机过程所生成的无限总体中的一个样本实现）。时

间序列的本质特征是相邻观测值之间的依赖性，例如工厂每天的产量，医院每个月治愈某种疾病的人数，……。时间序列分析的对象就是这种依赖性，一旦我们了解了它就可以利用时间序列的历史数据对它的未来取值进行预测，或者根据输入输出的时间序列来研究动态系统和控制模型。考虑到未知因素的影响，我们不可能用一个确定模型去实现最优预测和控制，而是要用概率模型（或随机模型）。对于探索经济、商业、工程、气象……中以时间顺序出现的数据的内在规律，时间序列分析是一个基本工具。

当我们得到了 $\{X_1, X_2, \dots, X_t\}$ 的观察值，如何预测 X_{t+1} 的结果？预测值 \hat{X}_{t+1} 是一个有关 $\{X_1, X_2, \dots, X_t\}$ 的函数，由第 176 页的练习 2.26，为了使均方误差 $E(X_{t+1} - \hat{X}_{t+1})^2$ 最小， \hat{X}_{t+1} 应是条件期望 $E(X_{t+1}|X_1, \dots, X_t)$ ，称为最佳预测。然而，条件分布 $f(x_{t+1}|x_1, \dots, x_t)$ 一般是未知的，也很难从观察数据中估计出来。

事实上， X_{t+1} 和 $\{X_1, X_2, \dots, X_t\}$ 之间关系可以很复杂。作为研究对象，我们不妨从最简单的入手：假设存在线性关系 $\hat{X}_{t+1} = a_0 + a_1 X_1 + \dots + a_t X_t$ ，其中系数 a_0, a_1, \dots, a_t 使得均方误差 $E(X_{t+1} - \hat{X}_{t+1})^2$ 最小，该结果被称为最佳线性预测。这类模型的优点是简单，结果仅依赖于 EX_i 和 $E(X_i X_j), i, j = 1, 2, \dots$ 。

定义 11.1. 复数域上时间序列 $\{X_t : t \in \mathbb{Z}\}$ 的自协方差函数 (autocovariance function, ACVF) 和自相关函数 (autocorrelation function) 分别定义为

$$\begin{aligned}\gamma_X(t, s) &= E[(X_t - EX_t)(\overline{X_s} - \overline{EX_s})] \\ &= E(X_t \overline{X_s}) - EX_t \overline{EX_s} \\ \rho_X(t, s) &= \frac{\gamma_X(t, s)}{\sigma_X(t)\sigma_X(s)}\end{aligned}$$

定义 11.2 (平稳时间序列). 时间序列 $\{X_t : t \in \mathbb{Z}\}$ 如果满足下面的条件，则称之为弱平稳 (weakly stationary) 时间序列，简称平稳时间序列。

- $E(X_t) = \mu$ ，其中 μ 为常数。
- $\forall t \in \mathbb{Z}$ ，皆有 $V(X_t) < \infty$ 。
- $\forall r, s, t \in \mathbb{Z}$ ，皆有 $\gamma_X(r, s) = \gamma_X(r + t, s + t)$ 。

若对于任意时间点 $t_1, t_2, \dots, t_k \in \mathbb{Z}$ ，随机向量 $(X_{t_1}, X_{t_2}, \dots, X_{t_k})^\top$ 的分布与 $(X_{t_1+h}, X_{t_2+h}, \dots, X_{t_k+h})^\top$ 的相同，则称该时间序列是强平稳的 (strongly stationary)。

性质 11.1. 对于一个平稳时间序列 $\{X_t\}$ ，方差函数 $\sigma_X^2(t) = \gamma_X(t, t) = \gamma_X(0, 0)$ 与 t 无关，不妨记作 σ^2 。自协方差函数和自相关函数也与 t 无关，即

$$\begin{aligned}\gamma(h) &= E(X_{t+h} \overline{X_t}) - EX_{t+h} \overline{EX_t} = \gamma_X(h, 0) \\ \rho(h) &= \frac{\gamma(h)}{\sigma^2}\end{aligned}$$

性质 11.2. 自协方差函数 $\gamma(h)$ 满足下面的性质:

$$\begin{aligned}\gamma(0) &= \sigma^2 \\ \gamma(-h) &= \gamma(h) \\ |\gamma(h)| &\leq \gamma(0), \forall h\end{aligned}$$

证明. 令 $\tau = t - h$, 利用自协方差函数不依赖于时间点, 我们有

$$\begin{aligned}\gamma(-h) &= E[(X_t - \mu)(X_{t-h} - \mu)] \\ &= E[(X_{t-h} - \mu)(X_t - \mu)] \\ &= E[(X_\tau - \mu)(X_{\tau+h} - \mu)] \\ &= \gamma(h)\end{aligned}$$

利用 Cauchy-Schwarz 不等式, 我们有

$$|\gamma(h)| \leq E|(X_t - \mu)(X_{t+h} - \mu)| \leq \sigma^2 = \gamma(0)$$

□

练习 11.1. 请证明平稳时间序列的自相关函数满足以下性质:

$$\begin{aligned}\rho(0) &= 1 \\ \rho(-h) &= \rho(h) \\ |\rho(h)| &\leq 1, \forall h\end{aligned}$$

性质 11.3. 自协方差函数 $\gamma(h)$ 是半正定的, 即对于任意时间点 t_1, t_2, \dots, t_k 和任意实数 b_1, b_2, \dots, b_k 皆有

$$\sum_{i,j=1}^k \gamma(t_i - t_j) b_i g_j \geq 0$$

证明. 定义 $Y = b_1 X_{t_1} + \dots + b_k X_{t_k}$, 显然 $V(Y) \geq 0$, 即

$$\begin{aligned}V(Y) &= E \left[\sum_{j=1}^k b_j (X_{t_j} - \mu) \right]^2 \\ &= \sum_{i,j=1}^k \gamma(t_i - t_j) b_i b_j \geq 0\end{aligned}$$

□

例 11.1. 独立同分布的噪声时间序列 $\{Z_t\}$ 若满足 $EZ_t = \mu$ 和 $EZ_t^2 = \sigma^2 < +\infty$, 则

$\{Z_t\}$ 是平稳的, 记作 $\{Z_t\} \sim \text{IID}(\mu, \sigma^2)$ 。事实上, 根据独立性假设, 我们有

$$\gamma(h) = \begin{cases} \sigma^2 & \text{若 } h = 0 \\ 0 & \text{其他} \end{cases}$$

特别地, 噪声 $\{Z_t\} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ 是平稳的, 被称为白噪声 (white noise), 记作 $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ 。

定义 11.3 (线性时间序列). 一个平稳时间序列 $\{X_t\}$ 如果有如下的表示, 则称之为线性时间序列 (linear time series)。

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad \text{其中 } \{Z_t\} \sim \text{IID}(0, \sigma^2)$$

定义 11.4. 时间序列 $\{W_t : t = 0, 1, 2, \dots\}$ 被称为随机游动 (random walk), 如果 $W_0 = 0$ 且 W_t 是由 t 个独立同分布的随机变量叠加而成, 即

$$W_t = X_1 + X_2 + \dots + X_t, \quad \text{其中 } t = 1, 2, \dots \quad (11.1)$$

例 11.2. 令噪声时间序列 $\{X_t\} \sim \text{IID}(0, \sigma^2)$, 则随机游动 (11.1) 满足 $EW_t = 0, EW_t^2 = t\sigma^2 < +\infty$, 但 $\{W_t\}$ 不是平稳的, 因为协方差函数 $\gamma_W(t+h, t)$ 与 t 有关。

$$\begin{aligned} \gamma_W(t+h, t) &= \text{Cov}(W_{t+h}, W_t) \\ &= \text{Cov}(W_t + X_{t+1} + \dots + X_{t+h}, W_t) \\ &= \text{Cov}(W_t, W_t) \\ &= t\sigma^2 \end{aligned}$$

定理 11.1. 分别记 $\gamma(j), \rho(j)$ 为 γ_j, ρ_j , 其中 $j = 0, 1, 2, \dots, k$ 。如果 $\gamma_0 > 0$ 且 $\lim_{j \rightarrow \infty} \gamma_j = 0$, 则 $\forall k > 0$ 下面的两个矩阵 Γ_k 和 R_k 都是正定的。

$$\Gamma_k = \begin{pmatrix} \gamma_0 & c_s & \cdots & \gamma_k \\ c_s & \gamma_0 & \cdots & \gamma_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_k & \gamma_{k-1} & \cdots & \gamma_0 \end{pmatrix} \quad R_k = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_k \\ \rho_1 & 1 & \cdots & \rho_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k & \rho_{k-1} & \cdots & 1 \end{pmatrix}$$

定义 11.5 (谱密度). 如果 $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, 下面的函数称为时间序列 $\{X_t : t \in \mathbb{Z}\}$

的谱密度 (spectral density)。

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h), \text{ 其中 } -\pi \leq \lambda \leq \pi \quad (11.2)$$

显然, ACVF 和谱密度有下面的关系:

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda \quad (11.3)$$

练习 11.2. 白噪声 $\text{WN}(\mu, \sigma^2)$ 的谱密度是常数 $f(\lambda) = \sigma^2/(2\pi)$ 。

定义 11.6. 如果随机过程 $\{X_t : t \in \mathbb{Z}\}$ 能表示成白噪声的线性组合, 则称之为线性过程 (linear process)。即,

$$X_t = \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k}, \text{ 其中 } \{Z_t\} \sim \text{WN}(0, \sigma^2) \text{ 且 } \sum_{k=-\infty}^{\infty} |\psi_k| < \infty \quad (11.4)$$

练习 11.3. 线性过程 (11.4) 的自协方差函数是

$$\gamma(h) = \sigma^2 \sum_{k=-\infty}^{\infty} \psi_k \psi_{k+h}$$

11.1 ARMA 模型

自 回归滑动平均 (autoregressive moving average, ARMA) 过程常用于研究随季节变化的价格、销售量等波动规律，我们可以把它视作自回归 (autoregressive, AR) 过程和滑动平均 (moving average, MA) 过程的混合体，或者 AR 过程和 MA 过程是 ARMA 过程的特例。

定义 11.7. 一个平稳过程 $\{X_t : t \in \mathbb{Z}\}$ 称为一个 ARMA(p, q) 过程，如果它满足

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad (11.5)$$

其中， $\{Z_t\} \sim WN(0, \sigma^2)$

特别地，ARMA($p, 0$) 过程称为 AR(p) 过程，ARMA($0, q$) 过程称为 MA(q) 过程。

定义 11.8. 对于过程 $\{X_t\}$ ，滞后算子 (lag operator) L 和 k 阶后移算子 L^k 分别定义为

$$\begin{aligned} LX_t &= X_{t-1} \\ L^k X_t &= X_{t-k}, \text{ 其中 } k = 1, 2, \dots \end{aligned}$$

利用后移算子，式 (11.5) 可以简化为

$$\varphi(L)X_t = \theta(L)Z_t \quad (11.6)$$

其中， $\{Z_t\} \sim WN(0, \sigma^2)$

$$\varphi(z) = 1 - \varphi_1 z - \cdots - \varphi_p z^p \quad (11.7)$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \quad (11.8)$$

定义 11.9. 对于 ARMA(p, q) 过程 (11.5)，如果存在常数 $\psi_k, k = 0, 1, 2, \dots$ 满足下面两个条件，则称该 ARMA(p, q) 过程是因果的 (causal)。

$$\sum_{k=0}^{\infty} |\psi_k| < \infty$$

$$X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k}, \text{ 其中 } t \in \mathbb{Z} \quad (11.9)$$

定理 11.2. 考虑 ARMA(p, q) 过程 (11.6)，如果多项式 (11.7) 和 (11.8) 没有公共零点，则该 ARMA(p, q) 过程是因果的当且仅当对于所有的 $|z| \leq 1$ 皆有 $\varphi(z) \neq 0$ 。另

外, 式 (11.9) 中的常数 $\psi_k, k = 0, 1, 2, \dots$ 就是下述多项式 $\psi(z)$ 的系数。

$$\psi(z) = \frac{\theta(z)}{\varphi(z)} = \sum_{k=0}^{\infty} \psi_k z^k, \text{ 其中 } |z| \leq 1$$

定义 11.10. 对于 ARMA(p, q) 过程 (11.5), 如果存在常数 $\pi_k, k = 0, 1, 2, \dots$ 满足下面两个条件, 则称该 ARMA(p, q) 过程是可逆的 (invertible)。

$$\begin{aligned} & \sum_{k=0}^{\infty} |\pi_k| < \infty \\ & Z_t = \sum_{k=0}^{\infty} \pi_k X_{t-k}, \text{ 其中 } t \in \mathbb{Z} \end{aligned} \quad (11.10)$$

定理 11.3. 考虑 ARMA(p, q) 过程 (11.6), 如果多项式 (11.7) 和 (11.8) 没有公共零点, 则该 ARMA(p, q) 过程是可逆的当且仅当对于所有的 $|z| \leq 1$ 皆有 $\theta(z) \neq 0$ 。另外, 式 (11.10) 中的常数 $\pi_k, k = 0, 1, 2, \dots$ 就是下述多项式 $\pi(z)$ 的系数。

$$\pi(z) = \frac{\varphi(z)}{\theta(z)} = \sum_{k=0}^{\infty} \pi_k z^k, \text{ 其中 } |z| \leq 1$$

11.1.1 趋势性和季节性

假设时间序列数据由以下随机过程产生,

$$X_t = m_t + s_t + Y_t, \text{ 满足 } EY_t = 0, s_{t+d} = s_t, \sum_{j=1}^d s_j = 0, t = 1, 2, \dots, n \quad (11.11)$$

其中, m_t 是一个缓慢变化的函数, 称作趋势分量 (trend component); s_t 是一个周期函数 (不妨设其周期为 d), 称作季节分量 (seasonal component); Y_t 是白噪声。

(11.11) 被称为经典分解模型 (classical decomposition model), 它的一个简化版本就是下面的非季节模型 (nonseasonal model)。

$$X_t = m_t + Y_t, \text{ 满足 } EY_t = 0, t = 1, 2, \dots, n \quad (11.12)$$

令 h 是一个非负整数, 则式 (11.12) 所定义的过程 $\{X_t\}$ 的双侧滑动平均 (two-sided moving average) 定义为

$$\begin{aligned} W_t &= \frac{1}{2h+1} \sum_{j=-h}^h X_{t-j} \\ &= \frac{1}{2h+1} \sum_{j=-h}^h m_{t-j} + \frac{1}{2h+1} \sum_{j=-h}^h Y_{t-j}, \text{ 其中 } h+1 \leq t \leq n-h \end{aligned}$$

上式中, 假设误差项在时间区间 $[t-h, t+h]$ 里的均值接近零, 假设 m_t 在该时间区间里近似线性, 则对趋势分量 m_t 的估计是

$$\hat{m}_t = W_t = \frac{1}{2h+1} \sum_{j=-h}^h X_{t-j}, \text{ 其中 } h+1 \leq t \leq n-h \quad (11.13)$$

如果 $m_t = \sum_{j=0}^k c_j t^j$, 用算子 ∇^k 作用于 (11.12) 两侧, 不难得到一个均值为 $k!c_k$ 的平稳过程。事实上,

$$\nabla^k X_t = k!c_k + \nabla^k Y_t$$

对于经典分解模型 (11.11), 季节分量的周期 d 若为奇数, 即 $d = 2h+1$, 则趋势分量的估计为 (11.13)。否则, 设 $d = 2h$, 趋势分量的估计为

$$\hat{m}_t = \frac{1}{d} \left(\frac{X_{t-h}}{2} + \sum_{j=1-h}^{h-1} X_{t-j} + \frac{X_{t+h}}{2} \right), \text{ 其中 } h < t \leq n-h$$

11.2 预测

已知随机变量 X_1, X_2, \dots, X_n 和 Y 具有有限期望和方差: $EX_1 = \mu_1, EX_2 = \mu_2, \dots, EX_n = \mu_n, EY = \mu$, 随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的协方差矩阵记作 $\Gamma = (\gamma_{ij})_{n \times n}$, \mathbf{X} 与 Y 的协方差向量定义为

$$\boldsymbol{\gamma} = \begin{pmatrix} c_s \\ \vdots \\ \gamma_n \end{pmatrix} = \begin{pmatrix} \text{Cov}(X_1, Y) \\ \vdots \\ \text{Cov}(X_n, Y) \end{pmatrix}$$

在所有形如 $\hat{Y} = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$ 的随机变量中, 作为 Y 的最佳近似, 未知参数 $\beta_0, \beta_1, \dots, \beta_n$ 要使得均方误差 $E(\hat{Y} - Y)^2$ 最小。为简化计算, 我们考虑如下随机变量

$$\hat{Y} = \mu + a_1(X_1 - \mu_1) + \dots + a_n(X_n - \mu_n) \quad (11.14)$$

显然, $E\hat{Y} = \mu$ 。为了最小化 $f(a_1, \dots, a_n) = E(\hat{Y} - Y)^2$, 只需要

$$\frac{\partial f}{\partial a_j} = 2E[(\hat{Y} - Y)X_j] = 0, \text{ 其中 } j = 1, \dots, n$$

换句话说, 只需要保证 $\text{Cov}(\hat{Y} - Y, X_j) = 0, j = 1, \dots, n$ 即可最小化 $E(\hat{Y} - Y)^2$ 。

练习 11.4. 若 Γ 可逆, (11.14) 作为 Y 的最佳线性预测, 未知参数 $(a_1, \dots, a_n)^\top = \Gamma^{-1}\boldsymbol{\gamma}$ 。

定理 11.4. 若平稳时间序列 $\{X_t\}$ 满足 $EX_t = 0, \gamma(0) > 0$ 且 $\lim_{h \rightarrow \infty} \gamma(h) = 0$, 则对于 X_{n+1} 的线性预测 $\hat{X}_{n+1} = a_1 X_n + \dots + a_n X_1$,

- ① 参数为 $(a_1, \dots, a_n)^\top = \Gamma^{-1}\boldsymbol{\gamma}$, 其中协方差矩阵 $\Gamma = (\gamma_{ij})_{n \times n}$ 中 $\gamma_{ij} = \gamma(i-j)$, 并且 $\boldsymbol{\gamma} = (\gamma(1), \dots, \gamma(n))^\top$ 。
- ② 均方误差 $E(\hat{X}_{n+1} - X_{n+1})^2 = \gamma(0) - \boldsymbol{\gamma}^\top \Gamma^{-1} \boldsymbol{\gamma}$ 。

11.2.1 ARMA 模型的预测

11.3 参数估计

考虑平稳时间序列 $\{X_t\}$, 设 $E(X_t) = \mu$ 未知, 它的无偏估计是

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$$

定理 11.5. 已知平稳时间序列 $\{X_t\}$ 有均值 μ 和自协方差函数 $\gamma(\cdot)$, 当 $n \rightarrow \infty$ 时,

- ① 若 $\gamma(n) \rightarrow 0$, 则

$$V(\bar{X}_n) = E(\bar{X}_n - \mu)^2 \rightarrow 0$$

- ② 若 $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, 则

$$nV(\bar{X}_n) \rightarrow \sum_{h=-\infty}^{\infty} \gamma(h) = 2\pi f(0)$$

11.3.1 谱密度的估计

11.3.2 ARMA 模型的估计

11.4 习题

11.1. 建设中……

第十二章

统计决策理论与贝叶斯分析概要

千江有水千江月，万里无云万里天。

雷庵正受《嘉泰普灯录》

决策问题如果受随机因素或缺失信息的影响，则需要通过统计决策理论 (statistical decision theory) 寻求最优的非确定性的结果。该理论的历史可追溯到 Pascal 在《思想录》中提出的期望损失的概念（见本书第 169 页的例 2.52）。对任何一个统计问题，如果解题者要为他们的错误结论付出相应的代价，错误越大代价越大，理性的选择应该是使得期望损失最小的决策。

1939 年，罗马尼亚裔美国统计学家 Abraham Wald (1902-1950) 撰文指出参数估计和假设检验都是统计决策问题，重新激发起人们对决策理论的兴趣。Wald 提出了损失函数 (loss function)、风险函数 (risk function)、可容决策规则 (admissible decision rule)、Bayes 风险原则 (Bayes risk principle)、极小极大原则 (minmax principle) 等重要概念和方法 [156]，他是现代统计决策理论的奠基人。



损失函数是统计决策的起点，它反映了专家知识和特定需求。给定了损失函数，贝叶斯学派始终如一地选择期望损失最小的决策，有或没有观测数据的时候都是如此。频率派则需要预先制定决策规则，基于损失函数和样本定义一个风险函数，然后根据某些原则（如极大极小原则、Bayes 风险原则等）来选择最优的决策。

从操作流程看，贝叶斯学派的期望损失原则显得更简洁。在无数据的情况下，频率派的 Bayes 风险原则与贝叶斯学派的期望损失原则取得同样的结果。

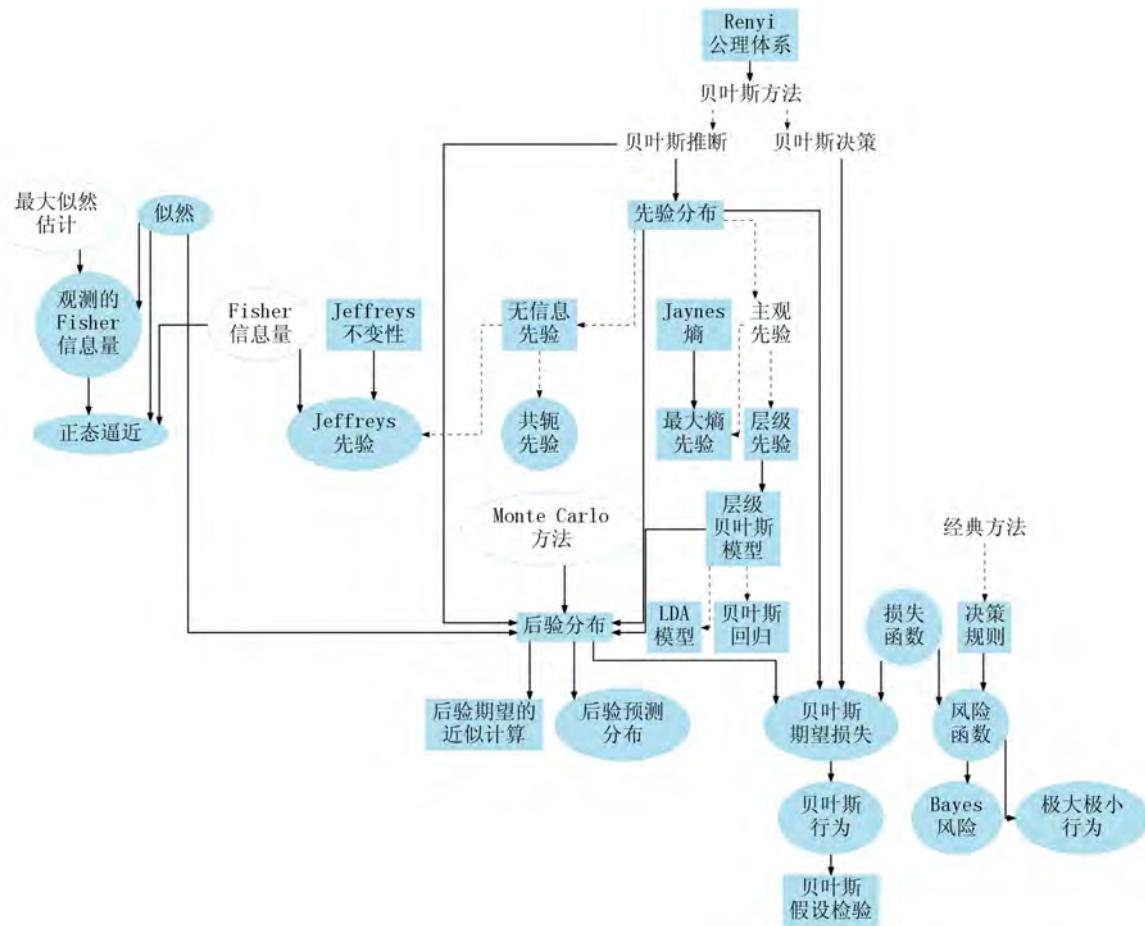
贝叶斯分析与统计决策理论有着千丝万缕的联系。例如，在没有观察样本的情况下，人们依然可以根据先验知识或以往的经验给出统计决策的结果（见例 2.52），

而贝叶斯学派恰恰主张把先验知识纳入到统计推断中，在这一点上贝叶斯分析与统计决策理论不谋而合。

经过英国数学家 Frank Plumpton Ramsey (1903-1930), 意大利概率统计学家 Bruno de Finetti (1906-1985), 美国数学家与统计学家 Leonard Jimmie Savage (1917-1971) 等贝叶斯学者的努力, 贝叶斯决策理论已经发展成为统计决策理论中的一个活跃的分支 [9], 其经典著作首推美国统计学家 Morris Herman DeGroot (1931-1989) 的《最优统计决策》(Optimal Statistical Decisions, 1970)。

有一个区分频率派和贝叶斯学派的“试金石”，就是它们对未知参数的理解。频率派认为未知参数是未知的固定值，而贝叶斯学派则坚信未知参数是随机变量。在不同信仰基础上发展起来的两类统计学无所谓谁对谁错，实用主义的作法是看具体问题需用哪种方法来解决更合适些就用哪种方法。

第十二章的主要内容及其关系



12.1 统计决策理论中的基本概念

怎样的统计决策才是最优的？如何寻找它？首先必须明确所有决策行动组成的集合 \mathbb{A} ，称之为行动空间 (action space)。例如对参数 $\theta \in \Theta$ 进行点估计，行动空间 \mathbb{A} 就是整个参数空间 Θ ，其中每个行动 $a \in \mathbb{A}$ 就是对 θ 的点估计。其次，得有一个事先约定好了的评估标准。在统计决策中常用“损失”来衡量一个具体的决策行动，这个想法可追溯至 Laplace，他曾提出参数估计就像统计学家参与一场与自然规律之间的赌博游戏，猜得愈差输得愈惨。

定义 12.1 (损失函数). 令 Θ 是参数空间， \mathbb{A} 为行动空间。损失函数 (loss function) 定义为 $L : \Theta \times \mathbb{A} \rightarrow \mathbb{R}$ ，负的损失就是“收益”。

例 12.1. 估计某新药对未来的占有率为 $\theta \in \Theta = [0, 1]$ ，行动空间为 $\mathbb{A} = [0, 1]$ 。低估了占有率将给药厂带来损失，高估了占有率将导致产品过剩而给药厂带来更大的损失，可以定义损失函数如下。

$$L(\theta, a) = \begin{cases} \theta - a & \text{若 } a < \theta \\ 2(a - \theta) & \text{若 } a \geq \theta \end{cases}$$

本节内容

损失函数是理性决策的基础，由领域专家具体给出。第一小节介绍贝叶斯学派利用期望损失来寻找最优行动——贝叶斯行动，数据通过更新参数的后验分布来影响决策结果，无数据的情况下利用参数的先验分布也是可行的。频率派只从样本中“挖掘”未知信息，第二小节介绍如何通过决策规则先把损失函数“加工”成风险函数，然后再利用极小极大原则或 Bayes 风险原则寻找最优的决策规则。

关键知识

(1) 会求贝叶斯期望损失和贝叶斯行动；(2) 掌握频率派的决策原则：极小极大原则和 Bayes 风险原则。

12.1.1 贝叶斯学派的期望损失原则

贝叶斯学派把未知参数 θ 视为随机变量，认为信息有三个来源：样本、损失函数和参数的先验信息。回顾第 169 页的例 2.52，决策行动是 $a_1 = \text{投资}$ 和 $a_2 = \text{存银行}$ ，参数 $\theta \sim 0.9\langle 1 \rangle + 0.1\langle 0 \rangle$ ，其中 1, 0 分别代表投资成功和失败。对于某一具体的决策行动 a ，损失 $L(\theta, a)$ 是由 θ 定义的随机变量，例 2.52 利用了期望损失来评估决策行动的优劣，下面明确给出贝叶斯期望损失的定义。

定义 12.2 (贝叶斯期望损失). 若已知参数 θ 的分布^{*}，不妨设其分布函数为 $F(\theta)$ ，是当前对 θ 的认识，则行动 a 的贝叶斯期望损失 (Bayesian expected loss) 或期望损失为

$$\rho(a) = E[L(\theta, a)] = \int_{\Theta} L(\theta, a) dF(\theta) \quad (12.1)$$

贝叶斯学派的期望损失原则寻找具有最小贝叶斯损失的行动 a_* ，称之为贝叶斯行动，即 $a_* = \underset{a \in A}{\operatorname{argmin}} \{\rho(a)\}$ 。

例 12.2. 接着例 12.1，在没有任何观测数据的情况下，参考同种类其他药品的占有率为，假设 θ 的先验分布 $\theta \sim U(0.1, 0.2)$ ，即 θ 的密度函数为 $\pi(\theta) = 10I_{(0.1, 0.2)}(\theta)$ ，则

$$\begin{aligned} \rho(a) &= \int_0^1 L(\theta, a) \pi(\theta) d\theta \\ &= \int_0^a 20(a - \theta) I_{(0.1, 0.2)}(\theta) d\theta + \int_a^1 10(\theta - a) I_{(0.1, 0.2)}(\theta) d\theta \\ &= \begin{cases} 0.15 - a & \text{当 } a \leq 0.1 \\ 15a^2 - 4a + 0.3 & \text{当 } 0.1 \leq a \leq 0.2 \\ 2a - 0.3 & \text{当 } a \geq 0.2 \end{cases} \end{aligned}$$

求得贝叶斯行动 $a_* = 2/15$ ，即该新药的市场占有率为 $2/15$ 。

例 12.3 (贝叶斯检验). 考虑假设检验问题 $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1 = \Theta - \Theta_0$ ，行动 a_0 和 a_1 分别表示“接受 H_0 ”和“接受 H_1 ”，不妨设损失函数为

$$L(\theta, a_j) = \begin{cases} 0 & \text{若 } \theta \in \Theta_j, \text{ 其中 } j = 0, 1 \\ L_j > 0 & \text{若 } \theta \notin \Theta_j, \text{ 即 } \theta \in \Theta_{1-j} \end{cases}$$

在有数据 $\mathbf{X} = \mathbf{x}$ 的情况下，设 θ 的后验分布为 $F(\theta|\mathbf{x})$ ，则行动 $a_j, j = 0, 1$ 的贝

*若无观测数据，则用 θ 的先验分布。若有观测数据，则用 θ 的后验分布，此时式 (12.1) 得到的结果称为后验期望损失。观察样本通过 θ 的后验分布影响决策行动的选择，而那些未观测到的数据则丝毫不影响贝叶斯行动。

叶斯期望损失为

$$\begin{aligned}\rho(a_j) &= \int_{\Theta} L(\theta, a_j) dF(\theta|x), \text{ 其中 } j = 0, 1 \\ &= L_j \int_{\Theta_{1-j}} dF(\theta|x) \\ &= L_j P(\Theta_{1-j}|x)\end{aligned}$$

当 $\rho(a_0) > \rho(a_1)$ 时, a_1 是贝叶斯行动, 即贝叶斯检验拒绝零假设 H_0 当且仅当

$$\frac{L_0}{L_1} > \frac{P(\Theta_0|x)}{P(\Theta_1|x)} \quad (12.2)$$

因为 $P(\Theta_0|x) = 1 - P(\Theta_1|x)$, 拒绝 H_0 的条件 (12.2) 也可等价地为

$$P(\Theta_1|x) > \frac{L_1}{L_0 + L_1}$$

因此, 贝叶斯检验的拒绝域为

$$R = \left\{ x : P(\Theta_1|x) > \frac{L_1}{L_0 + L_1} \right\}$$

练习 12.1. 设未知参数 θ 的先验分布是 $\theta \sim p\langle\theta_0\rangle + (1-p)\langle\theta_1\rangle$, 请根据例 12.3, 给出简单假设 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$ 的贝叶斯检验的拒绝域。

12.1.2 频率派的决策方法

频率派把未知参数 θ 视为待定的某个值，而不是随机变量，所以不承认贝叶斯期望损失 (12.1)。为了得到未知参数 θ 的信息，频率派将设计合理的试验，譬如对例 12.1 中的占有率 θ ，在随机调查的 n 个人中有 X 个人有意买此药，则 $X \sim B(n, \theta)$ 。频率派认为有关 θ 的所有信息都隐藏在观察样本之中，应该得到观测数据到后再做决策。

定义 12.3 (决策规则和风险函数). 一个决策规则 (decision rule) 或规则就是从 $X \sim F_\theta(x)$ 的样本空间 \mathcal{X} 到行动空间 \mathbb{A} 的可测函数，即 $\delta : \mathcal{X} \rightarrow \mathbb{A}$ ，其风险函数 (risk function) 定义为

$$R(\theta, \delta) = E[L(\theta, \delta(X))] = \int_{\mathcal{X}} L(\theta, \delta(x)) dF_\theta(x) \quad (12.3)$$

特别地，在没有随机试验进而没有任何观测数据的情况下，一个决策规则就是一个决策行动，此时风险函数即是损失函数。

定义 12.4. 决策规则 δ_1 不次于决策规则 δ_2 当且仅当对任意的 $\theta \in \Theta$ 皆有 $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ ，记作 $\delta_1 \leq \delta_2$ 。显然， \leq 是定义在规则集合上的一个偏序关系。若没有其他规则不次于规则 δ ，则称 δ 是可容决策规则 (admissible decision rule)，否则就称 δ 是非可容决策规则。

 非可容决策规则一般不会被采用，但可容决策规则也不见得是最优的。给定了决策规则 δ ， $R(\cdot, \delta)$ 是 $\Theta \rightarrow \mathbb{R}$ 的函数，所以决策规则的优劣比较是有关 θ 的函数之间的比较。按照这种标准通常很难找到最优的决策规则，因为很多时候函数 $R(\cdot, \delta_1)$ 和 $R(\cdot, \delta_2)$ 是不可比较的。以函数为评估指标往往难于操作，例如功效函数在备择假设的参数空间 Θ_1 上常常也是不可比较的。

例 12.4 (引自 [9]). 已知总体 $X \sim N(\theta, 1)$ ，为了估计未知参数 θ 考虑平方误差损失函数 $L(\theta, a) = (\theta - a)^2$ ，其中 $a \in \mathbb{A} = \mathbb{R}$ 。定义决策规则 $\delta_c(x) = cx$ ，其中 $c \in \mathbb{R}$ ，它所对应的风险函数为

$$\begin{aligned} R(\theta, \delta_c) &= E(\theta - cX)^2 \\ &= E[c(\theta - X) + (1 - c)\theta]^2 \\ &= c^2 + (1 - c)^2\theta^2 \end{aligned}$$

特别地， $R(\theta, \delta_1) = 1$ 。显然，若 $c > 1$ ，则决策规则 δ_1 优于 δ_c ；若 $0 \leq c < 1$ ，则决策规则 δ_c 和 δ_1 不可比较。

为了弥补无法通过风险函数比较决策规则孰优孰劣的遗憾，频率派提出了极小极大原则和 Bayes 风险原则，它们都确保了在各自原则之下“最优解”的存在性，但牺牲掉了些许合理性和原本的信仰。

定义 12.5 (极小极大原则). 对决策规则 $\delta: \mathcal{X} \rightarrow \mathbb{A}$ 而言, 以其最大风险 $\sup_{\theta \in \Theta} R(\theta, \delta)$ 来评估 δ 的优劣, 任意两个决策规则之间就可以进行比较。我们把如下定义的决策规则称为极小极大决策规则。

$$\delta_* = \operatorname{argmin}_{\delta} \left\{ \sup_{\theta \in \Theta} R(\theta, \delta) \right\} \quad (12.4)$$

在无数据的情况下, 也称 $\operatorname{argmin}_{a \in \mathbb{A}} \sup_{\theta \in \Theta} L(\theta, a)$ 为极小极大行动。

在例 12.4 中, δ_1 是极小极大决策规则, 这是因为

$$\sup_{\theta} R(\theta, \delta_c) = \sup_{\theta} [c^2 + (1 - c)^2 \theta^2] = \begin{cases} 1 & \text{若 } c = 1 \\ \infty & \text{若 } c \neq 1 \end{cases}$$

 一个决策规则 δ 可能几乎取不到最大风险, 换句话说, 最坏的情况可能几乎不发生。因为极小极大原则有时杞人忧天, 所以我们也把它称作“谨慎者的选择”。如例 2.52, 这是没有数据的例子, 极小极大决策行动是 $a_2 =$ 存银行, 这是因为

$$\begin{aligned} \sup L(\theta, a_2) &= \max\{-200, -200\} = -200 \\ \sup L(\theta, a_1) &= \max\{-2000, 10000\} = 10000 \end{aligned}$$

不同的原则反映了决策者的性格与好恶, 一般无法形式地论证孰对孰错, 所以不能根据极小极大原则在某个具体问题上的结论不合我意而判定该原则是错的。另外, 因为原则本身就是评价标准, 所谓的最优决策都是针对某一标准而言的, 没有普适的最优决策。

定义 12.6 (Bayes 风险原则). 风险函数 $R(\theta, \delta)$ 是关于未知参数 θ 和决策规则 δ 的函数, 为了能比较不同决策规则之间的优劣, 还有一个方法就是把未知参数 θ 视为随机变量, 则 $R(\theta, \delta)$ 是由 θ 定义的随机变量, 决策规则 δ 的 Bayes 风险定义为关于 θ 的平均风险, 即

$$r(\delta) = E[R(\theta, \delta)]$$

Bayes 规则定义为 Bayes 风险最小的规则, 即

$$\delta_* = \operatorname{argmin}_{\delta} \{r(\delta)\}$$

显然, 在无数据的情况下, Bayes 规则就是贝叶斯行动, 两个学派的决策方法取得了相同的结果。譬如, 对于例 2.52 的决策问题, 按照 Bayes 风险原则, 应该选择行动 $a_1 =$ 投资。

例 12.5. 接着例 12.4, 令 $\theta \sim N(0, \tau^2)$, 则规则 δ_c 的 Bayes 风险为

$$\begin{aligned} r(\delta_c) &= E[R(\theta, \delta_c)] \\ &= E[c^2 + (1 - c)^2 \theta^2] \\ &= c^2 + (1 - c)^2 \tau^2 \end{aligned}$$

在 $c_0 = \tau^2/(1 + \tau^2)$ 处取到最小, 其 Bayes 风险为 $\tau^2/(1 + \tau^2)$ 。

12.2 贝叶斯分析

所有科学研究的目标都是为了揭示客观规律。频率派认为概率是随机事件本身固有的物理属性，通过大量可重复的随机试验能够对它进行了解，而贝叶斯学派则认为概率描述的是信念度 (belief degree)，是个体对不确定性的主观认识。Laplace 说，“概率论归根结底就是简化为计算的常识。”

例 12.6. 某球迷观看一场重播的足球比赛，若此人并不知道比赛的结果，比赛对他就依然充满着不确定性，他还可以预测比赛的结果，还可以表达对某队赢球的信念度，甚至可以和同样不知道结果的朋友打赌。对球迷来说，得知结果再看比赛是无趣的，绿茵场上的不确定性正是足球的魅力所在。

这个例子有助于我们理解客观概率和主观概率的关系：明知某随机事件的概率客观存在，但在无法计算它的情况下^{*}，主观地赋予它一个信念度是再自然不过的事情了。没人会反对用 0,1 之间的某个数字表达一下自己对某事情的遗憾程度，贝叶斯学派问道，为什么就不能用类似的手段表达一下对结果不确定事件的信念度呢？

例如，根据当前的天气状况，张三认为“明天下雨”的概率是 90%，李四却认为只有 25%。两人经验阅历不同，对“明天下雨”这一未来事件的预测出现分歧也是很正常的，我们周围每天都在发生这样或那样的分歧。

然而需要强调的是，个体在经验上的差异不是贝叶斯统计学的研究对象，推断模式本身的规律性才是贝叶斯学派真正关注的。

1955 年，匈牙利数学家 Alfréd Rényi (1921-1970) 撰长文《论概率的一个新公理体系》，在 Kolmogorov 概率公理体系的基础之上提出了一个贝叶斯概率公理体系。在 Kolmogorov 的体系里，条件概率是由公理导出的概念（见第 77 页的定理 1.7）；而在 Rényi 的体系里，公理直接描述条件概率。对 Rényi 的体系感兴趣的读者也可参阅 Rényi 的著作《概率论》[132]。



定义 12.7 (Rényi, 1955). 已知样本空间 (Ω, \mathcal{S}) 并且 \mathcal{S} 的非空子集 \mathcal{C} 不包含空集 \emptyset ，如果实值函数 $P\{\cdot|\cdot\} : \mathcal{S} \times \mathcal{C} \rightarrow \mathbb{R}$ 满足以下三个属性，则称 P 为条件概率，称 $(\Omega, \mathcal{S}, \mathcal{C}, P)$ 为条件概率空间。如果没有特殊说明， $P\{A|B\}$ 已暗含 $A \in \mathcal{S}, B \in \mathcal{C}$ 。

① 凸性：对任意事件 $A \in \mathcal{S}, B \in \mathcal{C}$ 皆有 $P\{A|B\} \geq 0$ 并且 $P\{B|B\} = 1$ 。

^{*}人类不是万能的，科技再发达也有局限性。争论任何仪器都探测不到的空间里是否住着神仙是没有意义的，因为它的存在不具有可观测的效应。在哲学和科学上，人类认清自身主观意识的局限性是本分之事，丝毫不减它的尊严。

② 可列可加性：若事件 $A_1, A_2, \dots \in \mathcal{S}$ 两两不交，则对任意 $B \in \mathcal{C}$ 皆有

$$P\left\{\bigcup_{j=1}^{\infty} A_j \mid B\right\} = \sum_{j=1}^{\infty} P\{A_j \mid B\}$$

③ 乘法法则：对任意事件 $A, B \in \mathcal{S}, C \in \mathcal{C}$ ，如果 $BC \in \mathcal{C}$ ，则

$$P\{A|BC\}P\{B|C\} = P\{AB|C\}$$



与凸性和可列可加性一样，乘法法则也反映了主观认识的客观规律——给定信息 C 之后，条件概率 $P\{B|C\}$ 是对 B 的信念度，此时对 AB 的信念度 $P\{AB|C\}$ 可分解为 $P\{A|BC\}P\{B|C\}$ ，其中 $P\{A|BC\}$ 是暂时接受信息 B 时对 A 的信念度。

性质 12.1. 已知条件概率空间 $(\Omega, \mathcal{S}, \mathcal{C}, P)$ ，如果 $\Omega \in \mathcal{C}$ ，则 $(\Omega, \mathcal{S}, P\{\cdot|\Omega\})$ 可解释为 Kolmogorov 意义下的概率空间。

性质 12.2. 条件概率空间 $(\Omega, \mathcal{S}, \mathcal{C}, P)$ 满足下面的性质：(1) $P\{A|B\} = P\{AB|B\}$ ；(2) $P\{A|B\} \leq 1$ ；(3) $P\{\emptyset|B\} = 0$ ；(4) 若 $AB = \emptyset$ ，则 $P\{A|B\} = 0$ ；(5) $P\{A|BC\}P\{B|C\} = P\{B|AC\}P\{A|C\}$ ；(6) 如果 $A \subseteq A' \subseteq B \subseteq B'$ ，则 $P\{A|B'\} \leq P\{A'|B\}$ 。 (7) Bayes 公式^{*}：如果 $P\{B\} \neq 0$ ，则

$$P\{A|B\} = \frac{P\{B|A\}P\{A\}}{P\{B\}}, \text{ 其中 } P\{B\} = P\{B|A\}P\{A\} + P\{B|A^c\}P\{A^c\}$$

证明. 性质 (1) 至 (6) 的证明留作习题。由可列可加性和乘法法则可得 $P\{B\} = P\{BA\} + P\{BA^c\} = P\{B|A\}P\{A\} + P\{B|A^c\}P\{A^c\}$ ，请读者根据性质 (5) 推得 Bayes 公式。 \square

Rényi 在论文 [131] 中证明了他给出的公理体系涵盖了 Kolmogorov 的体系，从而贝叶斯学派对概率的理解更为宽泛。如果需要，贝叶斯主义者也可以赋予概率以频率派的解释。

贝叶斯学派曾乐衷于在哲学层面批判经典统计学以便坚定对贝叶斯主义的信仰和奠定贝叶斯理论的基础 [9, 10, 60]，譬如，贝叶斯学派乐此不疲地批评频率派不承认先验知识 [9]，只会“一本正经”地基于观测数据作推断，而且有时候这一做法还未贯彻到底。

^{*}在《概率的分析理论》中 Laplace 重新发现了 Bayes 公式并主张将它用作统计推断模式。Bayes 和 Laplace 都是主观贝叶斯学派的开山鼻祖。

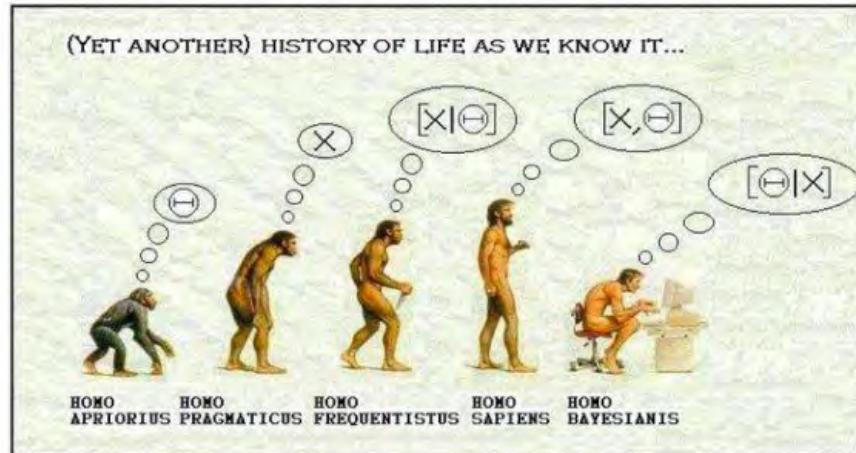


图 12.1: 人类思维方式最终走向贝叶斯主义: 通过观察获得未知参数的后验知识。

贝叶斯学派认为, 频率派统计推断的模式显然未反映出正常的人的思考方式, 因为在日常生活中常会遇到无据可查的窘况, 然而实践中没有什么能妨碍人们对不确定性的探索。例如, 有没有观测数据都不妨碍人们结合个人经验始终一贯地按照贝叶斯期望损失原则进行决策。

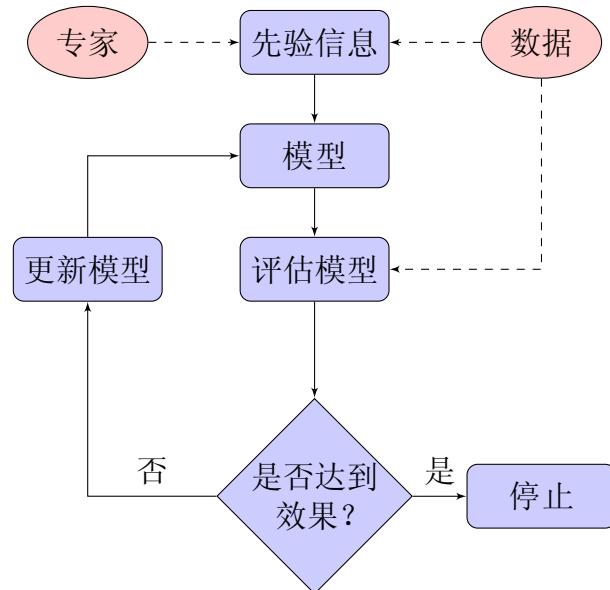


图 12.2: 构建贝叶斯统计模型的一般流程。

二十世纪, 贝叶斯学派发展出了两个重要支流——主观贝叶斯学派和客观贝叶斯学派, 它们在某些观点上有分歧。

- 主观贝叶斯学派坚决主张“概率”是对某命题的主观信念程度, 代表人物有英国数学家 Frank Plumpton Ramsey (1903-1930), 意大利概率统计学家 Bruno

de Finetti (1906-1985) 和美国统计学家 Leonard Jimmie Savage (1917-1971) 及其学生 Morris Herman DeGroot (1931-1989)。

- 客观贝叶斯学派认为统计分析仅依赖于预先假定的模型和需分析的数据，其中没有任何主观决策的成分，代表人物有英国统计学家、地球物理学家 Harold Jeffreys (1891-1989) 和 Dennis Victor Lindley (1923-)，美国物理学家 Edwin Thompson Jaynes (1922-1998) 等。客观贝叶斯推断受经典统计学的影响而更愿意通过遵循一些客观标准来减少主观性，譬如利用最大熵原则选择参数的无信息先验等。

二十世纪八十年代以后，得益于计算机科学和随机模拟技术的发展，贝叶斯统计学克服了数值计算的困难致力于算法设计与实际应用 [35, 55, 104, 123, 149] 而被越来越多的实践者关注，也逐渐渗透至机器学习、模式识别、信号处理等领域。

侧重于实践的贝叶斯数据分析除了需要 R 语言这样的工具，还有两个基于 Gibbs 抽样的贝叶斯推断工具值得推荐，它们是 BUGS 和 JAGS，常用于实现层级贝叶斯模型。R 语言提供了与 BUGS 和 JAGS 的接口。

- BUGS (Bayesian inference Using Gibbs Sampling) 是基于 Markov 链 Monte Carlo 方法（简称 MCMC 方法，详见第 15 章）的贝叶斯分析工具，早期在类 UNIX 环境下研发，后移植到 Windows 环境更名为 WinBUGS，目前仍是免费的，但未开源。
- 免费的开源软件 JAGS (Just Another Gibbs Sampler) 是 BUGS 的 C++ 重写，与 BUGS 完全兼容。

另外，加拿大统计学家 Radford M. Neal (1956-) 的开源软件 FBM (Flexible Bayesian Modeling) 也实现了很多 MCMC 方法和贝叶斯模型，如贝叶斯神经网络、高斯过程等。读者在互联网上能找到这些软件的详细介绍和用户手册，本书不再赘述。

本节内容

第一小节谈论如何选择参数的先验分布，如无信息先验（共轭先验、Jeffreys 先验）、最大熵先验、层级先验、ML-II 先验等，先验的选择依具体的需求而定。第二节介绍了几个近似求解后验期望的计算方法，用到了附录 E 中介绍的 Laplace 近似积分法（见第 773 页的定理 E.13）。第三节是似然函数的解析逼近。第四节是层级贝叶斯模型的初探，通过实例讲解如何用 BUGS 或 JAGS 实现层级贝叶斯模型。

关键知识

(1) 会根据具体问题选择参数的先验分布；(2) 了解似然函数与后验分布的近似计算；(3) 能设计简单的层级贝叶斯模型并通过 BUGS 或 JAGS 实现它。

12.2.1 后验 \propto 似然 \times 先验

站在贝叶斯主义者的立场上，个体的先验知识反映在统计模型中就是未知参数 θ 的先验分布 $\pi(\theta)$ 。

定义 12.8. 已知随机变量 X 的密度函数为 $p(x|\theta)$ ，其中参数 θ 的概率/密度函数为 $\pi(\theta)$ ，分布函数为 $F^\pi(\theta)$ 。定义 X 的预测分布 (predictive distribution) 为

$$m^\pi(x) = \int_{\Theta} p(x|\theta) dF^\pi(\theta) = \begin{cases} \sum_{\theta \in \Theta} p(x|\theta)\pi(\theta) & \text{当 } \theta \text{ 为离散型} \\ \int_{\Theta} p(x|\theta)\pi(\theta)d\theta & \text{当 } \theta \text{ 为连续型} \end{cases}$$

显然，预测分布 $m^\pi(x)$ 即为 X 的边缘密度函数。

例 12.7. 设 k -分类试验 $X \sim p_1\langle 1 \rangle + p_2\langle 2 \rangle + \cdots + p_k\langle k \rangle$ ，其中未知参数 $\mathbf{p} = (p_1, \dots, p_k)^\top \sim \text{Dirichlet}(\boldsymbol{\alpha})$ ， $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)^\top$ 。有时，简记为 $X \sim \text{Dirichlet-Multin}(1; \mathbf{p}; \boldsymbol{\alpha})$ 。试求预测分布 $P(X_1 = x_1, \dots, X_n = x_n | \boldsymbol{\alpha})$ ，其中 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Dirichlet-Multin}(1; \mathbf{p}; \boldsymbol{\alpha})$ ，其图模型见图 12.3。

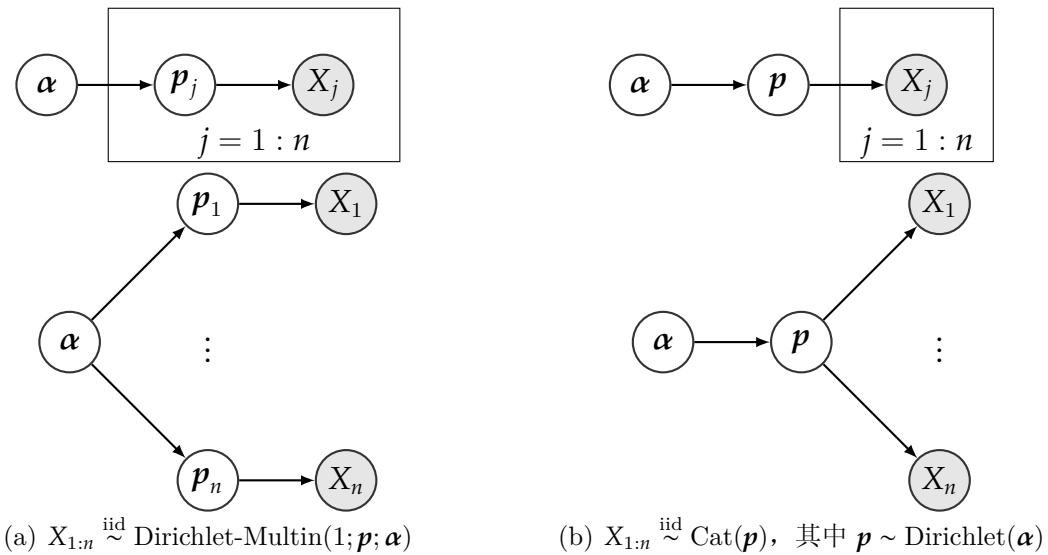


图 12.3: (a) 和 (b) 中上下两个图模型是等价的。图中的矩形框又称面板 (plate)，表示对框内的过程进行多次独立的重复。

解. 根据习题 4.30 的结果，我们得到

$$P(X_1 = x_1, \dots, X_n = x_n | \boldsymbol{\alpha}) = \frac{B(\alpha_1 + n_1, \dots, \alpha_j + n_j, \dots, \alpha_k + n_k)}{B(\alpha_1, \dots, \alpha_k)}$$

其中, n_j 是 x_1, \dots, x_n 中 j 的个数, $B(\alpha_1, \dots, \alpha_k) = \Gamma(\alpha_1) \cdots \Gamma(\alpha_k) / \Gamma(\alpha_1 + \cdots + \alpha_k)$ 。特别地, $X \sim \text{Dirichlet-Multin}(1; \mathbf{p}; \boldsymbol{\alpha})$ 的预测分布是

$$P(X = j | \boldsymbol{\alpha}) = \frac{\alpha_j}{\alpha_1 + \cdots + \alpha_k}, \text{ 其中 } j = 1, 2, \dots, k$$

例 12.8. 设样本点 $X_j \sim p(x_j | \theta_j)$, $j = 1, 2, \dots, n$ 相互独立, 已知未知参数 $\theta_1, \theta_2, \dots, \theta_n \stackrel{\text{iid}}{\sim} \pi_0(\theta)$, 则随机向量 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n)^\top$ 的先验分布是 $\pi(\boldsymbol{\theta}) = \prod_{j=1}^n \pi_0(\theta_j)$ 。记 X_j 的边缘密度为 $m^{\pi_0}(x_j)$, 记 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$, 则样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的边缘密度 $m^\pi(\mathbf{x})$ 为

$$\begin{aligned} m^\pi(\mathbf{x}) &= \int_{\Theta^n} p(\mathbf{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\Theta^n} \left[\prod_{j=1}^n p(x_j | \theta_j) \right] \left[\prod_{j=1}^n \pi_0(\theta_j) \right] d\boldsymbol{\theta} \\ &= \prod_{j=1}^n \int_{\Theta} p(x_j | \theta_j) \pi_0(\theta_j) d\theta_j \\ &= \prod_{j=1}^n m^{\pi_0}(x_j) \end{aligned}$$

众所周知, 知识的积累是一个不断更新的过程。在获得观测数据 $\mathbf{X} = \mathbf{x}$ 之后, 通过参数的条件分布 $\pi(\theta | \mathbf{x})$, 个体可以更新对参数 θ 的主观认识。

定义 12.9. 已知样本 $\mathbf{X} = (X_1, \dots, X_n)^\top \sim p(\mathbf{x} | \theta)$, 其中 $\mathbf{x} = (x_1, \dots, x_n)^\top$ 。假设未知参数 θ 的先验分布为 $\theta \sim \pi(\theta)$, 参数 θ 的后验分布 (posterior distribution) 定义为

$$\pi(\theta | \mathbf{x}) = \frac{p(\mathbf{x} | \theta) \pi(\theta)}{p(\mathbf{x})} \propto p(\mathbf{x} | \theta) \pi(\theta) \quad (12.5)$$

其中, $p(\mathbf{x}) = \sum_{\theta \in \Theta} p(\mathbf{x} | \theta) \pi(\theta)$, 或者 $\int_{\Theta} p(\mathbf{x} | \theta) \pi(\theta) d\theta$

我们把 $p(\mathbf{x} | \theta)$ 视为有关 θ 的函数, 记作 $\mathcal{L}(\theta; \mathbf{x})$, 称为似然函数或似然。式 (12.5) 也称作 Bayes 公式, 其中 $\pi(\theta | \mathbf{x})$ 就是 θ 的条件分布。Bayes 公式是贝叶斯推断 (Bayesian inference) 的核心, 常常通俗地简记作“后验 \propto 似然 \times 先验”。

例 12.9. 接着**例 12.7**, 在观察到数据 $\mathbf{X}_{\text{obs}} = \mathbf{x}_{\text{obs}} = (x_1, \dots, x_n)^\top$ 之后, 计算未知参数 \mathbf{p} 的后验分布。

解. 令 n_j 是 x_1, \dots, x_n 中 j 的个数,

$$\begin{aligned}\pi(\boldsymbol{p} | \mathbf{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}, \boldsymbol{\alpha}) &\propto P(\mathbf{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}} | \boldsymbol{p}, \boldsymbol{\alpha}) \pi(\boldsymbol{p} | \boldsymbol{\alpha}) \\ &\propto \prod_{j=1}^k p_j^{n_j} \prod_{j=1}^k p_j^{\alpha_j-1} \\ &= \prod_{j=1}^k p_j^{\alpha_j+n_j-1}\end{aligned}$$

于是, $\boldsymbol{p} | \mathbf{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}, \boldsymbol{\alpha} \sim \text{Dirichlet}(\alpha_1 + n_1, \dots, \alpha_j + n_j, \dots, \alpha_k + n_k)$ 。

例 12.10. 设样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\theta, \tau^2)$, 其中参数 θ 未知, τ^2 已知。样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 具有下面的分解性质。

$$\sum_{i=1}^n (X_i - \theta)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \theta)^2$$

给了样本值 $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^\top$ 后, 似然函数 $\mathcal{L}(\theta; \boldsymbol{x})$ 为

$$\begin{aligned}\mathcal{L}(\theta; \boldsymbol{x}) &= \exp \left\{ \frac{-\sum_{i=1}^n (x_i - \bar{x})^2 - n(\bar{x} - \theta)^2}{2\tau^2} \right\} \\ &\propto \exp \left\{ \frac{-(\bar{x} - \theta)^2}{2\tau^2/n} \right\}\end{aligned}$$

如果 θ 的先验分布 $\pi(\theta)$ 是正态分布, 则 $\mathcal{L}(\theta; \boldsymbol{x})\pi(\theta)$ 通过配方法总能整理成某个正态分布的密度函数, 即 θ 的后验分布 $\pi(\theta|\boldsymbol{x})$ 依然是正态分布。具体见第 656 页的例 12.19。

性质 12.3. 已有样本 \mathbf{X} 与未知的可观察样本 \mathbf{X}_{new} 关于参数 $\theta \in \Theta$ 条件独立, 即 $p(\mathbf{x}_{\text{new}}|\theta, \mathbf{x}) = p(\mathbf{x}_{\text{new}}|\theta)$, 在观测到数据 $\mathbf{X} = \boldsymbol{x}$ 之后, 可给出新样本 \mathbf{X}_{new} 的后验预测分布 (posterior predictive distribution) $p(\mathbf{x}_{\text{new}}|\boldsymbol{x})$ 。

$$p(\mathbf{x}_{\text{new}}|\boldsymbol{x}) = \begin{cases} \sum_{\theta \in \Theta} p(\mathbf{x}_{\text{new}}|\theta)\pi(\theta|\boldsymbol{x}) & \text{参数 } \theta \text{ 是离散型} \\ \int_{\Theta} p(\mathbf{x}_{\text{new}}|\theta)\pi(\theta|\boldsymbol{x})d\theta & \text{参数 } \theta \text{ 是连续型} \end{cases} \quad (12.6)$$

证明. 令 $F(\theta|\boldsymbol{x})$ 是给定观测到数据 $\mathbf{X} = \boldsymbol{x}$ 之后 θ 的后验分布函数, 则

$$p(\mathbf{x}_{\text{new}}|\boldsymbol{x}) = \int_{\Theta} p(\mathbf{x}_{\text{new}}|\theta, \boldsymbol{x})dF(\theta|\boldsymbol{x}) = \int_{\Theta} p(\mathbf{x}_{\text{new}}|\theta)dF(\theta|\boldsymbol{x}) \quad \square$$

例 12.11. 接着**例 12.9**, 根据未知参数的后验分布 $p|\mathbf{X}_{\text{obs}} = \mathbf{x}_{\text{obs}}, \boldsymbol{\alpha} \sim \text{Dirichlet}(\alpha_1 + n_1, \dots, \alpha_k + n_k)$, 可得后验预测分布 $P(X_{\text{new}} = j|\mathbf{X}_{\text{obs}} = \mathbf{x}_{\text{obs}}, \boldsymbol{\alpha}), j = 1, 2, \dots, k$ 如下。

$$P(X_{\text{new}} = j|\mathbf{X}_{\text{obs}} = \mathbf{x}_{\text{obs}}, \boldsymbol{\alpha}) = \frac{\alpha_j + n_j}{\alpha_1 + \dots + \alpha_k + n}$$

其中, n_j 是 x_1, \dots, x_n 中 j 的个数, 并且 $n = n_1 + \dots + n_k$ 。

算法 12.1. 在观测到新的数据 $\mathbf{X}_{\text{new}} = \mathbf{x}_{\text{new}}$ 之后, 利用 Bayes 公式 (12.5) 可以得到下述迭代算法, 对参数的后验分布进行更新。

$$\pi(\theta|\mathbf{x}, \mathbf{x}_{\text{new}}) = \frac{p(\mathbf{x}_{\text{new}}|\theta)\pi(\theta|\mathbf{x})}{p(\mathbf{x}_{\text{new}}|\mathbf{x})} \propto p(\mathbf{x}_{\text{new}}|\theta)\pi(\theta|\mathbf{x}) \quad (12.7)$$

 在得到数据 \mathbf{x} 而尚未得到 \mathbf{x}_{new} 之际, 对参数 θ 的认知就是后验分布 $\pi(\theta|\mathbf{x})$ 。参数分布之“先验”与“后验”是相对而言的: 今天的后验可能就是明天的先验。在得到数据 \mathbf{x}_{new} 之后, 式 (12.7) 就是 Bayes 公式 (12.5) 的变种, 这个基于中间结果 $\pi(\theta|\mathbf{x})$ 的更新算法符合人们对参数 θ 的自然认知过程。

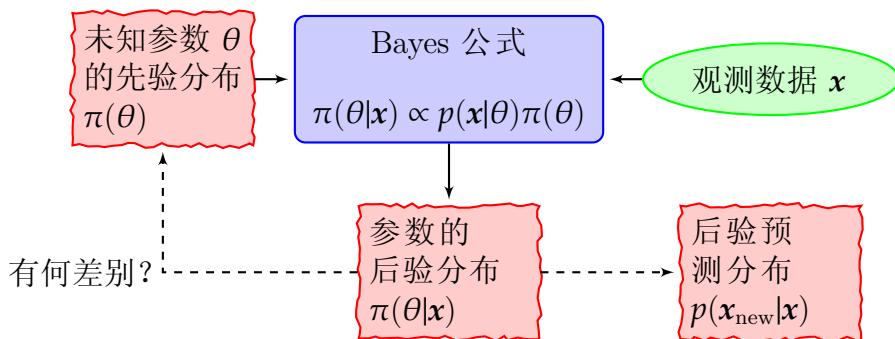


图 12.4: 在得到观测数据之前, 贝叶斯方法允许通过参数 θ 的先验分布来表达对 θ 的以往的经验。在得到观测数据之后, Bayes 公式则以固定的模式把观测数据和以往经验转化为对 θ 的新的认识。

例 12.12. 血友病是 X 染色体隐性基因病, 男性患者从母亲那里继承了一条染病的 X 染色体, 女性患者则继承了两条染病的 X 染色体, 其中一条来自患病的父亲。女性如果只有一条 X 染色体带有血友病基因, 则不表现出病症而只是血友病基因携带者。

已知某女性 F 不是血友病患者, 她的父母也不是血友病患者, 但 F 有一个患血友病的兄弟, 若观察到 F 的两个儿子皆无血友病, 问 F 是血友病基因携带者的概率? 若还观察到 F 的第三个儿子也无血友病, 问情况又将如何?

解. 因为 F 有一个患血友病的兄弟, 所以 F 有可能是血友病基因携带者。用随机变量 $\theta = 1$ 或 0 来表示 F 是否为血友病基因携带者, 通过遗传学知识得到参数 θ 的先

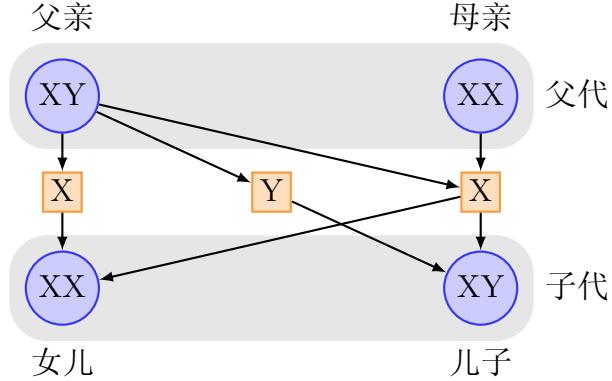


图 12.5: 人类有一对性染色体——女性的两条 X 染色体分别来自于父母；男性的 X 染色体来自于母亲，而 Y 染色体来自于父亲。

验分布是 $\theta \sim \frac{1}{2}\langle 1 \rangle + \frac{1}{2}\langle 0 \rangle$ 。观察样本 $\mathbf{X} = (X_1, X_2)^\top$ 中，0-1 分布的随机变量 X_1, X_2 分别表示 F 的两个儿子是否患病，则 $X_1 \perp\!\!\!\perp X_2 | \theta$ ，即 $p(x_1, x_2 | \theta) = p(x_1 | \theta)p(x_2 | \theta)$ 。

在得到样本值 $\mathbf{x} = (x_1, x_2)^\top$ 后， $p(\mathbf{x} = \mathbf{0} | \theta = 0) = p(x_1 = 0 | \theta = 0)p(x_2 = 0 | \theta = 0) = 1$ ，且 $p(\mathbf{x} = \mathbf{0} | \theta = 1) = p(x_1 = 0 | \theta = 1)p(x_2 = 0 | \theta = 1) = 0.25$ 。利用式 (12.5) 可得，

$$\begin{aligned}\pi(\theta = 1 | \mathbf{x} = \mathbf{0}) &= \frac{p(\mathbf{x} = \mathbf{0} | \theta = 1)\pi(\theta = 1)}{p(\mathbf{x} = \mathbf{0} | \theta = 1)\pi(\theta = 1) + p(\mathbf{x} = \mathbf{0} | \theta = 0)\pi(\theta = 0)} \\ &= \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = 0.2 \\ \pi(\theta = 0 | \mathbf{x} = \mathbf{0}) &= 1 - \pi(\theta = 1 | \mathbf{x} = \mathbf{0}) = 0.8\end{aligned}$$

观察到第三个儿子也不是血友病患者（即 $x_{\text{new}} = 0$ ）之后，利用式 (12.7) 可算得 F 是血友病基因携带者的后验概率更小了，符合常理。

$$\pi(\theta = 1 | \mathbf{x} = \mathbf{0}, x_{\text{new}} = 0) = \frac{0.5 \times 0.2}{0.5 \times 0.2 + 1 \times 0.8} = \frac{1}{9}$$

12.2.2 参数的先验分布

参数的先验分布反映了对参数的先验认识，大致分为无信息先验和主观先验两大类。在一些原则之下，每个类里又细分出一些具体选择先验分布的方法。

- 无信息先验：非正常先验、Jeffreys 先验、共轭先验等。
- 主观先验：最大熵先验、层级先验、类型 II 最大似然先验等。

定义 12.10. 当对参数的先验分布一无所知时，为了使用贝叶斯推断模式“后验 \propto 似然 \times 先验”，在某些合理的假设之下采用的先验分布称为无信息先验 (noninformative prior) 或非主观先验 (non-subjective prior)。

绝对意义的“无信息”是没有的，至少已知参数是否连续、参数空间等信息。譬如，总体 $X \sim N(\theta, \tau^2)$ 中 θ 未知，参数空间为 $\Theta = \mathbb{R}$ ，任何实数都等可能地被选中的那个分布所含参数的信息最少，最能体现“无信息”。然而，我们没办法表达这样的先验分布，通常迫不得已采用的无信息先验为 $\pi(\theta) = 1$ ，但它违背了归一性 $\int_{-\infty}^{+\infty} \pi(\theta) d\theta = 1$ ，称为非正常先验 (improper prior)。

Bayes 推荐使用均匀分布作为参数的无信息先验。1812 年，Laplace 首次使用非正常先验 $\pi(\theta) = 1$ 来表示 \mathbb{R} 上的均匀分布。选用非正常先验并不影响贝叶斯推断，譬如例 12.10 中，若选择无信息先验 $\pi(\theta) = 1$ ，则参数 θ 的后验分布为

$$\theta | \mathbf{X} = \mathbf{x} \sim N\left(\bar{x}, \frac{\tau^2}{n}\right)$$

定义 12.11 (位置参数). 如果随机向量 $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ 的密度函数具有形式 $g(\mathbf{x} - \boldsymbol{\theta})$ ，其中 $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^n$ ，则称参数 $\boldsymbol{\theta}$ 为位置参数，称 $g(\mathbf{x} - \boldsymbol{\theta})$ 为位置密度函数。例如正态分布 $N(\boldsymbol{\theta}, \Sigma)$ (其中 Σ 固定) 具有位置密度函数。

例 12.13 (位置参数的无信息先验). 假设随机向量 \mathbf{X} 具有位置密度函数 $g(\mathbf{x} - \boldsymbol{\theta})$ ，其中 $\mathcal{X} = \Theta = \mathbb{R}^n$ 。在位置变换之下，随机向量 $\mathbf{Y} = \mathbf{X} + \mathbf{c}$ 也具有位置密度函数 $g(\mathbf{y} - \boldsymbol{\eta})$ ，其中 $\boldsymbol{\eta} = \boldsymbol{\theta} + \mathbf{c}$ 。

从结构上看， $\mathbf{X} \sim g(\mathbf{x} - \boldsymbol{\theta})$ 与 $\mathbf{Y} \sim g(\mathbf{y} - \boldsymbol{\eta})$ 的样本空间和参数空间都是一样的，位置参数 $\boldsymbol{\theta}$ 和 $\boldsymbol{\eta}$ 应该具有相同的无信息先验，不妨设它为 π 。对任意的 $A \subseteq \mathbb{R}^n$ 有 $P(\boldsymbol{\theta} \in A) = P(\boldsymbol{\eta} \in A) = P(\boldsymbol{\theta} \in A - \mathbf{c})$ ，其中 $A - \mathbf{c} = \{\mathbf{a} - \mathbf{c} : \mathbf{a} \in A\}$ 。于是，

$$\begin{aligned} \int_A \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} &= \int_{A - \mathbf{c}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_A \pi(\boldsymbol{\theta} - \mathbf{c}) d\boldsymbol{\theta} \end{aligned}$$

由 A 的任意性可得 $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} - \mathbf{c})$ ，特别地 $\pi(\mathbf{c}) = \pi(\mathbf{0})$ 。再由 \mathbf{c} 的任意性，所以 π 一定是常数函数。一般地，我们选择 $\boldsymbol{\theta}$ 的无信息先验为 $\pi(\boldsymbol{\theta}) = 1$ 。

定义 12.12. 如果似然函数具有形式 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = g[\psi(\boldsymbol{\theta}) - t(\mathbf{x})]$, 数据只作用于 $t(\mathbf{x})$, 即不同的数据只导致 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ 的位置不同, 则称 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ 是数据转换似然 (data-translated likelihood, [20])。

令 $\eta = \psi(\boldsymbol{\theta})$, 则 $g[\eta - t(\mathbf{x})]$ 是一个位置密度函数。选择位置参数 η 的无信息先验为 $\pi(\eta) = 1$, 则参数 $\boldsymbol{\theta}$ 的无信息先验可设为

$$\pi(\boldsymbol{\theta}) \propto \left| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|$$

例 12.14. 已知简单随机样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 来自总体 $N(\mu, \sigma^2)$,

□ 假设参数 σ^2 已知, μ 未知。令 $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$, 令似然函数为

$$\mathcal{L}(\mu; \mathbf{x}) = \exp \left\{ \frac{-n}{2\sigma^2} (\mu - \bar{x})^2 \right\}$$

由**定义 12.12**, $\mathcal{L}(\mu; \mathbf{x})$ 是数据转换似然, 其中 $t(\mathbf{x}) = \bar{x}$ 且 $\psi(\mu) = \mu$ 。于是, 参数 μ 的无信息先验可设为 $\pi(\mu) = 1$ 。

□ 假设参数 μ 已知, σ^2 未知。令 $s^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \mu)^2$, 令似然函数为

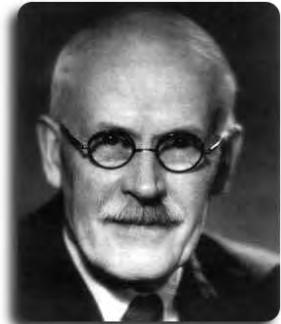
$$\begin{aligned} \mathcal{L}(\sigma; \mathbf{x}) &= \frac{\sigma^{-n}}{s^{-n}} \exp \left\{ \frac{-ns^2}{2\sigma^2} \right\} \\ &= \exp \left\{ -n \ln \left(\frac{\sigma}{s} \right) \right\} \times \exp \left\{ \frac{-n}{2} \exp \ln \left(\frac{s^2}{\sigma^2} \right) \right\} \\ &= \exp \left\{ -n(\ln \sigma - \ln s) - \frac{n}{2} \exp [-2(\ln \sigma - \ln s)] \right\} \end{aligned}$$

显然, $\mathcal{L}(\sigma; \mathbf{x})$ 是数据转换似然, 其中 $t(\mathbf{x}) = \ln s$ 且 $\psi(\sigma) = \ln \sigma$ 。于是, 参数 σ 的无信息先验可设为 $\pi(\sigma) \propto \sigma^{-1}, \sigma > 0$ 。

□ 假设参数 μ, σ^2 皆未知且相互独立, 则 $\pi(\mu, \sigma) \propto \sigma^{-1}, \sigma > 0$ 。

对参数 θ 一无所知就定义其无信息先验为 “ $\pi(\theta) \propto$ 常数” 的做法并不合理, 因为它不满足一一变换下的不变性, 例如对参数 $\eta = \exp \theta$ 同样一无所知, 但 $\pi(\eta) \propto 1/\eta$ 并非正比于常数。为解决这一矛盾, 英国统计学家、客观贝叶斯学派代表人物 Harold Jeffreys (1891-1989) 建议任一决定先验分布的方法 \mathcal{M} 都应该满足以下的原则。

定义 12.13 (Jeffreys 不变性原则). 按照方法 \mathcal{M} 得到模型 $p(\mathbf{x}|\boldsymbol{\theta})$ 中未知参数 $\boldsymbol{\theta} \in \mathbb{R}^k$ 的先验分布 $\pi(\boldsymbol{\theta})$ 。设一一映射



$g : \mathbb{R}^k \rightarrow \mathbb{R}^k$ 可微, 以 $\boldsymbol{\eta} = g(\boldsymbol{\theta}) \in \mathbb{R}^k$ 为参数, 按照同一方法 \mathcal{M} 给出参数 $\boldsymbol{\eta}$ 的先验分布 $\pi_g(\boldsymbol{\eta})$ 应该满足下面的条件。

$$\pi(\boldsymbol{\theta}) = \left| \det \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \right) \right| \pi_g(\boldsymbol{\eta})$$

换句话说, 通过方法 \mathcal{M} 得到的 $\boldsymbol{\eta}$ 的先验 $\pi_g(\boldsymbol{\eta})$ 恰是根据第 154 页的定理 2.15 所导出的 $\boldsymbol{\eta}$ 的分布 $|\det(\partial\boldsymbol{\theta}/\partial\boldsymbol{\eta})| \pi(\boldsymbol{\theta})$ 。满足 Jeffreys 不变性原则的先验 $\pi(\boldsymbol{\theta})$ 使得从模型 $p(\mathbf{x}|\boldsymbol{\theta})$ 得到的后验分布 $\pi(\boldsymbol{\theta}|\mathbf{x})$ 与从模型 $p(\mathbf{x}|\boldsymbol{\eta})$ 得到的后验分布 $p(\boldsymbol{\eta}|\mathbf{x})$ 依然保证具有关系

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \left| \det \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \right) \right| p(\boldsymbol{\eta}|\mathbf{x})$$

基于第 497 页的例 8.4 所揭示的 Fisher 信息矩阵的性质, Jeffreys 巧妙地提出了模型 $p(\mathbf{x}|\boldsymbol{\theta})$ 中未知参数 $\boldsymbol{\theta}$ 的所谓 “Jeffreys 先验” [81]。

定义 12.14 (Jeffreys 先验, 1961). 令 $\mathcal{I}(\boldsymbol{\theta})$ 是 $\boldsymbol{\theta}$ 的信息矩阵 (见定义 8.2), Jeffreys 先验定义为

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\theta})}$$

例 12.15. 已知总体 $X \sim B(n, \theta)$, 其中 n 已知, 未知参数 $\theta \in (0, 1)$ 的 Fisher 信息量为 $\mathcal{I}(\theta) = n/[\theta(1-\theta)]$ (见第 495 页的例 8.1), 于是 θ 的 Jeffreys 先验为

$$\pi(\theta) \propto [\theta(1-\theta)]^{-1/2}, \text{ 即 } \theta \sim \text{Beta}(1/2, 1/2)$$

例 12.16. 已知总体 $X \sim N(\mu, \sigma^2)$, 其中参数 μ, σ^2 皆未知, 则 $\det \mathcal{I}(\mu, \sigma) = 2/\sigma^4$ (见例 8.2), 于是 $\pi(\mu, \sigma) \propto \sigma^{-2}$ 。

练习 12.2. 求例 12.14 中各种情况之下参数的 Jeffreys 先验。答案: 与例 12.14 的相同。

定义 12.15. 已知总体分布 $X \sim f(x|\boldsymbol{\theta})$, 出于简化计算之目的, 无信息先验可以选择这样的分布——数据不改变参数的先验分布 $\pi(\boldsymbol{\theta})$ 和后验分布 $\pi(\boldsymbol{\theta}|\mathbf{x})$ 的分布族, 如此的先验分布 $\pi(\boldsymbol{\theta})$ 称为 $f(x|\boldsymbol{\theta})$ 的共轭先验 (conjugate prior)。

例 12.17. 假设总体 $\mathbf{N} = (N_1, N_2, \dots, N_k)^\top \sim \text{Multin}(n; p_1, p_2, \dots, p_k)$, 其中未知参数 $\mathbf{p} = (p_1, p_2, \dots, p_k)^\top$ 的共轭先验分布为 $\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k)$, 这是因为在得到观察数据 $\mathbf{N} = \mathbf{n} = (n_1, n_2, \dots, n_k)^\top$ 之后, \mathbf{p} 的后验分布仍是 Dirichlet 分布。具体来说,

$$\mathbf{p}|\mathbf{N} = \mathbf{n} \sim \text{Dirichlet}(\alpha_1 + n_1, \alpha_2 + n_2, \dots, \alpha_k + n_k)$$

上述事实的推导过程很简单，就是看一下 $\pi(\mathbf{p}|\mathbf{N} = \mathbf{n})$ 正比于的那个关于 \mathbf{p} 的函数属于哪种分布类型。

$$\begin{aligned}\pi(\mathbf{p}|\mathbf{N} = \mathbf{n}) &\propto \pi(\mathbf{p})f(\mathbf{n}|\mathbf{p}) \\ &\propto \prod_{j=1}^k p_j^{\alpha_j-1} \prod_{j=1}^k p_j^{n_j} \\ &= \prod_{j=1}^k p_j^{\alpha_j+n_j-1}\end{aligned}$$

例 12.18. 已知总体 X 的分布为指数分布族，即 X 的密度函数形如

$$f(x|\boldsymbol{\theta}) = h(x)\eta(\boldsymbol{\theta}) \exp\left\{\sum_{j=1}^k \lambda_j(\boldsymbol{\theta})T_j(x)\right\}$$

未知参数 $\boldsymbol{\theta}$ 的共轭先验分布为

$$\pi(\boldsymbol{\theta}) \propto [\eta(\boldsymbol{\theta})]^b \exp\left\{\sum_{j=1}^k \lambda_j(\boldsymbol{\theta})a_j\right\}$$

这是因为，给了样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 的观测值 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ ，未知参数 $\boldsymbol{\theta}$ 的后验分布依然是指数分布族，具体是

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{X} = \mathbf{x}) &\propto [\eta(\boldsymbol{\theta})]^{b'} \exp\left\{\sum_{j=1}^k \lambda_j(\boldsymbol{\theta})a'_j\right\} \\ \text{其中, } b' &= b + n \text{ 且 } a'_j = a_j + \sum_{i=1}^n T_j(x_i)\end{aligned}$$

练习 12.3. 请读者验证下面有关共轭先验的结果。

表 12.1: 常见总体 X 的分布及其共轭先验分布：数据不改变参数的分布族。

参数 θ 的共轭先验	总体 X 的分布	参数 θ 的后验分布
$N(\mu, \sigma^2)$	$N(\theta, \tau^2)$	$N\left(\frac{\tau^2}{\sigma^2 + \tau^2}\mu + \frac{\sigma^2}{\sigma^2 + \tau^2}x, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$
Beta(a, b)	$B(n, \theta)$	Beta($a + x, n - x + b$)
Gamma(α, β)	Expon(θ)	Gamma($\alpha + 1, \beta + x$)
Gamma(α, β)	Poisson(θ)	Gamma($\alpha + x, \beta + 1$)
inv-Gamma(α, β)	$N(\mu, \theta)$	inv-Gamma($\alpha + 1/2, \beta + (x - \mu)^2/2$)

例 12.19 (数据淹没先验). 已知总体 $X \sim N(\theta, \tau^2)$, 其中未知参数 θ 的先验分布为共轭先验 $\theta \sim N(\mu, \sigma^2)$, 参数 τ^2, μ, σ^2 均已知。由第 66 页的例 1.50, 则

$$\begin{aligned}\pi(\theta|x) &\propto \frac{\phi(x|\theta, \tau^2)\phi(\theta|\mu, \sigma^2)}{m(x)} \\ &\propto \phi\left(\theta \left| \frac{\tau^2}{\sigma^2 + \tau^2}\mu + \frac{\sigma^2}{\sigma^2 + \tau^2}x, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right.\right)\end{aligned}$$

已知样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 来自总体 X , 样本均值 $\bar{X} \sim N(\theta, \tau^2/n)$ 是充分统计量, 给定样本值 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 后 θ 的后验分布为 $\pi(\theta|\mathbf{x}) = \pi(\theta|\bar{x})$, 其中 $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$ 。进而得到参数 θ 的后验分布为

$$\theta|\mathbf{X} = \mathbf{x} \sim N\left(\frac{\tau^2/n}{\sigma^2 + \tau^2/n}\mu + \frac{\sigma^2}{\sigma^2 + \tau^2/n}\bar{x}, \frac{\sigma^2\tau^2}{n\sigma^2 + \tau^2}\right)$$

当样本容量 n 足够大时, 不难发现

$$\frac{\tau^2/n}{\sigma^2 + \tau^2/n}\mu + \frac{\sigma^2}{\sigma^2 + \tau^2/n}\bar{x} \rightarrow \bar{x}$$

即 $E(\theta|\mathbf{X} = \mathbf{x}) \approx \bar{x}$ 。不管 θ 的先验分布 $\theta \sim N(\mu, \sigma^2)$ 如何, 只要样本容量足够地大, 参数的后验分布与先验分布无关, 只由样本决定。我们把这种现象称为“数据淹没先验”, 即先验的作用几乎被数据掩盖, 弱化到可以忽略的程度。

第 182 页的定义 2.39 讨论过离散型随机变量的熵, 它刻画了随机变量取值不确定的程度。参数 $\theta \sim p_1\langle\theta_1\rangle + p_2\langle\theta_2\rangle + \dots + p_n\langle\theta_n\rangle$ 的熵在 $p_1 = \dots = p_n = 1/n$ 时最大, 此时有关 θ 取何值的信息最少, 因此 θ 的无信息先验可选为 $\theta \sim \frac{1}{n}\langle\theta_1\rangle + \frac{1}{n}\langle\theta_2\rangle + \dots + \frac{1}{n}\langle\theta_n\rangle$ 。

如果知道参数 θ 的部分信息, 如均值, 分布族, 超参数的信息, 以往的经验等, 此时应该如何选择 θ 的先验分布呢? 下面依次介绍最大熵先验, 层级先验和类型 II 最大似然先验。

定理 12.1. 已知离散型参数 $\theta \sim p_1\langle\theta_1\rangle + p_2\langle\theta_2\rangle + \dots + p_j\langle\theta_j\rangle + \dots$ 的部分信息由约束方程组 $E[g_k(\theta)] = \mu_k, k = 1, 2, \dots, m$ 描述, 则以下 p_j 使得随机变量 θ 的熵最大,

$$p_j = \frac{\exp\left\{\sum_{k=1}^m \lambda_k g_k(\theta_j)\right\}}{\sum_{j=1}^{\infty} \exp\left\{\sum_{k=1}^m \lambda_k g_k(\theta_j)\right\}} \quad (12.8)$$

其中系数 λ_k 由约束方程组决定。式 (12.8) 所定义的分布称为 θ 的最大熵先验。

证明. 利用 Lagrange 乘子法, 在约束条件 $E[g_k(\theta)] = \mu_k, k = 1, 2, \dots, m$ 下最大化目标函数 $-\sum_{j=1}^{\infty} p_j \ln p_j$ 等价于最大化下面的 Lagrange 函数,

$$\begin{aligned} f(\lambda_1, \dots, \lambda_m, p_1, p_2, \dots, p_j, \dots) \\ = -\sum_{j=1}^{\infty} p_j \ln p_j + \sum_{k=1}^m \lambda_k \left\{ \left[\sum_{j=1}^{\infty} p_j g_k(\theta_j) \right] - \mu_k \right\} \end{aligned}$$

其中, 待定参数 $\lambda_1, \dots, \lambda_m$ 称为 Lagrange 乘子。求解



$$\frac{\partial f}{\partial p_j} = -\ln p_j - 1 + \sum_{k=1}^m \lambda_k g_k(\theta_j) = 0$$

即得式 (12.8) (经过归一化)。□

例 12.20. 已知参数 $\theta \sim p_0 \langle 0 \rangle + p_1 \langle 1 \rangle + \dots + p_n \langle n \rangle + \dots$ 的期望 $E(\theta) = 2$, 由定理 12.1,

$$p_n = \frac{e^{\lambda n}}{\sum_{n=0}^{\infty} e^{\lambda n}} = (1 - e^{\lambda}) e^{\lambda n}$$

显然, $\theta \sim \text{Geom}(1 - e^{\lambda})$ 。由第 269 页的练习 4.11 知 $E(\theta) = e^{\lambda}/(1 - e^{\lambda}) = 2$, 解之得 θ 的最大熵先验分布为 $\theta \sim \text{Geom}(1/3)$ 。即,

$$\theta \sim \frac{1}{3} \langle 0 \rangle + \frac{2}{3^2} \langle 1 \rangle + \dots + \frac{2^n}{3^{n+1}} \langle n \rangle + \dots$$

定义 12.16 (Jaynes, 1968). 令 $\pi_0(\theta)$ 是连续型的参数 $\theta \in \Theta$ 的无信息先验, 连续型随机变量 $\theta \sim \pi(\theta)$ 的 Jaynes 熵定义为 $\pi_0(\theta)$ 到 $\pi(\theta)$ 的 Kullback-Leibler 散度 (见第 186 页的定义 2.43) 的相反数。

$$H(\theta) = -K(\pi/\pi_0) = -\int_{\Theta} \pi(\theta) \ln \frac{\pi(\theta)}{\pi_0(\theta)} d\theta \quad (12.9)$$

特别地, 如果 $\pi_0(\theta) = 1$, 则 θ 的 Jaynes 熵就是连续熵。

定理 12.2. 已知连续型参数 $\theta \in \Theta$ 的部分信息由约束方程组 $E[g_k(\theta)] = \mu_k, k = 1, 2, \dots, m$ 描述, 则以下分布 $\pi(\theta)$ 使得 θ 的 Jaynes 熵最大, 称为 θ 的最大熵先验。

$$\pi(\theta) = \frac{\pi_0(\theta) \exp \left\{ \sum_{k=1}^m \lambda_k g_k(\theta) \right\}}{\int_{\Theta} \pi_0(\theta) \exp \left\{ \sum_{k=1}^m \lambda_k g_k(\theta) \right\} d\theta} \quad (12.10)$$

其中系数 λ_k 由约束方程组决定。

例 12.21. 已知 $\theta \in \mathbb{R}$ 是位置参数, 其均值为 $E(\theta) = \mu$, 方差为 $V(\theta) = \sigma^2$, 若取 θ 的无信息先验 $\pi_0(\theta) = 1$, 则最大熵先验为

$$\pi(\theta) \propto \exp\{\lambda_1\theta + \lambda_2(\theta - \mu)^2\}$$

显然, θ 服从正态分布, 于是 $\theta \sim N(\mu, \sigma^2)$ 。特别注意, 若约束条件只有 $E(\theta) = \mu$, 则 θ 的最大熵先验不存在。但如果限制 $\theta \in (0, \infty)$, 则 θ 服从指数分布。

定义 12.17. 当参数 θ 的先验分布 $\pi_1(\theta|\xi)$ 中的超参数* $\xi \in \Xi$ 难以确定时, 可以通过超先验 $\pi_2(\xi)$ 给出 θ 的层级先验 (hierarchical prior) $\pi(\theta)$ 如下。

$$\pi(\theta) = \int_{\Xi} \pi_1(\theta|\xi) \pi_2(\xi) d\xi \quad (12.11)$$

例 12.22. 已知某疾病的患病率 θ 很小, 设其先验分布为 $\theta \sim U(0, \xi)$, 其中超参数 $\xi \sim U(0.01, 0.06)$ 。由式 (12.11), 参数 θ 的层级先验为

$$\begin{aligned} \pi(\theta) &= \frac{1}{0.06 - 0.01} \int_{0.01}^{0.06} \frac{I_{(0,\xi)}(\theta)}{\xi} d\xi \\ &= \begin{cases} 35.84 & \text{当 } 0 < \theta < 0.01 \\ 56.27 - 20 \ln \theta & \text{当 } 0.01 \leq \theta < 0.06 \\ 0 & \text{当 } 0.06 \leq \theta < 1 \end{cases} \end{aligned}$$

定义 12.18. 接着第 649 页的例 12.8, 若分布 $\pi_0(\theta)$ 不是完全确定的, 仅仅知道 $\pi_0(\theta)$ 属于参数 θ 的如下先验类

$$\varphi = \{\pi_0 : \pi_0(\theta) = \pi(\theta|\xi), \xi \in \Xi\}$$

其中 ξ 是将用边缘分布的信息来待定的超参数。满足下述条件的 $\hat{\pi} \in \varphi$ 被称为类型 II 最大似然先验, 简称 ML-II 先验, 它使得以往的观测数据 $\mathbf{X} = \mathbf{x}$ 以大概率出现。

$$\hat{\pi}_0 = \underset{\pi_0 \in \varphi}{\operatorname{argmax}} \prod_{j=1}^n m^{\pi_0}(x_j)$$

或者等价地, $\hat{\pi}_0(\theta) = \pi(\theta|\hat{\xi})$

$$\text{其中, } \hat{\xi} = \underset{\xi \in \Xi}{\operatorname{argmax}} \mathcal{L}(\xi; \mathbf{x}) = \underset{\xi \in \Xi}{\operatorname{argmax}} \prod_{j=1}^n m^{\pi(\theta|\xi)}(x_j)$$

*超参数 (hyperparameter) 即参数分布中的参数。例如, 总体 $X \sim N(\theta, 1)$, 其中参数 $\theta \sim N(0, \sigma^2)$, 而 σ^2 是一个未知的超参数。

例 12.23. 设总体 $X \sim N(\theta, \tau^2)$, 其中参数 τ^2 已知, 而 $\theta \sim N(\mu, \sigma^2)$, 按照[定义 12.18](#),

$$\varphi = \{\pi_0 : \pi_0(\theta) = \phi(\theta|\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \geq 0\}$$

设样本点 $X_j \sim p(x_j|\theta_j), j = 1, 2, \dots, n$ 相互独立, 由 X_j 的边缘密度 $m^{\pi_0}(x_j) = \phi(x_j|\mu, \tau^2 + \sigma^2)$ 得到似然函数 $\mathcal{L}(\mu, \sigma^2; \mathbf{x})$ 如下, 其中 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 。

$$\begin{aligned}\mathcal{L}(\mu, \sigma^2; \mathbf{x}) &= \prod_{j=1}^n \phi(x_j|\mu, \tau^2 + \sigma^2) \\ &= [2\pi(\tau^2 + \sigma^2)]^{-n/2} \exp\left\{\frac{-ns_n^2 - n(\mu - \bar{x})^2}{2(\tau^2 + \sigma^2)}\right\} \\ \text{其中, } \bar{x} &= \frac{1}{n} \sum_{j=1}^n x_j \text{ 且 } s_n^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2\end{aligned}$$

求解下面的方程组,

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mu} = 0 \\ \frac{\partial \mathcal{L}}{\partial \sigma^2} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \max(0, s_n^2 - \tau^2) \end{cases}$$

从过去的数据得到未知参数 θ 的 ML-II 先验 $\theta \sim N(\hat{\mu}, \hat{\sigma}^2)$ 。

12.2.3 后验分布及其期望的计算

已知简单随机样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 来自总体 $X \sim f(x|\boldsymbol{\theta})$, 其中 $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$ 是未知参数, 令其先验分布为 $\pi(\boldsymbol{\theta})$ 。设 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 是样本观测值, 有时候需要计算参数后验分布的某些数字特征, 如期望或各阶矩。在得到参数的后验分布 $\pi(\boldsymbol{\theta}|\mathbf{x})$ 之后, 不难计算这些数字特征, 如下例。

例 12.24. 在 n 重 Bernoulli 试验中, 事件 A 出现的次数 $X \sim B(n, \theta)$, 其中 n 已知, 未知参数 $\theta \in (0, 1)$ 的先验分别选为 Beta(1, 1) (即 U(0, 1)) 和 Beta(1/2, 1/2), 即 $\pi_1(\theta) = 1$ 和 $\pi_2(\theta) \propto [\theta(1-\theta)]^{-1/2}$ (见例 12.15)。若观察到 $X = x$, 求参数 θ 后验分布的期望。

解. 由表 12.1, 参数 θ 的后验分布分别为 Beta($x+1, n-x+1$) 和 Beta($x+0.5, n-x+0.5$), 基于练习 4.47 的结果得到参数的后验均值分别为

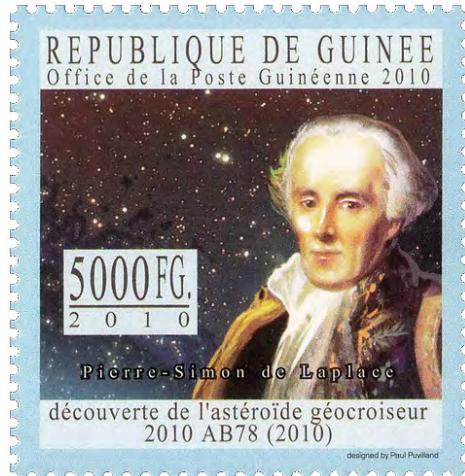
$$\begin{aligned} E_{\pi_1}(\theta|X=x) &= \frac{x+1}{n+2} \\ E_{\pi_2}(\theta|X=x) &= \frac{x+0.5}{n+1} \end{aligned} \quad (12.12)$$

在例 12.24 的条件之下, 式 (12.12) 被称作 Laplace 接续规则 (rule of succession)。即便极端地观察到 $x = 0$, 贝叶斯估计所得到的参数后验均值也是一个很小的值。同样 $x = n$ 时, 参数后验均值也仅仅是十分接近于 1 而已。请读者将之与频率派的点估计 $\hat{\theta} = x/n$ 对比一下看哪个看上去更合理。

对于接续规则, Laplace 解释说, “当某一事件发生的概率尚属未知时, 可以假定它是从 0 到 1 的任何一个值, 根据已知事件推断各种假设的概率是一个分数, 分子是对此假设下该事件发生的概率, 分母是在所有类似假设下该事件发生的概率之和。”在例 12.24 中, Laplace 所说的分数即是未知参数 θ 的后验分布 $\pi_1(\theta|X=x)$,

$$\begin{aligned} \pi_1(\theta|X=x) &\propto P(X=x|\theta)\pi_1(\theta) \\ &\propto \theta^x(1-\theta)^{n-x} \end{aligned}$$

Laplace 接着论述道, “如果在上述每个分数上再乘以相应假设下未来事件可能发生的概率, 然后再对所有可能假设得出的概率求和, 此时得到的结果将是根据已发生事件得到的该未来事件发生的概率。” Laplace 这段话说的就是计算后验预测分



布 $P(X_{\text{new}} = 1|X = x)$, 其中 $X_{\text{new}} \sim \theta \langle 1 \rangle + (1 - \theta) \langle 0 \rangle$ 。根据公式 (12.6),

$$\begin{aligned} P(X_{\text{new}} = 1|X = x) &= \int_0^1 P(X_{\text{new}} = 1|\theta) \pi_1(\theta|X = x) d\theta \\ &= \int_0^1 \theta \cdot \pi_1(\theta|X = x) d\theta \\ &= E_{\pi_1}(\theta|X = x) = \frac{x+1}{n+2} \end{aligned} \quad (12.13)$$

对于上式的解释是, 在 n 重 Bernoulli 试验中发现“正面”出现了 x 次, 若再做一次 Bernoulli 试验, “正面”出现的概率将是 $(x+1)/(n+2)$, 我们把式 (12.13) 也称作 Laplace 接续规则。

例 12.25. 请读者回顾第 75 页 Laplace 对“赌明天太阳照常升起”的胜算比例的那段话。太阳在过去的 1826213 天的试验里正常升起, 按照 Laplace 接续规则 (12.13), 太阳明天照常升起的概率是 $1826214/1826215 \approx 0.9999995$ 。

如果参数的后验分布难于计算, 可以通过一些近似的方法求得后验分布的期望, 例如下面的定理 12.3 和定理 12.4, 它们对于一类常见的参数后验分布都是适用的。

定理 12.3. 已知参数的后验分布 $\pi(\boldsymbol{\theta}|X = \mathbf{x}) = \mathcal{L}(\boldsymbol{\theta}; \mathbf{x})\pi(\boldsymbol{\theta}) = \exp\{nh(\boldsymbol{\theta})\}$, 其中 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 。设非负函数 $g(\boldsymbol{\theta})$ 和 $h(\boldsymbol{\theta})$ 满足 Laplace 近似积分法的条件 (见第 773 页的定理 E.13), 且 $h(\boldsymbol{\theta})$ 的极大值点为 $\hat{\boldsymbol{\theta}}$, 则

$$E\{g(\boldsymbol{\theta})|X = \mathbf{x}\} = g(\hat{\boldsymbol{\theta}}) \left\{ 1 + O\left(\frac{1}{n}\right) \right\} \quad (12.14)$$

证明. 参数 $g(\boldsymbol{\theta})$ 的后验分布的期望是

$$E\{g(\boldsymbol{\theta})|X = \mathbf{x}\} = \frac{\int_{\Theta} g(\boldsymbol{\theta}) \exp\{nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int_{\Theta} \exp\{nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}} \quad (12.15)$$

由定理 E.13, 不难推得结果 (12.14)。 \square

定理 12.4. 在定理 12.3 的条件下, 令 $nh_*(\boldsymbol{\theta}) = nh(\boldsymbol{\theta}) + \ln(g(\boldsymbol{\theta}))$, 若 $\hat{\boldsymbol{\theta}}_*$ 和 $\hat{\boldsymbol{\theta}}$ 分别是 $h_*(\boldsymbol{\theta})$ 和 $h(\boldsymbol{\theta})$ 的极大值点, 并且 $A_* = -\nabla^2 h(\hat{\boldsymbol{\theta}}_*)$ 和 $A = -\nabla^2 h(\hat{\boldsymbol{\theta}})$ 都是正定矩阵, 则

$$E\{g(\boldsymbol{\theta})|X = \mathbf{x}\} = \sqrt{\frac{\det(A_*)}{\det(A)}} \frac{\exp\{nh_*(\hat{\boldsymbol{\theta}}_*)\}}{\exp\{nh(\hat{\boldsymbol{\theta}})\}} \left\{ 1 + O\left(\frac{1}{n^2}\right) \right\} \quad (12.16)$$

证明. 利用定理 E.13 分别求 (12.15) 中分子和分母的 Laplace 近似, 便可得到结果 (12.16)。 \square

例 12.26. 接着**例 12.24**, 当试验重复多次, 即 n 取得很大的时候, 利用式 (12.14) 可得近似结果 $E_{\pi_1}(\theta|X=x) \approx x/n$, 与经典结果基本相同; 而 $E_{\pi_2}(\theta|X=x) \approx (x-0.5)/(n-1)$ 与精确解 $(x+0.5)/(n+1)$ 也相差无几。

例 12.27. 已知简单随机样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ 来自总体 $\text{Poisson}(\theta)$, 其中参数 θ 未知。假设先验分布 $\theta \sim \text{Gamma}(\alpha, \beta)$, 即 $\pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$ 。在观察到 $\mathbf{X} = \mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 之后, 参数 θ 的后验分布为

$$\theta|\mathbf{X}=\mathbf{x} \sim \text{Gamma}(\tilde{\alpha}, \tilde{\beta}), \text{ 其中 } \tilde{\alpha} = \alpha + \sum_{j=1}^n x_j \text{ 且 } \tilde{\beta} = \beta + n$$

由**性质 4.24** 得到参数的后验均值的精确解 $E(\theta|\mathbf{X}=\mathbf{x}) = \tilde{\alpha}/\tilde{\beta}$ 。下面分别用式 (12.14) 和 (12.16) 来近似地求 $E(\theta|\mathbf{X}=\mathbf{x})$, 第二个近似解比第一个近似解更精确一些。

□ 令 $nh(\theta) = \ln \mathcal{L}(\mathbf{x}; \theta) + \ln \pi(\theta)$, 其中似然函数 $\mathcal{L}(\theta; \mathbf{x}) = \theta^{\sum_{j=1}^n x_j} e^{-n\theta}$ 。显然, 函数 $\mathcal{L}(\theta; \mathbf{x})$ 及 $\pi(\theta)$ 中的常数因子并不影响求解 $h(\theta)$ 的极大值点。将 $\pi(\theta) = \theta^{\alpha-1} e^{-\beta\theta}$ 和 $\mathcal{L}(\theta; \mathbf{x})$ 代入, 求得

$$\begin{aligned} nh(\theta) &= \sum_{j=1}^n x_j \ln \theta - n\theta + (\alpha-1) \ln \theta - \beta\theta \\ &= (\tilde{\alpha}-1) \ln \theta - \tilde{\beta}\theta \end{aligned}$$

求得 $h(\theta)$ 的极大值点 $\hat{\theta} = (\tilde{\alpha}-1)/\tilde{\beta}$, 根据式 (12.14),

$$E(\theta|\mathbf{X}=\mathbf{x}) \approx \frac{\tilde{\alpha}-1}{\tilde{\beta}}$$

□ 令 $nh_*(\theta) = nh(\theta) + \ln \theta = \tilde{\alpha} \ln \theta - \tilde{\beta}\theta$, 先求得 $h_*(\theta)$ 的极大值点 $\hat{\theta}_* = \tilde{\alpha}/\tilde{\beta}$, 进而求得 $A_* = -h''_*(\hat{\theta}_*) = \tilde{\beta}^2/(n\tilde{\alpha})$ 。根据 $h(\theta)$ 的极大值点 $\hat{\theta} = (\tilde{\alpha}-1)/\tilde{\beta}$, 求得 $A = -h''(\hat{\theta}) = \tilde{\beta}^2/(n(\tilde{\alpha}-1))$ 。将所求的 $A, \hat{\theta}, A_*, \hat{\theta}_*$ 代入到公式 (12.16) 得到,

$$\begin{aligned} E(\theta|\mathbf{X}=\mathbf{x}) &\approx \sqrt{\frac{\tilde{\alpha}}{\tilde{\alpha}-1}} \frac{\left(\frac{\tilde{\alpha}}{\tilde{\beta}}\right)^{\tilde{\alpha}} \exp\{-\tilde{\alpha}\}}{\left(\frac{\tilde{\alpha}-1}{\tilde{\beta}}\right)^{\tilde{\alpha}-1} \exp\{-(\tilde{\alpha}-1)\}} \\ &= \frac{\tilde{\alpha}}{\tilde{\beta}} \left(\frac{\tilde{\alpha}}{\tilde{\alpha}-1}\right)^{\tilde{\alpha}-1/2} e^{-1} \end{aligned}$$

例 12.28 (贝叶斯回归模型). 考虑线性模型 $\mathbf{X} = A\boldsymbol{\beta} + \boldsymbol{\epsilon}$, 其中 \mathbf{X} 是 n 维观察向量,

β 是 k 维回归系数, A 是 $n \times k$ 常数矩阵, $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^\top$ 是误差向量, 其中 $\epsilon_1, \epsilon_2, \dots, \epsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ 且 σ^2 未知。在得到观察数据 $X = x$, 令

$$\begin{aligned}\hat{\beta} &= (A^\top A)^{-1} A^\top x \\ \hat{x} &= A\hat{\beta} \\ s^2 &= \frac{1}{m}(x - \hat{x})^\top(x - \hat{x}), \text{ 其中 } m = n - k\end{aligned}$$

利用推论 10.1, X 的密度函数 (10.7) 可整理为

$$f(x|\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{ms^2 + (\beta - \hat{\beta})^\top A^\top A(\beta - \hat{\beta})}{2\sigma^2}\right\}$$

令 $\pi(\sigma^2, \beta) \propto \sigma^{-2}$ 为参数的先验分布, 参数的后验分布为

$$\begin{aligned}\pi(\beta, \sigma^2 | X = x) &\propto f(x|\beta, \sigma^2) \times \pi(\sigma^2, \beta) \\ &\propto (\sigma^2)^{-k/2} \exp\left\{-\frac{(\beta - \hat{\beta})^\top A^\top A(\beta - \hat{\beta})}{2\sigma^2}\right\} (\sigma^2)^{-m/2-1} \exp\left\{-\frac{ms^2}{2\sigma^2}\right\} \\ &\propto \pi(\beta|\hat{\beta}, \sigma^2) \times \pi(\sigma^2|s^2)\end{aligned}$$

由逆 χ^2 分布的定义 4.18,

$$\begin{aligned}\beta|\hat{\beta}, \sigma^2 &\sim N\left(\hat{\beta}, \sigma^2(A^\top A)^{-1}\right) \\ \frac{\sigma^2}{ms^2} \Big| s^2 &\sim \chi_m^{-2}\end{aligned}$$

线性模型中回归参数的后验分布为

$$\begin{aligned}\pi(\beta|X = x) &= \frac{\Gamma\left(\frac{m+k}{2}\right) \sqrt{|A^\top A|} s^{-k}}{\Gamma\left(\frac{m}{2}\right) \sqrt{(m\pi)^k}} \left[\frac{(\beta - \hat{\beta})^\top A^\top A(\beta - \hat{\beta})}{ms^2} + 1 \right]^{-\frac{m+k}{2}} \\ \beta|X = x &\sim t_m\left(\hat{\beta}, s^2(A^\top A)^{-1}\right), \text{ 多元 } t \text{ 分布见第 339 页的定义 4.39} \\ \frac{(\beta - \hat{\beta})^\top A^\top A(\beta - \hat{\beta})}{ks^2} &\sim F_{k,m}\end{aligned}$$

12.2.4 贝叶斯模型选择

模型选择 (model selection) 在经典统计学里是个棘手的问题：模型太简单解释不了观察数据，模型太复杂有可能对观察数据拟合过度而失去一般化的能力。举个极端的例子：对于 $k+1$ 个二维点 $(x_j, y_j)^\top \in \mathbb{R}^2, j = 0, 1, \dots, k$ ，总存在 k 次的 Lagrange 插值多项式 $L_k(x)$ 穿过这 $k+1$ 个点，其中

$$L_k(x) = \sum_{j=0}^k y_j \prod_{\substack{i=0 \\ i \neq j}}^k \frac{x - x_i}{x_j - x_i}$$

其实，穿过这给定 $k+1$ 个点的函数有无穷多， $L_k(x)$ 只是其中之一。这类函数在观察数据上的误差虽然为零，但往往不能很好地描述数据的产生机制，也不具备良好的预测能力，在实践中无法将模型的效果泛化。

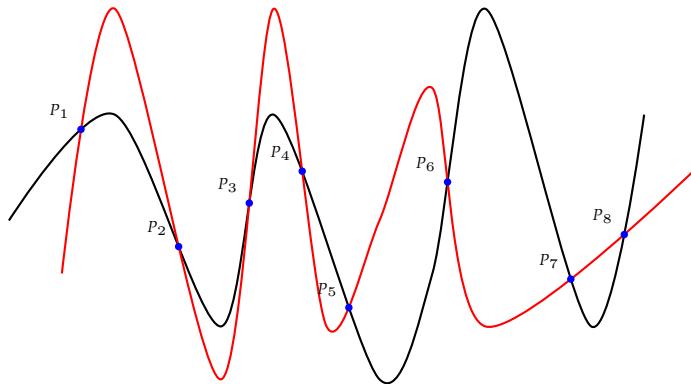


图 12.6: 穿过有限样本点的曲线有无穷多，到底哪个能揭示出数据的本质？

模型选择在贝叶斯统计学里却是直接了当的。令 m 是由参数 $\boldsymbol{\theta}_m$ 定义的一类模型，例如 $\beta_0 + \beta_1 x + \dots + \beta_k x^k + \epsilon$ ，其中 $\epsilon \sim N(0, \sigma^2)$ ，参数是 $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_k, \sigma^2)^\top$ 。在给了观察数据 D 之后，我们可以计算似然 $P(D|m)$ ，合理的模型应该使得该似然尽可能地大。

定义 12.19. 给定观察数据 D ，模型 m 的贝叶斯证据 (Bayesian evidence) 是如下定义的边缘似然，它是 $P(D|\boldsymbol{\theta}_m, m)$ 的加权平均。

$$P(D|m) = \int_{\Theta_m} P(D|\boldsymbol{\theta}_m, m) p(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m \quad (12.17)$$

12.2.5 层级贝叶斯模型

如果所考虑的统计问题中的若干参数（包括隐藏参数）之间有某种内在的联系，对这些参数的研究往往需要借助层级模型 (hierarchical model)。譬如，在实际应用中我们常会遇到这样的困难——观察样本并非来自于同一个总体，而是若干个总体，这些总体分布中的参数被某些条件制约着。制约条件经常通过参数的联合分布来描述，下面给出一个极端的例子，参数先验地独立同分布于某个分布。

例 12.29. 有 $N = 12$ 家医院，令 X_i 表示医院 i 在 n_i 次婴儿心脏手术中的死亡数（即手术失败的次数），其中 $i = 1, 2, \dots, 12$ ，则 $X_i \sim \text{B}(n_i, \theta_i)$ ，其中表示手术失败概率的参数 θ_i 未知。假设先验地有 $\theta_1, \theta_2, \dots, \theta_{12} \stackrel{\text{iid}}{\sim} \text{U}(0, 1)$ ，即各家医院手术失败概率相互独立。对 n_i, X_i 有如下的观察结果，请给出这些未知参数的后验分布。

表 12.2: 12 家医院的婴儿心脏手术的次数，以及手术失败的次数。

医院	1	2	3	4	5	6	7	8	9	10	11	12
n_i	47	148	119	810	211	196	148	215	207	97	256	360
X_i	0	18	8	46	8	13	9	31	14	8	29	24

解. 记 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{12})^\top, \mathbf{x} = (0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24)^\top$ ，则

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta}) \propto \prod_{i=1}^N \theta_i^{x_i} (1 - \theta_i)^{n_i - x_i}$$

显然 $\pi(\theta_i|\mathbf{X} = \mathbf{x}) \propto \theta_i^{x_i} (1 - \theta_i)^{n_i - x_i}$ ，所以 $\theta_i|\mathbf{X} = \mathbf{x} \sim \text{Beta}(x_i + 1, n_i - x_i + 1)$ 。参数 θ_i 的后验分布可以独立地进行更新，其后验均值和后验方差分别为

$$\begin{aligned} \text{E}(\theta_i|\mathbf{X} = \mathbf{x}) &= \frac{x_i + 1}{n_i + 2}, \text{ 即 Laplace 接续规则 (12.16)} \\ \text{V}(\theta_i|\mathbf{X} = \mathbf{x}) &= \frac{(x_i + 1)(n_i - x_i + 1)}{(n_i + 2)^2(n_i + 3)} \end{aligned}$$

例 12.29 中的独立同分布假设 $\theta_1, \theta_2, \dots, \theta_{12} \stackrel{\text{iid}}{\sim} \text{U}(0, 1)$ 意味着各家医院的治疗效果是固定的，之间没有关联。在实践中，一个更合理的假设是这些治疗效果是类似的。利用 logit 函数^{*}将 θ_i 映到 \mathbb{R} 上，再假设它们服从一个正态分布，见下面的例子。

^{*}logit 函数 $\text{logit}(p)$ 是 sigmoid 函数（见第 ?? 页）的逆函数，通常用来将 $(0, 1)$ 区间映射为整个实数域 \mathbb{R} ，该函数的具体定义如下（请读者利用 GnuPlot 绘制它的函数图像）。

$$\text{logit}(p) = \ln \frac{p}{1-p} = \ln p - \ln(1-p), \text{ 其中 } 0 < p < 1$$

例 12.30. 接着**例 12.29**, 假设 $b_i = \text{logit}(\theta_i) \sim N(\mu, \sigma^2)$, 其中 $i = 1, 2, \dots, 12$, 令 $\mu \sim N(0, 10^6)$, 令 $\tau = 1/\sigma^2, \tau \sim \text{Gamma}(0.001, 0.001)$ 。我们可以用一个草图来描绘这个框架, 这样做好处是直观性非常强。

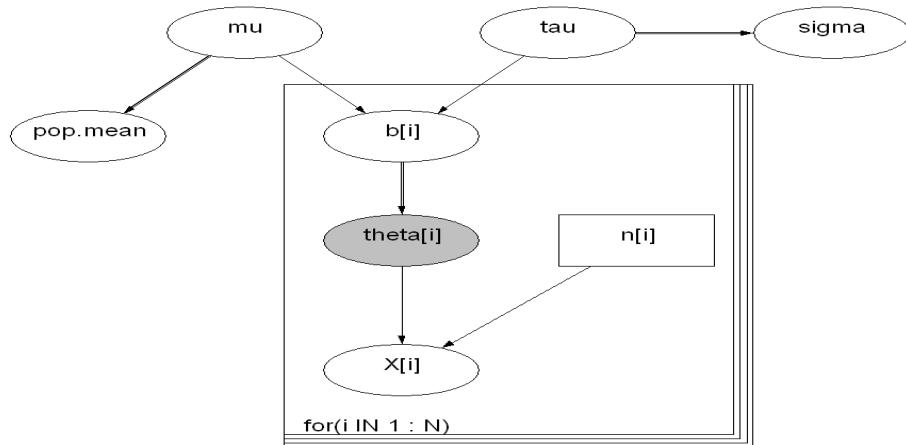


图 12.7: 例 12.30 的图模型 (graphical model) 以涂鸦方式描绘了变量间的关系。

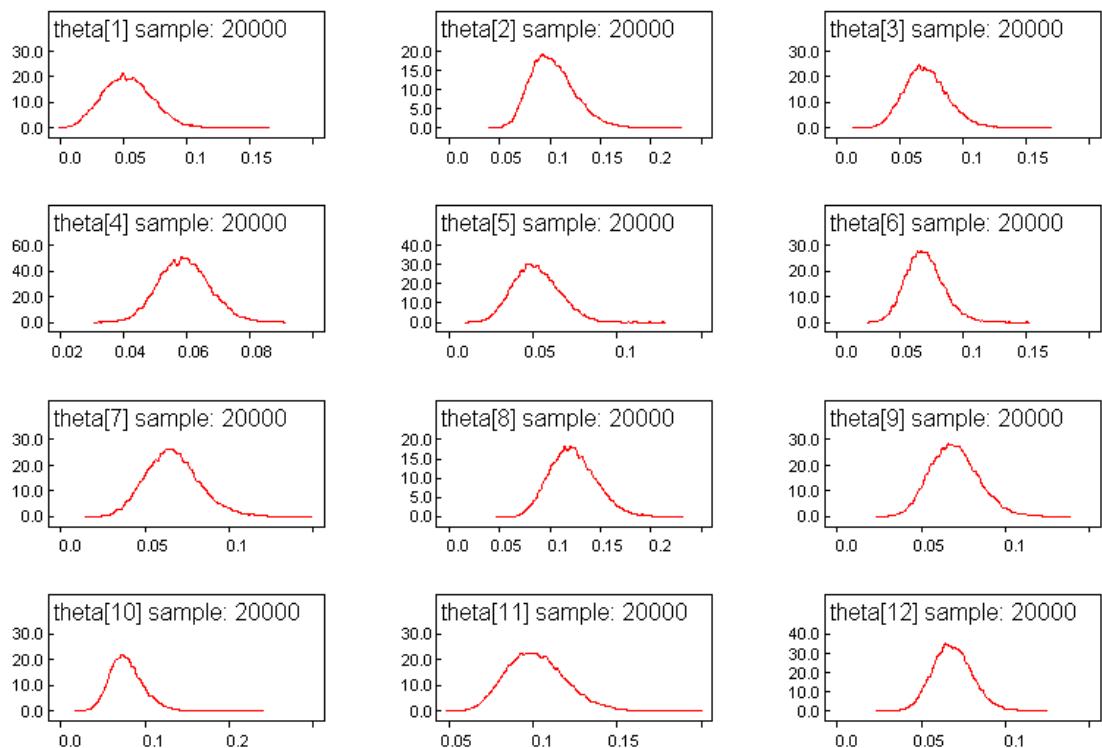


图 12.8: 利用 BUGS 工具模拟例 12.30 中的参数 $\theta_1, \dots, \theta_{12}$ 的后验分布。

例 12.31 (取自 BUGS 案例集). 观察 10 个电力设备, 其中设备 i 在长度为 t_i 的时间

段内发生故障的次数为 X_i , 满足

$$X_i \sim \text{Poisson}(\theta_i t_i), \text{ 其中 } i = 1, 2, \dots, 10$$

其中, 未知参数 θ_i 是设备 i 在单位时间内的平均故障数。现观察到在时间段 t_i 内的故障数 X_i 如下, 试给出 θ_i 的后验分布。

表 12.3: 电力设备 $i = 1, 2, \dots, 10$ 在时间段 t_i 内发生故障次数的观察结果。

设备	1	2	3	4	5	6	7	8	9	10
t_i	94.3	15.7	62.9	126	5.24	31.4	1.05	1.05	2.1	10.5
X_i	5	1	5	14	3	19	1	1	4	22

解. 用 BUGS 实现模型: $\theta_1, \dots, \theta_{10} \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$, 其中超参数 $\alpha \sim \text{Expon}(1)$, $\beta \sim \text{Gamma}(0.1, 1)$ 。

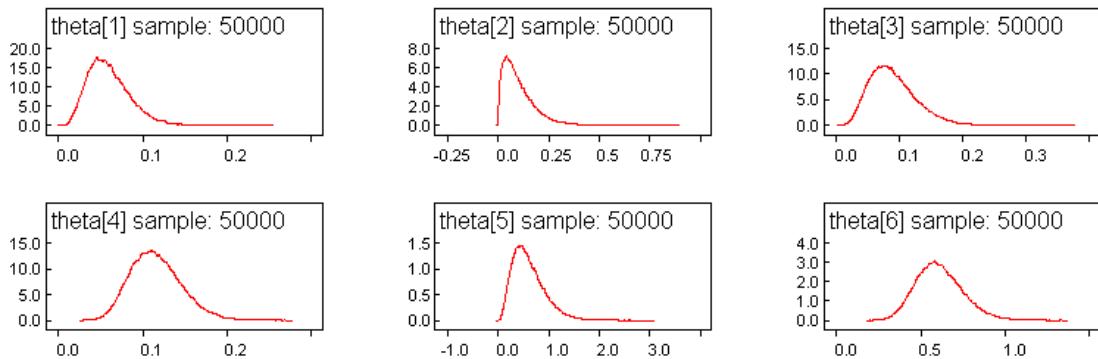


图 12.9: 例 12.31 中的部分参数经过 BUGS 工具模拟的后验分布。

把例 12.29、例 12.30、例 12.31 的问题一般化: 已知非简单随机样本 $X_i \sim p(x|\theta_i), i = 1, 2, \dots, N$, 其中未知参数 $\theta_1, \dots, \theta_N$ 不一定来自于同一个总体, 在信息缺乏的时候, 贝叶斯统计学总是假定它们是可交换的, 即 $\theta_1, \dots, \theta_N$ 的联合密度函数 $\pi(\theta_1, \dots, \theta_N)$ 是个对称函数, 这样的先验分布称为满足可交换性 (exchangeability)^{*}。最简单的情形是 $\theta_1, \dots, \theta_N \stackrel{\text{iid}}{\sim} \pi(\theta|\beta)$, 此时 $\theta = (\theta_1, \dots, \theta_N)^\top$ 的密度函数为

$$\pi(\theta|\beta) = \prod_{i=1}^N \pi(\theta_i|\beta)$$

^{*}显然, 独立同分布的随机变量是可交换的。由于贝叶斯推断以数据为中心, 而不是以模型为中心, 所以贝叶斯建模在很多情况下并不假设从同一个总体中抽样。在没有任何信息能断言 $\theta_1, \dots, \theta_N$ 中哪个更特殊些的时候, 可交换性就成为贝叶斯学派要求 $\theta_1, \dots, \theta_N$ 的先验分布必须满足的一个基本条件。

这里未知参数 β 称为超参数，假设它来自某个预定的分布 $\pi(\beta)$ 。

意大利数学家、贝叶斯学派统计学家 Bruno de Finetti (1906-1985)

在观察到样本 $X = x$ 之后，则贝叶斯模型为

$$\begin{aligned} p(\beta, \theta|x) &\propto p(\beta, \theta)p(x|\beta, \theta) \\ &= \pi(\beta)\pi(\theta|\beta)p(x|\theta) \end{aligned}$$

上式中 $p(x|\beta, \theta) = p(x|\theta)$ 是因为 $X = (X_1, \dots, X_N)^\top$ 的分布只依赖于 θ ，超参数 β 通过 θ 来影响 X 的观察结果。

下面介绍自然语言处理中两个主题模型 (topic model)，分别是 Thomas Hofmann 于 1999 年提出的概率潜在语义标引 (probabilistic latent semantic indexing, PLSI) 模型 [75]，David Blei 等人于 2003 年提出的潜在 Dirichlet 分配 (latent Dirichlet allocation, LDA) 模型 [18]，二者都是层级贝叶斯模型。

不妨设整个词集的规模为 V ，即我们只关注这 V 个词，其他的词都被忽略了。已知语料是由 M 篇文本构成，其中第 m 篇文本有 N_m 个词。不计较词的出现次序，一个文本可以抽象为一个“词袋”(a bag of words)^{*}，里面装着 N_m 个词（允许有重复的词）。假设共有 K 个主题 $1, \dots, K$ ，它们隐藏在文字背后，是无法观测的。

例 12.32 (PLSI 模型). 假设文档 $d \in \{d_1, \dots, d_M\}$ 是这样生成的：独立地重复以下两个步骤直至得到整个词袋。

- 从分布 $P(z|d)$ 产生一个主题 $z \in \{z_1, \dots, z_K\}$ ；
- 在主题 z 之下，从分布 $P(w|z)$ 产生词 w 。

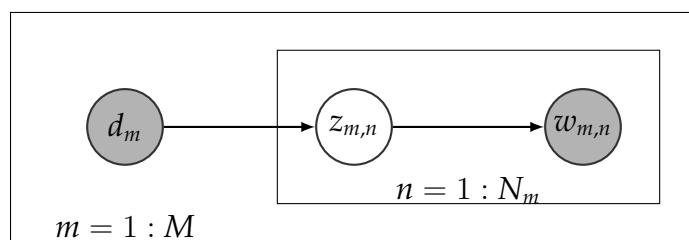


图 12.10: PLSI 模型是一个两层的贝叶斯模型：隐藏变量 $z_{m,n}$ 表示文档 d_m 中第 n 个词 $w_{m,n}$ 的主题。深色节点表示可观测变量，白色节点表示不可观测变量。

可观测变量 d, w 的联合分布是

$$\begin{aligned} P(d, w) &= P(d) \sum_z P(z|d)P(w|z) \\ &= \sum_z P(z)P(d|z)P(w|z), \text{ 因为 } d \perp\!\!\!\perp_z w \end{aligned}$$

^{*}词袋忽略了词序等信息，这种简化对某些应用是合理的，如文本分类、主题分析等。

上述结果可类比于奇异值分解 (E.2)。因为奇异值分解是潜在语义标引 (latent semantic indexing, LSI) 的基础，上述结果便是 PLSI 名字的由来。请注意：它不是矩阵的奇异值分解，仅仅是一个类比而已。

$$P(d = d_m, w = w_n) = \sum_{k=1}^K P(z_k) P(d_m|z_k) P(w_n|z_k)$$

即， $A = U\Sigma V^\top$

其中， $A_{mn} = P(d = d_m, w = w_n)$, $m = 1, \dots, M; n = 1, \dots, N$

$$U_{mk} = P(d_m|z_k), \quad i = 1, \dots, m; k = 1, \dots, K$$

$$\Sigma = \text{diag}(P(z_1), \dots, P(z_K))$$

$$V_{nk} = P(w_n|z_k), \quad n = 1, \dots, N; k = 1, \dots, K$$

由独立性假设，对数似然是 $\sum_{d,w} n(d,w)P(d,w)$ ，其中 $n(d,w)$ 表示文档 d 中 w 出现的次数。下面，我们利用类 EM 算法求 $P(d), P(z|d), P(w|z)$ 最大化该对数似然。

□ E 步骤：观察到 d, w ，主题 z 的后验概率是

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_z P(z)P(d|z)P(w|z)}$$

□ M 步骤：依次更新 $P(w|z), P(d|z), P(z)$ 如下。

$$P(w|z) = \frac{1}{S} \sum_d n(d,w)P(z|d,w), \quad \text{其中 } S = \sum_{d,w} n(d,w)P(z|d,w)$$

$$P(d|z) = \frac{1}{S} \sum_w n(d,w)P(z|d,w)$$

$$P(z) = \frac{1}{R} \sum_{d,w} n(d,w)P(z|d,w), \quad \text{其中 } R = \sum_{d,w} n(d,w) \text{ 是所有文档里的总词数}$$

例 12.33 (LDA 模型). LDA 主题模型认为，一个主题 k 就是给定词集上某个待定的类别分布，记作 $\text{Cat}(\boldsymbol{\varphi}_k)$ ，其中参数 $\boldsymbol{\varphi}_k \in \Delta_{V-1}, k = 1, 2, \dots, K$ 满足 $\boldsymbol{\varphi}_{1:K} \stackrel{\text{iid}}{\sim} \text{Dir}(\boldsymbol{\beta})$ 。

LDA 主题模型从观察数据（即语料）中“学习”出主题，并为每个词都标注一个主题。LDA 主题模型从语言生成的角度解释第 m 个“词袋”的产生过程（见图 12.11）：

- 随机抽取 $\boldsymbol{\theta}_m \sim \text{Dirichlet}(\boldsymbol{\alpha})$;
- 随机产生 N_m 个主题 $z_{m,1:N_m} \stackrel{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\theta}_m)$;
- 独立产生单词 $w_{m,n} \sim \text{Cat}(\boldsymbol{\varphi}_{z_{m,n}})$, 其中 $n = 1, 2, \dots, N_m$ 。

因为主题次数这一 K 维随机向量服从 Dirichlet-多项分布 $\text{Dirichlet-Multin}(N_m; \boldsymbol{\theta}_m; \boldsymbol{\alpha})$ (见定义 4.37), 所以“词袋”也可以通过下面的方法生成:

- 产生 $(n_1, \dots, n_K)^\top \sim \text{Dirichlet-Multin}(N_m; \boldsymbol{\theta}_m; \boldsymbol{\alpha})$;
- 对于主题 $k = 1, \dots, K$, 从 $\text{Cat}(\boldsymbol{\varphi}_k)$ 独立产生 n_k 个样本。

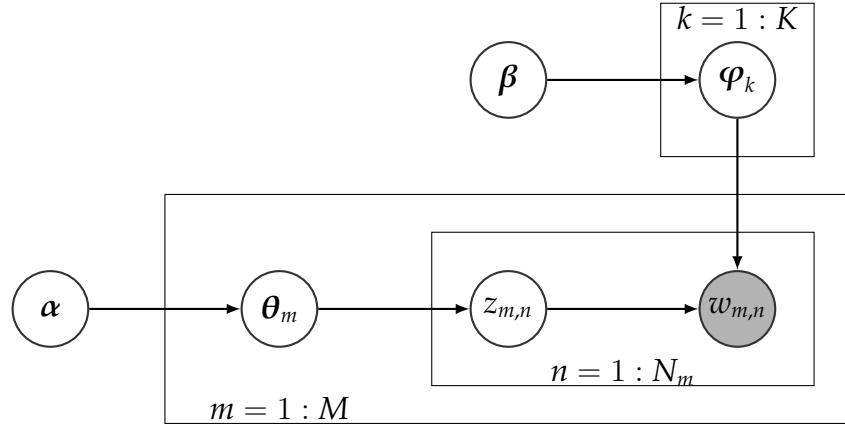


图 12.11: LDA 主题模型是一个三层的贝叶斯模型, 其中 $\boldsymbol{\theta}_{1:M} \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\alpha})$, $\boldsymbol{\varphi}_{1:K} \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\beta})$, N_m 个词的主题 $z_{m,1:N_m} \stackrel{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\theta}_m)$, 第 n 个词 $w_{m,n} \sim \text{Cat}(\boldsymbol{\varphi}_{z_{m,n}})$ 。

12.3 习题

- 12.1. 试证明第 645 页的性质 12.2。
- 12.2. 设总体 $X \in \mathcal{X} = \mathbb{R}$ 的密度函数具有形式 $\sigma^{-1}f(\sigma^{-1}x)$, 其中 $\sigma > 0$, 则称 σ 为尺度参数, 称 $\sigma^{-1}f(\sigma^{-1}x)$ 为尺度密度函数, 例如 $N(0, \sigma^2)$, 参数 α 固定的 $\text{Gamma}(\alpha, \beta)$ 等。试证明: 在尺度变换 $Y = cX, c > 0$ 的不变性之下, 尺度参数 σ 的无信息先验分布为非正常先验 $\pi(\sigma) = \sigma^{-1}$ 。
- 12.3. 设总体 $\mathbf{X} \sim \text{Multin}(\theta_1, \theta_2, \dots, \theta_k)$, 求未知参数 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)^\top$ 的 Jeffreys 先验分布。
- 12.4. 设总体 $X \sim \text{Expon}(\beta)$, 取参数 β 的无信息先验为 $\pi(\beta) = \beta^{-1}$ 。已知来自总体 X 的简单随机样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$, 求参数的后验分布 $\beta|\mathbf{X} = \mathbf{x}$ 。

第三部分

概率统计中的一些实用算法

概率统计中的计算

横看成岭侧成峰，远近高低各不同。不识庐山真面目，只缘身在此山中。

苏轼《题西林壁》

计算是数学和科学中必不可少的，无论是数值计算还是符号计算，都是实际应用离不开的手段。从里耶秦简上刻着的乘法九九口诀、古老的算盘到计算机再到云计算，计算工具的革命带来了社会生产力的解放，计算成本越来越小、计算效率越来越高让很多理论得以实现。



在计算的机械化和自动化的历史中，英国数学家、机械工程师 Charles Babbage (1791-1871) 是计算机的先驱，他提出的差分机和分析机虽然没有最终完成，但机械设计之精巧令人叹为观止，也提出了条件语句、循环语句、寄存器等重要想法。英国数学家、逻辑学家、计算机科学家、密码学家 Alan Turing (1912-1954) 的人生更具传奇色彩，他被视为计算机科学和人工智能之父。1936 年，Turing 在论文《论可计算数及其在判定问题上的应用》中提出 Turing 机的概念，论证了没有一个通用的算法能够判定任意一个 Turing 机是否会停机，证明了判定问题是无解的。Turing 机是理论计算机模型，有许多变种，在计算能力上是等价的。Turing 机可识别的形式语言是递归可枚举语言 (recursively enumerable language)，是 Chomsky 层级中的 0 型语言。

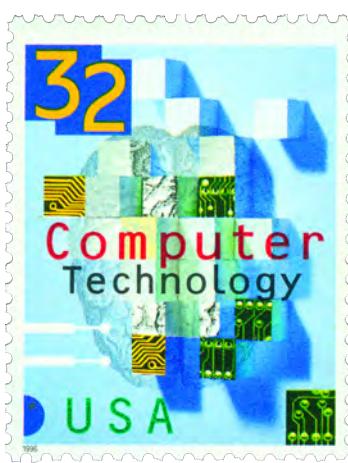


天才的美籍匈牙利裔数学家、物理学家、计算机科学家 John von Neumann (1903-1957) 是计算机之父和博弈论之父。1945 年，他奠定了现代计算机逻辑结构设计基础，遵循他的设计原则的计算机统称为 von Neumann 机。

计算机最初用于数值分析，目标在于最优化、求解线性方程组、矩阵分解、函数求值、曲线拟合、积分计算、解微分方程等，按照求解方式，有直接法（如，高斯消去法、单纯形法等）和迭代法（如，Newton-Raphson 法、共轭梯度法等）之分。

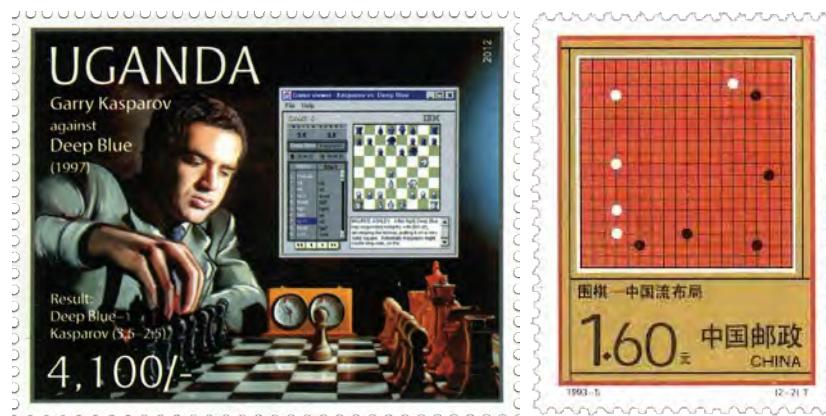
当找不到或难以找到确定性的算法的时候，人们不得不求助于确定性的近似算法和随机算法。前者不断地逼近真实解，后者以某个概率逼近真实解。确定性的算法和随机算法是一对姊妹，它们对于计算来说就像硬币的两面。有些实际问题，如旅行推销员问题（见第 741 页的例 15.21），用确定性的算法复杂度太高，用随机算法反倒“柳暗花明又一村”。在计算领域，“看，现在是妹妹要救姐姐，等一会那个姐姐一定会救妹妹的。”（《大话西游》）我们要重视随机算法，它们往往有意想不到的效果，甚至可以把计算能力拓展到极致。

另外，在计算机图形学、计算机辅助设计、计算机视觉、人工智能等都需要各式各样的计算。计算技术的进步促进了数学和自然科学的发展，反之亦然。



John von Neumann 说，“如果你能明确地告诉我机器做不了什么，我总能造出一台机器就做这个。”人工智能在很多领域已经超过了人类：1997 年，IBM 的超

级计算机“深蓝”(Deep Blue)打败了俄罗斯国际象棋特级大师、世界冠军 Garry Kasparov (1963-)。2016 年, DeepMind 的程序 AlphaGo 打败了人类围棋的所有顶级大师, 其核心技术包括强化学习和 Monte Carlo 树搜索。随机模拟技术在人工智能的发展过程中起到了至关重要的作用, 其重要性将逐渐被更多的研究者意识到。John von Neumann 坚信, “真相太复杂, 只允许近似”。



在本书的第三部分, 我们将介绍概率统计在实际应用中常用的一些算法。具体包括隐 Makrov 模型的三个经典算法, 一般图模型的团树算法, 参数估计的期望最大化算法, 随机模拟的 Markov 链 Monte Carlo (MCMC) 算法、Gibbs 抽样、模拟退火算法、数据增扩算法等。

第十三章

概率图模型

天网恢恢，疏而不失。

《老子》

利用图来刻画随机变量之间的条件独立性，以及利用图论来计算条件概率的模型称为概率图模型，简称图模型 (graphical model)，例如第 152 页的例 2.34。很多概率模型可以归类到图模型，例如第 43 页的例 1.29 提到的隐 Markov 模型，它是一种图模型——Bayes 网络。

我们用图来描述随机变量之间的依赖关系，有助于直观地理解概率模型。

13.1 隐 Markov 模型及其算法

现 在我们为第 43 页的例 1.29 所描述的问题构建数学模型。有 n 个状态（即盒子）构成的 Markov 链，状态是不可观察的，标号为 $S = \{1, 2, \dots, n\}$ ；有 m 个不同的观察结果（即不同的颜色），标号为 $\Omega = \{\omega_1, \omega_2, \dots, \omega_m\}$ 。设随机变量 Z_t, X_t 分别为 t 时刻系统所处的状态和观察到的结果。模型的参数描述如下，

- 状态的初始分布为 $Z_1 \sim \pi_1 \langle 1 \rangle + \dots + \pi_n \langle n \rangle$ ，记 $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^\top$ 。
- 从状态 i 转移到状态 j 的概率为 $P(z_t = j | z_{t-1} = i) = a_{ij}$ 。显然， a_{ij} 满足

$$\sum_{j=1}^n a_{ij} = 1, \text{ 其中 } i = 1, 2, \dots, n$$

称矩阵 $A = (a_{ij})_{n \times n}$ 为状态转移矩阵 (transition matrix)，它是一个 Markov 矩阵。

- 在状态 i 观察到第 k 个结果的概率为 $P(x_t = \omega_k | z_t = i) = b_{ik}$ ，满足

$$\sum_{k=1}^m b_{ik} = 1, \text{ 其中 } i = 1, 2, \dots, n$$

称矩阵 $B = (b_{ik})_{n \times m}$ 为发射矩阵 (emission matrix)，其第 i 行就是状态 i 下观察结果的分布 $b_{i1} \langle \omega_1 \rangle + \dots + b_{im} \langle \omega_m \rangle$ 。有时为了强调观察结果，也把 b_{ik} 记作 $b_i(k)$ 。

上述模型称为隐 Markov 模型 (hidden Markov model, HMM)，记作 $\mathcal{M} = \langle S, \Omega, A, B, \boldsymbol{\pi} \rangle$ ，其中 $\boldsymbol{\theta} = (A, B, \boldsymbol{\pi})$ 称为隐 Markov 模型的参数。

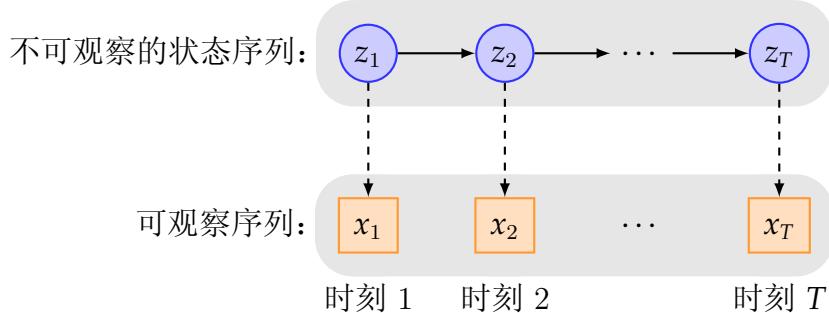


图 13.1: 不可观察的状态序列 $z_1 z_2 \dots z_T$ 是一个 Markov 过程。

例 13.1. 彝族姑娘阿诗玛的内心状态分为“愉快”和“忧郁”，是不可观察的。然而，她的外在行为“读书”、“发呆”和“弹琴”是可观察的。譬如，在“愉快”的状态下，我们观察到阿诗玛“读书”、“发呆”、“弹琴”的概率分别为 0.5, 0.1 和 0.4。

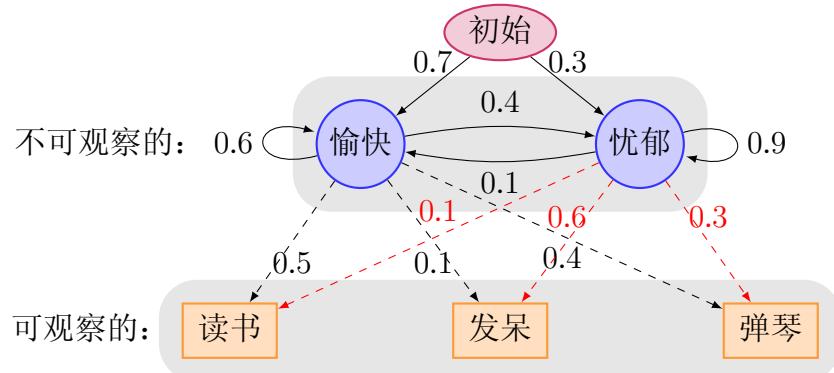


图 13.2: 阿诗玛的内心状态和外在行为的隐 Markov 模型。

图 13.2 所示的隐 Markov 模型中，令 $\omega_1 = \text{“读书”}$, $\omega_2 = \text{“发呆”}$, $\omega_3 = \text{“弹琴”}$ 。为方便起见，约定用 R 表示“读书”，用 M 表示“发呆”，用 G 表示“弹琴”。用 1 来表示状态“愉快”，用 2 来表示状态“忧郁”，则初始分布为 $0.7\langle 1 \rangle + 0.3\langle 2 \rangle$ ，状态转移矩阵 A 和发射矩阵 B 分别为

$$A = \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix} \text{ 和 } B = \begin{pmatrix} 0.5 & 0.1 & 0.4 \\ 0.1 & 0.6 & 0.3 \end{pmatrix}$$

设到 T 时刻为止观察到的结果的序列为 $x_{1:T} = x_1 x_2 \cdots x_T$ ，对隐 Markov 模型 $\mathcal{M} = \langle S, \Omega, A, B, \pi \rangle$ ，有如下几个问题有待解决。

① 观察结果的概率：给定了参数 θ ，求概率 $P(x_{1:T}|\theta)$ 。譬如，在例 13.1 中，求 $P(RRGM|\theta)$ 。解决了这个问题，我们就能明确回答对于给定的长度，哪个观察序列最有可能出现。例如，直观上阿诗玛一旦进入“忧郁”状态便很难走出来，而“发呆”是处于“忧郁”状态最有可能的结果，所有长度为 4 的观察序列中 $MMMM$ 的概率是不是最大呢？

② 状态序列的概率：给定了参数 θ ，求最有可能的状态序列 $\hat{z}_{1:T}$ ，使得

$$\hat{z}_{1:T} = \underset{z_{1:T}}{\operatorname{argmax}} P(z_{1:T}|x_{1:T}, \theta), \text{ 其中 } z_{1:T} = z_1 z_2 \cdots z_T \quad (13.1)$$

这个问题具有非常广泛的应用背景，例如自然语言处理中的词性标注。对机器

而言，观察到的是词序列^{*}，未知的是对应的词性序列。如果隐 Markov 模型的参数真实地反映出该语言的统计规律，还有什么比 (13.1) 更合理的标准呢？

③ 参数的训练：若参数 θ 未知，求解 θ 的最大似然估计 $\hat{\theta}$ ，使得

$$\hat{\theta} = \operatorname{argmax}_{\theta} P(x_{1:T} | \theta)$$

如果机器能够在观测数据的基础上为取得更好的效果而动态地修改模型参数，外观上就似乎具备了学习的能力，问题三正是有关隐 Markov 模型参数更新的。

本节内容

隐 Markov 模型有着广泛的应用背景，二十世纪七十年代中期用于语音识别，八十年代末用于生物序列分析。本节重点介绍有关隐 Markov 模型的向前向后算法、Viterbi 算法和 Baum-Welch 算法，其中向前向后算法是 Baum-Welch 算法的基础，而后者又是第 14 章即将介绍的期望最大化算法的特例。

关键知识

(1) 隐 Markov 模型；(2) 向前算法、向后算法；(3) Viterbi 算法；(4) Baum-Welch 算法。

^{*}对某些语言，譬如中文，词的切分 (segmentation) 和词性 (part-of-speech, POS) 判定是捆绑在一起的，甚至要借助句法分析才能得到正确的结论。

13.1.1 观察序列的概率：向前算法与向后算法

已知隐 Markov 模型 $\mathcal{M} = \langle S, \Omega, A, B, \pi \rangle$, 参数为 $\theta = (A, B, \pi)$ 。本节试图解决隐 Markov 模型的问题一。令 $z_{1:T} = z_1 z_2 \cdots z_T$ 是不可观察的状态序列, 则

$$P(z_{1:T}|\theta) = \pi_{z_1} a_{z_1 z_2} a_{z_2 z_3} \cdots a_{z_{T-1} z_T} \quad (13.2)$$

假设观察结果是独立的, 即

$$P(x_{1:T}|z_{1:T}, \theta) = \prod_{t=1}^T P(x_t|z_t, \theta) = b_{z_1}(x_1) b_{z_2}(x_2) \cdots b_{z_T}(x_T) \quad (13.3)$$

由全概率公式, 隐 Markov 模型的问题一的答案如下,

$$P(x_{1:T}|\theta) = \sum_{z_{1:T}} P(x_{1:T}|z_{1:T}, \theta) P(z_{1:T}|\theta) \quad (13.4)$$

例 13.2. 接着**例 13.1**, 问观察到序列 $x_{1:4} = RRGM$ 的概率?

解. 本例中状态序列 $z_{1:T} = z_1 z_2 z_3 z_4$ 共有 16 种可能, 即 1111, 2111, ⋯, 2222。依次利用式 (13.2) 和式 (13.3) 计算, 再代入式 (13.4)。譬如,

$$P(2111|\theta) = \pi_2 a_{21} a_{11} a_{11} = 0.3 \times 0.1 \times 0.6 \times 0.6 = 0.0108$$

$$P(x_{1:4}|2111, \theta) = b_{21} b_{11} b_{13} b_{12} = 0.1 \times 0.5 \times 0.4 \times 0.1 = 0.002$$

根据公式 (13.4), 求得 $P(x_{1:4}|\theta) \approx 0.01738$ 。请读者自行验证这一结果。

若像**例 13.2** 用蛮力 (brute force) 方法把式 (13.2) 和式 (13.3) 代入式 (13.4) 老老实实地计算 $P(x_{1:T}|\theta)$, 因为 S 遍历了所有可能的状态序列, 所以复杂度为 $O(Tn^T)$ 。试想一下两个状态序列仅仅最后一个不同, 蛮力方法独立地对待它们而使得中间的计算步骤白白地重复多次, 岂能不浪费时间? 下面, 我们设计动态规划算法*来改进算法复杂度。

算法 13.1 (向前算法). 为了讨论的方便, 我们把 t 时刻处于状态 i 且观察到 $x_1 x_2 \cdots x_t$ 的概率 $\alpha_t(i) = P(x_1 x_2 \cdots x_t, z_t = i|\theta)$ 称为向前变量。

■ 初始赋值: 初始时刻从状态 i 观察到 x_1 的概率为

$$\alpha_1(i) = \pi_i b_i(x_1), 1 \leq i \leq n$$

*动态规划算法的重点就是在计算过程中寻找递归关系, 尽可能减少中间结果的重复计算。感兴趣的读者可参阅 T. H. Cormen 等人的《算法导论》[27]。

□ 递归步骤：向前变量 $\alpha_t(i), 1 \leq i \leq n$ 与 $\alpha_{t+1}(j)$ 的递归关系如下，

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^n \alpha_t(i) a_{ij} \right] b_j(x_{t+1}), \text{ 其中 } 1 \leq j \leq n, 1 \leq t \leq T-1$$

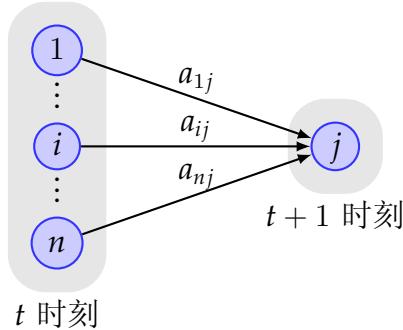


图 13.3: $\sum_{i=1}^n \alpha_t(i) a_{ij}$ 是 $t+1$ 时刻即将处于状态 j 并且已观察到 $x_1 x_2 \cdots x_t$ 的概率。

□ 计算结果：到 T 时刻为止观察到 $x_1 x_2 \cdots x_T$ 的概率为

$$P(x_{1:T}|\theta) = \sum_{i=1}^n \alpha_T(i) \quad (13.5)$$

例 13.3. 接着**例 13.1**，若观察到序列 $x_{1:7} = RRGMGRR$ ，下面列出所有的 $\alpha_t(i)$ 。利用公式 (13.5)，观察序列 RRGMGRR 的概率是 0.00024350，即第 7 列之和，而**例 13.2** 所求的概率即是第 4 列之和。

状态	时刻	1	2	3	4	5	6	7
1		0.35	0.107	0.0262	0.00175	0.001044	0.0005348	0.00018247
2		0.03	0.017	0.0173	0.01563	0.004430	0.0004405	0.00006103

练习 13.1. 如果观察序列 $x_{1:T}$ 和 $x'_{1:T}$ 仅仅次序不同，是否 $P(x_{1:T}|\theta) = P(x'_{1:T}|\theta)$ ？

答案：否。接着上例， $x'_{1:7} = GRRRMGR$ 的概率是 0.00019632。

由递归关系知，计算 $\alpha_{t+1}(j)$ 的复杂度为 $O(n)$ ，所以对每个固定的 t ，计算 $\{\alpha_t(i) : 1 \leq i \leq n\}$ 需要 $O(n^2)$ 步。因为 $t = 1, 2, \dots, T$ ，所以向前算法的时间复杂度为 $O(Tn^2)$ 。类似地，计算 $P(x_{1:T}|\theta)$ 也可采用向后算法。

算法 13.2 (向后算法). 若当前为 t 时刻且处于状态 i ，我们把即将观察到 $x_{t+1} x_{t+2} \cdots x_T$ 的概率 $\beta_t(i) = P(x_{t+1} x_{t+2} \cdots x_T | z_t = i, \theta)$ 称为向后变量。

□ 初始赋值： $\beta_T(i) = 1, 1 \leq i \leq n$

■ 递归步骤：向后变量 $\beta_{t+1}(j), 1 \leq j \leq n$ 与 $\beta_t(i)$ 的递归关系如下，

$$\beta_t(i) = \sum_{j=1}^n a_{ij} b_j(x_{t+1}) \beta_{t+1}(j), \text{ 其中 } 1 \leq i \leq n, 1 \leq t \leq T-1$$

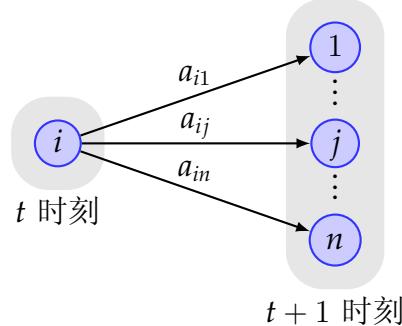


图 13.4: $a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)$ 是 t 时刻处于状态 i , $t+1$ 时刻即将处于状态 j 并且在未来观察到 $x_{t+1} x_{t+2} \cdots x_T$ 的概率。

■ 计算结果：到 T 时刻为止观察到 $x_1 x_2 \cdots x_T$ 的概率为

$$P(x_{1:T}|\theta) = \sum_{i=1}^n \pi_i \beta_1(i)$$

算法 13.3. 为解决隐 Markov 模型的问题一，向前算法和向后算法可按下面的方法并行地执行，时间复杂度依然是 $O(Tn^2)$ 。

$$P(x_{1:T}|\theta) = \sum_{i=1}^n \sum_{j=1}^n \alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j), \text{ 其中 } 1 \leq t \leq T \quad (13.6)$$

证明. 对于任意固定的 t , 皆有

$$\begin{aligned} P(x_{1:T}|\theta) &= \sum_{i=1}^n P(x_1 \cdots x_T, z_t = i | \theta) \\ &= \sum_{i=1}^n P(x_1 \cdots x_t, z_t = i, x_{t+1} \cdots x_T | \theta) \\ &= \sum_{i=1}^n P(x_1 \cdots x_t, z_t = i | \theta) P(x_{t+1} \cdots x_T | z_t = i, \theta) \\ &= \sum_{i=1}^n \alpha_t(i) \beta_t(i) \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j) \end{aligned} \quad \square$$

13.1.2 状态序列的概率: Viterbi 算法

考虑隐 Markov 模型的问题二, 就是寻找满足条件 (13.1) 的状态序列 $\hat{z}_{1:T}$, 即

$$\begin{aligned}\hat{z}_{1:T} &= \operatorname{argmax}_{z_{1:T}} P(z_{1:T}|x_{1:T}, \theta) \\ &= \operatorname{argmax}_{z_{1:T}} P(z_{1:T}|x_{1:T}, \theta)P(x_{1:T}|\theta) \\ &= \operatorname{argmax}_{z_{1:T}} P(z_{1:T}, x_{1:T}|\theta)\end{aligned}$$

换句话说, 隐 Markov 模型的问题二等同于找 $z_{1:T}$ 最大化 $P(z_{1:T}, x_{1:T}|\theta)$ 。1967 年, 美籍意大利裔电子工程师 Andrew James Viterbi (1935-) 提出了 Viterbi 算法, 解决了隐 Markov 模型的问题二。Viterbi 算法也是一个动态规划算法。



算法 13.4 (Viterbi 算法). 我们把 t 时刻处于状态 i 且观察到 $x_1 x_2 \cdots x_t$ 的最有可能的状态子序列 z_1, \dots, z_{t-1} 称为 Viterbi 路径, 其中 z_{t-1} 所示的状态记作 $\Delta_t(i)$ 。该 Viterbi 路径的概率 $\delta_t(i)$ 称为 Viterbi 变量, 即

$$\delta_t(i) = \max_{z_{1:(t-1)}} P(z_{1:(t-1)}, z_t = i, x_{1:t}|\theta)$$

- 初始赋值: $\delta_1(i) = \pi_i b_i(x_1), 1 \leq i \leq n$, 并且 $\Delta_1(i) = 0$
- 递归步骤: Viterbi 变量 $\delta_t(i), 1 \leq i \leq n$ 与 $\delta_{t+1}(j)$ 之间的递归关系如下,

$$\delta_{t+1}(j) = \left[\max_{1 \leq i \leq n} \delta_t(i) a_{ij} \right] b_j(x_{t+1}), \text{ 其中 } 1 \leq t \leq T-1, 1 \leq j \leq n$$

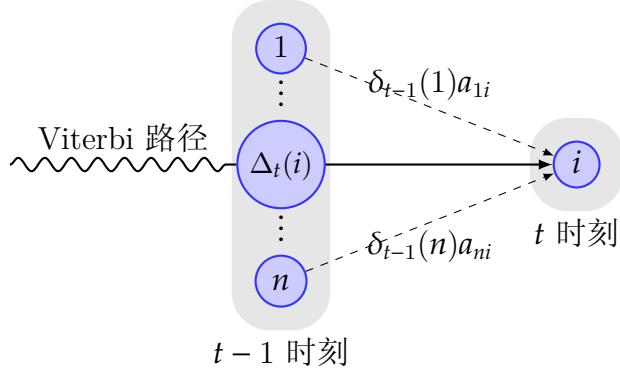


图 13.5: 记录下 t 时刻状态 i 的 Viterbi 路径在 $t-1$ 时刻的状态 $\Delta_t(i)$ 。

Viterbi 路径最后一个状态是

$$\Delta_t(i) = \operatorname{argmax}_{1 \leq k \leq n} [\delta_{t-1}(k)a_{ki}] b_i(x_t)$$

■ 状态回溯：计算 $P(\hat{z}_{1:T}, x_{1:T} | \theta) = \max_{1 \leq i \leq n} [\delta_T(i)]$ ，并回溯得到状态序列

$$\begin{aligned}\hat{z}_T &= \operatorname{argmax}_{1 \leq i \leq n} [\delta_T(i)] \\ \hat{z}_t &= \Delta_{t+1}(\hat{z}_{t+1}), \text{ 其中 } t = T-1, \dots, 1\end{aligned}$$

例 13.4. 接着**例 13.1**，观察序列 $x_{1:4} = RRGM$ 对应着状态序列 1122，观察序列 $x_{1:6} = RRGMGR$ 对应着状态序列 112222，而观察序列 $x_{1:7} = RRGMGRR$ 对应着状态序列 1111111。显然，更多的观察可能导致不同的结论。

练习 13.2. 请读者验证 Viterbi 算法 13.4 的算法复杂度为 $O(Tn^2)$ 。

13.1.3 模型参数的训练: Baum-Welch 算法

隐 Markov 模型 $\mathcal{M} = \langle S, \Omega, A, B, \pi \rangle$ 的问题三就是求解未知参数 $\theta = (A, B, \pi)$ 的最大似然估计 $\hat{\theta}$ 。二十世纪六十年代末, 美国计算机科学家 Leonard E. Baum 和 Lloyd Richard Welch 基于向前算法和向后算法提出了 Baum-Welch 算法从而解决了该问题。

Baum-Welch 算法其实是第 14 章即将介绍的期望最大化算法的特例, 届时我们将给出 Baum-Welch 算法的理论基础, 本节只简单描述 Baum-Welch 算法。

性质 13.1. 若给定了观察结果 $x_{1:T} = x_1 x_2 \cdots x_T$ 和参数 θ , 我们能够计算下面的概率。

□ 利用式 (13.6), 算得 t 时刻从状态 i 到状态 j 的条件转移概率 $P(z_t = i, z_{t+1} = j | x_{1:T}, \theta)$, 称之为 Baum-Welch 变量, 并记作 $\xi_t(i, j)$ 。

$$\begin{aligned} P(z_t = i, z_{t+1} = j | x_{1:T}, \theta) &= \frac{P(z_t = i, z_{t+1} = j, x_{1:T} | \theta)}{P(x_{1:T} | \theta)}, \text{ 其中 } 1 \leq t \leq T - 1 \\ &= \frac{\alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^n \sum_{j=1}^n \alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)} \end{aligned}$$

□ t 时刻出现状态 i 的条件概率 $P(z_t = i | x_{1:T}, \theta)$, 记作 $\gamma_t(i)$ 。

$$\begin{aligned} P(z_t = i | x_{1:T}, \theta) &= \sum_{j=1}^n P(z_t = i, z_{t+1} = j | x_{1:T}, \theta) \\ &= \sum_{j=1}^n \xi_t(i, j) \end{aligned}$$

□ 发生状态转移 $i \rightarrow j$ 的条件概率是

$$P(i \rightarrow j | x_{1:T}, \theta) = \sum_{t=1}^{T-1} \xi_t(i, j)$$

□ 基于上面的结果，不难算得出现状态 i 的条件概率 $P(i|x_{1:T}, \theta)$ ，

$$\begin{aligned} P(i|x_{1:T}, \theta) &= \sum_{j=1}^n P(i \rightarrow j|x_{1:T}, \theta) \\ &= \sum_{j=1}^n \sum_{t=1}^{T-1} \xi_t(i, j) \\ &= \sum_{t=1}^{T-1} \gamma_t(i) \end{aligned}$$

算法 13.5 (Baum-Welch 算法). 用户设定一个接近 0 的正数 $\epsilon > 0$ 用作阈值来判定参数是否需要继续更新。设当前的参数设置为 $\hat{\theta}$ ，下面的算法给出了参数更新的过程。

■ 更新参数：新参数 $\tilde{\theta} = (\tilde{A}, \tilde{B}, \tilde{\pi})$ 按如下方式计算，

$$\begin{aligned} \tilde{a}_{ij} &= \frac{P(i \rightarrow j|x_{1:T}, \hat{\theta})}{P(i|x_{1:T}, \hat{\theta})} \\ \tilde{b}_i(k) &= \frac{\sum_{t=1}^{T-1} \gamma_t(i) I_k(x_t)}{P(i|x_{1:T}, \hat{\theta})}, \text{ 其中, 指示函数 } I_k(x_t) = \begin{cases} 1 & \text{若 } x_t = \omega_k \\ 0 & \text{否则} \end{cases} \\ \tilde{\pi}_i &= c_s(i) \end{aligned}$$

■ 条件判定：若 $|\log P(x_{1:T}|\tilde{\theta}) - \log P(x_{1:T}|\hat{\theta})| < \epsilon$ ，则返回 $\hat{\theta}$ 并结束更新；否则，令 $\hat{\theta} = \tilde{\theta}$ 并重返更新参数的步骤。

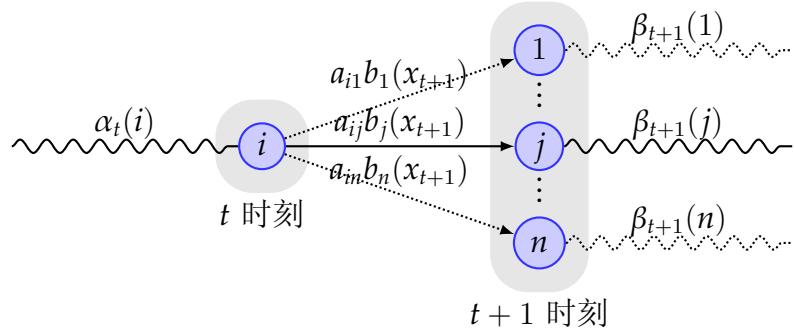


图 13.6: t 时刻处于状态 i , $t + 1$ 时刻处于状态 j , 概率 $P(z_t = i, z_{t+1} = j, x_{1:T}|\theta) = \alpha_t(i)a_{ij}b_j(x_{t+1})\beta_{t+1}(j)$ 由向前算法和向后算法求得。

13.2 无向图模型与 Bayes 网络

考虑概率图模型 G , 它首先是一个无向图或有向图 $G = (V, E)$, 其中的节点代表随机变量。在不引起歧义的时候, 我们有时不区分节点和它代表的随机变量。并且, 这些变量的联合分布与图 G 的某个图论性质有关。例如, Markov 网络 (或 Markov 随机场) 是一类图模型, 它与图论中团^{*}的概念相关。

^{*}无向图 $G = (V, E)$ 的子图 C 是一个团 (clique), 当且仅当 C 是一个完全图。如果不存在包含 C 的更大的团, 则称 C 是一个极大团 (maximal clique)。

13.2.1 Markov 网络与条件随机场

Markov 网络源于对 Ising 模型（见第 727 页的例 15.13）的研究，后来在图像处理、计算机视觉、机器学习等领域找到应用，并影响了 Bayes 网络的研究。当随机变量之间依赖关系存在循环的时候，Markov 网络是一个可选的建模工具。

定义 13.1. 若概率图模型 $G = (V, E)$ 满足下面的条件，则称之为 Markov 网络 (Markov network) 或 Markov 随机场 (Markov random field, MRF)。

① 非邻接节点的条件独立性：

$$X_u \perp\!\!\!\perp X_v | \mathbf{X}_{V - \{u, v\}}$$

其中， X_v 表示节点 v 对应的随机变量， $\mathbf{X}_{V - \{u, v\}}$ 表示 $V - \{u, v\}$ 那些节点对应的随机向量，有时简记作 $\mathbf{X}_{-\{u, v\}}$ 。

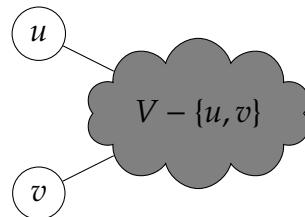


图 13.7：若节点 u, v 不邻接，则 $V - \{u, v\}$ 阻断了节点 u, v 之间的通讯。其概率含义是 X_u 与 X_v 在给定 $\mathbf{X}_{-\{u, v\}}$ 的条件下是独立的。

② 局部 Markov 性质：

$$X_v \perp\!\!\!\perp \mathbf{X}_{-\{v\}-N(v)} | \mathbf{X}_{N(v)}$$

其中， $N(v)$ 表示节点 v 的邻域，即与 v 邻接的节点的集合。

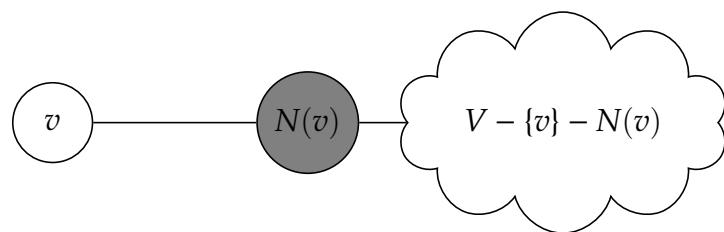


图 13.8：显然， $N(v)$ 阻断了节点 v 和 $V - \{v\} - N(v)$ 的通讯。即，在给定 $\mathbf{X}_{-\{v\}-N(v)}$ 的条件下， X_v 和 $\mathbf{X}_{-\{v\}-N(v)}$ 独立。

性质 13.2. 满足以下全局 Markov 条件的概率图模型也是一个 Markov 网络。

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_C, \text{ 其中, } A \text{ 中的点到 } B \text{ 中的点的任意路径都经过 } C \quad (13.7)$$

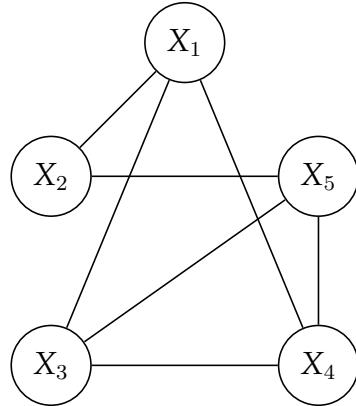
证明. 容易验证全局 Markov 性质 (13.7) 比定义 13.1 的条件都要强。 \square

定理 13.1 (团分解). 对于 Markov 网络 $G = (V, E)$, 随机向量 $\mathbf{X} = (X_v)_{v \in V}^\top$ 的概率密度函数 $p(\mathbf{x})$ 具有以下分解:

$$p(\mathbf{x}) = \prod_{C \in \text{mcl}(G)} p_C(x_C), \text{ 其中 } \text{mcl}(G) \text{ 是 } G \text{ 的极大团的集合} \quad (13.8)$$

证明. 令 C 是一个极大团, 对于任意 $v \notin C$, 总存在 C 内一个节点 u 与 v 不邻接。 \square

例 13.5. 考虑下面的 Markov 网络, 显然 $X_2 \perp\!\!\!\perp X_3, X_4 | X_1, X_5$ 。请读者给出更多的条件独立关系。



根据定理 13.1, X_1, \dots, X_5 的联合分布是

$$p(x_1, \dots, x_5) = p(x_1, x_3, x_4)p(x_3, x_4, x_5)p(x_1, x_2)p(x_2, x_5)$$

2001 年, J. Lafferty 等人提出条件随机场 (conditional random field, CRF) 模型, 它是 Markov 随机场的一个变种。

定义 13.2. 考虑图模型 $G = (V, E)$, 其中 V 分为不交的两个子集 \mathbf{X} 和 \mathbf{Y} , 在给定 \mathbf{X} 的条件下, \mathbf{Y} 满足局部 Markov 性质。即

$$Y_v \perp\!\!\!\perp Y_{-v-N(v)} | \mathbf{X}, Y_{N(v)}$$

13.2.2 Bayes 网络

定义 13.3. 如果从节点 X_i 到节点 X_j 有一条有向路径, 则称 X_j 是 X_i 的后辈节点 (descendant)。我们把 X_i 的所有后辈节点的全体记作 $\text{de}(X_i)$ 。例如, 图 2.22 中 $\text{de}(X_3) = \{X_4, X_5\}$ 。特别地, 如果 X_i 是 X_j 的父辈节点, 则 X_j 称为 X_i 的儿辈节点。

定义 13.4 (Markov 毯). 在贝叶斯网络中, 一个节点 X 的父辈节点、儿辈节点以及儿辈节点的其他父辈节点的集合称为 Markov 毛 (Markov blanket) 或者 Markov 边界 (Markov boundary), 记作 ∂X 。例如, 图 2.22 中, 节点 X_3 的 Markov 边界是

$$\begin{aligned}\partial X_3 &= \text{pa}(X_3) \cup \{X_4, X_5\} \cup \text{pa}(X_4) \cup \text{pa}(X_5) - \{X_3\} \\ &= \{X_1, X_2, X_4, X_5\}\end{aligned}$$

13.3 习题

13.1. 接着例 13.1, 请读者验证所有长度为 4 的观察序列中 $MMMM$ 的概率最大。

第十四章

期望最大化算法

不畏浮云遮望眼，自缘身在最高层。

王安石《登飞来峰》

前人工作的基础上，1977 年美国统计学家 Arthur P. Dempster (1929-)，Nan M. Laird 和 Donald B. Rubin (1943-) 发表论文 [33] 明确提出了不完全数据问题中用于最大似然估计的迭代算法——期望最大化算法 (Expectation-Maximization 算法，简称 EM 算法)。DLR 论文是统计学中引用率最高的论文之一，已有数万次之多，EM 算法自然成为统计计算中的著名算法。

EM 算法主要包括两个步骤：求期望的 E 步骤和求最大化的 M 步骤。该算法以理论的简洁性和收敛的稳定性著称，另外它还具有启发性，经过多年的发展，EM 算法衍生出许多的变种。同时，研究者也发现 EM 算法与 MCMC 方法有着密切的联系，二者都已经成为统计计算不可或缺的有力工具。目前对 EM 算法介绍得最全面的参考书是 G. J. McLachlan 和 T. Krishnan 合著的《EM 算法及其扩展》[109]，本书只作简单介绍。为了直观地理解 EM 算法，先举两个简单的例子。

例 14.1. DLR 文章引用了 Rao 曾举过的一个最大似然估计的例子（见第 516 页的例 8.25）：令样本 $\mathbf{X} = (X_1, X_2, X_3, X_4)^\top \sim \text{Multin}(n; \frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}, \frac{1}{4} - \frac{1}{4}\theta, \frac{1}{4} - \frac{1}{4}\theta, \frac{1}{4}\theta)$ ，已知样本值 $\mathbf{x} = (x_1, x_2, x_3, x_4)^\top = (125, 18, 20, 34)^\top$ ，求参数 θ 的最大似然估计。

解. 对数似然函数 $\ell(\theta; \mathbf{x}) = x_1 \ln(2 + \theta) + (x_2 + x_3) \ln(1 - \theta) + x_4 \ln \theta$ 是单峰函数，参数 θ 的最大似然估计存在且唯一，答案见例 8.25。下面提供另一种解法，对该具体问题而言并非是最简便的，但有抛砖引玉的作用。

令可观测的随机变量 X_1 是由两个不可观测的隐性 (latent) 随机变量 Y_1, Y_2 构成，即 $X_1 = Y_1 + Y_2$ ，并且

$$(Y_1, Y_2, X_2, X_3, X_4)^\top \sim \text{Multin}\left(n; \frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4} - \frac{1}{4}\theta, \frac{1}{4} - \frac{1}{4}\theta, \frac{1}{4}\theta\right)$$

由性质 4.39, $Y_2 \sim \text{B}(n, \theta/4)$ 。观测数据 \mathbf{x} 和隐性变量 $\mathbf{y} = (y_1, y_2)^\top$ 构成了完全数据, 其中 $y_1 + y_2 = x_1 = 125$ 。此时我们可以得到形式上更简单的似然函数和对数似然函数

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}, \mathbf{y}) &= \theta^{y_2+x_4} (1-\theta)^{x_2+x_3}, \text{ 其中 } y_2, \theta \text{ 都未知} \\ \ell(\theta; \mathbf{x}, \mathbf{y}) &= (y_2 + x_4) \ln \theta + (x_2 + x_3) \ln(1-\theta)\end{aligned}$$

不难得得到 θ 的最大似然估计

$$\hat{\theta} = \frac{y_2 + x_4}{y_2 + x_2 + x_3 + x_4}$$

上式中, y_2 是一个隐性变量, 猜想可用 Y_2 的后验期望来替换它。DLR 文章演示, 通过下面的迭代步骤可以得到 θ 的最大似然估计。

□ 不妨设当前对 θ 的估计是 $\theta^{(t)}$, 则 $Y_2|\mathbf{X} = \mathbf{x} \sim \text{B}(x_1, p^{(t)})$, 其中

$$p^{(t)} = \frac{\frac{1}{4}\theta^{(t)}}{\frac{1}{2} + \frac{1}{4}\theta^{(t)}} = \frac{\theta^{(t)}}{\theta^{(t)} + 2}$$

进而, y_2, y_1 的点估计分别是

$$y_2^{(t)} = \mathbb{E}_{\theta^{(t)}}(Y_2|\mathbf{X} = \mathbf{x}) = x_1 p^{(t)} = \frac{x_1 \theta^{(t)}}{\theta^{(t)} + 2} \text{ 和 } y_1^{(t)} = x_1 - y_2^{(t)} \quad (14.1)$$

它们都是 \mathbf{x} 和 $\theta^{(t)}$ 的函数。记 $\mathbf{y}^{(t)} = (y_1^{(t)}, y_2^{(t)})^\top$ 为当前对 \mathbf{y} 的估计。

□ 令 $Q(\theta, \theta^{(t)}) = \ell(\theta; \mathbf{x}, \mathbf{y}^{(t)}) = (y_2^{(t)} + x_4) \ln \theta + (x_2 + x_3) \ln(1-\theta)$, 通过 $dQ(\theta, \theta^{(t)})/d\theta = 0$ 得到对 θ 的最大似然估计, 并令其为 $\theta^{(t+1)}$, 即

$$\theta^{(t+1)} = \frac{y_2^{(t)} + x_4}{y_2^{(t)} + x_2 + x_3 + x_4} \quad (14.2)$$

若令初值 $\theta^{(0)} = 0.95$, 通过递归关系 (14.1) 和 (14.2) 可以得到 $\theta^{(t)}$ 的序列 (稍后的性质 14.1 将保证这是一个“良性循环”) 0.6614827, 0.6313092, 0.6274154, 0.6269003, 0.6268320, 0.6268229, 0.6268217, 0.6268215, …, 收敛于 θ 的最大似然估计 $\hat{\theta} = 0.6268215$ 。

例 14.2 (数据缺失时的参数估计). 随机向量 $(X, Y)^\top \sim N_2(\boldsymbol{\mu}, \Sigma)$ 的观测数据如下, 其中 * 表示该数据缺失了。

$$\begin{array}{ccccccccc} x & 0 & 2 & 1 & -1 & * & 3 & 1 \\ y & 1 & 0 & 3 & 1 & 0 & * & * \end{array} \Rightarrow \begin{array}{ccccccccc} x & 0 & 2 & 1 & -1 & \color{red}{1} & 3 & 1 \\ y & 1 & 0 & 3 & 1 & 0 & \color{red}{1} & \color{red}{1} \end{array}$$

观测数据

第一轮最大似然估计

参数 $\mu = (\mu_1, \mu_2)^\top$ 的最大似然估计为

$$\mu_1^{(1)} = \frac{1}{6}(0 + 2 + 1 - 1 + 3 + 1) = 1$$

$$\mu_2^{(1)} = \frac{1}{5}(1 + 0 + 3 + 1 + 0) = 1$$

记 $\mu^{(1)} = (\mu_1^{(1)}, \mu_2^{(1)})^\top$, 相应地用它们暂时替换*处的值。未知参数 $\Sigma = (\sigma_{11}, \sigma_{12}; \sigma_{21}, \sigma_{22})$ 的最大似然估计为

$$\sigma_{11}^{(1)} = \frac{1}{7}[(0-1)^2 + (2-1)^2 + (1-1)^2 + (-1-1)^2 + (1-1)^2 + (3-1)^2 + (1-1)^2] = 10/7$$

类似地, $\sigma_{22}^{(1)} = 6/7$ (请读者验证)

$$\begin{aligned}\sigma_{12}^{(1)} = \sigma_{21}^{(1)} &= [(0-1)(1-1) + (2-1)(0-1) + (1-1)(3-1) \\ &\quad + (-1-1)(1-1) + (1-1)(0-1) + (3-1)(1-1) \\ &\quad + (1-1)(1-1)]/7 = -1/7\end{aligned}$$

记 $\Sigma^{(1)} = (\sigma_{11}^{(1)}, \sigma_{12}^{(1)}; \sigma_{21}^{(1)}, \sigma_{22}^{(1)})$ 。由第 338 页的定理 4.13 知条件分布 $X|Y=y$ 和 $Y|X=x$ 分别如下，

$$X|Y=y \sim N(\mu_2 + \sigma_{12}\sigma_{22}^{-1}(y - \mu_2), \sigma_{11} - \sigma_{12}^2\sigma_{22}^{-1}) \quad (14.3)$$

$$Y|X=x \sim N(\mu_1 + \sigma_{12}\sigma_{11}^{-1}(x - \mu_1), \sigma_{22} - \sigma_{12}^2\sigma_{11}^{-1}) \quad (14.4)$$

在当前状态 $(X, Y)^\top \sim N_2(\mu^{(1)}, \Sigma^{(1)})$ 之下，利用式 (14.3) 和式 (14.4) 分别对缺失的 x 分量和 y 分量进行更新。譬如，若 x 分量缺失，可用条件期望 $E(X|Y = y)$ 对它赋值，其中

$$E(X|Y=y) = \mu_2^{(1)} + \frac{\sigma_{12}^{(1)}}{\sigma_{22}^{(1)}}(y - \mu_2^{(1)})$$

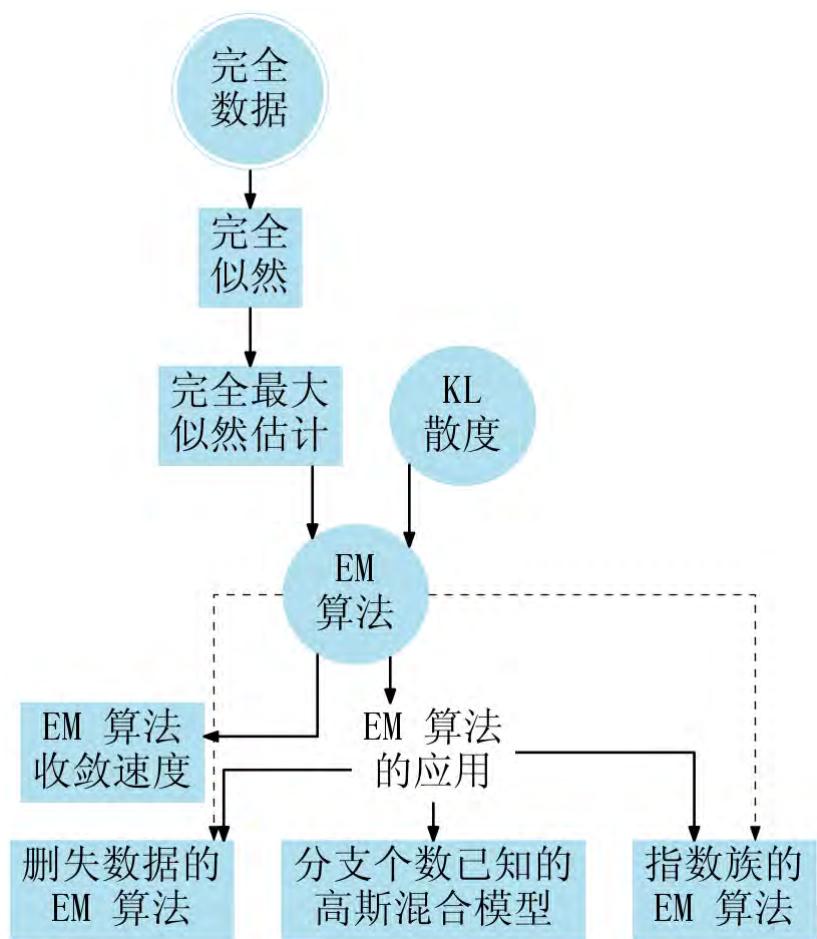
经过对缺失数据赋值，我们得到更新后的数据。

$$\begin{array}{ccccccccc} x & 0 & 2 & 1 & -1 & \textcolor{red}{7/6} & 3 & 1 \\ y & 1 & 0 & 3 & 1 & 0 & \textcolor{red}{4/5} & 1 \end{array}$$
 第二轮最大似然估计 \Rightarrow 新一轮求 μ 和 Σ 的 MLE
 以及条件期望的过程

基于这些数据，又可以求 μ 最大似然估计： $\mu_1^{(2)} = 43/42, \mu_2^{(2)} = 34/35$ 。类似地，又可以求 Σ 的最大似然估计 $\Sigma^{(2)}$ （留作练习），再用条件期望对缺失分量进行更新。如此反复，直至参数不再更新（或差异小于给定的阈值）为止，缺失数据也得到了补全。

$$\boldsymbol{\theta}^{(t+1)} = h(\mathbf{x}, \mathbf{y}^{(t)}) \quad \mathbf{y}^{(t)} = \text{E}_{\boldsymbol{\theta}^{(t)}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$$

第十四章的主要内容及其关系



14.1 完全数据与最大似然估计

像 例 14.1 和例 14.2, 观察样本 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 连同隐性数据或缺失数据等不可观测数据 $\mathbf{Y} \in \mathcal{Y}$ 扩充而得的数据 $\mathbf{Z} \in \mathcal{Z}$ 称为完全数据 (complete data) 或扩充数据, 引入 \mathbf{Y} 的目的或者为了简化了似然函数, 或者为了在缺失数据的情况下使得最大似然估计得以进行。

完全似然: 令 $f_\theta(\mathbf{x}, \mathbf{y})$ 是 \mathbf{X} 和 \mathbf{Y} 的联合密度函数, 其中参数 $\boldsymbol{\theta}$ 未知, 完全似然函数定义为 $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) = f_\theta(\mathbf{x}, \mathbf{y})$ 。很多时候, 完全对数似然函数 $\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) = \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y})$ 更常用。

完全最大似然估计: 令 $h_\theta(\mathbf{y}|\mathbf{x})$ 为给定 $\mathbf{X} = \mathbf{x}$ 条件下的 \mathbf{Y} 的条件密度函数, 则 \mathbf{X} 的密度函数为 $g_\theta(\mathbf{x}) = f_\theta(\mathbf{x}, \mathbf{y})/h_\theta(\mathbf{y}|\mathbf{x})$, 对数似然函数为

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln g_\theta(\mathbf{x}) = \ln f_\theta(\mathbf{x}, \mathbf{y}) - \ln h_\theta(\mathbf{y}|\mathbf{x}) \quad (14.5)$$

进而, 基于完全似然函数的 $\boldsymbol{\theta}$ 的完全最大似然估计是

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}) = \operatorname{argmax}_{\boldsymbol{\theta}} [\ln f_\theta(\mathbf{x}, \mathbf{y}) - \ln h_\theta(\mathbf{y}|\mathbf{x})] \quad (14.6)$$

式 (14.5) 的一个特点是, $\ln f_\theta(\mathbf{x}, \mathbf{y}) - \ln h_\theta(\mathbf{y}|\mathbf{x})$ 中的 \mathbf{y} 被“内耗”掉了, 这个特点被 EM 算法利用来计算 (14.6), 详见第一小节。

本节内容

第一小节介绍 EM 算法的理论基础, 并考察它的收敛速度。

关键知识

(1) 掌握 EM 算法的理论基础; (2) 了解 EM 算法的常见变种。

14.1.1 EM 算法及其收敛速度

考虑到式 (14.5) 的特点, $\ln f_{\theta}(x, Y)$ 和 $\ln h_{\theta}(Y|x)$ 都是由 Y 定义的随机变量。然而, $\ln f_{\theta}(x, Y) - \ln h_{\theta}(Y|x)$ 的表达式 $\ell(\theta; x)$ 中却不含 Y 。不妨假设对未知参数 θ 的当前估计为 $\theta^{(t)}$, 在给定条件 $X = x$ 之下, 恒有

$$\begin{aligned}\ell(\theta; x) &= E_{\theta^{(t)}}[\ell(\theta; x)|X = x] \\ &= E_{\theta^{(t)}}[\ln f_{\theta}(x, Y)|X = x] - E_{\theta^{(t)}}[\ln h_{\theta}(Y|x)|X = x] \\ &= Q(\theta, \theta^{(t)}) - H(\theta, \theta^{(t)})\end{aligned}\quad (14.7)$$

其中, $H(\theta, \theta^{(t)}) = E_{\theta^{(t)}}[\ln h_{\theta}(Y|x)|X = x]$, 并且

$$\begin{aligned}Q(\theta, \theta^{(t)}) &= E_{\theta^{(t)}}[\ln f_{\theta}(x, Y)|X = x] \\ &= \int_{\mathcal{Y}} h_{\theta^{(t)}}(y|x) \ln f_{\theta}(x, y) dy \\ &= \int_{\mathcal{Y}} \frac{f_{\theta^{(t)}}(x, y)}{g_{\theta^{(t)}}(x)} \ell(\theta; x, y) dy \\ &\propto \int_{\mathcal{Y}} f_{\theta^{(t)}}(x, y) \ell(\theta; x, y) dy\end{aligned}\quad (14.8)$$

算法 14.1 (EM 算法). 为求得 (14.6) 的结果, 参数的更新采用如下的迭代算法。

$$\theta^{(t+1)} \leftarrow \operatorname{argmax}_{\theta} Q(\theta, \theta^{(t)}) = \operatorname{argmax}_{\theta} \int_{\mathcal{Y}} f_{\theta^{(t)}}(x, y) \ell(\theta; x, y) dy \quad (14.9)$$

性质 14.1. 算法 14.1 保证了迭代序列 $\{\theta^{(t)} : t = 0, 1, 2, \dots\}$ 总是往增加似然的方向收敛*, 即

$$\ell(\theta^{(t+1)}; x) \geq \ell(\theta^{(t)}; x), \text{ 其中 } t = 0, 1, 2, \dots \quad (14.10)$$

显然, 等号成立当且仅当 $Q(\theta^{(t+1)}, \theta^{(t)}) = Q(\theta^{(t)}, \theta^{(t)})$ 。

证明. 根据式 (14.9), 显然 $Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) \geq 0$ 。另外, 由附录 F 的定理 2.22 知 $h_{\theta^{(t)}}(y|x)$ 与 $h_{\theta^{(t+1)}}(y|x)$ 的 Kullback-Leibler 散度非负, 即

$$\begin{aligned}H(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)}) &= \int_{\mathcal{Y}} h_{\theta^{(t)}}(y|x) \ln \frac{h_{\theta^{(t)}}(y|x)}{h_{\theta^{(t+1)}}(y|x)} dy \\ &= K(h_{\theta^{(t)}}, h_{\theta^{(t+1)}}) \geq 0\end{aligned}$$

于是, $\ell(\theta^{(t+1)}; x) - \ell(\theta^{(t)}; x) = [Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})] + [H(\theta^{(t)}, \theta^{(t)}) -$

*Dempster 把满足式 (14.10) 的参数更新算法称为广义 EM 算法 (简记作 GEM 算法), 并不见得一定要是式 (14.9) 的样子。

$H(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)})] \geq 0$, 得证。 \square

 若已知 $\ell(\boldsymbol{\theta}; \mathbf{x})$ 是单峰的, 则 EM 算法收敛至 $\boldsymbol{\theta}$ 的最大似然估计。若 $\ell(\boldsymbol{\theta}; \mathbf{x})$ 是多峰的, 初值 $\boldsymbol{\theta}^{(0)}$ 的选择将影响到算法的结果。为了避免 EM 算法收敛到局部极值点, 一般采用并行策略选择多个初值。

令 Θ 为参数空间, EM 算法定义了映射 $M : \Theta \rightarrow \Theta$ 使得 $\boldsymbol{\theta}^{(t+1)} = M(\boldsymbol{\theta}^{(t)})$, 则 EM 算法的结果 $\boldsymbol{\theta}^* = M(\boldsymbol{\theta}^*)$ 即是 M 的不动点。根据式 (14.7) 不难得得到观察信息

$$\begin{aligned} I(\boldsymbol{\theta}; \mathbf{x}) &= -\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}^2} \\ &= E_{\boldsymbol{\theta}^*} \left[-\frac{\partial^2 \ln f_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Y})}{\partial \boldsymbol{\theta}^2} \middle| \mathbf{X} = \mathbf{x} \right] - E_{\boldsymbol{\theta}^*} \left[-\frac{\partial^2 \ln h_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\partial \boldsymbol{\theta}^2} \middle| \mathbf{X} = \mathbf{x} \right] \\ &= -\frac{\partial^2 Q(\boldsymbol{\theta}, \tau)}{\partial \boldsymbol{\theta}^2} \Big|_{\tau=\boldsymbol{\theta}^*} + \frac{\partial^2 H(\boldsymbol{\theta}, \tau)}{\partial \boldsymbol{\theta}^2} \Big|_{\tau=\boldsymbol{\theta}^*} \end{aligned}$$

① DLR 证明了

$$\left[\frac{\partial M(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}^*} \right] \left[\frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}^*} \right] = \left[\frac{\partial^2 H(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}^*} \right] \quad (14.11)$$

进而, 在 $\boldsymbol{\theta}^*$ 的某邻域内, EM 算法的收敛速度为

$$\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2} \left(\frac{\partial^2 Q}{\partial \boldsymbol{\theta}^2} \right)^{-1}$$

② 1982 年, T. A. Louis 进一步证明了

$$\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2} = -V \left[\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{Y})}{\partial \boldsymbol{\theta}} \right] \quad (14.12)$$

例 14.3. 接着例 14.2, 令当前对参数 θ 的估计是 $\theta^{(t)}$, 于是 $Y_2|\mathbf{X} = \mathbf{x} \sim B(x_1, p^{(t)})$, 其中 $p^{(t)} = \theta^{(t)} / (\theta^{(t)} + 2)$ 。进而,

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E_{\theta^{(t)}}[(Y_2 + x_4) \ln \theta + (x_3 + x_4) \ln(1 - \theta) | \mathbf{X} = \mathbf{x}] \\ &= [E_{\theta^{(t)}}(Y_2 | \mathbf{X} = \mathbf{x}) + x_4] \ln \theta + (x_3 + x_4) \ln(1 - \theta) \end{aligned}$$

通过 $dQ(\theta, \theta^{(t)})/d\theta = 0$ 我们得到参数更新的迭代公式,

$$\theta^{(t+1)} = \frac{E_{\theta^{(t)}}(Y_2 | \mathbf{X} = \mathbf{x}) + x_4}{E_{\theta^{(t)}}(Y_2 | \mathbf{X} = \mathbf{x}) + x_2 + x_3 + x_4} = \frac{\frac{x_1 \theta^{(t)}}{\theta^{(t)} + 2} + x_4}{\frac{x_1 \theta^{(t)}}{\theta^{(t)} + 2} + x_2 + x_3 + x_4}$$

选择初始值 $\theta^{(0)} = 0.5$, 经过 5 次更新后得到 $\theta^{(5)} \approx 0.6268$, 可近似地当作 θ 的最大似然估计 $\hat{\theta}$ 。

$$\begin{aligned} -\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta^2} \Big|_{\hat{\theta}} &= \frac{E_{\theta^{(t)}}(Y_2 | \mathbf{X} = \mathbf{x}) + x_4}{\hat{\theta}^2} + \frac{x_3 + x_4}{(1 - \hat{\theta})^2} \\ &= \frac{29.83 + 34}{0.6268^2} + \frac{38}{(1 - 0.6268)^2} = 435.2 \\ V \left[\frac{\partial \ell(\theta; \mathbf{x}, \mathbf{Y})}{\partial \theta} \right]_{\hat{\theta}} &= \frac{V_{\hat{\theta}}(Y_2 | \mathbf{X} = \mathbf{x})}{\hat{\theta}^2} \\ &= \left(\frac{125}{\theta^2} \right) \left(\frac{\hat{\theta}}{\hat{\theta} + 2} \right) \left(\frac{2}{\hat{\theta} + 2} \right) = 57.8 \\ -\frac{\partial^2 \ell(\theta; \mathbf{x})}{\partial \theta^2} \Big|_{\hat{\theta}} &= 435.3 - 57.8 = 377.5 \end{aligned}$$

因此, $\hat{\theta}$ 的标准误差是 $\sqrt{1/377.5} = 0.05$ 。

14.1.2 EM 算法的若干变种

如果 E 步骤, 即式 (14.8) 计算困难, 可用 Monte Carlo 方法近似求解: 从分布 $h_{\theta^{(t)}}(\mathbf{y}|\mathbf{x})$ 里抽样得到 $\mathbf{y}_1, \dots, \mathbf{y}_m$, 令 $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \frac{1}{m} \sum_{j=1}^m \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}_j)$ 。

例 14.4. 在例 14.1 中, 从 $B(x_1, p^{(t)})$ 里抽样得到 y_1, \dots, y_m , 则 $E_{\theta^{(t)}}(Y_2 | \mathbf{X} = \mathbf{x}) \approx \sum_{j=1}^m y_j / m$ 。令 $m = 10, \theta^{(0)} = 0.4$, 经过 9-12 次迭代便可得到 $\hat{\theta} \approx 0.627$ 。

14.2 期望最大化算法的应用

缺失数据的统计分析需要 EM 算法, 例 14.2 介绍了用 EM 算法处理缺失数据: 用期望值替代缺失值, 并基于补缺后的数据估计参数, 然后在此参数设定之下再计算缺失值的条件期望, 如此反复……。在 DLR 论文之前, 已有很多学者利用这样的迭代过程填充缺失值, 例如隐 Markov 模型中的 Baum-Welch 算法。

令 $\boldsymbol{\theta}^*$ 最大化 $\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln g_{\boldsymbol{\theta}}(\mathbf{x})$, 基于正态分布的统计推断, $\boldsymbol{\theta}|\mathbf{X} = \mathbf{x} \sim N_d(\boldsymbol{\theta}^*, \Sigma_{\text{obv}})$, 其中

$$\begin{aligned}\Sigma_{\text{obv}}^{-1} &= - \left. \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}^*} \\ &= - \left. \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}^*} + \left. \frac{\partial^2 H(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}^*} \\ &= - \int_{\mathcal{Y}} \left. \frac{\partial^2 \ln \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y})}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}^*} h_{\boldsymbol{\theta}^*}(\mathbf{y}|\mathbf{x}) d\mathbf{y} + \int_{\mathcal{Y}} \left. \frac{\partial^2 \ln h_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}^*} h_{\boldsymbol{\theta}^*}(\mathbf{y}|\mathbf{x}) d\mathbf{y}\end{aligned}$$

14.2.1 分支个数已知的高斯混合模型

已知简单随机样本 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \in \mathbb{R}^d$ 来自高斯混合 (Gaussian mixture) 总体 $\sum_{i=1}^k p_i \phi(\mathbf{x}|\boldsymbol{\mu}_i, \Sigma_i)$, 该总体有时也简记作 $\sum_{i=1}^k p_i N_d(\boldsymbol{\mu}_i, \Sigma_i)$, 其中参数 $\boldsymbol{\mu}_i, \Sigma_i, 0 < p_i < 1, i = 1, 2, \dots, k$ 都是未知的且 $\sum_{i=1}^k p_i = 1$, 而分支个数 k 是有限的且是已知的。已知样本值为 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, 根据[算法 14.1](#) 设计出具体的 EM 算法给出参数 $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k, \Sigma_1, \dots, \Sigma_k, p_1, \dots, p_k)$ 的最大似然估计。

令每个样本点 \mathbf{X}_j 伴随着一个 k 维的隐性随机向量 $\mathbf{Y}_j = (Y_{j1}, \dots, Y_{jk})^\top$, 其中只有一个分量为 1, 其他分量都为 0。显然, $\sum_{i=1}^k Y_{ji} = 1$ 。若分量 $Y_{ji} = 1$, 则表示样本 \mathbf{X}_j 抽取自分支 $N_d(\boldsymbol{\mu}_i, \Sigma_i)$ 。设隐性数据为 $\mathbf{y}_1, \dots, \mathbf{y}_n$, 则完全对数似然函数为

$$\ell(\boldsymbol{\theta}; \mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{y}_1, \dots, \mathbf{y}_n) = \sum_{j=1}^n \sum_{i=1}^k [y_{ji} \ln p_i + y_{ji} \ln \phi(\mathbf{x}_j | \boldsymbol{\mu}_i, \Sigma_i)] \quad (14.13)$$

对比似然函数 $\ell(\boldsymbol{\theta}; \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{j=1}^n \ln [\sum_{i=1}^k p_i \phi(\mathbf{x}_j | \boldsymbol{\mu}_i, \Sigma_i)]$, 式 (14.13) 已有了些简化。为了简单陈述起见, [例 14.5](#) 考虑了二分支二维高斯混合模型的参数估计问题, 读者可尝试将之推广到 k 分支的情形。

例 14.5. 已知样本来自二分支高斯混合总体 $p\phi(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1) + (1-p)\phi(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2)$, 利用 EM 算法对参数 $\boldsymbol{\theta} = (p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ 进行估计。按下面的方法产生测试数据。

解. 令伴随着样本点 \mathbf{X}_j 的隐性变量 $Y_j \in \{1, 0\}$ 表示 \mathbf{X}_j 是否来自分支 $N_d(\boldsymbol{\mu}_1, \Sigma_1)$, 不妨设 $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ 为隐性数据, 则完全对数似然函数为

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{y}) = & \sum_{j=1}^n [y_j \ln p + y_j \ln \phi(\mathbf{x}_j | \boldsymbol{\mu}_1, \Sigma_1) \\ & + (1 - y_j) \ln(1 - p) + (1 - y_j) \ln \phi(\mathbf{x}_j | \boldsymbol{\mu}_2, \Sigma_2)] \end{aligned}$$

令 $\partial \ell(\boldsymbol{\theta}; \mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{y}) / \partial \boldsymbol{\theta} = \mathbf{0}$, 利用[定理 E.11](#) 和[定理 E.14](#) 可得参数 $\boldsymbol{\theta}$ 的最大似然估计如下。

$$\begin{aligned} \hat{p} &= \frac{1}{n} \sum_{j=1}^n y_j \\ \hat{\boldsymbol{\mu}}_1 &= \frac{\sum_{j=1}^n y_j \mathbf{x}_j}{\sum_{j=1}^n y_j} & \hat{\boldsymbol{\mu}}_2 &= \frac{\sum_{j=1}^n (1 - y_j) \mathbf{x}_j}{n - \sum_{j=1}^n y_j} \\ \hat{\Sigma}_1 &= \frac{1}{n} \sum_{j=1}^n y_j (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_1)^\top & \hat{\Sigma}_2 &= \frac{1}{n} \sum_{j=1}^n (1 - y_j) (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)^\top \end{aligned}$$

按照 EM 算法, 初始的参数估计 $\boldsymbol{\theta}^{(0)} = (p^{(0)}, \boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \Sigma_1^{(0)}, \Sigma_2^{(0)})$ 由用户设定, 求

解参数 θ 的最大似然估计可转化为下述 E 步骤和 M 步骤的迭代过程。

E 步骤: 设 $\theta^{(t)} = (p^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}, \Sigma_1^{(t)}, \Sigma_2^{(t)})$ 是当前的参数估计。

$$\begin{aligned} E_{\theta^{(t)}}(Y_j | \mathbf{X}_j = \mathbf{x}_j) &= P_{\theta^{(t)}}(Y_j = 1 | \mathbf{X}_j = \mathbf{x}_j) \\ &= \frac{P_{\theta^{(t)}}(Y_j = 1) P_{\theta^{(t)}}(\mathbf{X}_j = \mathbf{x}_j | Y_j = 1)}{P_{\theta^{(t)}}(\mathbf{X}_j = \mathbf{x}_j)} \\ &= \frac{p^{(t)} \phi(\mathbf{x}_j | \mu_1^{(t)}, \Sigma_1^{(t)})}{p^{(t)} \phi(\mathbf{x}_j | \mu_1^{(t)}, \Sigma_1^{(t)}) + (1 - p^{(t)}) \phi(\mathbf{x}_j | \mu_2^{(t)}, \Sigma_2^{(t)})} \end{aligned}$$

置 $y_j \leftarrow E_{\theta^{(t)}}(Y_j | \mathbf{X}_j = \mathbf{x}_j)$, 即用 $Y_j | \mathbf{X}_j = \mathbf{x}_j$ 在当前参数设定下的条件期望替代隐性数据 $y_j, j = 1, 2, \dots, n$ 。

M 步骤: 将参数按照最大似然估计之结果 $\hat{p}, \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2$ 从 $\theta^{(t)}$ 更新至 $\theta^{(t+1)}$ 。继续重复 E 步骤和 M 步骤直至达到精度要求。

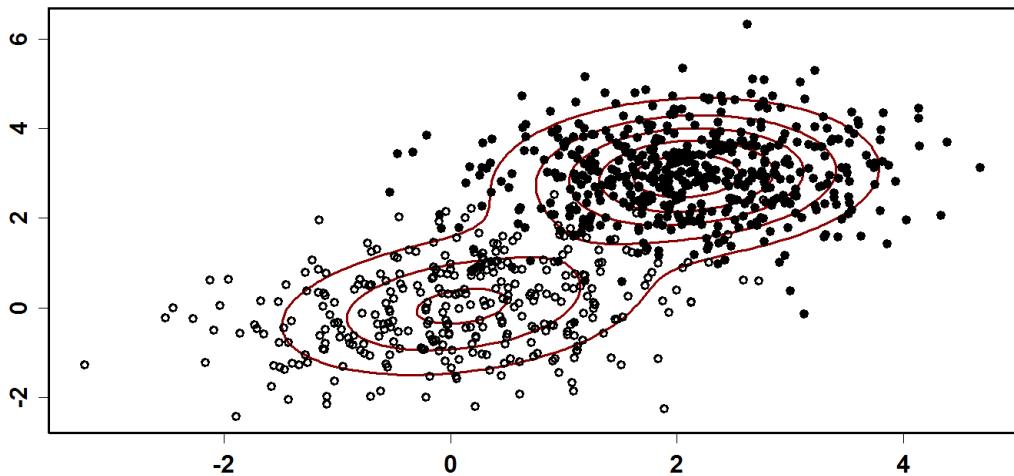


图 14.1: EM 算法经过 $N = 1000$ 次迭代得到参数的最大似然估计: $\hat{p} = 0.398$, $\hat{\mu}_1 = (0.015, -0.008)^\top$, $\hat{\mu}_2 = (1.928, 2.997)^\top$, $\hat{\Sigma}_1 = (1.115, 0.281; 0.281, 0.973)$, $\hat{\Sigma}_2 = (0.722, -0.006; -0.006, 0.795)$ 。在此参数之下, 二分支高斯混合总体的密度函数曲面的等高线如图所示。

14.2.2 针对删失数据的 EM 算法

在工程实践或生物医学中，对产品或生物的寿命进行预测和评估是可靠性或生存分析研究的内容，如今已发展成为统计学的重要分支——可靠性与生存分析理论，简称生存分析 (survival analysis)。

在生存分析中，寿命数据往往有删失的特点。例如检验一件电子产品的寿命，在试验结束时它仍然未寿终正寝，我们并不知道它的确切寿命，只知道它的寿命大于 r ，此时称该产品的寿命在 r 上是（右）删失的，并称 r 为（右）删失数据 (censored data)^{*}。

例 14.6. 已知观察样本 $X_1, X_2, \dots, X_m, Y_{m+1}, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\theta, 1)$ ，其中不可观察样本 Y_{m+1}, \dots, Y_n 在 r 上是右删失的，用 EM 算法给出参数 θ 的最大似然估计？

解. 完全对数似然函数是

$$\ell(\theta; \mathbf{x}, \mathbf{y}) = -\frac{1}{2} \sum_{i=1}^m (x_i - \theta)^2 - \frac{1}{2} \sum_{i=m+1}^n (y_i - \theta)^2$$

从 $d\ell(\theta; \mathbf{x}, \mathbf{y})/d\theta = 0$ 解得参数 θ 的最大似然估计如下，

$$\hat{\theta} = \frac{1}{n}(x_1 + \dots + x_m + y_{m+1} + \dots + y_n)$$

设当前的参数设置为 $\theta^{(t)}$ ，缺失数据 Y_{m+1}, \dots, Y_n 来自密度函数为 $\phi(y - \theta^{(t)})/[1 - \Phi(r - \theta^{(t)})]$ 的截尾正态总体 Y ，其中 $y > r$ 。把 y_{m+1}, \dots, y_n 都设置为 $E_{\theta^{(t)}}(Y)$ ，参数更新方式如下，

$$\begin{aligned}\theta^{(t+1)} &= \frac{x_1 + \dots + x_m + (n-m)E_{\theta^{(t)}}(Y)}{n} \\ &= \frac{x_1 + \dots + x_m}{n} + \frac{n-m}{n} \left[\theta^{(t)} + \frac{\phi(r - \theta^{(t)})}{1 - \Phi(r - \theta^{(t)})} \right]\end{aligned}$$

*类似地，如果只知道它的寿命小于 l ，则称该产品的寿命在 l 上是左删失的，并称 l 为左删失数据。在实际应用中，右删失数据的情形更常见些。

14.2.3 指数族的 EM 算法

假设扩充数据 $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ 的密度函数 $f_{\theta}(\mathbf{z})$ 如下,

$$f_{\theta}(\mathbf{z}) = \frac{b(\mathbf{z})}{\alpha(\boldsymbol{\theta})} \exp\{\boldsymbol{\theta}^T s(\mathbf{z})\}$$

其中, 向量值函数 $s = s(\mathbf{z}) = (s_1(\mathbf{z}), \dots, s_k(\mathbf{z}))^\top$, 即 \mathbf{Z} 的密度函数属于正则指数分布族 (见定义 7.15)。设当前的参数估计是 $\boldsymbol{\theta}^{(t)}$, 由式 (14.8),

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \int_{\mathcal{Y}} h_{\boldsymbol{\theta}^{(t)}}(\mathbf{y}|\mathbf{x}) \ln b(\mathbf{z}) d\mathbf{y} + \boldsymbol{\theta}^T \int_{\mathcal{Y}} s(\mathbf{z}) h_{\boldsymbol{\theta}^{(t)}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} - \ln \alpha(\boldsymbol{\theta})$$

上式右端第一项不含 $\boldsymbol{\theta}$, 所以在 E 步骤不必考虑它。E 步骤只需计算上式中第二项中的积分, 即

$$\int_{\mathcal{Y}} s(\mathbf{z}) h_{\boldsymbol{\theta}^{(t)}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}(s(\mathbf{Z})|\mathbf{X} = \mathbf{x})$$

不妨将上式的结果记为 $\mathbf{s}^{(t)}$, 则在 EM 算法的 M 步骤里只需考虑最大化 $\boldsymbol{\theta}^T \mathbf{s}^{(t)} - \ln \alpha(\boldsymbol{\theta})$ 即可, 即求解方程 $\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) / \partial \boldsymbol{\theta} = \mathbf{0}$ 。

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} &= \frac{\partial [\boldsymbol{\theta}^T \mathbf{s}^{(t)} - \ln \alpha(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \\ &= \mathbf{s}^{(t)} - \frac{\partial \ln \alpha(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \mathbf{s}^{(t)} - \frac{1}{\alpha(\boldsymbol{\theta})} \int_{\mathcal{Z}} b(\mathbf{z}) \frac{\partial \exp\{\boldsymbol{\theta}^T s(\mathbf{z})\}}{\partial \boldsymbol{\theta}} d\mathbf{z} \\ &= \mathbf{s}^{(t)} - \int_{\mathcal{Z}} s(\mathbf{z}) \frac{b(\mathbf{z})}{\alpha(\boldsymbol{\theta})} \exp\{\boldsymbol{\theta}^T s(\mathbf{z})\} d\mathbf{z} \\ &= \mathbf{s}^{(t)} - \mathbb{E}_{\boldsymbol{\theta}}(s(\mathbf{Z})) \end{aligned}$$

问题最终化归为解有关 $\boldsymbol{\theta}$ 的方程

$$\mathbb{E}_{\boldsymbol{\theta}}(s(\mathbf{Z})) = \mathbf{s}^{(t)}$$

第十五章

随机模拟技术

夫尺有所短，寸有所长，物有所不足。智有所不明，数有所不逮，神有所不通。

屈原《卜居》

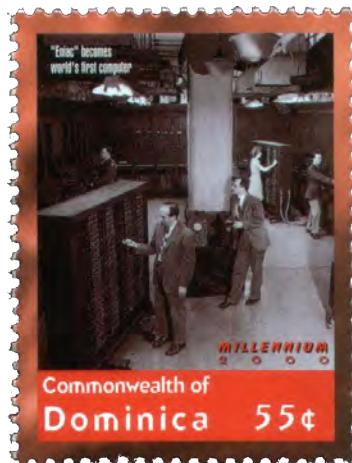
我们在第 1 章介绍了最早的随机模拟试验——Buffon 投针试验（见第 36 页），这种利用随机抽样的数值方法（即 Monte Carlo 方法）还可以用来计算定积分，特别是带有复杂边界条件的多元定积分。

Monte Carlo 方法的收敛速度与样本数目有关，而与样本所在空间的维数无关，这样的数值计算方法处理高维数据非常适合。随着计算机科学与技术的飞速发展，随机模拟方法得到了广泛的应用（如核物理、流体力学、计算统计学等），特别是在一些无法利用确定性算法得到精确解的问题上，随机模拟取得了令人瞩目的成绩。

本章将简介一些经典的随机模拟技术，如 von Neumann 舍选法、复合抽样、重要度抽样、Metropolis 算法、Metropolis-Hastings 算法、Gibbs 抽样、切片抽样、可逆跳 MCMC 方法等抽样算法，还有模拟退火算法、数据增扩算法等应用技术。其中，von Neumann 舍选法是最简单、最基本的抽样方法，它具有一定的启发性。



说到 von Neumann 在随机模拟上的贡献，美籍波兰裔数学家 Stanislaw Marcin Ulam (1909-1984) 回忆道，“1946 年我刚从疾病中康复，玩纸牌时遇到的一个问题促使我首次有想法和企图尝试 Monte Carlo 方法。这个问题是：Canfield 纸牌游戏中将 52 张牌都摆好的机会是多大？利用纯组合计算，我花了大量时间来估算，我不知道比抽象思维更可行的方法是不是把纸牌摆个一百遍，然后简单地观察并数一数成功的次数就好。那时已



经可以预知快速计算机新时代的到来，我立刻想到中子扩散问题和数学物理中的其他问题，甚至更一般地，如何将某些微分方程描述的过程转换成可解释为一连串随机操作的等价形式。后来，我把这个想法告诉了 John von Neumann，我们开始着手实际的计算。”

von Neumann 被 Ulam 的想法吸引，他很快意识到它的重要性，并首次在计算机 ENIAC (Electronic Numerical Integrator And Computer) 上实现了随机抽样。可以说，电子计算机为随机模拟从玩具模型（如例 1.27 的 Buffon 投针试验）到实用模型的飞跃提供了物质基础，并且大大地促进了对随机模拟方法的研究，可谓“水到渠成”。

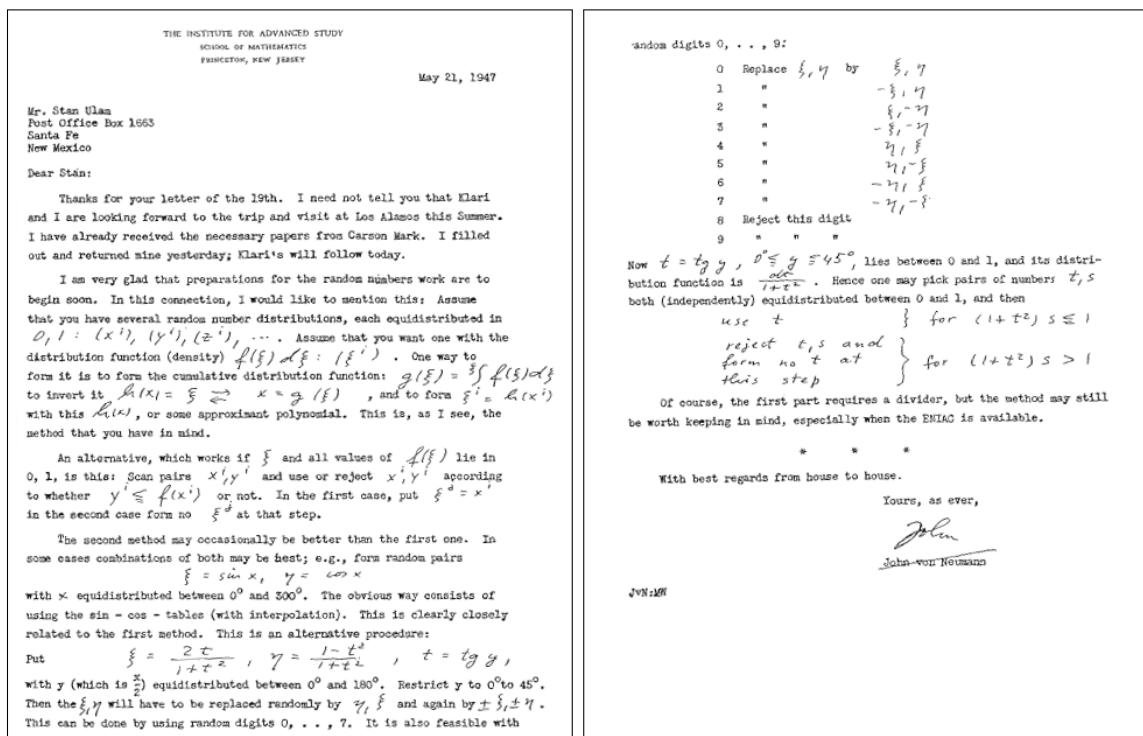


图 15.1: von Neumann 于 1947 年写给 Ulam 的信，讨论随机数的产生算法，包括逆 CDF 法，以及 von Neumann 舍选法的雏形等。

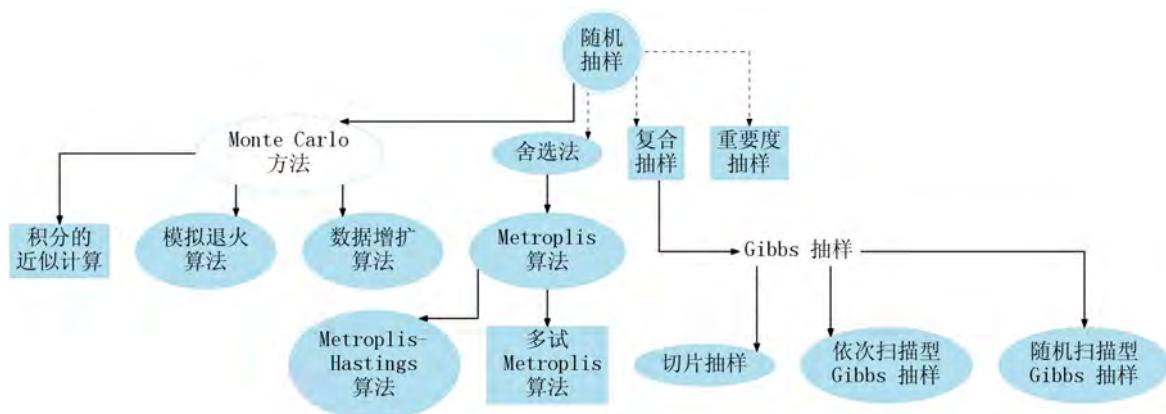
当时，Ulam 和 von Neumann 在 Los Alamos 国家实验室的同事，美籍希腊裔物理学家 Nicholas Metropolis (1915-1999) 等人也参与了 Monte Carlo 方法的研究。1953 年，Metropolis 及同事 A. W. Rosenbluth, M. N. Rosenbluth (1927-2003), A. H. Teller, E. Teller (1908-2003) 发表论文提出了著名的 Metropolis 算法，标志着随机模拟技术进入突飞猛进的发展期。

Metropolis 算法是 Markov 链 Monte Carlo (Markov chain Monte Carlo, 简称 MCMC) 方法的鼻祖，是人类历史上最为重要的算法之一，它的诞生是物理学、数学、统计学、计算机科学与技术的共同结晶。

之后出现的 Metropolis-Hastings 算法、Gibbs 抽样、切片抽样、可逆跳 MCMC 方法等抽样算法也都是应具体应用的需求而发展起来的。考虑到篇章的限制，本章对那些较新的随机模拟技术只能作“粗线条”的描绘。

随机模拟技术常用的工具软件有 BUGS 和与其兼容的衍生或重写软件，如 WinBUGS，OpenBUGS，JAGS 等。另外，R 语言的某些工具包实现了 M-H 算法及其变种，还提供了 BUGS 语言的接口。

第十五章的主要内容及其关系



15.1 产生随机数的传统方法

在第 4 章我们学习到很多分布，连同它们的随机数产生算法。除了这些常见的分布，在实践中还会遇到各式各样不常见的分布，如何去产生它们的随机数呢？我们需要一些通用的算法来解决这个问题。一般要求 RNG 算法简洁明了、效率高且产生的随机数质量好（即其经验分布非常接近总体分布）。

逆 CDF 法（见第 280 页的 [算法 4.10](#)）是一个通用的 RNG，但 CDF 常常不是显式的，更何况它的逆。所以，逆 CDF 法不是一个非常实用的 RNG。

为什么我们要对随机数的产生算法那么感兴趣？一个最主要的原因是为了高效的（近似）计算，下面举几个例子来说明（更多的例子见 [§15.3](#)）。

例 15.1. 令一元函数 $h(x) = [\sin(30x) + \cos(2x)]^2$ ，用随机模拟方法近似计算定积分

$$q = \int_0^1 h(x)dx$$

解. 抽取 n 个均匀分布 $U[0, 1]$ 的随机数 x_1, x_2, \dots, x_n ，则 q 的近似值为

$$\hat{q}_n = \frac{1}{n} \sum_{j=1}^n h(x_j)$$

显然，对于那些在非负函数 h 下取值很小的随机数，它们对结果的影响也不大。当取 $n = 2000, 2001, \dots, 20000$ 时，函数 $h(x)$ 的图像和定积分 $q = \int_0^1 h(x)dx$ 的近似值 \hat{q}_n 的情况如下图所示。

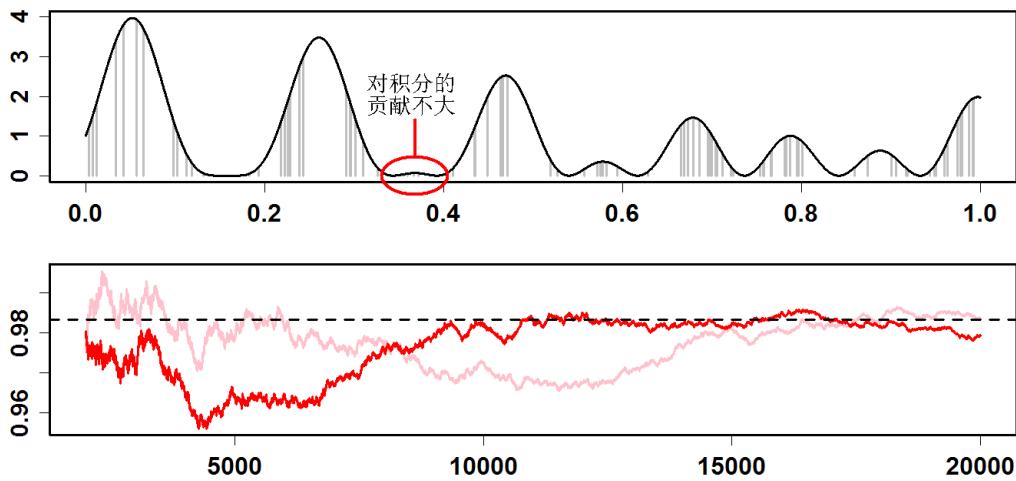


图 15.2: 上半图是函数 $h(x) = [\sin(30x) + \cos(2x)]^2$ 的图像和 $U[0, 1]$ 的随机数。下半图是用 Monte Carlo 方法求得 q 的近似解 \hat{q}_n ，其中横坐标为抽样次数 n ，纵坐标是模拟的结果，虚线为该积分的“精确解” 0.98321336323392。

 利用例 15.2 所示的简单 Monte Carlo 方法计算定积分 $\int_a^b h(x)dx$, 若 $h(x) = \text{常数}$, 则不论抽样多少都可以得到正确的结果; 若 $h(x)$ 是 delta 函数 (见第 219 页的例 3.3), 则抽样再多也是“竹篮打水一场空”。一般来说, $h(x)$ 的变化越“剧烈”, 简单 Monte Carlo 方法算得的定积分越不稳定。

例 15.2. 已知 $f(\mathbf{x})$ 是随机向量 \mathbf{X} 的密度函数, 其中 $\mathbf{x} \in \mathbb{R}^d$, 实值函数 $g(\mathbf{x}, \mathbf{y})$ 满足 $\text{Ex}[g(\mathbf{X}, \mathbf{y})] < \infty$, 近似地求解下面的定积分。

$$q(\mathbf{y}) = \int_{\mathbb{R}^d} g(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) d\mathbf{x}$$

解. 简单 Monte Carlo 方法把积分的问题转化为抽样的问题: 令 $\mathbf{X}_1, \dots, \mathbf{X}_n$ 是来自总体 $f(\mathbf{x})$ 的简单随机样本, 利用强大数律有

$$\frac{1}{n} \sum_{j=1}^n g(\mathbf{X}_j, \mathbf{y}) \xrightarrow{a.s.} q(\mathbf{y})$$

取 $f(\mathbf{x})$ 的随机数 $\mathbf{x}_1, \dots, \mathbf{x}_n$, 当 n 很大时, $q(\mathbf{y})$ 的近似解为

$$\hat{q}_n(\mathbf{y}) = \frac{1}{n} \sum_{j=1}^n g(\mathbf{x}_j, \mathbf{y})$$

特别地, 当例 15.2 中的 $g(\mathbf{x}, \mathbf{y})$ 为 $h(\mathbf{x})$ 时,

$$\int_{\mathbb{R}^d} h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \approx \frac{1}{n} \sum_{j=1}^n h(\mathbf{x}_j)$$

本节内容

关键知识

- (1) von Neumann 舍选法; (2) 复合抽样; (3) 重要度抽样。

15.1.1 von Neumann 舍选法

美籍匈牙利裔数学家、物理学家、计算机科学家 John von Neumann (1903-1957) 于 1951 年提出一种借助均匀分布 $U[0, 1]$ 从任意给定的密度函数 $\pi(x)$ 中抽样的方法, von Neumann 的这项工作激发了一系列 Monte Carlo 方法的产生。

von Neumann 的方法基于下面浅显的基本事实: 已知 $\pi(x)$ 是一个密度函数, 区域 $D_\pi = \{(x, u) : x \in \mathbb{R}, 0 \leq u \leq \pi(x)\}$ 上的均匀分布 $(X, U)^\top \sim U(D_\pi)$ 的密度函数在区域 D_π 上等于 1, 在区域 D_π^c 上等于 0。进而, 随机变量 X 的边缘密度函数为 $\pi(x)$, 这是因为

$$\pi(x) = \int_0^{\pi(x)} du \quad (15.1)$$



若分布 $\pi(x)$ 难于抽样, 为了得到 $\pi(x)$ 的随机数, 可以往区域 D_π 上均匀地投钉 (见下图), 落点的横坐标即 $\pi(x)$ 的随机数——这个过程能够用计算机来模拟实现。

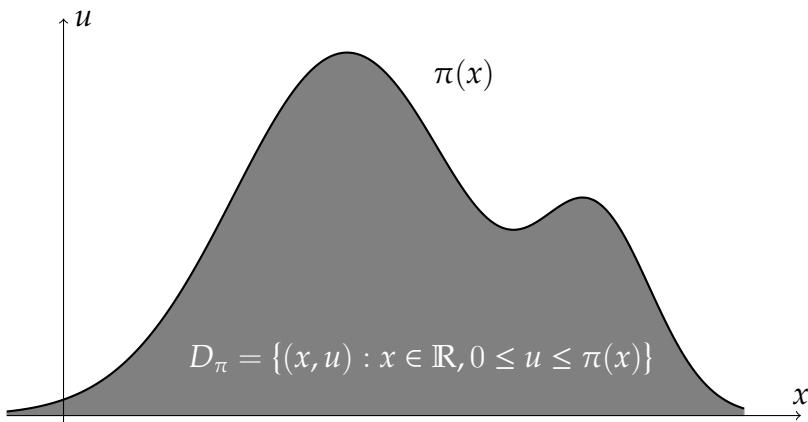


图 15.3: 二维随机向量 $(X, U)^\top$ 服从区域 D_π 上的均匀分布。如果 $\pi(x)$ 有界, 仅需用到均匀分布的随机数产生算法就可产生 $\pi(x)$ 的随机数。

例 15.3. 下面通过一个实例讲解如何产生分布 $Beta(\alpha, \beta)$ (其中 $\alpha > 1, \beta > 1$) 的随机数。譬如 $Beta(5, 3)$, 因为 $X \sim Beta(5, 3)$ 的密度函数 $\pi(x) = b_{5,3}(x)$ 落于矩形 $\Omega = [0, 1] \times [0, 2.5]$ 内, 它的随机数可用下面的方法产生:

- 先产生 $U(\Omega)$ 的随机数 $(x^*, y^*)^\top$;
- 如果 $y^* \leq \pi(x^*)$, 则 x^* 是 $\pi(x)$ 的随机数。

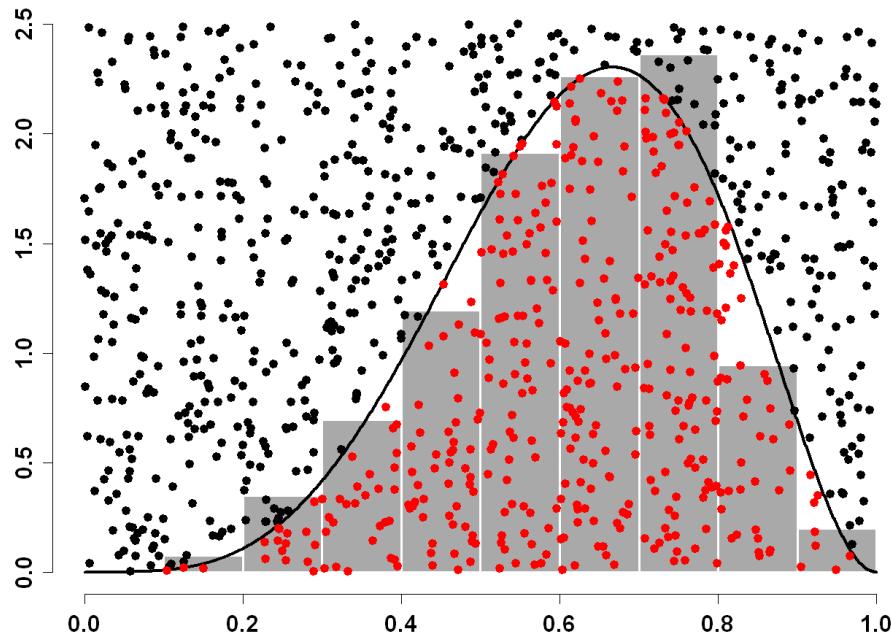


图 15.4: 往矩形 $\Omega = [0, 1] \times [0, 2.5]$ 上均匀地投钉 1000 枚, 落于密度函数 $b_{\alpha, \beta}(x)$ 之下的钉的横坐标就是分布 $\text{Beta}(\alpha, \beta)$ 的随机数, 图中显示了它们的直方图。

 **例 15.3** 中, 很多 $U(\Omega)$ 的随机数被浪费掉了, 特别是那些横坐标在密度函数 $b_{5,3}(x)$ 尾部的随机数 $(x^*, y^*)^\top$ 。若密度函数 $\pi(x)$ 无上界, 譬如至少有一个参数小于 1 的分布 $\text{Beta}(\alpha, \beta)$ (见第 302 页的图 4.26), 例 15.3 的方法将失效。

针对 $\pi(x)$ 难于抽样的情况, von Neumann 提出了一个巧妙的算法, 为抽样提供了一般的解决途径, 被称为 von Neumann 舍选法 (reject-accept method, 见[算法 15.1](#)), 其直观解释见图 15.5。舍选法的大致思路是先借助某个易于抽样的密度函数 $f(x)$ 产生一些随机数, 然后从这些随机数中筛选出 $\pi(x)$ 的随机数。

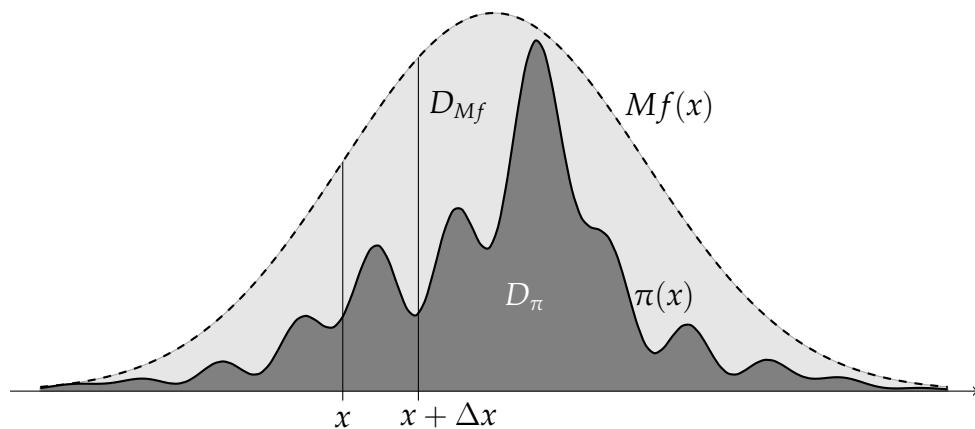


图 15.5: 密度函数 $f(x)$ 易于抽样且能“罩”住 $\pi(x)$, 即存在 $M \geq 1$, 满足 $\pi(x) \leq Mf(x)$ 。如何从 $f(x)$ 的随机数里筛选出 $\pi(x)$ 的随机数? 这是舍选法的“亮点”。

想象区域 $D_{Mf} = \{(x, u) : x \in \mathbb{R}, 0 \leq u \leq Mf(x)\}$ 上均匀地散落了很多钉，其中一些落在子区域 $D_\pi = \{(x, u) : x \in \mathbb{R}, 0 \leq u \leq \pi(x)\}$ 上。现在，考虑 D_{Mf} 上介于 $x, x + \Delta x$ 之间的窄条，当 $\Delta x \rightarrow 0$ 时，仅有比例为 $\pi(x)/[Mf(x)]$ 的钉的横坐标可作为 $\pi(x)$ 的随机数。

算法 15.1 (von Neumann 舍选法, 1951). 已知密度函数 $\pi(x), f(x)$ 满足 $\pi(x) \leq Mf(x)$ ，其中 $M \geq 1$ ，我们称 $f(x)$ 为 $\pi(x)$ 的罩函数。下面的算法产生 $X \sim \pi(x)$ 的随机数 x^* 。

□ 独立地产生 $Y \sim f(y)$ 的随机数 y^* 和 $U[0, 1]$ 的随机数 u^* 。

□ 若 $u^* \leq \pi(y^*)/[Mf(y^*)]$ ，置 $x^* \leftarrow y^*$ ；否则返回上一步骤。

证明. 由**算法 15.1**，独立的随机变量 $Y \sim f(y)$ 和 $U \sim U[0, 1]$ 定义了新的随机变量 X ，满足 $X = Y$ 当且仅当 $U \leq \pi(Y)/[Mf(Y)]$ 。下面，求随机变量 X 的分布函数。

$$\begin{aligned} P\{X \leq x\} &= P\left\{Y \leq x \middle| U \leq \frac{\pi(Y)}{Mf(Y)}\right\} \\ &= P\left\{Y \leq x, U \leq \frac{\pi(Y)}{Mf(Y)}\right\} \div P\left\{U \leq \frac{\pi(Y)}{Mf(Y)}\right\} \\ &= \int_{-\infty}^x \left\{ \int_0^{\frac{\pi(y)}{Mf(y)}} f(y) du \right\} dy \div \int_{-\infty}^{+\infty} \left\{ \int_0^{\frac{\pi(y)}{Mf(y)}} f(y) du \right\} dy \\ &= \int_{-\infty}^x \pi(y) dy \end{aligned}$$

于是， $X \sim \pi(x)$ ，即**算法 15.1** 给出了 $\pi(x)$ 的随机数。 □

 在实践中，为提高**算法 15.1** 的效率，不仅要求 $f(x)$ 容易抽样，最好使得 $Mf(x)$ 与目标密度函数 $\pi(x)$ 比较接近，这样就能以大概率 $\pi(y^*)/[Mf(y^*)]$ 选 y^* 为 $\pi(x)$ 的随机数，保证**算法 15.1** 的效率。总体来说，**算法 15.1** 使 $Y \sim f(y)$ 的随机数以 $1/M$ 的概率被选为 $X \sim \pi(x)$ 的随机数，这是因为

$$P\{X = Y\} = P\left\{U \leq \frac{\pi(Y)}{Mf(Y)}\right\} = \frac{1}{M}$$

算法 15.1 的一个极端的例子是 $Y \sim \pi(y)$ ，取 $M = 1$ ，于是 $P\{X = Y\} = 1$ 。此时，若取 $M = 2$ ，**算法 15.1** 依然可行，但效率降低了。

例 15.4. 已知函数 $g(x) = \exp(-x^2/2)\{[\sin(6x)]^2 + (\cos x)^2 \exp(\sin(4x)) + 0.6\}$ ，利用**算法 15.1** 产生分布 $\pi(x) \propto g(x)$ 的随机数。

解. 密度函数 $\pi(x)$ 的曲线见图 15.5。取 $f(x) = \phi(x)$ ，总存在 $M \geq 1$ 使得比例 $\pi(x)/[Mf(x)] = \{[\sin(6x)]^2 + (\cos x)^2 \exp(\sin(4x)) + 0.6\}/5$ 。

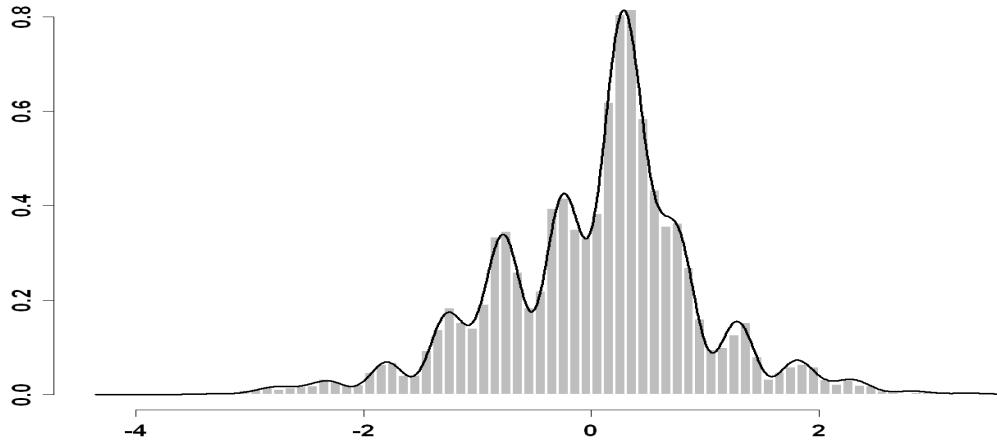


图 15.6: 例 15.4 利用 von Neumann 舍选法 (算法 15.1) 产生了一人为构造的、奇形怪状的分布 $\pi(x) \propto \exp(-x^2/2)\{[\sin(6x)]^2 + (\cos x)^2 \exp(\sin(4x)) + 0.6\}$ 的随机数, 其直方图和曲线 $\pi(x)$ 如图所示。

练习 15.1. 利用算法 15.1 产生分布 $\text{Laplace}(\mu, \sigma)$ 的随机数。

算法 15.2. 若密度函数 $f(x) \propto g(x)h(x)$, 其中 $g(x)$ 是一个易于抽样的分布, $h(x)$ 的值域是 $[0, 1]$ 。产生 $g(x)$ 的随机数 x^* 和 $U[0, 1]$ 的随机数 u^* 直至 $u^* \leq h(x^*)$, 输出 x^* 作为 $f(x)$ 的随机数。

例 15.5. 利用算法 15.1 来产生分布 $X \sim \text{Gamma}(\alpha, 1)$ 的随机数, 其中 $\alpha \geq 1$ 已知, 并用它们模拟计算 $E(X^2) = \alpha^2 + \alpha$ 。

解. 与 $\text{Gamma}(\alpha, 1)$ 接近且比较容易抽样的分布是 $\text{Gamma}(\lfloor \alpha \rfloor, \lfloor \alpha \rfloor / \alpha)$, 利用第 296 页的算法 4.15 先产生 $Y \sim \text{Gamma}(\lfloor \alpha \rfloor, \lfloor \alpha \rfloor / \alpha)$ 的随机数 y^* 。由 Gamma 分布的密度函数 (4.15),

$$\frac{g_{\alpha,1}(y)}{g_{\lfloor \alpha \rfloor, \lfloor \alpha \rfloor / \alpha}(y)} = \frac{\Gamma(\lfloor \alpha \rfloor)}{\Gamma(\alpha)} \left(\frac{\alpha}{\lfloor \alpha \rfloor} \right)^{\lfloor \alpha \rfloor} \left(y \exp \left\{ -\frac{y}{\alpha} \right\} \right)^{\alpha - \lfloor \alpha \rfloor}$$

由不等式 $\exp(y/\alpha) > ey/\alpha$, 上式显然存在上界。于是, 总能找到一个合适的 M 使得

$$\frac{g_{\alpha,1}(y)}{M g_{\lfloor \alpha \rfloor, \lfloor \alpha \rfloor / \alpha}(y)} = \left(\frac{ey}{\alpha} \exp \left\{ -\frac{y}{\alpha} \right\} \right)^{\alpha - \lfloor \alpha \rfloor} < 1$$

利用算法 15.2, 以概率 $[ey \exp(-y/\alpha)/\alpha]^{\alpha - \lfloor \alpha \rfloor}$ 置 $x^* \leftarrow y^*$ 便得到 $\sim \text{Gamma}(\alpha, 1)$ 的随机数 x^* 。产生足够多的随机数来模拟计算 $E(X^2) = \alpha^2 + \alpha$ (见下图)。

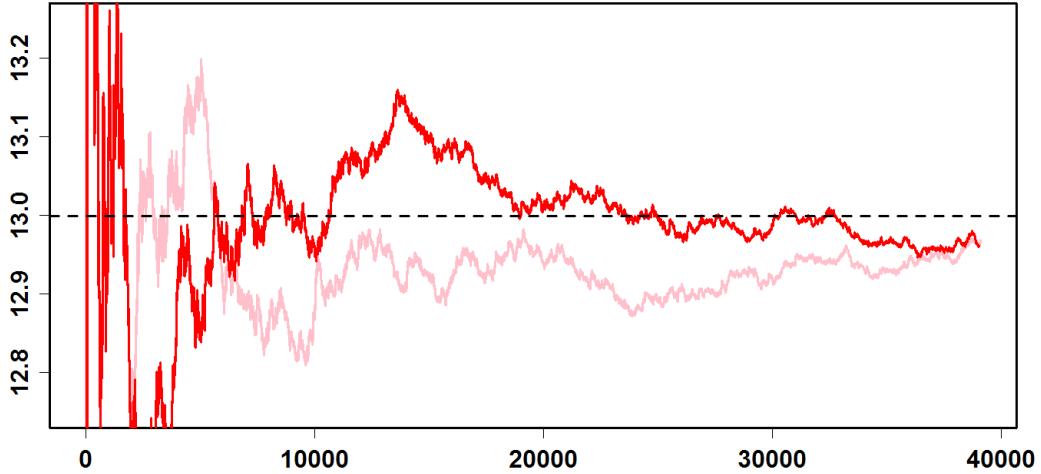


图 15.7: 利用例 15.5 描述的算法, 产生 $X \sim \text{Gamma}(3.14, 1)$ 的随机数, 用它们模拟计算 $E(X^2) = 12.9996$ 。

算法 15.3. 已知密度函数 $f_1(x), f_2(x)$ 和非负实数 p_1, p_2 满足 $p_1 + p_2 = 1$, 有一个易于抽样的密度函数 $g(x) \geq p_1 f_1(x)$, 则 $p_1 f_1(x) + p_2 f_2(x)$ 的随机数 x^* 可以这样产生:

- 产生 $g(y)$ 的随机数 y^* 和 $\text{U}[0, 1]$ 的随机数 u^* 。
- 若 $u^* \leq p_1 f_1(y^*)/g(y^*)$, 则 $x^* \leftarrow y^*$ 。否则, 产生 $f_2(x)$ 的随机数 x^* 。

练习 15.2. 请读者论证上述算法, 并将它与第 278 页的算法 4.9 比较优劣。

算法 15.4 (Devroye [34]). 若密度函数 $f(x)$ 的二阶绝对矩存在, 则其特征函数 $\varphi(t)$ 的一阶和二阶导数存在。在抽样之前, 我们先求以下两个值:

$$a \leftarrow \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\varphi(t)| dt \quad b \leftarrow \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\varphi''(t)| dt$$

- 产生 $U, V \stackrel{\text{iid}}{\sim} \text{U}[-1, 1]$ 的随机数 u_*, v_* 。

$$\begin{array}{ll} \text{若 } u_* < 0, \quad x_* \leftarrow v_* \sqrt{\frac{b}{a}} & t_* \leftarrow |u_*|a \\ \text{否则, } x_* \leftarrow \frac{1}{v_*} \sqrt{\frac{b}{a}} & t_* \leftarrow |u_*|av_*^2 \end{array}$$

- 若 $t_* \leq f(x_*)$, 则输出 x_* , 否则返回第一步。

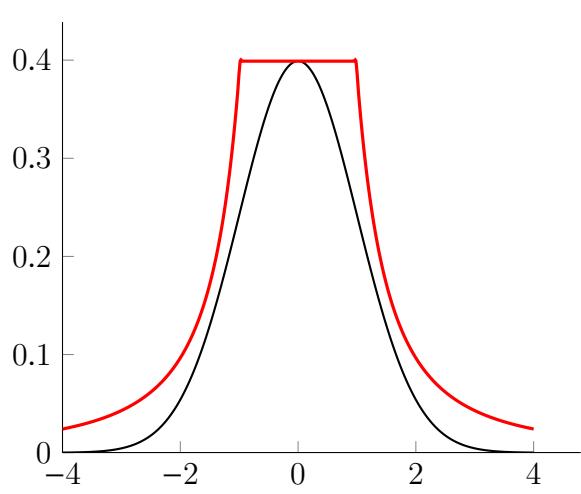
例 15.6. 利用算法 15.4 来产生 $X \sim N(0, 1)$ 的随机数。首先求得 $\varphi(t) = \exp(-t^2/2)$ 的二阶导数,

$$\varphi''(t) = (t^2 - 1) \exp(-t^2/2)$$

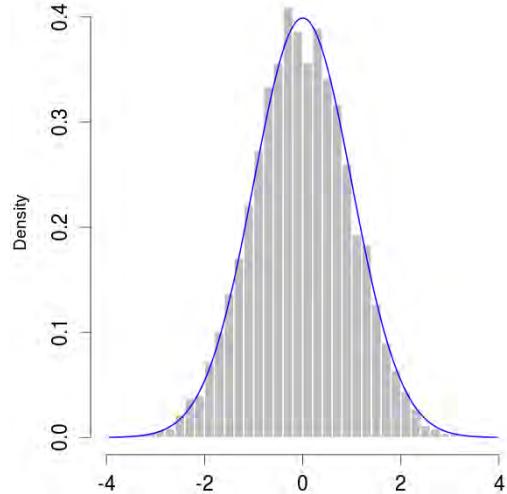
在抽样之前，计算算法 15.4 中的 a, b 。

$$a = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\varphi(t)| dt = \frac{1}{\sqrt{2\pi}}$$

$$b = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\varphi''(t)| dt = \frac{2}{\pi\sqrt{e}}$$



(a) 罩函数 $g(x) = \min(a, bx^{-2})$



(b) 利用算法 15.4 产生 $N(0, 1)$ 的随机数

图 15.8: 习题 3.15 是算法 15.4 的理论基础：这个平头的凹函数 $g(x) = \min(a, bx^{-2})$ 罩住 $f_X(x) = \phi(x)$ ，正比于一个易于抽样的密度函数，并且它贴近 $f_X(x)$ ，对舍选法来说是高效的。

15.1.2 复合抽样和重要度抽样

已知密度函数 $f(\mathbf{x})$ 和条件密度函数 $g(\mathbf{y}|\mathbf{x})$, 其中 $\mathbf{x} \in \mathbb{R}^d$, 则随机向量 \mathbf{Y} 的(边缘)密度函数为

$$q(\mathbf{y}) = \int_{\mathbb{R}^d} g(\mathbf{y}|\mathbf{x})f(\mathbf{x})d\mathbf{x} \quad (15.2)$$

针对式 (15.2), 本节的主要目标是回答下面的两个问题, 其中, 第一个问题在应用中尤为重要。

- 如何从 $q(\mathbf{y})$ 中抽样?
- 如何求得 $q(\mathbf{y})$ 的解析表达式?

算法 15.5 (复合抽样). 可按以下步骤产生式 (15.2) 所定义的分布 $q(\mathbf{y})$ 的随机数 \mathbf{y}^* 。

- 先从分布 $f(\mathbf{x})$ 产生随机数 \mathbf{x}^* , 然后
- 再从条件分布 $g(\mathbf{y}|\mathbf{x}^*)$ 产生随机数 \mathbf{y}^* 。

另外, 从 $f(\mathbf{x})$ 中产生随机数 $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$, 我们得到 $q(\mathbf{y})$ 的近似如下,

$$\hat{q}_n(\mathbf{y}) = \frac{1}{n} \sum_{j=1}^n g(\mathbf{y}|\mathbf{x}_j^*)$$

例 15.7. 利用**算法 15.5** 和第 144 页的**例 2.27** 的结果设计二元正态分布 $(X, Y)^\top \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的随机数 $(x^*, y^*)^\top$ 的产生算法。

解. 先产生 $N(\mu_X, \sigma_X^2)$ 的随机数 x^* , 然后再产生 $N(\mu_Y + \rho(x^* - \mu_X)\sigma_Y/\sigma_X, (1 - \rho^2)\sigma_Y^2)$ 的随机数 y^* 。

练习 15.3. 依据**例 15.7** 所描述的算法, 请读者用 R 语言产生第 137 页的**图 2.17** 所示的二元正态分布的随机数。

例 15.8. 已知 $X \sim N(0, 1)$ 和 $Y|X = x \sim N(x, 1)$, 利用**算法 15.5** 近似地求随机变量 Y 的密度函数。

解. 由第 66 页的**例 1.50** 知, $Y \sim N(0, 2)$ 。不难发现, 随着 n 的增大, $\hat{q}_n(y)$ 越来越接近 Y 的密度函数(见**图 15.9**)。

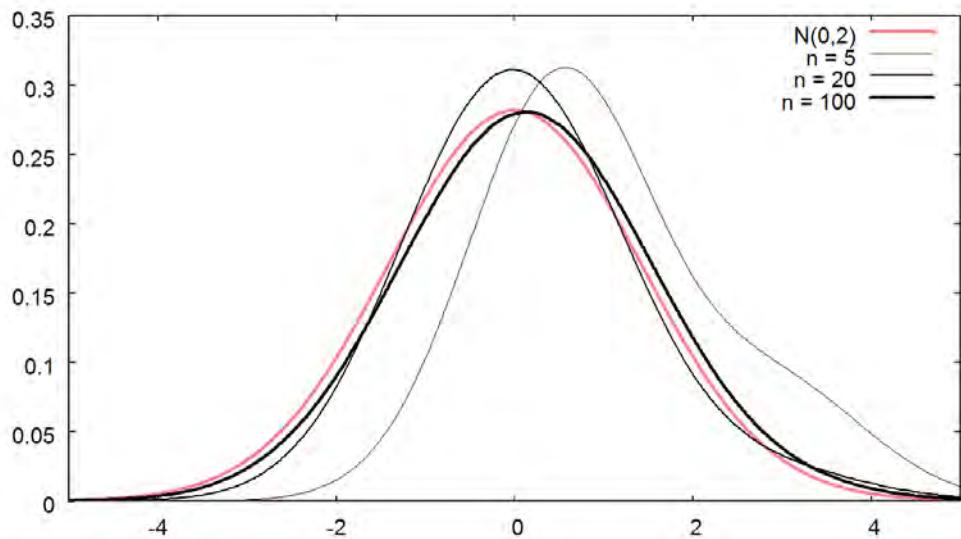


图 15.9: 在例 15.8 中, 取 $n = 5, 20, 100$, 算法 15.5 给出的 $\hat{q}_n(y)$, 随着 n 的增大越来越接近 $Y \sim N(0, 2)$ 的密度函数 $q(y)$ 。

如果在算法 15.5 中直接从 $f(\mathbf{x})$ 产生随机数比较困难, 则可以考虑另一个容易抽样的密度函数 $p(\mathbf{x}) \neq 0$, 式 (15.2) 具有如下的等价形式。

$$q(\mathbf{y}) = \int_{\mathbb{R}^d} g(\mathbf{y}|\mathbf{x}) \frac{f(\mathbf{x})}{p(\mathbf{x})} p(\mathbf{x}) d\mathbf{x}$$

经过“偷梁换柱”, 我们把是算法 15.5 稍作改动, 便得到下面的重要度抽样 (importance sampling) 算法, 简称 IS 算法。

算法 15.6 (重要度抽样). 从 $p(\mathbf{x})$ 中产生随机数 $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$, 我们得到 $q(\mathbf{y})$ 的近似如下,

$$\hat{q}_n(\mathbf{y}) = \frac{1}{n} \sum_{j=1}^n w_j g(\mathbf{y}|\mathbf{x}_j^*), \text{ 其中 } w_j = \frac{f(\mathbf{x}_j^*)}{p(\mathbf{x}_j^*)} \text{ 是权重}$$

例 15.9. 接着例 15.1, 如果 $X \sim \text{Gamma}(\alpha, \beta)$, 譬如 $X \sim \text{Gamma}(3, 2)$, 用 IS 算法 15.6 近似地计算 $q = E[h(X)]$ 。

解. 因为均匀分布比 Gamma 分布易于抽样, 由结果 (5.9), 抽取 $U(0, 1)$ 的随机数 x_1, x_2, \dots, x_n , 则 q 的近似值为

$$\hat{q}_n = \frac{1}{n} \sum_{j=1}^n h(x_j) g_{\alpha, \beta}(x_j)$$

其中, $g_{\alpha, \beta}(x)$ 是 $\text{Gamma}(\alpha, \beta)$ 的密度函数。

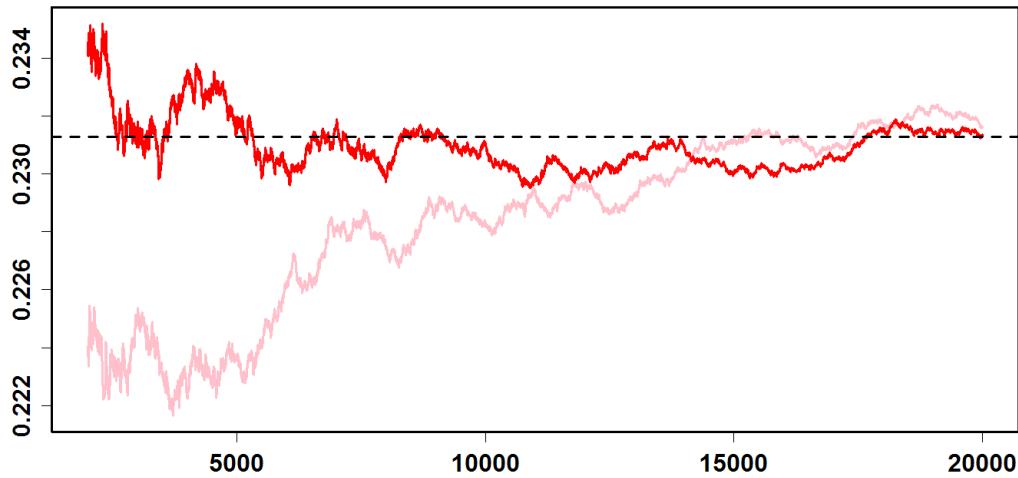


图 15.10: 例 15.9 利用重要度抽样方法求得的 $q = E[h(X)]$ 的近似解 \hat{q}_n , 其中横坐标为抽样次数 n , 纵坐标是模拟的结果, 虚线为“精确解” 0.2312729。

练习 15.4. 接着例 15.8, 利用 Maxima 实现算法 15.6 来近似地求随机变量 Y 的密度函数。提示: 取均匀分布 $U(-m, m)$ 的随机数 x_1, \dots, x_n , 其中 m 是很大的正实数, 则 Y 的密度函数近似为

$$\hat{q}_n(y) = \frac{2m}{n} \sum_{j=1}^n \phi(x_j) \phi(y|x_j, 1)$$

15.2 Markov 链 Monte Carlo 方法

受计算机科学技术的影响，当产生随机数的效率极大提高而成本极大降低时，随机模拟技术就进入了实用阶段。问题又回到了抽样算法上，尤其是高维分布的抽样，需要有适合计算机工作特点的简单迭代算法。用什么样的数学工具来“制造”理想的算法呢？

本节内容 MCMC 方法都是 Metropolis 算法的变种和发展，通过构造以目标分布 $\pi(x)$ 为其平稳分布的 Markov 链来实现 $\pi(x)$ 的抽样。第一小节介绍了模拟退火算法，它是 Metropolis 算法在组合最优化问题上的一个应用。第二小节介绍了经典的 Metropolis-Hastings 算法，它是对 Metropolis 算法的改进。

关键知识 (1)。

15.2.1 Metropolis 算法

Monte Carlo 方法经常用于模拟带随机性的物理系统，为那些用确定算法不可行或不可能解决的问题带来希望。

1953 年，美籍希腊裔物理学家、数学家、计算机科学家 Nicholas Metropolis (1915-1999, 照片见右)，及其同事 A. W. Rosenbluth, M. N. Rosenbluth (1927-2003), A. H. Teller, E. Teller (1908-2003)，为研究粒子系统的平稳性质，考虑了物理学中常见的 Boltzmann 分布 (4.2.10) 的抽样问题，他们合作撰文《利用快速计算机求解状态方程组》首次提出了一类迭代的 Monte Carlo 方法 [110]，即 Metropolis 算法，并在 MANIAC (Mathematical and Numerical Integrator and Computer) 计算机上加以实现。这篇论文被收录在《统计学中的重大突破》第三卷 [98] 中，Metropolis 算法也被遴选为二十世纪的十个最重要的算法之一。



然而，Metropolis 算法的真正推导者是美国物理学家 M. N. Rosenbluth (1927-2003, 照片见左)，所以该算法应该被称作 Rosenbluth 算法。

Metropolis 算法的精妙之处在于把分布 $\pi(\mathbf{x})$ 的抽样问题转化为构造一个 Markov 链 P_π ，使其平稳分布就是 $\pi(\mathbf{x})$ 。Metropolis 算法的好处是，不管 $\mathbf{x} \in \mathbb{R}^n$ 的维数多高，不管初始状态如何，总可以通过 Markov 链 P_π 以迭代的方式实现 $\pi(\mathbf{x})$ 的抽样——这正是计算机的长项。

Metropolis 算法是首个普适的抽样方法，并启发了一系列 Markov 链 Monte Carlo (MCMC) 方法，所以人们把它视为随机模拟技术腾飞的起点。

例 15.10 (Boltzmann 分布的抽样). 令 E_s 是状态 s 的能量，考虑 Boltzmann 分布 $\pi_1\langle 1 \rangle + \cdots + \pi_s\langle s \rangle + \cdots + \pi_n\langle n \rangle$ 的抽样问题，其中

$$\pi_s = \frac{1}{Z} \exp\left(-\frac{E_s}{k_B T}\right), \quad \text{其中 } Z = \sum_{s=1}^n \exp\left(-\frac{E_s}{k_B T}\right)$$

Metropolis 等人在 [110] 中定义从状态 i 转移到状态 j 的概率 p_{ij} 为

$$p_{ij} = \begin{cases} 1 & \text{如果 } E_j \leq E_i \\ \exp\left(-\frac{E_j - E_i}{k_B T}\right) & \text{如果 } E_j > E_i \end{cases} = \min\left\{1, \frac{\pi_j}{\pi_i}\right\}$$

接着，Metropolis 等人说明，如此定义的 Markov 链 $P_{n \times n} = (p_{ij})$ 以该 Boltzmann 分布为平稳分布。随便指定一个初始状态，按照 Markov 链 P 迭代地产生下一个状

态。经过足够多次的迭代后，当前状态似乎“遗忘”了初始状态。为了使得随机数序列看起来与初始状态无关，往往把序列的前段预烧掉，而只取剩下的作为抽样结果，即该 Boltzmann 分布的“随机数”。

算法 15.7 (Metropolis 算法的离散版). 令 $Q_{n \times n} = (q_{ij})$ 是用户建议的对称转移矩阵，下面构造一个 Markov 链 $P_{n \times n} = (p_{ij})$ ，使其平稳分布是 $\pi_1\langle 1 \rangle + \pi_2\langle 2 \rangle + \cdots + \pi_n\langle n \rangle$ 。令

$$\alpha_{ij} = \min \left\{ 1, \frac{\pi_j}{\pi_i} \right\} = \begin{cases} 1 & \text{如果 } \pi_j/\pi_i \geq 1 \\ \pi_j/\pi_i & \text{如果 } \pi_j/\pi_i < 1 \end{cases}$$

其中， π_j/π_i 称作 Metropolis 比， α_{ij} 称作接受率。定义转移矩阵 $P_{n \times n} = (p_{ij})$ ，其中

$$p_{ij} = \begin{cases} \alpha_{ij}q_{ij} & \text{如果 } i \neq j \\ 1 - \sum_{j \neq i} p_{ij} & \text{如果 } i = j \end{cases}$$

则 Markov 链 P 的平稳分布就是 $\pi_1\langle 1 \rangle + \pi_2\langle 2 \rangle + \cdots + \pi_n\langle n \rangle$ 。

证明. 由性质 6.14，只需验证 Markov 链 $P = (p_{ij})$ 是可逆的，事实上

$$\pi_i p_{ij} = \pi_i \min \left\{ 1, \frac{\pi_j}{\pi_i} \right\} q_{ij} = \min\{\pi_i, \pi_j\} q_{ij} = \pi_j \min \left\{ 1, \frac{\pi_i}{\pi_j} \right\} q_{ji} = \pi_j p_{ji} \quad \square$$

 **算法 15.7** 基于建议的状态转移矩阵 $Q = (q_{ij})$ 构造 Markov 链 $P = (p_{ij})$ ，可以理解为先以概率 q_{ij} 建议从当前状态 i 转移到另一状态 j ，再以概率 α_{ij} 决定是否接受 j 作为 i 的下一个状态。若建议状态 j 未被接受，则下一个状态仍是 i 。从算法 15.7 不难看出，如果状态 j 的概率 π_j 很小，即便它成为建议状态，也会因为低的接受率而难被抽取到。

另外，理论上预烧的序列越长越好，实践中有时凭经验选择即可，譬如预烧长度设为 10^4 。有时，还需用一些判定 Markov 链是否收敛到平稳分布的方法来设置预烧的长度，但是这个话题超出本书讨论的范围，感兴趣的读者可参阅 [136] 的第七章。

例 15.11. 用 Metropolis 算法 15.7 实现二项分布 $X \sim B(n, p)$ 的抽样。

解. 因为共有 $n+1$ 个状态 $0, 1, \dots, n$ ，不妨建议 $(n+1) \times (n+1)$ 的状态转移矩阵 Q 中每个元素都是 $1/(n+1)$ 。设当前状态是 $X^{(t)} = i$ ，根据算法 15.7，先随机选取建议状态 j ，再在 $\{i, j\}$ 中随机选出一个作为下一个状态 $X^{(t+1)}$ ，选到 j 的概率是 α_{ij} 。

$$\alpha_{ij} = \min \left\{ 1, \frac{P(X=j)}{P(X=i)} \right\} = \min \left\{ 1, \frac{C_n^j p^j (1-p)^{n-j}}{C_n^i p^i (1-p)^{n-i}} \right\} = \min \left\{ 1, \frac{C_n^j}{C_n^i} p^{j-i} (1-p)^{i-j} \right\}$$

以二项分布 $B(20, 0.5)$ 为例, 初始状态分别设为 0 和 20, 利用算法 15.7 产生两列长度为 5000 的伪随机数序列, 预烧掉前 1000 个结果。

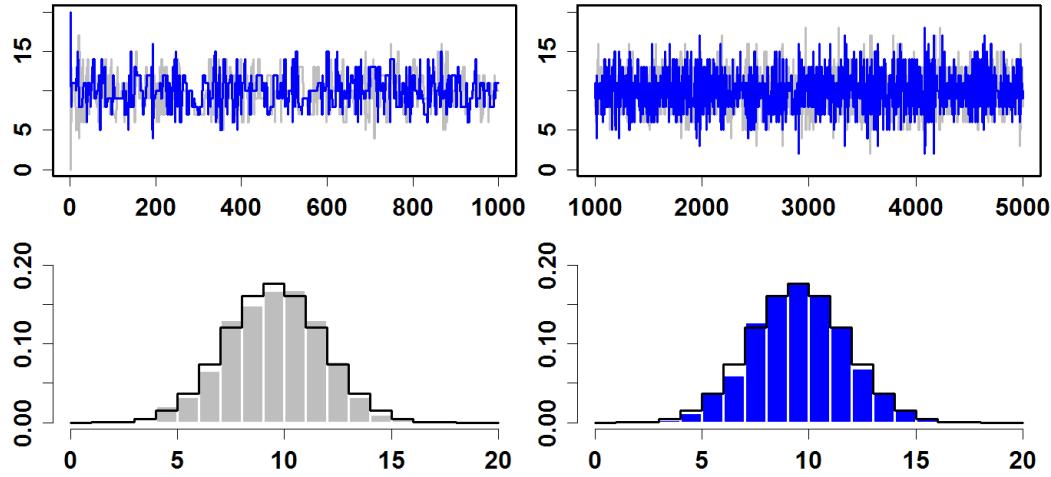


图 15.11: 第一行是利用 Metropolis 算法产生二列 $B(20, 0.5)$ 的伪随机数序列, 横坐标是迭代次数, 初始状态分别是 0 和 20。第二行是预烧掉前 10^3 个结果后伪随机数的直方图, 折线是 $B(20, 0.5)$ 的概率函数。

把离散版的 Metropolis 算法稍作推广, 便得到下面一般形式的 Metropolis 算法, 来产生分布 $\pi(\mathbf{x})$ 的伪随机数。

算法 15.8 (Metropolis, 1953). 设目标分布 $\pi(\mathbf{x})$ 当前所产生的随机数是 $\mathbf{x}^{(t)}$, 初始状态 $\mathbf{x}^{(0)}$ 由用户指定。已知概率函数 $q(\mathbf{y}; \mathbf{x})$ 关于 \mathbf{x}, \mathbf{y} 对称并且比较容易抽样, 譬如多元正态分布 $\phi(\mathbf{y}|\mathbf{x}, I)$, 我们称之为建议分布 (proposal distribution)。下面两个步骤利用建议分布 $q(\mathbf{y}; \mathbf{x})$ 来产生分布 $\pi(\mathbf{x})$ 的下一个随机数 $\mathbf{x}^{(t+1)}$ 。

- 从当前的建议分布 $q(\mathbf{y}; \mathbf{x}^{(t)})$ 产生备选的随机数 \mathbf{y}^* ;
- 产生分布 $U[0, 1]$ 的随机数 u^* , 下一个随机数 $\mathbf{x}^{(t+1)}$ 定义为

$$\mathbf{x}^{(t+1)} = \begin{cases} \mathbf{y}^* & \text{如果 } u^* \leq \alpha(\mathbf{x}^{(t)}, \mathbf{y}^*) \\ \mathbf{x}^{(t)} & \text{否则} \end{cases} \quad (15.3)$$

其中, $\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ 1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})} \right\}$ 称为接受函数

我们把 $r(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$ 称作 Metropolis 比, 把式 (15.3) 称作 Metropolis 规则。该规则的直观含义是 $\mathbf{x}^{(t+1)}$ 以概率 $\alpha(\mathbf{x}^{(t)}, \mathbf{y}^*)$ 选择 \mathbf{y}^* , 以概率 $1 - \alpha(\mathbf{x}^{(t)}, \mathbf{y}^*)$ 选择 $\mathbf{x}^{(t)}$ 。

例 15.12. 利用算法 15.8, 产生分布 $\pi(x) \propto \exp\{-E(x)/T\}$ 的伪随机数, 其中 $E(x) = 10 \cos(x) \sin(x^2/30) + 10^{-5}x^4 + 0.3x$, 温度 $T = 10$ 。

解. 选取建议分布 $q(y; x) = \phi(y|x, 5^2)$ 。依照[算法 15.8](#), 若当前状态是 $x^{(t)}$, 从当前的建议分布 $N(x^{(t)}, 5^2)$ 产生随机数 y^* , 根据 Metropolis 规则对 $x^{(t+1)}$ 进行赋值。

y^* 以大概率落在 $x^{(t)}$ 的附近。如果 Metropolis 比 $\pi(y^*)/\pi(x^{(t)}) \geq 1$, 则 $x^{(t+1)}$ 几乎必然是 y^* 。即便 $\pi(y^*)/\pi(x^{(t)}) < 1$, y^* 依然有可能成为 $x^{(t+1)}$ 。利用[算法 15.8](#), 产生足够长的序列, 只留下最后的 500 个结果, 见下图。

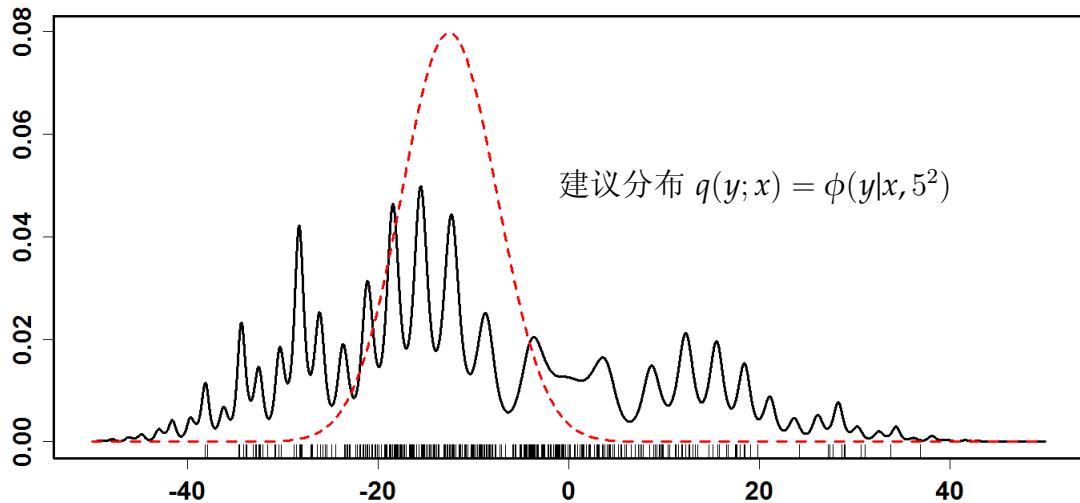


图 15.12: 实线是例 15.12 中的密度函数 $\pi(x) \propto \exp\{-E(x)/10\}$, 虚线是当前的建议分布 $N(x^{(t)}, 5^2)$ 。产生 500 个 $\pi(x)$ 的随机数, 见横轴上的小竖线。

例 15.13 (一维 Ising 模型^{*}). 考虑排成一行的 d 个粒子的自旋状态, 用随机向量 $\mathbf{X} = (X_1, \dots, X_s, \dots, X_d)^\top$ 来描述它, 其中 $X_s = \pm 1$ 表示第 s 个粒子的自旋方向, 只有相邻的自旋有相互作用。例如,

$$\begin{array}{cccccc} X_1 & \cdots & X_s & \cdots & X_d \\ \oplus 1 & \cdots & \ominus 1 & \cdots & \ominus 1 \end{array}$$

假设状态 \mathbf{X} 服从 Boltzmann 分布, 其中状态 $\mathbf{x} = (x_1, \dots, x_s, \dots, x_d)^\top$ 具有能量 $E(\mathbf{x}) = -k_B T \sum_{s=1}^{d-1} x_s x_{s+1}$, 则状态 \mathbf{X} 的概率函数是

$$\pi(\mathbf{x}) = \frac{1}{Z} \exp \left\{ \sum_{s=1}^{d-1} x_s x_{s+1} \right\}$$

利用 Metropolis 算法 15.8 模拟磁化强度 $M = \sum_{s=1}^d X_s$ 的分布。

解. 设当前的状态是 $\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_d^{(t)})^\top$ 。

^{*}Ernst Ising (1900-1998) 是德国物理学家, 于 1925 年给出了一维 Ising 模型的精确解并揭示了一维 Ising 模型无相变。

随机地选择位置 s , 令 $\mathbf{y}^* = (x_1^{(t)}, \dots, -x_s^{(t)}, \dots, x_d^{(t)})^\top$ 。

利用 Metropolis 规则 (15.3), 其中 Metropolis 比是

$$r(\mathbf{x}, \mathbf{y}) = \exp \left\{ -2x_s^{(t)} (x_{s-1}^{(t)} + x_{s+1}^{(t)}) \right\}$$

例如, $d = 60$, 分别取初始状态 $\mathbf{x}^{(0)} = (1, \dots, 1)$ 和 $(-1, \dots, -1)$, 利用 Metropolis 算法迭代 10^5 次, 预烧掉 10^4 个结果后剩下的子序列作为 M 的抽样结果。

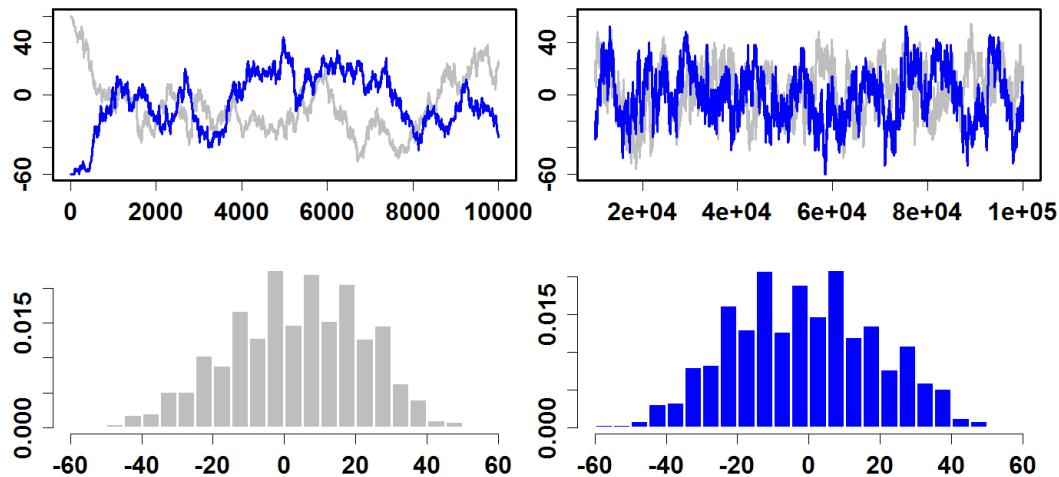


图 15.13: 第一行是利用 Metropolis 算法模拟一维 Ising 模型中磁化强度的两个抽样序列, 横坐标是迭代次数。第二行是预烧掉前 10^4 个结果后的直方图。

15.2.2 Metropolis-Hastings 算法

1970 年, 加拿大统计学家 W. Keith Hastings (1930-) 发表论文《利用 Markov 链的 Monte Carlo 抽样方法及其应用》(该文也被收录在《统计学中的重大突破》第三卷 [98] 中), 将 Metropolis 算法 15.8 中的建议分布从“对称的概率函数”推广到满足 “ $q(\mathbf{y}; \mathbf{x}) > 0 \Leftrightarrow q(\mathbf{x}; \mathbf{y}) > 0$ ” 的一般概率函数, Hastings 提出了下面的 Metropolis-Hastings 算法, 简称 M-H 算法。

算法 15.9 (Metropolis-Hastings, 1970). 设分布 $\pi(\mathbf{x})$ 当前的随机数是 $\mathbf{x}^{(t)}$, 下面基于建议分布 $q(\mathbf{y}; \mathbf{x})$ 产生 $\pi(\mathbf{x})$ 的下一个随机数 $\mathbf{x}^{(t+1)}$ 。

- 从 $q(\mathbf{y}; \mathbf{x}^{(t)})$ 产生备选的随机数 \mathbf{y}^* ;
- 产生分布 $U[0, 1]$ 的随机数 u^* , 下一个随机数 $\mathbf{x}^{(t+1)}$ 定义为

$$\mathbf{x}^{(t+1)} = \begin{cases} \mathbf{y}^* & \text{如果 } u^* \leq \alpha(\mathbf{x}^{(t)}, \mathbf{y}^*) \\ \mathbf{x}^{(t)} & \text{否则} \end{cases}$$

其中, $\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ 1, \frac{\pi(\mathbf{y})q(\mathbf{x}; \mathbf{y})}{\pi(\mathbf{x})q(\mathbf{y}; \mathbf{x})} \right\}$ 是 M-H 算法的接受函数

证明. 只需说明由 M-H 算法产生的 Markov 链是可逆的即可。此 Markov 链的转移函数如下,

$$A(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}; \mathbf{x})\alpha(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}; \mathbf{x}) \min \left\{ 1, \frac{\pi(\mathbf{y})q(\mathbf{x}; \mathbf{y})}{\pi(\mathbf{x})q(\mathbf{y}; \mathbf{x})} \right\} \quad (15.4)$$

进而, $\pi(\mathbf{x})A(\mathbf{x}, \mathbf{y}) = \min\{\pi(\mathbf{x})q(\mathbf{y}; \mathbf{x}), \pi(\mathbf{y})q(\mathbf{x}; \mathbf{y})\}$

易见, $\pi(\mathbf{x})A(\mathbf{x}, \mathbf{y})$ 是有关 \mathbf{x}, \mathbf{y} 的对称函数, $A(\mathbf{x}, \mathbf{y})$ 满足细致平衡条件 (6.7), 即 M-H 算法产生的 Markov 链是可逆的。 \square

 M-H 算法中的接受函数不是唯一的。为使得 (15.4) 定义的 $A(\mathbf{x}, \mathbf{y})$ 满足细致平衡条件, 接受函数也可以定义为

$$\alpha(\mathbf{x}, \mathbf{y}) = \frac{\pi(\mathbf{y})q(\mathbf{x}; \mathbf{y})}{\pi(\mathbf{y})q(\mathbf{x}; \mathbf{y}) + \pi(\mathbf{x})q(\mathbf{y}; \mathbf{x})}$$

$$\alpha(\mathbf{x}, \mathbf{y}) = \frac{\delta(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{x})q(\mathbf{y}; \mathbf{x})} \leq 1, \text{ 其中 } \delta(\mathbf{x}, \mathbf{y}) \text{ 是 } \mathbf{x}, \mathbf{y} \text{ 的对称函数}$$

例 15.14. 接着**例 15.12**, 为了演示**算法 15.9**, 选取偏正态分布(见第 288 页的**定义 4.13**) $Y \sim SN(x, 100; 0.1)$ 为建议分布, 其密度函数 $q(y; x) = 2\phi(0.1(y - x))\Phi(0.01(y - x))$ 不是 y, x 的对称函数。

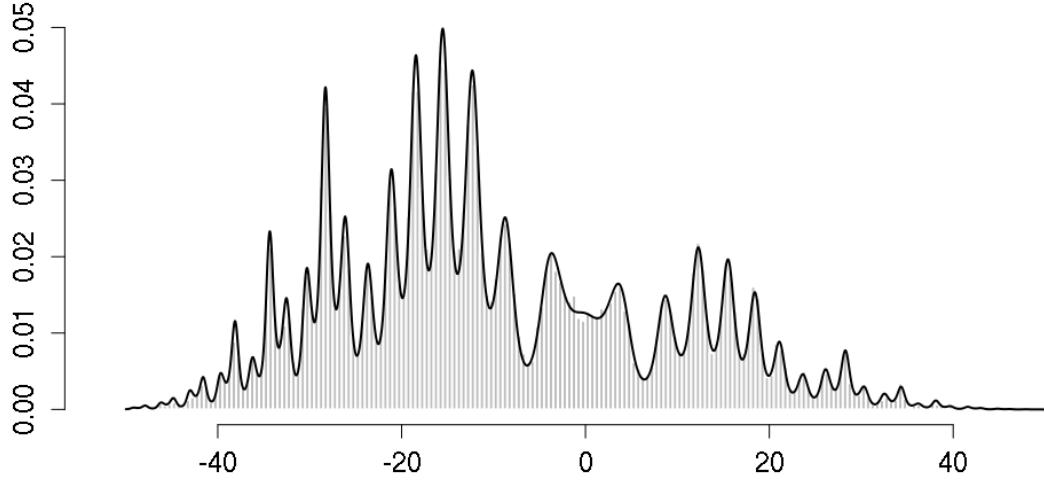


图 15.14: 利用[算法 15.9](#) 和建议分布 $\text{SN}(x, 100; 0.1)$ 来产生例 15.12 所示分布 $\pi(x)$ 的伪随机数, 预烧掉某段后, 绘制剩余结果的直方图。

[算法 15.9](#) 中建议分布的选取多少有些任意性。实践中, 建议分布选取得是否合适, 会影响到 M-H 算法的抽样效果。

练习 15.5. 请读者尝试在[例 15.14](#) 中取 $Y \sim \text{SN}(x, 1; 1)$ 为建议分布, 看看 M-H 算法的效果如何。

2000 年, 刘军等人提出多试 Metropolis 算法 (multiple-try Metropolis 算法, 简称 MTM 算法) [106], 它是 Metropolis-Hastings 算法的变种。

算法 15.10 (多试 Metropolis 算法). 令 $q(\mathbf{x}; \mathbf{y})$ 是满足条件 $q(\mathbf{x}; \mathbf{y}) > 0 \Leftrightarrow q(\mathbf{y}; \mathbf{x}) > 0$ 的任一建议分布, $\lambda(\mathbf{x}; \mathbf{y})$ 是用户定义的有关 \mathbf{x}, \mathbf{y} 的非负对称函数。定义权重函数 $w(\mathbf{x}; \mathbf{y}) = \pi(\mathbf{x})q(\mathbf{x}; \mathbf{y})\lambda(\mathbf{x}; \mathbf{y})$, 设当前状态是 $\mathbf{x}^{(t)}$ 。

- 从建议分布 $q(\mathbf{x}^{(t)}; \mathbf{y})$ 中独立地抽取 m 个建议状态 $\mathbf{y}_1, \dots, \mathbf{y}_m$, 并计算权重 $w(\mathbf{y}_j, \mathbf{x}^{(t)})$, $j = 1, \dots, m$ 。
- 正比于权重 $w(\mathbf{y}_j, \mathbf{x}^{(t)})$, 从 $\mathbf{y}_1, \dots, \mathbf{y}_m$ 抽出候选状态 \mathbf{y}^* 。
- 从建议分布 $q(\mathbf{y}^*; \mathbf{x})$ 中独立地抽取 $m-1$ 个建议状态 $\mathbf{x}_1, \dots, \mathbf{x}_{m-1}$, 令 $\mathbf{x}_m = \mathbf{x}^{(t)}$ 。
- 以概率 $\alpha(\mathbf{x}^{(t)}, \mathbf{y}^*)$ 接受 \mathbf{y}^* 为下一个状态, 其中接受函数 $\alpha(\mathbf{x}, \mathbf{y})$ 定义为

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ 1, \frac{w(\mathbf{y}_1; \mathbf{x}) + \dots + w(\mathbf{y}_m; \mathbf{x})}{w(\mathbf{x}_1; \mathbf{y}) + \dots + w(\mathbf{x}_m; \mathbf{y})} \right\}$$

特别地, 当 $q(\mathbf{x}; \mathbf{y})$ 是对称函数, 选 $\lambda(\mathbf{x}; \mathbf{y}) = 1/q(\mathbf{x}; \mathbf{y})$, 则权重函数 $w(\mathbf{x}; \mathbf{y}) = \pi(\mathbf{x})$, 且接受函数简化为

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ 1, \frac{\pi(\mathbf{y}_1) + \cdots + \pi(\mathbf{y}_m)}{\pi(\mathbf{x}_1) + \cdots + \pi(\mathbf{x}_m)} \right\}$$

本书第 14 章讨论了用 EM 算法解决分支个数已知的高斯混合模型的参数估计问题。如果分支个数也是未知的, 例如样本来自总体 $M_k = \sum_{j=1}^k p_j N(\mu_j, \sigma_j^2)$, 其中分支个数 k , 混合比例 $p_j, j = 1, 2, \dots, k$ (满足 $0 \leq p_j \leq 1$ 且 $\sum_{j=1}^k p_j = 1$), 以及每个分支的参数 $\rho_j = (\mu_j, \sigma_j^2)$ 都是未知的, EM 算法就力不从心了, 这是因为分支个数也成为有待估计的参数——分支个数不同, 其他未知参数 $\boldsymbol{\theta}^{(k)} = (\mathbf{p}^{(k)}, \boldsymbol{\rho}^{(k)}) = (p_1, \dots, p_k, \rho_1, \dots, \rho_k)$ 的维数也不同。

1995 年, 英国统计学家 Peter J. Green (1950-) 提出了一类扩展的 MCMC 方法——可逆跳 MCMC 方法 (reversible jump MCMC, 简称为 RJMCMC) 解决了变维数空间上参数的后验分布的模拟 [64], 也包括分支个数未知的高斯混合模型的参数估计 [134, 147]。

对于不同的维数 $k \neq k'$, RJMCMC 方法的基本想法是分别给 $\boldsymbol{\theta}^{(k)}$ 和 $\boldsymbol{\theta}^{(k')}$ 补充两个合适的模拟结果 $\boldsymbol{\vartheta}^{(k)} \sim f_k(\boldsymbol{\vartheta}^{(k)})$ 和 $\boldsymbol{\vartheta}^{(k')} \sim f_{k'}(\boldsymbol{\vartheta}^{(k')})$, 使得

$$(\boldsymbol{\theta}^{(k')}, \boldsymbol{\vartheta}^{(k')}) = T_{k \rightarrow k'}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\vartheta}^{(k)})$$

是一个双射。由 Metropolis-Hastings 算法, 从模型 M_k 到模型 $M_{k'}$ 的概率为 $\min(1, A_{k \rightarrow k'})$, 其中 $A_{k \rightarrow k'}$ 为

$$A_{k \rightarrow k'} = \underbrace{\frac{\pi(k', \boldsymbol{\theta}^{(k')})}{\pi(k, \boldsymbol{\theta}^{(k)})}}_{\text{模型比}} \times \underbrace{\frac{\pi_{k' \rightarrow k} f_{k'}(\boldsymbol{\vartheta}^{(k')})}{\pi_{k \rightarrow k'} f_k(\boldsymbol{\vartheta}^{(k)})}}_{\text{建议比}} \times \underbrace{\left| \frac{\partial T_{k \rightarrow k'}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\vartheta}^{(k)})}{\partial (\boldsymbol{\theta}^{(k)}, \boldsymbol{\vartheta}^{(k)})} \right|}_{\text{雅可比行列式 } |J|} \quad (15.5)$$

15.2.3 Gibbs 抽样与切片抽样

1984 年, 美国统计学家 Stuart Geman 和 Donald Geman (1943-) 两兄弟利用统计力学方法研究图像处理时提出一类特殊的 MCMC 方法, 因为用到了 Gibbs 分布, 该方法被命名为 Gibbs 抽样 (Gibbs sampling) [56], 适用于多元分布 $\mathbf{X} = (X_1, \dots, X_d)^\top \sim \pi(\mathbf{x})$ 的抽样。我们把实现 Gibbs 抽样的算法或程序称为 Gibbs 抽样器 (Gibbs sampler)。

Gibbs 抽样的理论基础是下面的结果, 由若干 (一维) 条件分布来定义 Markov 链。高维分布的 Gibbs 抽样, 正是因为下面的结果而变为若干低维分布的抽样。

定理 15.1. 已知概率分布 $\pi(\mathbf{x})$, 如果从 \mathbf{x} 到 \mathbf{x}' 的转移概率 $A(\mathbf{x}, \mathbf{x}')$ 为

$$A(\mathbf{x}, \mathbf{x}') = \pi(x_1|x_2, \dots, x_d)\pi(x_2|x'_1, x_3, \dots, x_d) \cdots \pi(x_d|x'_1, \dots, x'_{d-1})$$

则由 $A(\mathbf{x}, \mathbf{x}')$ 定义的 Markov 链的平稳分布是 $\pi(\mathbf{x})$ 。

算法 15.11 (依次扫描型 Gibbs 抽样). 抽取分布 $\pi(\mathbf{x})$ 的随机数: 初始值 $\mathbf{x}^{(0)}$ 由用户给定, 设 $\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_d^{(t)})^\top$ 是第 t 轮 Gibbs 抽样的结果。为得到第 $t+1$ 轮抽样的结果 $\mathbf{x}^{(t+1)} = (x_1^{(t+1)}, \dots, x_d^{(t+1)})^\top$, 依次从下面的条件分布中抽取随机数 $x_j^{(t+1)}$,

$$X_j^{(t+1)} \sim \pi(x_j|x_1^{(t+1)}, \dots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \dots, x_d^{(t)}), \text{ 其中 } j = 1, \dots, d$$

例 15.15. 已知 $\mathbf{X} \sim N_2(\mathbf{0}, \Sigma)$, 协方差矩阵 $\Sigma = (1, \rho; \rho, 1)$, 其中 $|\rho| < 1$ 。设初始值为 $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)})^\top$, 利用**算法 15.11**, Gibbs 抽样的过程是

$$\begin{aligned} X_1^{(t+1)} | X_2^{(t)} &= x_2^{(t)} \sim N(\rho x_2^{(t)}, 1 - \rho^2) \\ X_2^{(t+1)} | X_1^{(t+1)} &= x_1^{(t+1)} \sim N(\rho x_1^{(t+1)}, 1 - \rho^2) \end{aligned}$$

按照上述迭代规则, 我们得到 $\mathbf{X}^{(t)} = (X_1^{(t)}, X_2^{(t)})^\top$ 的分布

$$\begin{pmatrix} X_1^{(t)} \\ X_2^{(t)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \rho^{2t-1} x_2^{(0)} \\ \rho^{2t} x_2^{(0)} \end{pmatrix}, \begin{pmatrix} 1 - \rho^{4t-2} & \rho - \rho^{4t-1} \\ \rho - \rho^{4t-1} & 1 - \rho^{4t} \end{pmatrix} \right)$$

当 t 很大时, $\mathbf{X}^{(t)} = (X_1^{(t)}, X_2^{(t)})^\top$ 近似地服从分布 $N(0, 0, 1, 1, \rho)$ 。下图显示了 $N(0, 0, 1, 1, -0.8)$ 的 Gibbs 抽样的全过程以及预烧后所得抽样结果的散点图。

例 15.16. 接着习题 4.30, 下面给出 $\mathbf{X} = (X_1, \dots, X_n)^\top$ 的 Gibbs 抽样: 令 \mathbf{x}_{-j} 是向量

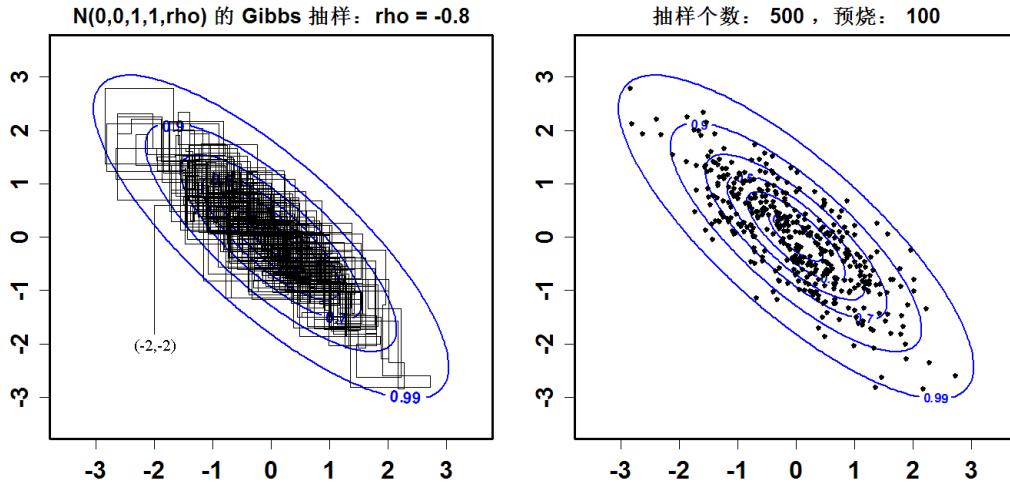


图 15.15: 当 $\rho = -0.8$, 初始值取 $(-2, -2)^\top$ 时, 二元正态分布 $N(0, 0, 1, 1, \rho)$ 的 Gibbs 抽样过程如左图所示, 右图是预烧掉 100 个最初抽样结果后的散点图。

$\mathbf{x} = (x_1, \dots, x_n)^\top$ 去掉第 j 个分量而得到的向量, 则

$$\begin{aligned} P(X_j = i | \mathbf{X}_{-j} = \mathbf{x}_{-j}, \boldsymbol{\alpha}) &= \frac{P(X_j = i, \mathbf{X}_{-j} = \mathbf{x}_{-j} | \boldsymbol{\alpha})}{P(\mathbf{X}_{-j} = \mathbf{x}_{-j} | \boldsymbol{\alpha})}, \text{ 其中 } i = 1, 2, \dots, k \\ &\propto \frac{\Gamma(\alpha_i + n^{(i)})}{\Gamma(\alpha_i + n_{-j}^{(i)})}, \text{ 其中 } \begin{cases} n^{(i)} \text{ 是 } i, \mathbf{x}_{-j} \text{ 中 } i \text{ 的个数} \\ n_{-j}^{(i)} \text{ 是 } \mathbf{x}_{-j} \text{ 中 } i \text{ 的个数} \end{cases} \\ &= \frac{\Gamma(\alpha_i + n_{-j}^{(i)} + 1)}{\Gamma(\alpha_i + n_{-j}^{(i)})} = \alpha_i + n_{-j}^{(i)} \end{aligned}$$

因为 $\sum_{i=1}^k P(X_j = i | \mathbf{X}_{-j} = \mathbf{x}_{-j}, \boldsymbol{\alpha}) = 1$, 所以

$$P(X_j = i | \mathbf{X}_{-j} = \mathbf{x}_{-j}, \boldsymbol{\alpha}) = \frac{\alpha_i + n_{-j}^{(i)}}{\sum_{i=1}^k (\alpha_i + n_{-j}^{(i)})} = \frac{\alpha_i + n_{-j}^{(i)}}{\sum_{i=1}^k n^{(i)} - 1 + \sum_{i=1}^k \alpha_i} = \frac{\alpha_i + n_{-j}^{(i)}}{n - 1 + \sum_{i=1}^k \alpha_i}$$

在算法 15.11 中, 从 $\mathbf{x}^{(t)}$ 到 $\mathbf{x}^{(t+1)}$ 的更新也可以不那么死板地依分量的次序而行, 下面的算法就是按给定的分布列随机地挑选出一个分量加以更新, 效果与算法 15.11 等同。

算法 15.12 (随机扫描型 Gibbs 抽样). 令 $\mathbf{X}_{-j} = (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_d)^\top$, 从 $\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_d^{(t)})$ 得到 $\mathbf{x}^{(t+1)}$ 分为下面两个步骤。

- 根据某分布列 $p_1\langle 1 \rangle + \dots + p_j\langle j \rangle + \dots + p_d\langle d \rangle$ 选择坐标 j ; 然后
- 从分布 $X_j \sim \pi(x_j | \mathbf{x}_{-j}^{(t)})$ 抽得随机数 $x_j^{(t+1)}$; 并置 $\mathbf{x}_{-j}^{(t+1)} = \mathbf{x}_{-j}^{(t)}$, 即除了第 j 个分量外, $\mathbf{x}^{(t+1)}$ 与 $\mathbf{x}^{(t)}$ 相同。

练习 15.6. 论证随机扫描型 Gibbs 抽样的合理性。提示：根据算法 15.12 中 $x^{(t+1)}$ 的构造方式，说明 $X^{(t+1)}$ 与 $X^{(t)}$ 的分布相同。

练习 15.7. 利用算法 15.12 来实现例 15.15 的抽样问题。

切片抽样 (slice sampling) 是一类特殊的 Gibbs 抽样方法 [113]，它的理论依据是第 714 页所描述的有关公式 (15.1) 的基本事实：目标分布 $X \sim \pi(x)$ 是均匀分布 $(X, Y)^\top \sim U\{(x, y) : 0 \leq y \leq \pi(x)\}$ 的边缘分布。为得到 $\pi(x)$ 的抽样，先搞定 $U\{(x, y) : 0 \leq y \leq \pi(x)\}$ 的 Gibbs 抽样。

算法 15.13 (切片抽样). 设当前的抽样结果是 $(x^{(t)}, y^{(t)})^\top$ ，下一步为产生 $(x^{(t+1)}, y^{(t+1)})^\top$ ，Gibbs 抽样的过程是

- 先抽取均匀分布 $U[0, \pi(x^{(t)})]$ 的随机数 $y^{(t+1)}$ ，然后
- 再抽取均匀分布 $U\{D^{(t+1)}\}$ 的随机数 $x^{(t+1)}$ ，其中区域 $D^{(t+1)} = \{x : \pi(x) \geq y^{(t+1)}\}$ 。

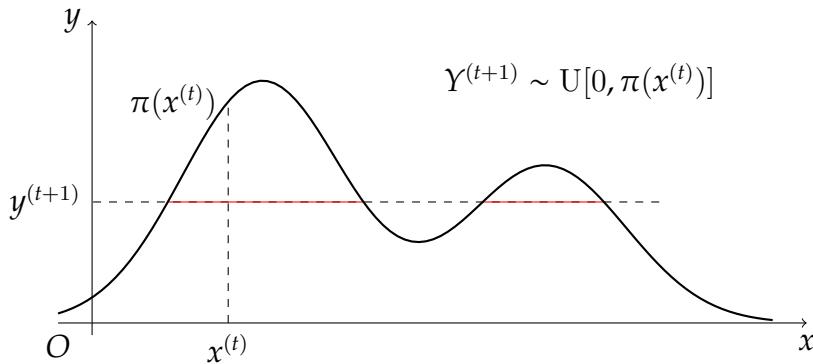


图 15.16: 切片抽样的示意图。若 $\pi(x)$ 不是单峰的，则算法 15.13 有可能导致区域 $D^{(t+1)} = \{x : \pi(x) \geq y^{(t+1)}\}$ 不连通，使得 $X^{(t+1)} \sim U\{D^{(t+1)}\}$ 的抽样变得困难。

例 15.17. 利用切片抽样算法 15.13 产生标准正态分布 $X \sim N(0, 1)$ 的随机数：选择初始值 $x^{(0)} = 1.14$ ，设当前的抽样结果是 $(x^{(t)}, y^{(t)})^\top$ 。

- 先产生 $U[0, \phi(x^{(t)})]$ 的随机数 $y^{(t+1)}$ ，然后
- 再产生 $U[-a, a]$ 的随机数 $x^{(t+1)}$ ，其中 $a = \sqrt{-2 \ln(y^{(t+1)}) \sqrt{2\pi}}$ 。

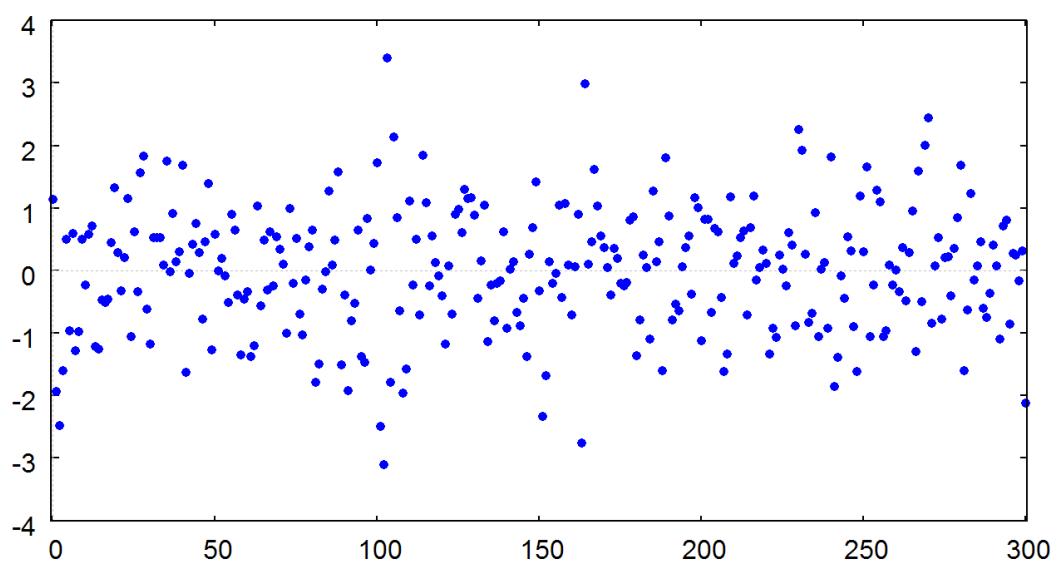


图 15.17: 例 15.17 利用切片抽样产生 $N(0, 1)$ 的随机数。横坐标是函数 `slice` 的迭代次数, 纵坐标是迭代结果——这是一个非确定性的动力系统。

15.3 随机模拟技术的应用

15.3.1 模拟退火算法

在冶金中，将金属加热至高温，金属原子的位置被打乱，缓慢地降温（或称退火）使得金属比快速地降温要更加坚硬，因为金属原子有更多的机会找到低能量的位置。1983年，S. Kirkpatrick 等人提出模拟退火 (simulated annealing, SA) 算法 [88]，通过模拟金属退火的过程来解决组合最优化问题，该算法是 MCMC 技术的一个成功应用。

假设最优化问题是寻找函数 $E(x)$ 的最小值，则它等价于寻找 $\exp\{-E(x)/T\}$ 的最大值，其中 T 表示任意给定的“温度”。

考虑温度的单调下降序列 $T_0 > T_1 > \dots > T_k > \dots$ ，其中 T_0 很大，且 $\lim_{k \rightarrow \infty} T_k = 0$ ，分布 $\pi_k(x) \propto \exp\{-E(x)/T_k\}$ 随着时刻 k 的增大越来越“高瘦”（见图 15.18），密度越来越集中在 $E(x)$ 的最小值点，即 $\pi_k(x)$ 的最大值点附近——这意味着，对分布 $\pi_k(x)$ 进行抽样，也越来越有可能取到最优解或其邻近的点。

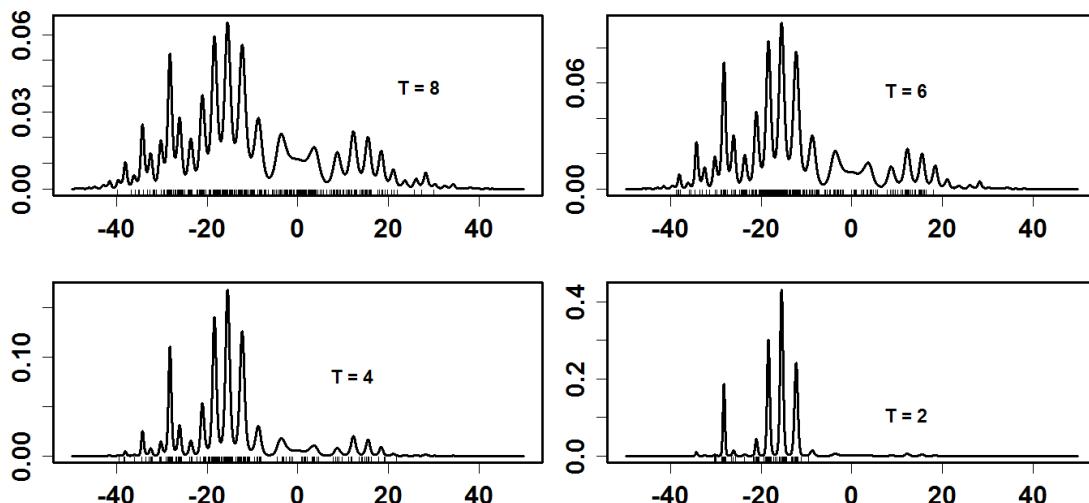


图 15.18：接着例 15.12，产生 $\pi(x) \propto \exp\{-E(x)/T\}$ 的随机数 500 个，其中温度 $T = 8, 6, 4, 2$ 。随着温度的降低，随机数越来越集中在 $E(x)$ 的最小值附近。

算法 15.14 (模拟退火，1983). 为寻找函数 $E(x)$ 的最小值，

- 给定初始点 x_0 和温度序列 $T_0 > T_1 > \dots > T_k > \dots$ 。
- 以 x_k 为起点，利用 MCMC 技术从目标分布 $\pi_k(x) \propto \exp\{-E(x)/T_k\}$ 产生 n_k 个随机数，预烧掉 $n_k - 1$ 个结果，将最后一个记作 x_{k+1} 。
- 置 $k \leftarrow k + 1$ ，重复上一个步骤。



虽然模拟退火**算法 15.14**的结果是不确定的，依赖于温度序列的设置和 $\pi_k(x)$ 的抽样，但瑕不掩瑜，该算法的优势依然是明显的——随着温度的降低，模拟

退火^{算法 15.14}利用 MCMC 技术总有可能逃离局部极值点。模拟退火算法所采用的 MCMC 技术也不限于 Metropolis 算法，只要能出色地完成 $\pi_k(x)$ 的抽样任务，任何 MCMC 方法都是适用的。模拟退火的过程简单地描述为

$$\begin{array}{ll} \text{温度:} & T_0 \rightarrow T_1 \rightarrow \cdots \rightarrow T_k \rightarrow \cdots \\ \text{抽样:} & \downarrow \quad \downarrow \quad \downarrow \\ \text{近似解:} & x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_k \rightarrow \cdots \end{array}$$

1984 年，美国统计学家 Stuart Geman 和 Donald Geman (1943-) 两兄弟证明了，只要温度变量 T_k 以足够慢的速度下降，模拟退火算法以概率 1 收敛到 $E(x)$ 的最小值点 [56]，这个性质使得模拟退火算法成为少有的全局最优化算法之一。尽管如此，对模拟退火算法，没有普适的最优降温策略。为了得到满意解，往往需要同时尝试几种不同的降温策略。

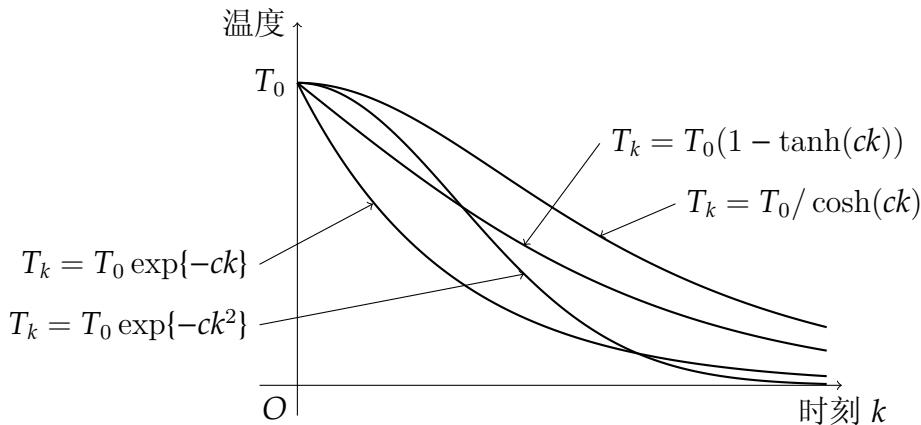


图 15.19: 几种不同的降温策略，其中， c 是某个正常数。

另外，模拟退火算法无法确保降温就一定能够得到更好的解。为了避免“捡了芝麻丢了西瓜”，建议每次更新状态时也顺便更新当前最优解。另外，在适当时机以当前最优解为初始点，重新从某高温处开始新一轮的模拟退火，有助于避免算法陷入局部最优解。总而言之，降温策略对模拟退火算法至关重要，这也是很多技巧的施展之地。

例 15.18. 接着^{例 15.12}，求 $E(x) = 10 \cos(x) \sin(x^2/30) + 10^{-5}x^4 + 0.3x$ 的最小值点。

解. (i) 给定初始点 $x_0 = 50$ ，设置抽样次数 $n_k = 100$ ，温度序列 $T_0 = 10 > T_1 = 6 > T_2 = 2 > T_3 = 1 > T_4 = 0.5$ 。(ii) 在温度 T_k 之下，以 $\pi_k(x) \propto \exp\{-E(x)/T_k\}$ 为目标分布，以 x_k 为初始点，利用 Metropolis 算法 15.14 产生 n_k 个随机数，预烧掉 $n_k - 1$ 个结果，把最后一个结果 x_{k+1} 作为冷却温度 T_{k+1} 之下的初始点。将步骤 (ii) 迭代数次后得到 x_1, \dots, x_5 ，如下图所示。

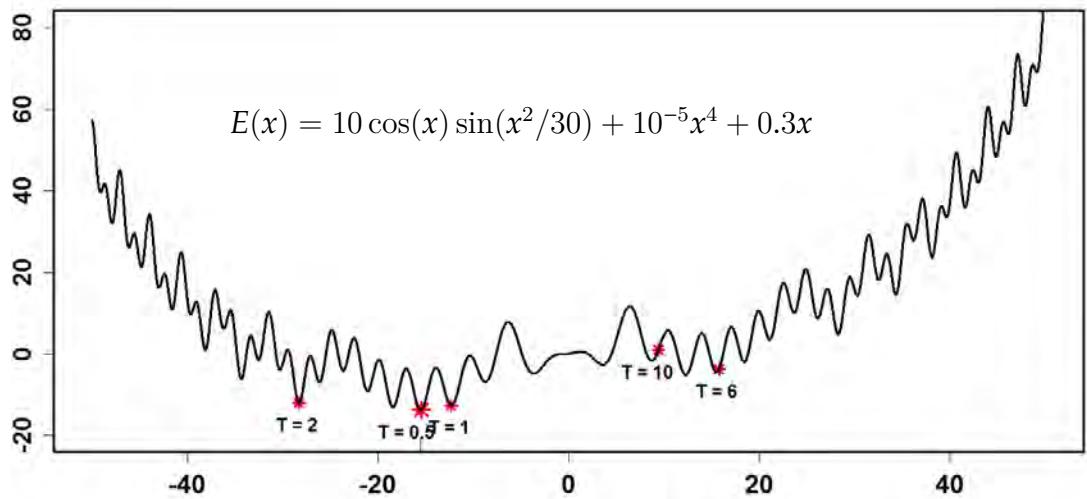


图 15.20: 利用模拟退火算法寻找函数 $E(x)$ 的最小值点。随着温度 T_k 的缓慢降低, 分布 $\pi_k(x) \propto \exp\{-E(x)/T_k\}$ 的随机数几乎必然收敛于 $E(x)$ 的最小值点。

例 15.19. 利用模拟退火算法求二元函数 $E(x, y) = -4(x-1)^2 \exp\{-x^2 - (y+1)^2\} + 6(xy - x^3 - y^5) \exp\{-x^2 - y^2\} + \exp\{-(x+1)^2 - y^2\} + 0.5 \ln(x^2 + y^2 + 1)$ 的最小值点。

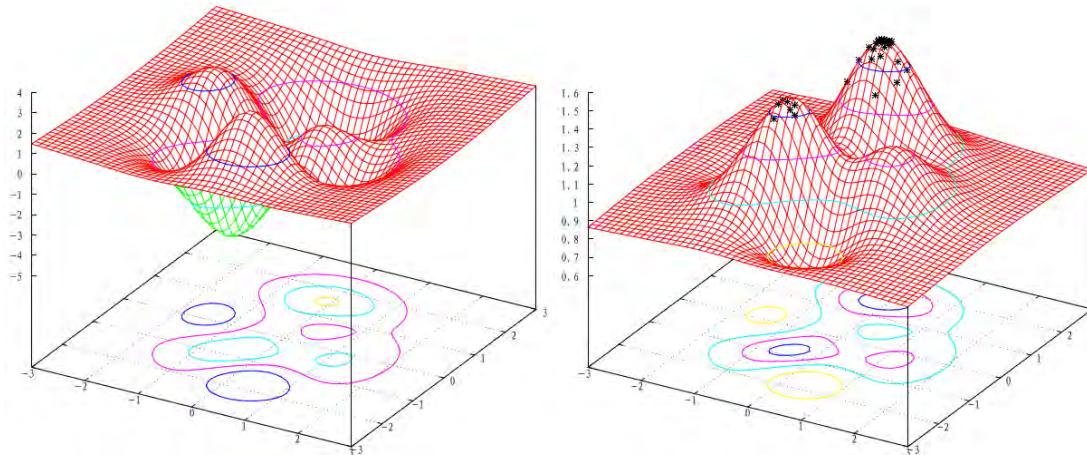


图 15.21: 函数 $E(x, y)$ (左图) 的最小值点, 即 $\exp\{-E(x, y)/10\}$ (右图) 的最大值点。即便从一个“坏点”出发, 模拟退火算法 15.14 也能很快收敛于 $E(x, y)$ 的最小值点 (右图曲面上的星点是中间结果)。

解. 曲面 $E(x, y)$ 和 $\exp\{-E(x, y)/10\}$ 见图 15.21。显然, 当 $x, y \rightarrow \pm\infty$ 时, 有 $E(x, y) \rightarrow +\infty$ 。为得到分布 $\pi_k(x, y) \propto \exp\{-E(x, y)/T_k\}$ 的随机数, 在 Metropolis 算法 15.8 中, 利用二元正态分布 $\phi(x, y|x^{(t)}, y^{(t)}, 1, 1, 0.2)$ 来产生建议状态 (x^*, y^*) , 其中 $(x^{(t)}, y^{(t)})$ 是当前状态。

设置初始点 $(-2, -2)$, 抽样个数 $n_k = 100$, 温度序列 $T_0 = 0.50, T_1 = 0.49, \dots, T_{48} = 0.02, T_{49} = 0.01$, 模拟退火 [算法 15.14](#) 寻找到 $E(x, y)$ 的最小值点是 $(-0.077188, 1.557944)$ 。

海客谈瀛洲，烟涛微茫信难求。越人语天姥，云霓明灭或可睹。
 天姥连天向天横，势拔五岳掩赤城。天台四万八千丈，对此欲倒东南倾。
 我欲因之梦吴越，一夜飞度镜湖月。湖月照我影，送我至剡溪。
 谢公宿处今尚在，绿水荡漾清猿啼。脚著谢公屐，身登青云梯。
 半壁见海日，空中闻天鸡。千岩万转路不定，迷花倚石忽已暝。
 熊咆龙吟殷岩泉，栗深林兮惊层巅。云青青兮欲雨，水澹澹兮生烟。
 列缺霹雳，丘峦崩摧。洞天石扇，訇然中开。
 青冥浩荡不见底，日月照耀金银台。霓为衣兮风为马，云之君兮纷纷而来下。
 虎鼓瑟兮鸾回车，仙之人兮列如麻。忽魂悸以魄动，恍惊起而长嗟。
 惟觉时之枕席，失向来之烟霞。世间行乐亦如此，古来万事东流水。
 别君去兮何时还？且放白鹿青崖间，须行即骑访名山。
 安能摧眉折腰事权贵，使我不得开心颜！

李白《梦游天姥吟留别》

例 15.20. 在约束条件 $y \geq 0, x^2 + 2y^2 - x + y + z = 43$ 和 $7x + y + z = 10$ 之下, 求目标函数 $f(x, y, z) = x + 3y + yz - z^2$ 的最大值 M 和最大值点。

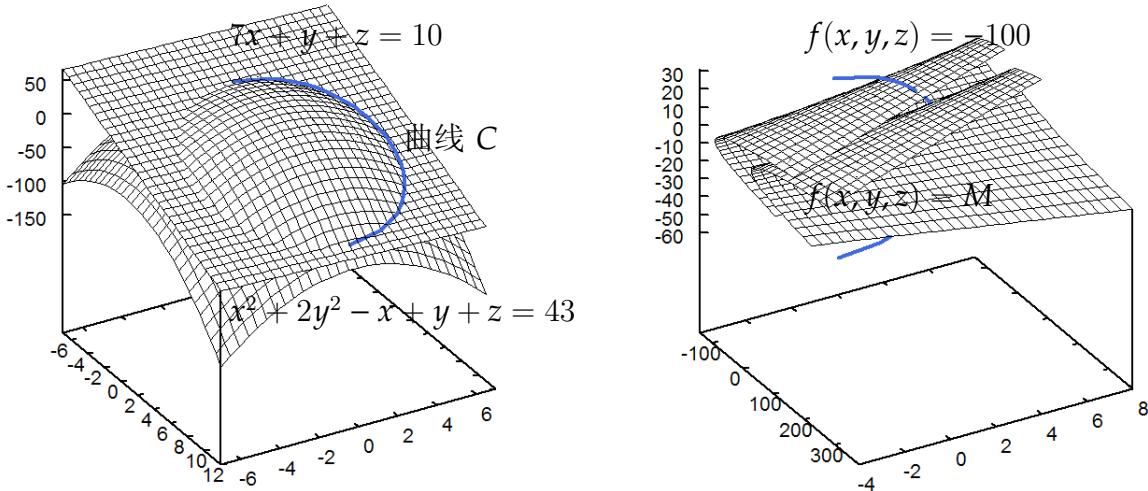


图 15.22: 左图直观显示了非线性约束条件, 即曲线 C 。右图是曲面 $f(x, y, z) = M$, 与曲线 C 仅交于一点, 交点即是 $f(x, y, z)$ 的最大值点。

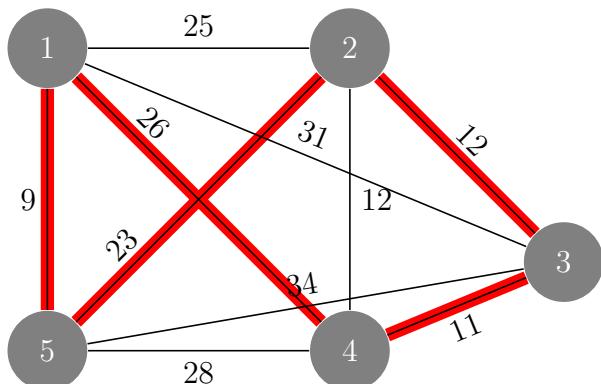
解. 一旦给出自由变量 $x \in [-3, 11]$ 的值, 由约束条件, 可以求出

$$y = \sqrt{\frac{49 - (x - 4)^2}{2}}, \quad z = 10 - 7x - y$$

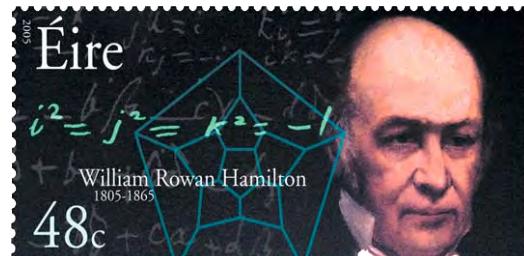
通常的作法是将上式代入 $f(x, y, z)$, 将之整理成有关 x 的函数 $g(x)$, 再利用数值计算的方法求得 $g(x)$ 的最大值 $M = 18.02316$, 最大值点是 $x = 0.535781, y = 4.301115, z = 1.948418$ 。

下面用模拟退火算法求解该问题。设置温度序列 $T_0 = 1, T_1 = 0.99, T_2 = 0.98, \dots$, 抽样个数 $n_k = 100$ 。利用 Metropolis 算法从分布 $\pi_k(x, y, z) \propto \exp\{f(x, y, z)/T_k\}$ 中抽样, 其中建议状态是这样产生的: 先从均匀分布 $U[-3, 11]$ 随机地产生 x 的值 x^* , 再由约束条件求出 y, z 的值 y^*, z^* 而得到建议状态 (x^*, y^*, z^*) 。模拟退火算法求得的最大值点是 $(0.5357719, 4.301112, 1.948485)$, 与正解的误差小于 10^{-4} 。

例 15.21 (旅行推销员问题, travelling salesman problem, 简称 TSP). 有 n 个城市 $1, 2, \dots, n$, 已知任意两个城市 i, j 间的距离 d_{ij} , 其中 $i, j = 1, 2, \dots, n$ 。推销员从某个城市 k 出发, 每个城市只许访问一次, 最后回到城市 k , 求走遍这 n 个城市的最短路径^{*}。例如,



(a) Hamilton 回路的例子



(b) W. M. Hamilton (1805-1865), 爱尔兰数学家、物理学家和天文学家。

图 15.23: 用无向加权完全图来给 TSP 建模: 城市表示为节点, 城市间的距离表示为边的权重。旅行推销员问题即寻找权重之和最小的 Hamilton 回路。

任意路径 $X = (X_1, X_2, \dots, X_n)$ 都可表示为 $(1, 2, \dots, n)$ 的某一置换。TSP 就是寻找路径 $x = (x_1, x_2, \dots, x_n)$, 使得下述目标函数达到最小。

$$E(x) = \sum_{j=1}^{n-1} d_{x_j, x_{j+1}} + d_{x_n, x_1}$$

*1972 年, 美国计算机科学家、数学家 Richard Manning Karp (1935-) 证明了 TSP 是 NP 难的。当 n 很小时, 穷举 $n!$ 种可能便可找出最优解。然而 n 很大时, 组合爆炸让穷举法不可行。Karp 曾给出 TSP 的动态规划算法, 算法复杂度为 $O(n^2 2^n)$ 。

解. 假设状态 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 服从 Boltzmann 分布, 则分布函数是

$$\pi(\mathbf{x}) \propto \exp\left\{-\frac{E(\mathbf{x})}{T}\right\}$$

利用模拟退火算法 15.14, 可以近似地求解 TSP。令初始状态为 $\mathbf{x}_0 = (1, 2, \dots, n)$, 假设当前温度是 T_k , 按下面的方法产生状态序列。

- 状态转移: 抽取 $U\{1, 2, \dots, n\}$ 的随机数 i , 再抽取 $U\{1, \dots, i-1, i+1, \dots, n\}$ 的随机数 j , 交换 \mathbf{x} 中的 x_i, x_j 得到建议状态 \mathbf{y} , 即

$$\begin{array}{ccccccccccccc} \mathbf{x}: & x_1 & \cdots & x_{i-1} & x_i & x_{i+1} & \cdots & x_{j-1} & x_j & x_{j+1} & \cdots & x_n \\ & & & & \downarrow & & & & \downarrow & & & \\ \mathbf{y}: & x_1 & \cdots & x_{i-1} & x_j & x_{i+1} & \cdots & x_{j-1} & x_i & x_{j+1} & \cdots & x_n \end{array}$$

- 利用 Metropolis 规则来定义下一个状态: 产生 $U[0, 1]$ 随机数 u^* , 若 $u^* \leq \exp\{-(E(\mathbf{y}) - E(\mathbf{x}))/T_k\}$, 则下一个状态是 \mathbf{y} , 否则还是 \mathbf{x} 。

对图 15.23 的 TSP, 令温度序列是 $T_0 = 10, T_1 = 2^{-1}T_0, \dots, T_k = 2^{-k}T_0, \dots$, 取 $n_k = 4$, 模拟退火算法 15.14 迅速地收敛到正解, 即是图 15.23 中粗线所示的路径。

 模拟退火算法的效果与温度变量 T_k 递减的速度有关。Geman 兄弟证明, 若温度 T_k 按 $O(\ln N_k^{-1})$ 递减, 其中 $N_k = n_1 + \dots + n_k$, 模拟退火算法几乎必然收敛于全局最优解 [56]。然而在实践中, 这样的退火速度过于缓慢, 为了效率人们往往像例 15.21 采用简单的退火策略 $T_k = c^k T_0$, 其中 $0 < c < 1$ 。该策略不能保证收敛于全局最优解, 有时需要通过多次尝试来调节参数 c 和 n_k 。例如,

例 15.22. 已知美国 48 个州府的直角坐标如下表所示, 用模拟退火算法求解这 48 个城市的 TSP。

表 15.1: 美国 48 个州府的直角坐标, 城市间距离采用欧氏距离。

6734	1453	2233	10	5530	1424	401	841	3082	1644	7608	4458
7573	3716	7265	1268	6898	1885	1112	2049	5468	2606	5989	2873
4706	2674	4612	2035	6347	2683	6107	669	7611	5184	7462	3590
7732	4723	5900	3561	4483	3369	6101	1110	5199	2182	1633	2809
4307	2322	675	1006	7555	4819	7541	3981	3177	756	7352	4506
7545	2801	3245	3305	6426	3173	4608	1198	23	2216	7248	3779
7762	4595	7392	2244	3484	2829	6271	2135	4985	140	1916	1569
7280	4899	7509	3239	10	2676	6807	2993	5185	3258	3023	1942

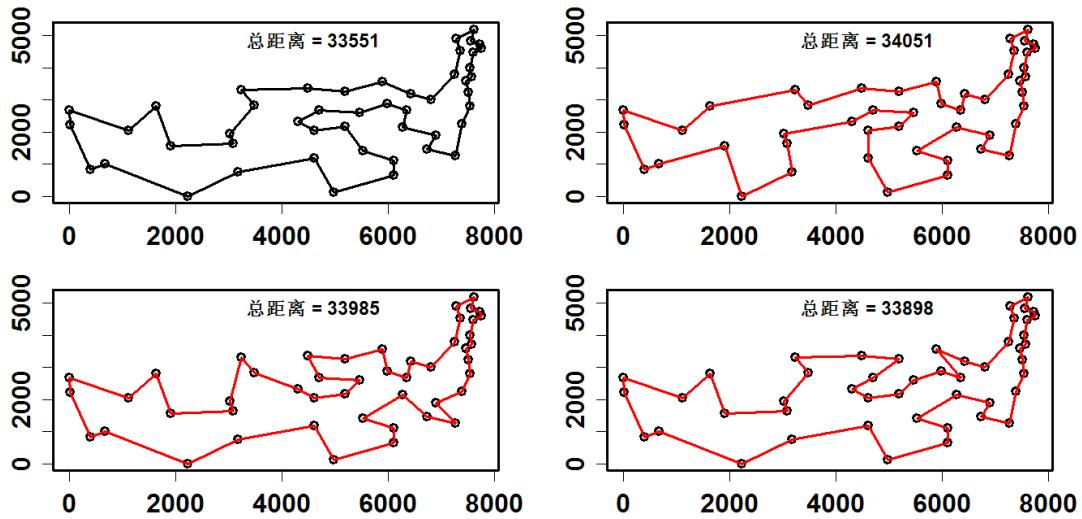


图 15.24: 美国 48 个州府的地图 (直角坐标系, 欧氏距离), 左上角的小图是 TSP 的正解。其他小图是利用类似例 15.21 的模拟退火算法求得的近似解。

解. 设置抽样个数 $n_k = 50$, 温度序列 $T_0 = 10^4, T_k = 0.999^k T_0$ 。仿照例 15.21 的方法, 求得该 TSP 的模拟退火解。图 15.24 显示了正解和三个模拟退火解。

练习 15.8. 请读者尝试不同的降温策略, 与例 15.22 的结果作比较, 看看在这个具体问题上哪种降温策略使得模拟退火算法更有效。

15.3.2 缺失数据的多重填补算法

缺失数据问题在数据分析里很常见，产生这个问题的原因各式各样：有的是因为某个变量的观测代价很大，有的是涉及隐私信息难以采集，有的是疏忽遗漏，有的是观测了错误的对象，等等。按照随机性特点，缺失数据可分为以下三种类型：

- 完全随机缺失 (missing completely at random, MCAR): 如果变量 Y 产生缺失数据的概率与 Y 的取值以及其他变量的取值无关，即

$$P(Y \text{ missing}|Y, X) = P(Y \text{ missing})$$

- 随机缺失 (missing at random, MAR): 如果控制好其他变量的取值， Y 产生缺失数据的概率与 Y 的取值无关，即

$$P(Y \text{ missing}|Y, X) = P(Y \text{ missing}|X)$$

- 非随机缺失 (missing not at random, MNAR): 有一个具体的原因或机制导致数据系统性地缺失。例如，

一个样本点 $\mathbf{x} = (x_1, \dots, x_n)^\top$ 中某个或某些分量是缺失的，我们则称这是一个有缺失数据的样本点。一个观察数据集里可能有很多这样有缺失数据的样本点，如何处理它们呢？有下述一些方法可供选择。

- 直接从观察数据集中删除那些含有缺失数据的样本点的作法称为成列删除 (listwise deletion)。它的优点是简单，对 MCAR 有效；缺点是当缺失数据为 MAR 时，可能会造成有偏估计。另外，仅仅因为少了个把分量而抛弃整个样本点会造成数据资源的浪费。
- 成对删除 (pairwise deletion) 试图克服成列删除浪费资源的缺点，利用尽可能多的样本点来计算统计量。例如，线性回归中仅用到样本均值和样本协方差矩阵，在计算两变量 X, Z 的样本协方差时，只需要删除那些在这两个变量上有缺失数据的样本点。这种方法的优点是尽可能多地利用了已有的数据资源，但也有一些缺点，譬如，如此构造的样本协方差矩阵可能不是正定的。
- 填补 (imputation) 就是以某种合理的猜测将缺失数据补充完整。譬如，利用边缘分布的均值，或者利用非缺失数据得到的回归曲线来填补缺失数据。
- 多重填补 (multiple imputation) 算法是一种基于随机模拟和并行技术的普适方法，见图 15.25。并行地执行下面的过程：随机地或以某种合理的猜测将缺失数据补充完整，然后利用该完整数据对未知参数进行估计。最后，将这些估计值 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 汇总成一个结果 $\hat{\theta}$ 。

多重填补算法的最初想法由美国统计学家 Donald Bruce Rubin (1943-) 于 1977 年提出，并在 *Multiple Imputation for Nonresponse in Surveys* (1987) 一书中详尽阐述。Rubin 在缺失数据、因果推断和贝叶斯分析上做了大量的工作，另外，他也是 EM 算法的提出者之一。多重填补的想法很简单，可以用下面的图来直观说明。

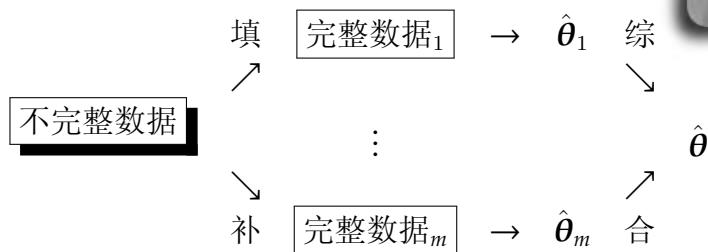


图 15.25: 多重填补算法示意图：如何填补缺失数据没有具体的要求，随机的也好，启发式的也好。该算法的亮点在于通过多次填补来减少参数估计的偏差，。

例 15.23. 总体 $(X, Y)^T$ 服从某正态分布，有一些观测值缺失。

15.3.3 数据增扩算法

观测数据 \mathbf{y}_{obs} 连同缺失数据 \mathbf{y}_{mis} 组成完全数据 $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$, 贝叶斯数据分析常需要计算未知参数 $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$ 的后验分布。1987 年, 美国统计学家 M. A. Tanner 和王永雄提出数据增扩 (data augmentation) 算法 (简称 DA 算法) 来计算参数的后验分布 [149, 150]。

$$p(\boldsymbol{\theta} | \mathbf{y}_{\text{obs}}) = \int p(\boldsymbol{\theta} | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) p(\mathbf{y}_{\text{mis}} | \mathbf{y}_{\text{obs}}) d\mathbf{y}_{\text{mis}} \quad (15.6)$$

上式可以通过 Monte Carlo 方法近似求解, 其中 $p(\mathbf{y}_{\text{mis}} | \mathbf{y}_{\text{obs}})$ 具体为

$$\begin{aligned} p(\mathbf{y}_{\text{mis}} | \mathbf{y}_{\text{obs}}) &= \int_{\Theta} p(\mathbf{y}_{\text{mis}} | \boldsymbol{\theta}, \mathbf{y}_{\text{obs}}) p(\boldsymbol{\theta} | \mathbf{y}_{\text{obs}}) d\boldsymbol{\theta}, \text{ 或经过变量替换为} \\ &= \int_{\Theta} p(\mathbf{y}_{\text{mis}} | \boldsymbol{\xi}, \mathbf{y}_{\text{obs}}) p(\boldsymbol{\xi} | \mathbf{y}_{\text{obs}}) d\boldsymbol{\xi} \end{aligned}$$

将之代入式 (15.6), 显然 $p(\boldsymbol{\theta} | \mathbf{y}_{\text{obs}})$ 即是下面积分方程的解。

$$g(\boldsymbol{\theta}) = \int_{\Theta} \kappa(\boldsymbol{\theta}, \boldsymbol{\xi}) g(\boldsymbol{\xi}) d\boldsymbol{\xi} \quad (15.7)$$

其中 $\kappa(\boldsymbol{\theta}, \boldsymbol{\xi}) = \int p(\boldsymbol{\theta} | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) p(\mathbf{y}_{\text{mis}} | \boldsymbol{\xi}, \mathbf{y}_{\text{obs}}) d\mathbf{y}_{\text{mis}}$ 是积分方程的核

算法 15.15 (数据增扩, 1987). 积分方程 (15.7) 的近似解, 可从初始解 $g_0(\boldsymbol{\theta})$ 出发, 利用下面的迭代方法求之。

$$g_{k+1}(\boldsymbol{\theta}) = \int_{\Theta} \kappa(\boldsymbol{\theta}, \boldsymbol{\xi}) g_k(\boldsymbol{\xi}) d\boldsymbol{\xi}, \text{ 其中 } k = 0, 1, 2, \dots$$

上式所刻画的从 $g_k(\boldsymbol{\theta})$ 到 $g_{k+1}(\boldsymbol{\theta})$ 的过程可用下面的数值方法实现。

- 填补步骤: 从分布 $g_k(\boldsymbol{\xi})$ 独立抽取 $\boldsymbol{\xi}^{(i)}$, 其中 $i = 1, 2, \dots, m$ 。
- 后验步骤: 从分布 $p(\mathbf{y}_{\text{mis}} | \boldsymbol{\xi}^{(i)}, \mathbf{y}_{\text{obs}})$ 独立抽取 $\mathbf{y}_{\text{mis}}^{(i)}$, 其中 $i = 1, 2, \dots, m$ 。
- 更新步骤: 将 $g_k(\boldsymbol{\theta})$ 按下述方式更新至 $g_{k+1}(\boldsymbol{\theta})$ 。

$$g_{k+1}(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^m p(\boldsymbol{\theta} | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}^{(i)})$$

在提出 DA 算法的同时, Tanner 和王永雄还证明了 DA 算法 15.15 的收敛性 [150]。

定理 15.2. 在某些正则条件之下，积分方程 (15.7) 有唯一的解 g ，并且

$$\lim_{k \rightarrow \infty} \|g_k - g\| = 0$$

如果 $g(\boldsymbol{\theta})$ 是 Lebesgue 可积函数（见附录 D），则 DA 算法每次更新都改善了近似解，即

$$\|g_{k+1} - g\| \leq \|g_k - g\|, \text{ 其中 } \|g\| = \int_{\Theta} |g(\boldsymbol{\theta})| d\boldsymbol{\theta}$$

例 15.24. 接着例 14.1，利用数据增扩算法求未知参数 θ 的后验分布 $p(\theta|\mathbf{y}_{\text{obs}})$ 。

■ 将下面的过程独立地重复 m 次，得到 $y_2^{(1)}, \dots, y_2^{(m)}$ 。

■ 按照当前对参数后验分布的估计 $p(\theta|\mathbf{y}_{\text{obs}})$ 产生随机数 θ ；

■ 从二项分布 $B(125, \theta/(\theta + 2))$ 产生随机数 y_2 。

■ 未知参数 θ 的后验分布为

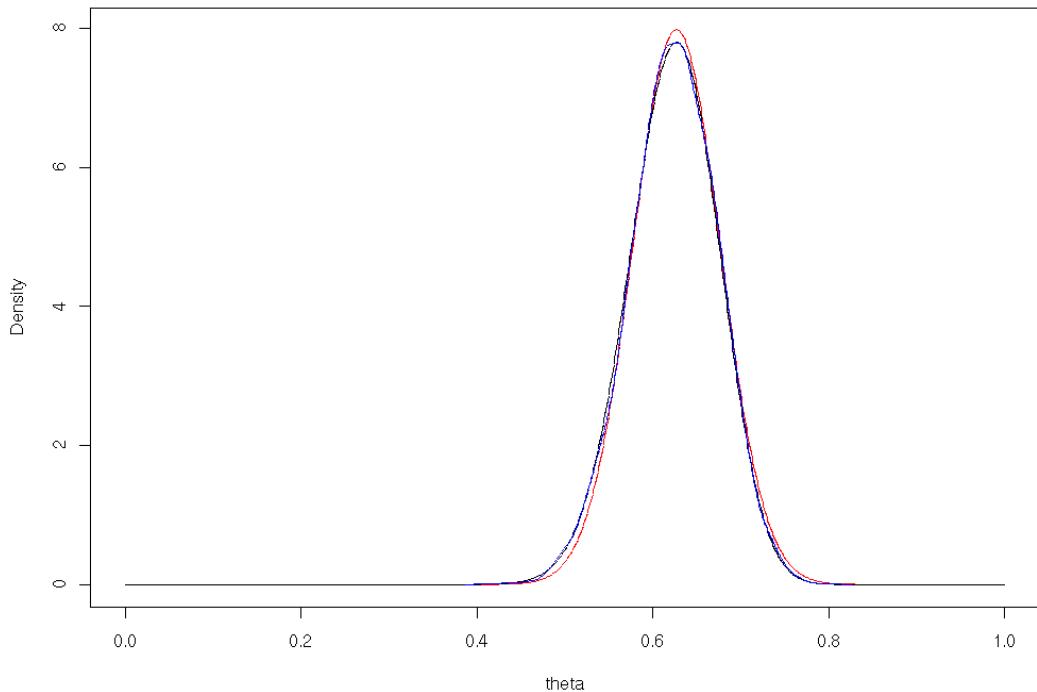
$$p(\theta|\mathbf{y}_{\text{obs}}) = \frac{1}{m} \sum_{i=1}^m b_{a_i, b_i}(\theta), \text{ 其中 } b_{a_i, b_i}(\theta) = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \theta^{a_i-1} (1-\theta)^{b_i-1}$$

此处 $a_i = y_2^{(i)} + x_4 + 1, b_i = x_2 + x_3 + 1$ ，并且 $b_{a_i, b_i}(\theta)$ 是 Beta 分布的密度函数（见第 302 页的定义 4.20）。

例 15.25. 黑线： $p(\theta|\mathbf{y}_{\text{obs}}) \propto (2 + \theta)^{125} (1 - \theta)^{38} \theta^{34}$

红线： $\theta \sim N(0.6268, 0.05^2)$

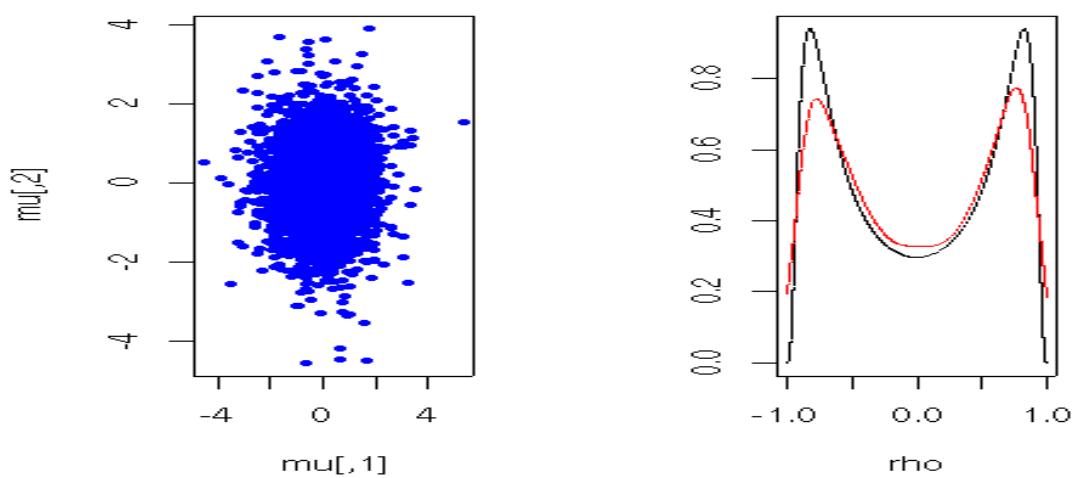
蓝线：由 DA 算法得到的 θ 的经验分布。



例 15.26 (二元正态分布的 DA 算法). 从二元正态总体 $N_2(\mu, \Sigma)$ 观察到

X_1	1	1	-1	-1	2	2	-2	-2	NA	NA	NA	NA
X_2	1	-1	1	-1	NA	NA	NA	NA	2	2	-2	-2

其中 NA 表示缺失数据。



例 15.27. 考虑二分类的贝叶斯模型 $P(Y_i = 1) = \Phi(x_i^\top \beta)$, 其中 $i = 1, 2, \dots, n$, 未知参数 β 是 d 维的列向量, Φ 是 $N(0, 1)$ 的分布函数。已知样本 X_1, X_2, \dots, X_n 的观测

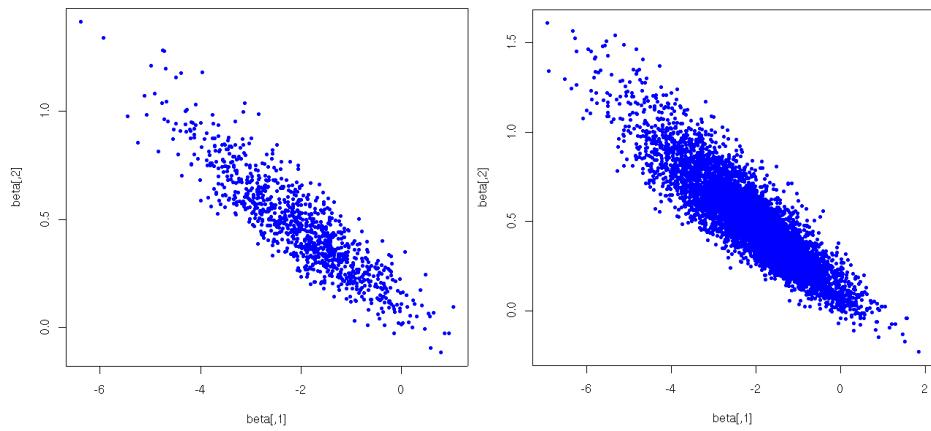
值 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, 设 $Y_i \in \{1, 0\}$ 是样本点 \mathbf{X}_i 的类标, 其观测结果为 $\mathbf{y} = (y_1, \dots, y_n)^\top$ 。引入 n 个独立的隐性变量 Z_1, \dots, Z_n 使得

$$Z_i | Y_i, \boldsymbol{\beta} \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, 1) \text{ 在 } 0 \text{ 处截尾, } Y_i = \begin{cases} 1 & \text{当 } Z_i > 0 \\ 0 & \text{当 } Z_i \leq 0 \end{cases}$$

$$\boldsymbol{\beta} | \mathbf{x}, z \sim N(\hat{\boldsymbol{\beta}}, (D^\top D)^{-1})$$

其中, $D^\top = (\mathbf{x}_1, \dots, \mathbf{x}_n)_{d \times n}$ 是数据矩阵, $\hat{\boldsymbol{\beta}} = (D^\top D)^{-1} D^\top z, z = (z_1, \dots, z_n)^\top$ 。

令 $\mathbf{y} = (0, 1, 0, 0, 1, 1, 1, 1, 1)^\top$ 和 $\mathbf{x}_i = (1, i)^\top, i = 1, 2, \dots, 10$ 。我们得到 $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ 的 1000 和 6000 个模拟结果。



15.4 习题

- 15.1. 试用模拟退火算法求解 0-1 背包问题：有 n 件物品，其中物品 j 的重量和价值分别为 w_j 和 c_j , $j = 1, 2, \dots, n$ 。选若干件物品放入一个承重为 W 的背包里，问如何选择使其价值之和最大？

第四部分

附录

附录 A

正态分布的由来

密度函数 $\phi(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x-\mu)^2/(2\sigma^2)\}$ 是如何得到的? C. F. Gauss (1809) 应误差分析和最小二乘法的研究曾经使用过正态分布的密度函数 $\phi(x|\mu, \sigma^2)$, 但历史上首位描绘它并揭开其神秘面纱的却是法国数学家 A. de Moivre。遗憾的是, de Moivre 没有明确给出正态分布的定义, 真正意识到这个重要概率分布的是 Laplace (1783) 和 Gauss (1809)。历史上, 正态分布常称作高斯分布 (Guassian distribution)。公正地讲, “正态分布之父”应该是 Laplace (见第 65 页)。



1718-1733 年, de Moivre 陆续发表了有关二项分布的研究成果, 他发现了当 n 很大时, 二项分布 $X \sim B(n, 1/2)$ 的概率函数 $P(X=k)$ 可用正态分布 $N(n/2, n/4)$ 的密度函数来近似, 具体说来,

$$2^{-n}C_n^k \approx \phi(k|n/2, n/4), \text{ 其中 } k = 0, 1, \dots, n$$

例 A.1. 为了解 $X \sim B(n, 1/2)$ 的概率函数与正态分布密度函数 $\phi(x|n/2, n/4)$ 在 $x = 0, 1, \dots, n$ 上的近似程度与 n 的关系, 现考察如下定义的误差函数 $\varepsilon(n)$, 它把所有的误差的绝对值加起来。

$$\varepsilon(n) = \sum_{k=0}^n |P(X=k|n, 1/2) - \phi(k|n/2, n/4)|$$

算得 $\varepsilon(5) \approx 0.04701, \varepsilon(50) \approx 0.00474, \varepsilon(150) \approx 0.00157$ 。不难发现 n 越大, 误差 $\varepsilon(n)$ 反而越小! 事实上, 由第 5 章的 de Moivre-Laplace 中心极限定理 (见第 367 页的定理 5.16) 可证得 $\lim_{n \rightarrow \infty} \varepsilon(n) = 0$ 。

后来 Laplace 将 de Moivre 的结果推广到 $B(n, p)$ 的情形, 证得了下面的结果, 这就是著名的 de Moivre-Laplace 中心极限定理的由来, 函数 $\phi(x|\mu, \sigma^2)$ 的重要地位

也逐步地在概率论中被认可。

定理 A.1. 令 $q = 1 - p, 0 < p < 1$, 对于满足条件 $|k - np| = o(npq)^{2/3}$ 的所有 k 皆有

$$P_n(k) = C_n^k p^k q^{n-k} \sim \frac{1}{\sqrt{2\pi npq}} \exp \left\{ -\frac{(k - np)^2}{2npq} \right\} \quad (\text{A.1})$$

证明. 利用 Stirling 公式* $n! \sim \sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n}$, 不难验证

$$C_n^k = \frac{n!}{k!(n-k)!} \sim \frac{1}{\sqrt{2\pi n} \left(\frac{k}{n}\right) \left(1 - \frac{k}{n}\right) \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k}}$$

令 $\tilde{p} = k/n$, 则有

$$\begin{aligned} C_n^k p^k q^{n-k} &\sim \frac{1}{\sqrt{2\pi n} \tilde{p} (1 - \tilde{p})} \left(\frac{p}{\tilde{p}}\right)^k \left(\frac{1-p}{1-\tilde{p}}\right)^{n-k} \\ &= \frac{1}{\sqrt{2\pi n} \tilde{p} (1 - \tilde{p})} \exp \left\{ -n \left[\frac{k}{n} \ln \frac{\tilde{p}}{p} + \left(1 - \frac{k}{n}\right) \ln \frac{1-\tilde{p}}{1-p} \right] \right\} \\ &= \frac{1}{\sqrt{2\pi n} \tilde{p} (1 - \tilde{p})} \exp \{-nh(\tilde{p})\} \end{aligned} \quad (\text{A.2})$$

其中函数 $h(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{1-p}$, $x \in (0, 1)$ 。于是,

$$h'(x) = \ln \frac{x}{p} - \ln \frac{1-x}{1-p} \text{ 且 } h''(x) = \frac{1}{x} + \frac{1}{1-x}$$

考虑 $h(x)$ 在 $x = p$ 点的 Taylor 展开

$$\begin{aligned} h(x) &= h(p) + h'(p)(x-p) + \frac{1}{2}h''(p)(x-p)^2 + o(|x-p|^2) \\ &= \frac{(x-p)^2}{2pq} + o(|x-p|^2) \end{aligned}$$

只要 n 足够地大, \tilde{p} 就能与 p 充分接近, 于是就有

$$h(\tilde{p}) \approx \frac{1}{2pq}(\tilde{p}-p)^2 = \frac{1}{2n^2pq}(k-np)^2$$

代入到 (A.2) 中便得到 (A.1)。 □

* de Moivre 本人恰是 Stirling 公式的真正发现者, 这种张冠李戴的混乱命名在数学史乃至科学史上都是司空见惯的事情, 好在不影响使用, 人们也就将错就错了。

附录 B

卷积的物理意义

考虑符合叠加原理的线性系统，该系统可抽象地表示为“输入 → 输出”的形式，即 $f(t) \rightarrow \psi(t)$ ，并且满足下述两个条件。

- ① 线性：如果 $f_i(t) \rightarrow \psi_i(t), i = 1, 2, \dots, n$ ，则 $\sum_{i=1}^n \alpha_i f_i(t) \rightarrow \sum_{i=1}^n \alpha_i \psi_i(t)$ 。
- ② 平移不变性：如果 $f(t) \rightarrow \psi(t)$ ，则 $f(t - \tau) \rightarrow \psi(t - \tau)$ 。

已知单位脉冲函数 $\delta(t)$ （定义见第 219 页的式 3.3）的输出为 $g(t)$ ，即 $\delta(t) \rightarrow g(t)$ 。在信号处理中， $g(t)$ 有时被称为系统函数。由平移不变性可知 $\delta(t - \tau) \rightarrow g(t - \tau)$ 。另外，对于任何输入 $f(t)$ ，由单位脉冲函数的性质，皆有

$$f(t) = \int_{-\infty}^{+\infty} f(\tau) \delta(t - \tau) d\tau$$

上式可直观地解释为 $f(t)$ 是一系列 $\delta(t - \tau)$ 的线性叠加，因此它的输出， $\psi(t)$ ，就是一系列 $g(t - \tau)$ 的线性叠加，即

$$\psi(t) = \int_{-\infty}^{+\infty} f(\tau) g(t - \tau) d\tau = f * g$$

也就是说，对于该线性系统，输出皆可表示为输入信号与系统函数的卷积。例如，回声可以用源声与一个反映各种反射效应的函数的卷积表示。卷积运算是这一类线性系统所固有的。

卷积的概念出自泛函分析，与 Fourier 变换有着密切的关系（见第 3 章的卷积定理 3.1）。卷积是一个非常重要的运算，在数学、物理、信号处理、机器学习等领域有着广泛的应用。

附录 C

Riemann-Stieltjes 积分

春有百花秋有月，夏有凉风冬有雪。

黄龙慧开《无门关》

积分理论中，Riemann 积分是最简单最基本的，但它无法统一表达离散型和连续型随机变量的数字特征。Riemann-Stieltjes 积分（常简称 R-S 积分，有时也称 Stieltjes 积分）是 Riemann 积分的自然推广，比 Riemann 积分更适合用于概率论。它是由荷兰数学家 Thomas Joannes Stieltjes (1856-1894) 于 1894 年在论文《连分数的研究》中提出的，刺激了对一般测度空间上的积分的研究，如 Lebesgue-Stieltjes 积分，它是 Lebesgue 积分的一般化（见附录 D）。

定义 C.1. 令 $g(x), u(x)$ 是定义在闭区间 $[a, b]$ 上的有界实函数。考虑 $[a, b]$ 的任意分割 $a = x_0 < x_1 < x_2 < \dots < x_n = b$ ，令 $\xi_j \in [x_j, x_{j+1}]$ ，如果下面的极限存在

$$\lim_{\max\{x_{j+1}-x_j\} \rightarrow 0} \sum_{j=0}^{n-1} g(\xi_j)[u(x_{j+1}) - u(x_j)] \quad (\text{C.1})$$



并且不论分割和介点 ξ_j 如何选择，极限都等于某个值 S ，则称该极限为在 $[a, b]$ 上 g 关于 u 的 Riemann-Stieltjes 积分。记作

$$S = \int_a^b g(x)du(x), \text{ 或简记为 } S = \int_a^b gdu$$

R-S 积分可以推广到无限区间或复值函数的情形。Riemann 积分是 R-S 积分的特例，即 $u(x) = x$ 。下面我们给 R-S 积分一个不严格的物理解释：在平面内，平行

于 y 轴存在力场 $g(x)$ 。沿曲线 $u(x)$ 把粒子从点 $(a, u(a))$ 移动到点 $(b, u(b))$ 所作的功, 见下图。

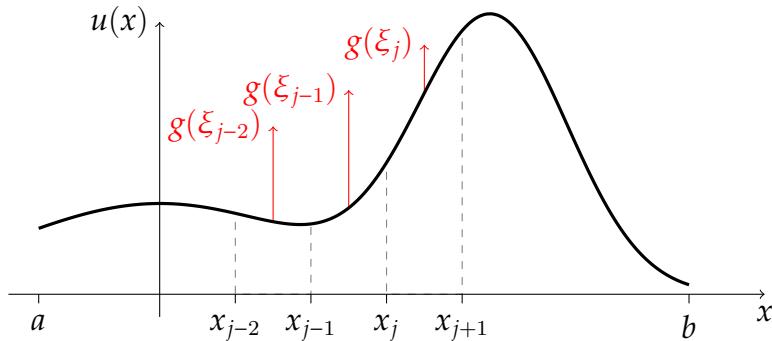


图 C.1: Riemann-Stieltjes 积分 (C.1) 的物理含义: 在平行于 y 轴的力场中, 点 (x, y) 处力的大小为 $g(x)$ 。将粒子从点 $(a, u(a))$ 沿着曲线 $u(x)$ 移动到点 $(b, u(b))$, 所作的功就是式 (C.1)。

定义 C.2. 设 $a = x_0 < x_1 < \dots < x_n = b$ 是对闭区间 $[a, b]$ 的任意分割, $[a, b]$ 上的实函数 $u(x)$ 的如下数字特征称为变差。

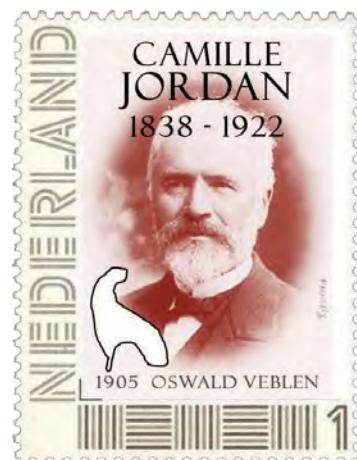
$$\bigvee_a^b(u) = \sup_{x_0, \dots, x_n} \sum_{j=1}^n |u(x_{j+1}) - u(x_j)|$$

如果 $\bigvee_a^b(u) < \infty$, 则称 u 具有有界变差。我们把区间 $[a, b]$ 上所有有界变差函数的全体记作 $\bigvee[a, b]$ 。如果对任意自然数 n , 实函数 $u(x) \in \bigvee[-n, n]$ 且 $\lim_{n \rightarrow \infty} \bigvee_{-n}^n(u) < \infty$, 则称 u 为 \mathbb{R} 上的有界变差函数。有界变差函数的不连续点至多可数, 并且都是第一类的。

1881 年, 法国数学家 Camille Jordan (1838-1922) 给出了变差的定义, 他还证明了 Jordan 分解定理: u 是有界变差函数当且仅当存在两个增函数 u_1, u_2 使得 $u = u_1 - u_2$ 。另外, $u(x) \in \bigvee[a, b]$ 当且仅当 $\forall c \in (a, b)$ 皆有 $u \in \bigvee[a, c]$ 且 $u \in \bigvee[c, b]$ 。分布函数 $F(x)$ 是 $[0, 1]$ 上的有界变差函数, 关于 F 的 R-S 积分具有一些好的性质。

例 C.1. 连续函数不一定是有界变差函数, 譬如, 在 $(0, 1]$ 上 $u(x) = x \sin(1/x)$ 且 $u(0) = 0$ 。

性质 C.1. 在有限区间 $[a, b]$ 上, Riemann-Stieltjes 积分具有以下性质。



① 若 g_1, g_2 关于 u 都是 R-S 可积的，则 g_1, g_2 的线性组合亦如此，且

$$\int_a^b (cg_1 + dg_2)du = c \int_a^b g_1 du + d \int_a^b g_2 du$$

② 若 g 关于 u_1, u_2 都是 R-S 可积的，则 g 关于 u_1, u_2 的线性组合亦如此，且

$$\int_a^b g d(cu_1 + du_2) = c \int_a^b g du_1 + d \int_a^b g du_2$$

③ 若 g 关于 u 在 $[a, b]$ 上 R-S 可积，则对 $\forall c \in (a, b)$, g 关于 u 在 $[a, c]$ 和 $[c, b]$ 上都是 R-S 可积的，且

$$\int_a^b g du = \int_a^c g du + \int_c^b g du$$

但该命题的逆命题不成立。

④ 若 g 关于 u 是 R-S 可积的，则 u 关于 g 也是 R-S 可积的，且有分部积分公式

$$\int_a^b g du = g(b)u(b) - g(a)u(a) - \int_a^b u dg$$

⑤ 若 g 在 $[a, b]$ 上是 Riemann 可积的， u 在 $[a, b]$ 上满足 Lipschitz 条件^{*}，则 g 关于 u 是 R-S 可积的。

⑥ 积分中值定理：如果 g 在 $[a, b]$ 上有界， $m \leq g(x) \leq M$ ，并且 u 在 $[a, b]$ 上单调增，则存在 $w \in [m, M]$ 使得

$$\int_a^b g du = w[u(b) - u(a)]$$

⑦ 若 $u(x)$ 的导数 $u'(x)$ 在 $[a, b]$ 上有界且 Riemann 可积，则 g 关于 u 的 R-S 积分可转化为 Riemann 积分（下面等式的右边部分）

$$\int_a^b g(x) du(x) = \int_a^b g(x) u'(x) dx$$

⑧ 若 x_0 同时是 g, u 的不连续点，则 g 关于 u 的 R-S 积分不存在。

^{*}存在常数 c ，使得 $\forall x_1, x_2 \in [a, b]$ 皆有 $|u(x_1) - u(x_2)| \leq c|x_1 - x_2|$ 。该条件是德国数学家 Rudolf Lipschitz (1832-1903) 于 1864 年提出来的，最初是作为判定 $u(x)$ 的 Fourier 级数收敛的充分条件。

- ⑨ 若 g 在 $[a, b]$ 上有界, u 在 $[a, b]$ 具有有界变差, 则 g 关于 u 是 R-S 可积的 (结果与 Lebesgue-Stieltjes 积分相同), 且

$$\left| \int_a^b g du \right| \leq \sup_{x \in [a, b]} |f(x)| \bigvee_a^b (u)$$

- ⑩ 设 u 是 $[a, b]$ 上有界变差函数, 如果 $\{g_n\}$ 是 $[a, b]$ 上一列关于 u 可积的函数, 并且在 $[a, b]$ 上一致收敛*于 g , 则

$$\lim_{n \rightarrow \infty} \int_a^b g_n du = \int_a^b g du$$

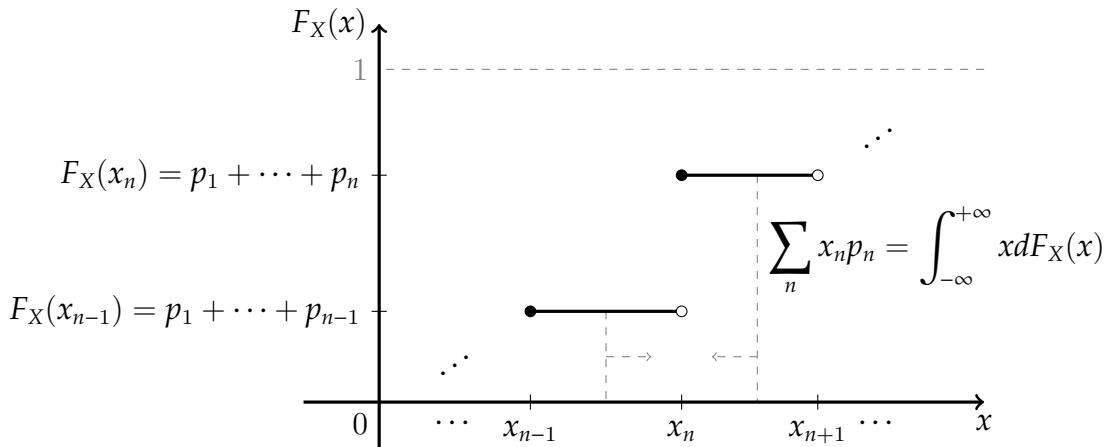


图 C.2: 离散型随机变量 X 的分布函数: 跳跃点 x 两侧的 F_X 值之差非零。

*一个函数序列 $g_1(t), g_2(t), \dots, g_n(t), \dots$ 在集合 T 上一致收敛 (亦称均匀收敛) 于 $g(t)$, 当且仅当 $\forall t \in T$, 对于任给的 $\epsilon > 0$, 总能找到 $N \in \mathbb{N}$ 使得当 $n > N$ 时有 $|g_n(t) - g(t)| < \epsilon$ 。它的几何直观含义是: 第 N 项以后所有的 $g_n(t)$ 都落于“带状区域” $(g(t) - \epsilon, g(t) + \epsilon)$ 之内。

附录 D

可测函数与 Lebesgue 积分

行到水穷处，坐看云起时。

王维《终南别业》

法国著名数学家 Henri Léon Lebesgue (1875-1941) 于 1902 年发表了名垂青史的论文《积分、长度和面积》，标志着古典分析过渡到现代分析 [68, 140, 141]。

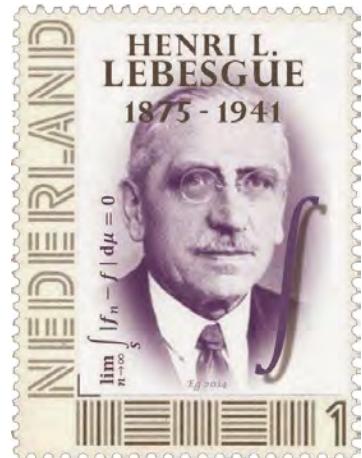
Lebesgue 在论文中定义的 Lebesgue 测度和 Lebesgue 积分如今已成为实变函数论研究的核心内容，同时也是概率论的严格数学基础 [6, 13]。本附录概要地介绍可测函数和 Lebesgue 积分的基本性质。

定理 D.1. 以下三种说法与第 109 页的**定义 2.2** 等价，都可作为可测空间 (Ω, \mathcal{S}) 上可测函数 g 的定义： $\forall r \in \mathbb{R}$,

① $\{\omega : g(\omega) > r\} \in \mathcal{S}$

② $\{\omega : g(\omega) \geq r\} \in \mathcal{S}$

③ $\{\omega : g(\omega) < r\} \in \mathcal{S}$



定理 D.2. 已知 g, h 都是可测空间 (Ω, \mathcal{S}) 上可测函数,

① 函数 $|g|^r, \max(g, h), \min(g, h), rg$ 也都是可测的，其中 $r \in \mathbb{R}$ 。

② 如果 F 是 \mathbb{R}^2 上的连续实值函数，则 $F(g, h)$ 是可测函数。特别地， $g \pm h, gh$ 是可测的；当 $h \neq 0$ 时， g/h 也是可测的。

③ 设 $g_n, n = 1, 2, \dots$ 是一列可测函数，(1) 则 $\limsup g_n$ 和 $\liminf g_n$ 都是可测的。
(2) 若 $\{g_n\}$ 在 Ω 上几乎处处收敛于 g ，则 g 是可测的。

④ 已知 f 是可测空间 (Ω, \mathcal{S}) 上的可测函数, g 是 \mathbb{R} 上的 Borel 函数, 则 $g(f)$ 是 (Ω, \mathcal{S}) 上的可测函数。

定义 D.1 (简单函数). 定义在可测空间 (Ω, \mathcal{S}) 上的可测函数 g 如果取值至多可数, 则称之为简单函数。令其取值为 r_1, r_2, \dots (两两不等), 显然其逆像 $\mathfrak{P} = \{A_j = g^{-1}(r_j) : j = 1, 2, \dots\}$ 是 Ω 的一个划分, 函数 g 可表示为

$$g(\omega) = \sum_{j=1}^{\infty} r_j I_{A_j}(\omega) \quad (\text{D.1})$$



定义 D.2 (可积的简单函数). 已知测度空间 $(\Omega, \mathcal{S}, \mu)$ (见**定义 1.13**) 和定义在 (Ω, \mathcal{S}) 上的简单函数 g , 见式(D.1)。如果级数 $\sum_{j=1}^{\infty} r_j \mu(A_j)$ 绝对收敛, 则称该简单函数是可积的。该级数之和即为 Lebesgue 积分

$$\int_{\Omega} g d\mu = \sum_{j=1}^{\infty} r_j \mu(A_j)$$

 简单函数的 Lebesgue 积分的过程好比计算大量各种面值硬币的总值, 先把硬币按面值 (如 r_1, r_2, \dots) 分堆, 相同面值 (取值 r_j) 的放在一起, 然后分别计算它们的个数 $\mu(A_j)$ 并将面值乘以个数得到每堆的值 $r_j \mu(A_j)$, 最后将各堆的值加起来得到总值 $\sum_{j=1}^{\infty} r_j \mu(A_j)$ 。而 Riemann 积分的过程则好比把不同面值的硬币都混在一起用依次累加的方法得到总值。



例 D.1. 简单随机变量 $X : \Omega \rightarrow \mathbb{R}$ 是一个简单函数。如果 $X = \sum_{j=1}^n x_j I_{A_j}$ 是定义在概率空间 (Ω, \mathcal{S}, P) 上的一个非负简单随机变量, 则

$$\int_{\Omega} X dP = \sum_{j=1}^n x_j P\{X = x_j\} = \sum_{j=1}^n x_j P(A_j)$$

例 D.2. 考虑定义在区间 $I = [0, 1]$ 上的 Dirichlet 函数 $D(x)$: 它在有理数上取值为 0, 在无理数上取值为 1。显然, $D(x)$ 是可积的简单函数, 但它的 Riemann 积分不存在, Lebesgue 积分等于 1 (因为 I 上无理数集合的 Lebesgue 测度为 1)。

定义 D.3 (Lebesgue 积分). 对于函数 $f : \Omega \rightarrow \mathbb{R}$, 若能找到一个可积的简单函数序列 $g_n, n = 1, 2, \dots$ 在 Ω 上 (可以除去一个零测集) 一致收敛于 f , 则 f 在 Ω 上的

Lebesgue 积分定义为

$$\int_{\Omega} f d\mu = \lim_{n \rightarrow \infty} \int_{\Omega} g_n d\mu, \text{ 如果 } \lim_{n \rightarrow \infty} \int_{\Omega} g_n d\mu < \infty \quad (\text{D.2})$$

此时, 称 f 是可积的, 记作 $f \in L^1(\Omega, \mu)$ 。这个定义不依赖于 g_n 的选取, 即如果还能找到别的可积的简单函数序列 $h_n, n = 1, 2, \dots$ 一致收敛于 f , 结果还是一样。当 $\Omega = \mathbb{R}^n$ 时, 若 μ 不是 Lebesgue 测度, 式 (D.2) 称作 Lebesgue-Stieltjes 积分。

定理 D.3. 随机变量 $X : (\Omega, \mathcal{S}, P) \rightarrow (\mathbb{R}, \mathcal{B}_1, F_X)$ 的期望定义为如下的 Lebesgue 积分。

$$E(X) = \int_{\Omega} X dP \quad (\text{D.3})$$

如果 $E(|X|) < \infty$, 则必有

$$E(X) = \int_{\mathbb{R}} x dF_X \quad (\text{D.4})$$

证明. 不妨设 $X \geq 0$, 令 n, k 为自然数。定义

$$A_{k,n} = \left\{ \omega \in \Omega : \frac{k-1}{n} \leq X(\omega) < \frac{k}{n} \right\}$$

对每个固定的 n , $\{A_{k,n} : k = 1, 2, \dots\}$ 是 Ω 的一个划分。定义简单随机变量 X_n 如下, 它的取值是 $\{k/n : k \in \mathbb{N}\}$ 。

$$X_n = \sum_{k=1}^{\infty} \frac{k}{n} I_{A_{k,n}}$$

于是, 我们得到 X 的双边控制, 进而得到 $E(X)$ 的双边控制。

$$\begin{aligned} X_n - \frac{1}{n} &\leq X \leq X_n \\ E(X_n) - \frac{1}{n} &\leq E(X) \leq E(X_n) \end{aligned}$$

上式两边分别为

$$E\left(X_n - \frac{1}{n}\right) = \sum_{k=1}^{\infty} \frac{k-1}{n} P(A_{k,n}) \quad E(X_n) = \sum_{k=1}^{\infty} \frac{k}{n} P(A_{k,n})$$

每个求和项都可以用来做如下的双边控制,

$$\frac{k-1}{n}P(A_{k,n}) \leq \int_{[\frac{k-1}{n}, \frac{k}{n})} xdF_X \leq \frac{k}{n}P(A_{k,n})$$

两边求和便得到

$$E\left(X_n - \frac{1}{n}\right) \leq \int_{[0,\infty)} xdF_X \leq E(X_n)$$

令 $n \rightarrow \infty$ 便可得到欲证的结果。对于一般情形, X 总可以被分解表示为 $X = X^+ - X^-$, 其中

$$X^+(\omega) = \max\{X(\omega), 0\} = \begin{cases} X(\omega) & \text{若 } X(\omega) > 0 \\ 0 & \text{否则} \end{cases}$$

$$X^-(\omega) = \max\{-X(\omega), 0\} = \begin{cases} -X(\omega) & \text{若 } X(\omega) < 0 \\ 0 & \text{否则} \end{cases}$$

显然, X^+, X^- 都是非负随机变量。请读者仿照已给出的证明补全一般情形。 \square

性质 D.1. Lebesgue 积分是 $L^1(\Omega, \mu)$ 上的线性泛函*, 是对积分概念的最重要的一般化。

① 如果 $f \in L^1(A, \mu)$, 其中 $A \subseteq \Omega$, 则 $\forall B \subseteq A$, 皆有 $f \in L^1(B, \mu)$ 。

② 如果 $f \in L^1(\Omega, \mu)$, 则 f 是 Ω 上可测的几乎处处有限的函数。

$$\int_A f d\mu = 0 \Rightarrow \begin{cases} \mu(A) = 0 & \text{若 } f \text{ 在 } A \text{ 上几乎处处为正} \\ f \stackrel{a.e.}{=} 0 & \text{若 } f \text{ 在 } A \text{ 上几乎处处非负} \end{cases}$$

③ 如果 $f, g \in L^1(\Omega, \mu)$ 只在一个零测集上不相等, 则

$$\int_{\Omega} f d\mu = \int_{\Omega} g d\mu$$

④ 如果 f 有界 (设 $m \leq f \leq M$) 且可测, 则 $f \in L^1(\Omega, \mu)$ 且

$$m\mu(\Omega) \leq \int_{\Omega} f d\mu \leq M\mu(\Omega)$$

*泛函 (functional) 就是把函数映为实数或复数的映射, 即函数的函数。

⑤ 如果 $f \in L^1(\Omega, \mu)$, $|g| \leq f$ 且 g 可测, 则 $g \in L^1(\Omega, \mu)$ 且

$$\left| \int_{\Omega} g d\mu \right| \leq \int_{\Omega} f d\mu$$

⑥ 如果 $f \in L^1(A, \mu)$, 可测集 $A_1, A_2, \dots, A_n, \dots$ 是 A 的划分, 则

$$\int_A f d\mu = \sum_{n=1}^{\infty} \int_{A_n} f d\mu$$

引理 D.1 (Fatou*, 1906). 若 $\{f_n : n = 1, 2, \dots\}$ 是非负可测函数序列, 则

$$\int_A (\liminf_{n \rightarrow \infty} f_n) d\mu \leq \liminf_{n \rightarrow \infty} \int_A f_n d\mu$$

$$\text{或者等价地, } \int_A (\limsup_{n \rightarrow \infty} f_n) d\mu \geq \limsup_{n \rightarrow \infty} \int_A f_n d\mu$$

例 D.3. 函数列 $f_n(x) = \begin{cases} 1/n & \text{当 } x \in [0, n] \\ 0 & \text{否则} \end{cases}$ 满足

$$\liminf_{n \rightarrow \infty} f_n = 0, \text{ 而 } \int_{\mathbb{R}} f_n(x) dx = 1$$

定理 D.4 (Lebesgue 控制收敛定理, 1909). 定义在测度空间 $(\Omega, \mathcal{S}, \mu)$ 上的实值可测函数序列 $\{f_n : n = 1, 2, \dots\}$ 在任一点 $\omega \in \Omega$ 上收敛于 $f(\omega)$, 并且存在可积函数 g 使得对于任意 n 和任意 $\omega \in \Omega$ 皆有 $|f_n(\omega)| \leq g(\omega)$, 即序列 $\{f_n\}$ 被 g 控制住, 则 f_n 和 f 在 Ω 上都是可积的, 并且还有

$$\lim_{n \rightarrow \infty} \int_{\Omega} f_n d\mu = \int_{\Omega} f d\mu$$

*Pierre Fatou (1878-1929) 是法国数学家和天文学家, 他在数学上的贡献主要在分析方面。

附录 E

矩阵计算的一些结果

矩阵理论是一个充分发展的数学分支，在很多领域都有着广泛的应用 [164]，尤其对多元统计分析、机器学习、信号处理等与数据分析相关的领域，它更是一件不可缺少的工具。有关矩阵的知识，读者可参阅 R. A. Horn 和 C. R. Johnson 合著的《矩阵分析》[76]，以及 G. H. Golub 和 C. F. van Loan 合著的《矩阵计算》[59]。该附录仅简介本书所需要的矩阵计算的一些基本知识。

在本书中我们约定：向量缺省地是列向量，如 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$ 。向量 \mathbf{x} 和 \mathbf{y} 的内积定义为 $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{j=1}^n x_j y_j$ 。零矩阵记为 O ， n 阶单位阵 (identity matrix) 记为 I_n 或者 I 。

为了行文的方便，在具体描述一个矩阵的时候约定以分号“;”表示“换行”，例如矩阵 $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$ 和分块矩阵 $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ 有时也记作 $A = (a_{11}, a_{12}, a_{13}; a_{21}, a_{22}, a_{23})$ 和 $B = (B_{11}, B_{12}; B_{21}, B_{22})$ 。

定理 E.1. 对于任意的 $\mathbf{x} \in \mathbb{R}^n$ 和 n 阶方阵 $A = (a_{ij})$ ，总有 $\mathbf{x}^\top A \mathbf{x} = \text{tr}(A \mathbf{x} \mathbf{x}^\top)$ ，其中 $\text{tr}(A) = \sum_{j=1}^n a_{jj}$ 是方阵 A 的迹 (trace)。

定义 E.1. 一般的教科书都是这样定义特征向量和特征值的：给定方阵 $A_{n \times n}$ ，非零向量 $\mathbf{x} \in \mathbb{R}^n$ 如果满足下述方程，则称 \mathbf{x} 是 A 的特征向量， λ 为对应的特征值。

$$A\mathbf{x} = \lambda\mathbf{x}, \text{ 其中 } \lambda \in \mathbb{C} \tag{E.1}$$

为什么我们要对 (E.1) 感兴趣？它源自十九世纪对二次曲面的研究。考虑实二次型 (quadratic form)

$$q(\mathbf{x}) = \sum_{1 \leq i \leq j \leq n} q_{ij} x_i x_j$$

人们问什么样的正交变换 $U_{n \times n} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$, 即坐标系旋转, 使得二次型 $q(\mathbf{x})$ 从形式上可以简化为以下标准形式?

$$q(U\mathbf{y}) = \lambda_1 y_1^2 + \dots + \lambda_n y_n^2, \text{ 其中 } \mathbf{y} = U^T \mathbf{x}$$

借助于矩阵这个工具, 总唯一存在一个实对称矩阵 $A_{n \times n}$, 称为二次型 $q(\mathbf{x})$ 的矩阵形式, 使得二次型 $q(\mathbf{x})$ 可以描述为

$$q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}, \text{ 其中 } a_{ii} = q_{ii} \text{ 且 } a_{ij} = \frac{1}{2}q_{ij}, i \neq j$$

于是, 上述问题也可以表述为什么样的正交矩阵 U 使得

$$\mathbf{x}^T A \mathbf{x} = \mathbf{y}^T U^T A U \mathbf{y} = \mathbf{y}^T \Lambda \mathbf{y}, \text{ 其中 } \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$$

因此, 人们对满足 $U^T A U = \Lambda$, 或者 $A U = U \Lambda$ 的矩阵 U, Λ 感兴趣, 即

$$A \mathbf{u}_j = \lambda_j \mathbf{u}_j, \text{ 其中 } j = 1, \dots, n$$

这便是特征向量和特征值的由来——任何数学概念的背后都有一个被抽象提炼出来的迫切需求, 迫切程度越高, 这概念就越重要。我们用 $\lambda_j(A), j = 1, \dots, n$ 来表示 $A_{n \times n}$ 的第 j 个特征值, 不失一般性, 约定 $\lambda_1(A) \geq \dots \geq \lambda_n(A)$ 。

性质 E.1. 设方阵 $A_{n \times n}$ 的特征值为 $\lambda_1, \dots, \lambda_n$, 则

$$\sum_{j=1}^n \lambda_j = \text{tr}(A)$$

$$\prod_{j=1}^n \lambda_j = \det(A)$$

证明. 特征多项式是

$$\det(\lambda I - A) = \prod_{j=1}^n (\lambda - \lambda_j)$$

比较等式两边 λ^{n-1} 的系数, 便证得第一式。令 $\lambda = 0$, 便证得第二式。 □

定义 E.2 (向量的 p -范数). 向量 $\mathbf{x} \in \mathbb{R}^n$ 的 p -范数 (p -norm) 定义为

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$$

显然, 2-范数 $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$ 就是向量 \mathbf{x} 的欧氏长度, 简记作 $\|\mathbf{x}\|$ 。另外,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$$

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

定义 E.3 (矩阵的 p -范数和 F -范数). 矩阵 $A = (a_{ij})_{m \times n}$ 的 p -范数和 Frobenius 范数 (简称 F -范数) 分别定义为

$$\begin{aligned} \|A\|_p &= \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \max_{\|\mathbf{x}\|_p=1} \|A\mathbf{x}\|_p \\ \|A\|_F &= \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} \end{aligned}$$

定理 E.2 (QR 分解). 已知矩阵 $X_{m \times n}$ 的秩为 n , 即列满秩, 则存在唯一的分解

$$X_{m \times n} = Q_{m \times n} R_{n \times n}$$

其中, Q 是正交矩阵, R 是对角线元素为正数的上三角矩阵。

证明. 向量 $\mathbf{y} \in \mathbb{R}^m$ 在向量 $\mathbf{x} \in \mathbb{R}^m$ 上的投影向量记作 $\text{proj}_{\mathbf{x}} \mathbf{y}$, 即

$$\text{proj}_{\mathbf{x}} \mathbf{y} = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\|\mathbf{x}\|^2} \mathbf{x}$$

特别地, 当 \mathbf{e} 是单位向量时, $\text{proj}_{\mathbf{e}} \mathbf{y} = \langle \mathbf{y}, \mathbf{e} \rangle \mathbf{e}$ 。下面, 利用 Gram-Schmidt 正交化 (orthogonalization process), 构造出 Q 和 R 。不妨设 $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, 则

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{x}_1, & \mathbf{q}_1 &= \frac{\mathbf{z}_1}{\|\mathbf{z}_1\|} \\ \mathbf{z}_2 &= \mathbf{x}_2 - \text{proj}_{\mathbf{q}_1} \mathbf{x}_2, & \mathbf{q}_2 &= \frac{\mathbf{z}_2}{\|\mathbf{z}_2\|} \\ &\vdots & & \\ \mathbf{z}_n &= \mathbf{x}_n - \sum_{j=1}^{n-1} \text{proj}_{\mathbf{q}_j} \mathbf{x}_j, & \mathbf{q}_n &= \frac{\mathbf{z}_n}{\|\mathbf{z}_n\|} \end{aligned}$$

不难验证如下定义的正交矩阵 Q 和上三角矩阵 R 使得 $X = QR$ 成立。

$$Q = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n), \quad R = \begin{pmatrix} \langle \mathbf{q}_1, \mathbf{x}_1 \rangle & \langle \mathbf{q}_1, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{q}_1, \mathbf{x}_n \rangle \\ 0 & \langle \mathbf{q}_2, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{q}_2, \mathbf{x}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \mathbf{q}_n, \mathbf{x}_n \rangle \end{pmatrix} \quad \square$$

定义 E.4. $\forall \mathbf{x} \neq \mathbf{0}$, 若方阵 $A_{n \times n}$ 满足 $\mathbf{x}^\top A \mathbf{x} > 0$ 或等价地, $\langle A \mathbf{x}, \mathbf{x} \rangle > 0$, 则称 A 为正定矩阵; 若 $\mathbf{x}^\top A \mathbf{x} \geq 0$, 则称 A 为半正定。

例 E.1. 给定无向图 $G = (V, E)$, 其中 $V = \{v_1, \dots, v_i, \dots, v_n\}$ 是顶点集合, E 是边的集合。定义矩阵 $D_{n \times n} = \text{diag}(d_1, \dots, d_i, \dots, d_n)$, 其中 d_i 是顶点 v_i 的度 (即, 连接顶点 v_i 的边的个数), 定义邻接矩阵 (adjacency matrix) $A_{n \times n} = (a_{ij})$, 其中 a_{ij} 取值 1 或 0, 表示边 $(i, j) \in E$ 与否。显然, 邻接矩阵是一个对称矩阵。我们称矩阵 $L = D - A$ 为拉普拉斯矩阵 (Laplacian matrix), 下面往证它是一个半正定矩阵。

$$\mathbf{x}^\top L \mathbf{x} = \mathbf{x}^\top (D - A) \mathbf{x} = \sum_{i=1}^n d_i x_i^2 - 2 \sum_{(i,j) \in E} x_i x_j = \sum_{(i,j) \in E} (x_i - x_j)^2 \geq 0$$

定义 E.5. 一个复值函数 $\varphi : \mathbb{R}^n \rightarrow \mathbb{C}$ 称为半正定函数, 如果对于任意的自然数 $m \in \mathbb{N}$, 对于任意的 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n, c_1, \dots, c_m \in \mathbb{C}$ 皆有 $\sum_{i,j=1}^m c_i \bar{c}_j \varphi(\mathbf{x}_i - \mathbf{x}_j) \geq 0$, 或等价地, $A = [\varphi(\mathbf{x}_i - \mathbf{x}_j)]_{m \times m}$ 为半正定矩阵。

定理 E.3. 对称矩阵 A 为正定矩阵当且仅当下列条件之一成立: (1) A 的所有特征值都大于零。 (2) 存在上三角矩阵 Q 使得 $A = Q^\top Q$, 称为 A 的 Cholesky 分解。当规定三角矩阵的对角元素皆取正值时, Cholesky 分解唯一。 (3) 存在可逆矩阵 C 使得 $A = C^\top C$ 。

定理 E.4 (谱分解). 对于任意对称矩阵 $S_{n \times n}$, 存在正交矩阵 $U = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ 和对角阵 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ 使得 S 具有以下的分解。

$$S = U \Lambda U^\top = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$$

其中 $\lambda_i = \lambda_i(S)$, \mathbf{u}_i 是对应的特征向量, $i = 1, 2, \dots, n$ 。

证明. 该分解等价于 $SU = U\Lambda$, 即 $S\mathbf{u}_i = \lambda_i \mathbf{u}_i$, 其中 \mathbf{u}_i 是 U 的第 i 列向量。 \square

定义 E.6. 给定矩阵 $A_{m \times n} = (\mathbf{a}_1, \dots, \mathbf{a}_n)$, 其秩为 $r \leq \min(m, n)$ 且 $m \geq n$ 。非负定矩阵 $A^\top A = (\mathbf{a}_i^\top \mathbf{a}_j)_{n \times n}$ 被称为 Gram 矩阵或内积矩阵, 不妨设其特征值 $\lambda_1 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_n = 0$ 。我们称 $\sigma_i = \sqrt{\lambda_i}$ 为矩阵 A 的第 i 个奇异值 (singular value), 记作 $\sigma_i(A)$, 其中 $i = 1, \dots, n$ 。

 因为矩阵 AA^\top 的非零特征值与 $A^\top A$ 的相同, 所以当我们谈论 A 的非零奇异值的时候, 不必计较它们定义自 $A^\top A$ 还是 AA^\top 。

定理 E.5 (奇异值分解). 给定矩阵 $A_{m \times n}$, 其秩为 r 且 $m \geq n$ 。存在列正交矩阵^{*} $U_{m \times r} = (\mathbf{u}_1, \dots, \mathbf{u}_r)$ 和 $V_{n \times r} = (\mathbf{v}_1, \dots, \mathbf{v}_r)$, 使得

$$A = U\Sigma V^\top = \sum_{j=1}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^\top \quad (\text{E.2})$$

其中, $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ 是 A 的非零奇异值构成的对角阵。矩阵分解 (E.2) 被称为奇异值分解 (singular value decomposition, SVD) 或紧凑 SVD。有的参考书采用分解形式 $A = U_{m \times n} \Sigma_{m \times n} V_{n \times n}^\top$, 其中 $U_{m \times n}$ 为列正交矩阵, $V_{n \times n}$ 为正交矩阵, 且

$$\Sigma_{m \times n} = \begin{pmatrix} \text{diag}(\sigma_1, \dots, \sigma_r) & O_{r \times (n-r)} \\ O_{(m-r) \times r} & O_{(m-r) \times (n-r)} \end{pmatrix}$$

证明. 根据谱分解定理 E.4, 存在正交矩阵 $V_{n \times n}$ 使得

$$V^\top A^\top A V = \begin{pmatrix} \Sigma^2 & O_{r \times (n-r)} \\ O_{(m-r) \times r} & O_{(m-r) \times (n-r)} \end{pmatrix}, \text{ 其中 } \Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$$

把 V 拆分成块矩阵 (V_1, V_2) , 其中 $V_1 = (\mathbf{v}_1, \dots, \mathbf{v}_r)$ 是 V 的前 r 列, 显然它满足

$$V_1^\top A^\top A V_1 = \Sigma^2$$

构造 $U_1 = AV_1\Sigma^{-1}$, 显然 $U_1\Sigma V_1^\top = A$ 。下面验证 U_1 是一个列正交矩阵。

$$U_1^\top U_1 = \Sigma^{-1} V_1^\top A^\top A V_1 \Sigma^{-1} = \Sigma^{-1} \Sigma^2 \Sigma^{-1} = I_r$$

即紧凑 SVD 成立。对于分解形式 $A = U_{m \times n} \Sigma_{m \times n} V_{n \times n}^\top$, 只需在 U_1 列向量的基础上增添一些单位正交向量使得 $U = (U_1, U_2)$ 为正交矩阵即可。 \square

算法 E.1. 基于奇异值分解定理 E.5 的证明, 矩阵 $A_{m \times n}$ 的紧凑 SVD 算法如下:

- 求 $A^\top A$ 的非零特征值 $\lambda_1 \geq \dots \geq \lambda_r > 0$ 及其对应的特征向量 $\mathbf{v}_1, \dots, \mathbf{v}_r$ 。定义 $V_{n \times r} = (\mathbf{v}_1, \dots, \mathbf{v}_r)$ 和 $\Sigma_{r \times r} = \text{diag}(\sigma_1, \dots, \sigma_r)$, 其中 $\sigma_1 = \sqrt{\lambda_1}, \dots, \sigma_r = \sqrt{\lambda_r}$ 。
- 定义 $U_{m \times r} = AV\Sigma^{-1}$, 则 U, Σ, V 满足紧凑 SVD (E.2)。

*一个矩阵 $A_{m \times n}$ 若满足 $A^\top A = I_n$, 则称为列正交矩阵; 若满足 $AA^\top = I_m$, 则称为行正交矩阵。二者统称半正交矩阵。

 奇异值分解的几何意义:对于秩为 r 的矩阵 $A_{m \times n}$, 存在单位正交向量 $\mathbf{v}_1, \dots, \mathbf{v}_r \in \mathbb{R}^n$ 和单位正交向量 $\mathbf{u}_1, \dots, \mathbf{u}_r \in \mathbb{R}^m$ 使得 $\mathbb{R}^n \rightarrow \mathbb{R}^m$ 的线性映射 $f(\mathbf{x}) = A\mathbf{x}$ 把正交基 $\mathbf{v}_1, \dots, \mathbf{v}_r$ 映为正交基 $\mathbf{u}_1, \dots, \mathbf{u}_r$, 即

$$\mathbf{u}_j = f(\mathbf{v}_j), \text{ 其中 } j = 1, \dots, r$$

推论 E.1. 接着定理 E.5, σ_k^2, \mathbf{v}_k 分别是 $A^\top A$ 的第 k 个特征值和特征向量; σ_k^2, \mathbf{u}_k 分别是 AA^\top 的第 k 个特征值和特征向量, $k = 1, 2, \dots, r$ 。另外,

$$\begin{aligned} A^\top A &= V\Sigma^2V^\top = (V\Sigma)(V\Sigma)^\top \\ AA^\top &= U\Sigma^2U^\top = (U\Sigma)(U\Sigma)^\top \end{aligned}$$

定理 E.6 (Eckart-Young, 1936). 接着定理 E.5, 构造 $A_k = U_k\Sigma_kV_k^\top = \sum_{j=1}^k \sigma_j \mathbf{u}_j \mathbf{v}_j^\top$, 其中 U_k, Σ_k, V_k 分别表示 U, Σ, V 的前 k 列构成的矩阵, 则

$$\begin{aligned} \min_{\text{rank}(B)=k} \|A - B\|_2 &= \|A - A_k\|_2 = \sigma_{k+1} \\ \min_{\text{rank}(B)=k} \|A - B\|_F &= \|A - A_k\|_F = \sqrt{\sum_{j=k+1}^r \sigma_j^2} \end{aligned}$$

Eckart-Young 定理保证了在所有秩为 k 的 $m \times n$ 矩阵中, A_k 是对 A 的最佳逼近, 我们称它为秩为 k 的 Eckart 近似。另外, 从 Eckart-Young 定理不难看出, 秩 k 越高, $\|A - A_k\|$ 越小, 即 A_k 对 A 的近似程度越好。

奇异值分解定理和 Eckart-Young 定理在矩阵论、多元统计、信号处理(包括图像处理)、自然语言处理等领域有着广泛而重要的应用。譬如, 多元统计里的主成分分析(principal component analysis, PCA)、自然语言处理中的潜在语义标引(latent semantic indexing, LSI)便是基于此。

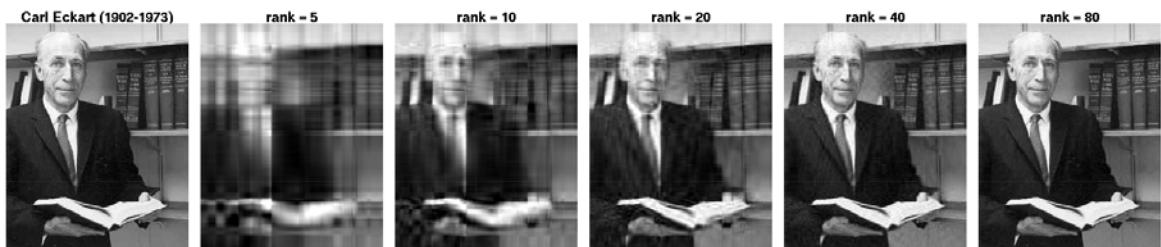


图 E.1: Carl Eckart (1902-1973), 美国物理学家。最左边的图片是 Eckart 的原始照片, 其余的是该图片的秩为 $5, 10, 20, 40, 80$ 的 Eckart 近似。秩越高, 越近似。

定义 E.7 (幂等矩阵). 如果方阵 A 满足 $AA = A$, 则称其为幂等矩阵(idempotent)。

性质 E.2. 不难验证幂等矩阵有下面的性质:

- 幂等矩阵的特征值要么是 1, 要么是 0。
- 若 A 是幂等矩阵, 则 $I - A$ 也是。
- 若 $A_{n \times n}$ 是对称的幂等矩阵, 不妨设其秩为 r , 则存在分解

$$A = UU^\top, \text{ 其中 } U_{n \times r} \text{ 是秩为 } r \text{ 的正交矩阵}$$

定理 E.7 (Perron, 1907). 若 $A_{n \times n} = (a_{ij})$ 是元素皆为正数的方阵, 则存在实特征值 $r > 0$, 恰为 A 的谱半径*, 其特征向量的分量皆为正数。 r 称作 A 的 Perron 根†, 它还满足

$$\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}$$

定理 E.8. 若分块方阵 $\Sigma = (A, B; C, D)$ 非奇异, 其中 A, D 为方阵且 D 非奇异, 则称矩阵 $S = A - BD^{-1}C$ 为矩阵 $(A, B; C, D)$ 关于 D 的 Schur 补 (Schur complement)‡。如果 S 可逆, 则 Σ^{-1} 可表示为下面的分块矩阵。

$$\Sigma^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}BD^{-1} \\ -D^{-1}CS^{-1} & D^{-1} + D^{-1}CS^{-1}BD^{-1} \end{pmatrix}$$

特别地, 若 Σ 是对称的正定矩阵, 则 Σ 关于 D 的 Schur 补 $S = A - BD^{-1}C$ 也是对称的正定矩阵。

定理 E.9 (Sherman-Morrison, 1949). 已知矩阵 $A_{n \times n}$ 可逆, 且 n 维向量 \mathbf{u}, \mathbf{v} 满足 $1 + \mathbf{v}^\top A^{-1} \mathbf{u} \neq 0$, 则矩阵 $A + \mathbf{u}\mathbf{v}^\top$ 可逆, 且

$$(A + \mathbf{u}\mathbf{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1} \mathbf{u}}$$

定义 E.8 (梯度). 已知函数 $f(\mathbf{x})$ 是关于向量 $\mathbf{x} = (x_1, \dots, x_n)^\top$ 的一个实值函数, 在 $\Omega \subseteq \mathbb{R}^n$ 上有定义且可微, $f(\mathbf{x})$ 相对于 \mathbf{x} 的梯度 (gradient) 定义为如下的 n 维列向量, 常记作 $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ 或 $\nabla_{\mathbf{x}} f$ 。在不强调 \mathbf{x} 时, 也简记作 $\text{grad}f$ 或 ∇f 。

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)^\top$$

*矩阵 $A_{n \times n}$ 的所有特征值 $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ 的最大模长 $\rho(A) = \max\{|\lambda_1|, \dots, |\lambda_n|\}$ 称为 A 的谱半径 (spectral radius)。

†该结果是德国数学家 Oskar Perron (1880-1975) 于 1907 年得到。

‡Issai Schur (1875-1941), 在德国工作的犹太数学家, 研究领域是群表示理论等。

梯度 $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ 的转置为

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^\top} = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)$$

定义 E.9. 令 m 维向量 $\mathbf{y} = (y_1, \dots, y_i, \dots, y_m)^\top$ 中的每个分量 y_i 都是 $\mathbf{x} \in \mathbb{R}^n$ 的实值函数, 即 $y_i = f_i(\mathbf{x})$, 并且函数 $f_i(\mathbf{x}), i = 1, 2, \dots, m$ 在 $\Omega \subseteq \mathbb{R}^n$ 上有定义且可微, 定义行向量值函数 $\mathbf{y}^\top = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$ 相对于 \mathbf{x} 的梯度为如下的 $n \times m$ 矩阵。

$$\frac{\partial \mathbf{y}^\top}{\partial \mathbf{x}} = (\nabla f_1, \dots, \nabla f_i, \dots, \nabla f_m) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_i}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_i}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_i}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

定义 E.10 (雅可比矩阵). 向量值函数 $\mathbf{y} = f(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^\top$ 的雅可比矩阵 (Jacobian matrix) 是一个 $m \times n$ 矩阵, 定义为

$$J_f = \frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \left(\frac{\partial \mathbf{y}^\top}{\partial \mathbf{x}} \right)^\top$$

显然, 矩阵 J_f 的 (i, j) 元素为 $\frac{\partial f_i}{\partial x_j}$ 。

性质 E.3. 对于任意 n 阶方阵 A , 皆有

$$\frac{\partial(A\mathbf{x})}{\partial \mathbf{x}^\top} = A$$

证明. 列向量 $f(\mathbf{x}) = A\mathbf{x}$ 的第 i 个元素为 $a_{i1}x_1 + \dots + a_{ij}x_j + \dots + a_{in}x_n$, 所以 J_f 的 (i, j) 元素为 a_{ij} , 得证。 \square

性质 E.4. 已知 $f(\mathbf{y})$ 和 $g(\mathbf{y})$ 都是实值函数, 其中 $\mathbf{y} \in \mathbb{R}^m$ 。已知 $\mathbf{y} = h(\mathbf{x}) = (h_1(\mathbf{x}), \dots, h_m(\mathbf{x}))^\top$ 是有关 \mathbf{x} 的向量值函数, 其中 $\mathbf{x} \in \mathbb{R}^n$ 。则

$$\begin{aligned} \text{乘法法则: } \frac{\partial(f(\mathbf{y})g(\mathbf{y}))}{\partial \mathbf{y}} &= g(\mathbf{y}) \frac{\partial f(\mathbf{y})}{\partial \mathbf{y}} + f(\mathbf{y}) \frac{\partial g(\mathbf{y})}{\partial \mathbf{y}} \\ \text{或者, } \nabla(fg) &= g\nabla f + f\nabla g \\ \text{链式法则: } \frac{\partial f(h(\mathbf{x}))}{\partial \mathbf{x}} &= \frac{\partial \mathbf{y}^\top}{\partial \mathbf{x}} \frac{\partial f(\mathbf{y})}{\partial \mathbf{y}} \end{aligned} \tag{E.3}$$

定理 E.10. 已知向量值函数 $f: \mathbf{x} \mapsto \mathbf{y}$ 在点 $\mathbf{p} \in \mathbb{R}^n$ 可微, 则在 \mathbf{p} 的足够小的邻域里

$f(\mathbf{x})$ 可由线性函数来近似, 即

$$f(\mathbf{x}) \approx f(\mathbf{p}) + J_f(\mathbf{p})(\mathbf{x} - \mathbf{p}) = f(\mathbf{p}) + (\mathbf{x} - \mathbf{p})^\top \nabla f(\mathbf{p})$$

定理 E.11. 令 $f(\mathbf{x})$ 如下表左列所定义, 其中 $\mathbf{c} = (c_1, \dots, c_n)^\top$ 为一个 n 维向量, 令 $S_{n \times n}$ 为任意 n 阶对称矩阵, 则有下面的结果。

$f(\mathbf{x})$	$\partial f(\mathbf{x}) / \partial \mathbf{x}$	$\partial^2 f(\mathbf{x}) / \partial \mathbf{x}^2$
$\mathbf{x}^\top \mathbf{c}$ or $\mathbf{c}^\top \mathbf{x}$	\mathbf{c}	O
$\mathbf{x}^\top \mathbf{x}$	$2\mathbf{x}$	$2I$
$\mathbf{x}^\top S \mathbf{x}$	$2S\mathbf{x}$	$2S$

证明. 记 S 的 (i, j) 元素为 s_{ij} , 往证 $\partial(\mathbf{x}^\top S \mathbf{x}) / \partial \mathbf{x} = 2S\mathbf{x}$ 如下, 其他留作练习。

$$\mathbf{x}^\top S \mathbf{x} = \sum_{i,j=1}^n x_i s_{ij} x_j \Rightarrow \frac{\partial(\mathbf{x}^\top S \mathbf{x})}{\partial x_k} = \sum_{j=1}^n s_{kj} x_j + \sum_{i=1}^n x_i s_{ik} = 2 \sum_{j=1}^n s_{kj} x_j, \quad k = 1, \dots, n \quad \square$$

定义 E.11. 对于多元实值函数 $f(\mathbf{x})$, 下面的方阵称为海森矩阵 (Hessian matrix):

$$\frac{\partial}{\partial \mathbf{x}^\top} \left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right] = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{n \times n}$$

海森矩阵是以德国数学家 Ludwig Otto Hesse (1811-1874) 命名的, 记作 $\nabla_x^2 f$ 或 $\partial^2 f(\mathbf{x}) / \partial \mathbf{x}^2$, 有时也简记作 $H(f)$ 或 $\nabla^2 f$ 。海森矩阵常用于近似计算

$$f(\mathbf{x} + \Delta \mathbf{x}) \approx f(\mathbf{x}) + \Delta \mathbf{x}^\top \nabla f + \frac{1}{2} \Delta \mathbf{x}^\top \nabla^2 f \Delta \mathbf{x}$$

定理 E.12. 令 $D \subseteq \mathbb{R}^n$ 为开凸集 (见**定义 F.1**), 若实值函数 $f(\mathbf{x})$ 在 D 上存在一阶和二阶偏导数且 $\forall \mathbf{x} \in D$ 皆有 $A(\mathbf{x}) = -\nabla^2 f$ 是正定矩阵, 则方程组 $\nabla f = \mathbf{0}$ 在 D 内至多只有一个解且若有解必是 f 的最大值点。

证明. 设方程组 $\nabla f = \mathbf{0}$ 在 D 内有两个解 $\mathbf{x}_1 \neq \mathbf{x}_2$, 则函数 $g(t) = f[t\mathbf{x}_1 + (1-t)\mathbf{x}_2]$ 在 $t \in [0, 1]$ 上二阶可导且 $g'(0) = g'(1) = 0$, 故存在 $t_0 \in (0, 1)$ 使得 $g''(t_0) = -(\mathbf{x}_1 - \mathbf{x}_2)^\top A[t_0 \mathbf{x}_1 + (1-t_0) \mathbf{x}_2] (\mathbf{x}_1 - \mathbf{x}_2) = 0$ 。由于 $\forall \mathbf{x} \in D, A(\mathbf{x})$ 是正定矩阵, 所以 $(\mathbf{x}_1 - \mathbf{x}_2)^\top A[t_0 \mathbf{x}_1 + (1-t_0) \mathbf{x}_2] (\mathbf{x}_1 - \mathbf{x}_2) > 0$, 矛盾! 于是至多有一个解。若 \mathbf{x}_0 是 $\nabla f = \mathbf{0}$ 的解, $\forall \mathbf{x} \in D$, 函数 $h(t) = f[t\mathbf{x}_0 + (1-t)\mathbf{x}_0]$ 在 $t \in [0, 1]$ 上二阶可导且 $h'(0) = 0, h''(t) < 0$, 其中 $t \in (0, 1]$, 进而 $h'(t) < h'(0) = 0$ 。于是, $f(\mathbf{x}) = h(1) < h(0) = f(\mathbf{x}_0)$ 。 \square

已知实值函数 $f(\mathbf{x})$ 在某开集中存在极值点, 如何求得它呢? 考虑 $f(\mathbf{x})$ 在点 $\mathbf{x}^{(t)}$

的某个小邻域内的 Taylor 级数展开，忽略其余项。

$$f(\mathbf{x}) \approx f(\mathbf{x}^{(t)}) + (\mathbf{x} - \mathbf{x}^{(t)})^\top \nabla f(\mathbf{x}^{(t)}) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(t)})^\top \nabla^2 f(\mathbf{x}^{(t)}) (\mathbf{x} - \mathbf{x}^{(t)})$$

令 $\nabla f(\mathbf{x}) = \mathbf{0}$ ，得到方程 $\nabla f(\mathbf{x}^{(t)}) + \nabla^2 f(\mathbf{x}^{(t)}) (\mathbf{x} - \mathbf{x}^{(t)}) \approx \mathbf{0}$ 。如果矩阵 $\nabla^2 f(\mathbf{x}^{(t)})$ 可逆，则可求得 $f(\mathbf{x})$ 的极值点

$$\mathbf{x} \approx \mathbf{x}^{(t)} - [\nabla^2 f(\mathbf{x}^{(t)})]^{-1} \nabla f(\mathbf{x}^{(t)})$$

算法 E.2 (Newton-Raphson). 设 $\mathbf{x}^{(t)}$ 是当前对 $f(\mathbf{x})$ 极值点的近似，则下一步的近似 $\mathbf{x}^{(t+1)}$ 可按下面的方法得到。

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - [\nabla^2 f(\mathbf{x}^{(t)})]^{-1} \nabla f(\mathbf{x}^{(t)})$$

初值 $\mathbf{x}^{(0)}$ 可随意指定。若极大（小）值点不唯一，从多个初值出发有益于判定某极大（小）值点不是最大（小）值点。

定理 E.13 (Laplace 近似积分法^{*}, 1774). 已知 $g(\mathbf{x}), h(\mathbf{x})$ 是定义在有界区域 $\Omega \subseteq \mathbb{R}^n$ 上的光滑实值函数，另外 $h(\mathbf{x})$ 还是有界的单峰函数，在内点 $\mathbf{x}_0 \in \Omega$ 处取得极大值，并且 $g(\mathbf{x}_0) \neq 0$ ，矩阵 $A = -\nabla^2 h(\mathbf{x}_0)$ 正定，若正数 $\lambda \rightarrow \infty$ ，则

$$\int_{\Omega} g(\mathbf{x}) \exp\{\lambda h(\mathbf{x})\} d\mathbf{x} = g(\mathbf{x}_0) \exp\{\lambda h(\mathbf{x}_0)\} \sqrt{\frac{(2\pi)^n}{\lambda^n |A|}} \left\{ 1 + O\left(\frac{1}{\lambda}\right) \right\} \quad (\text{E.4})$$

※证明. 详见 [160] 的第 495-500 页。 □

例 E.2. 单峰函数 $h(x) = -\cosh x$ 在内点 $x = 0$ 处 $h'(0) = 0, h''(0) < 0$ ，经验证 $h(x)$ 在 $x = 0$ 处取得极大值 -1 。令 $g(x) = \sin x/x$ ，则 $g(0) = 1$ 。利用式 (E.4)，当正数 λ 足够大时，近似地有

$$\int_{-1}^1 \frac{\sin x}{x} \exp\{-\lambda \cosh x\} dx \approx e^{-\lambda} \sqrt{\frac{2\pi}{\lambda}}$$

例 E.3. 已知非负的单峰实值函数 $g(\mathbf{x})$ 在 \mathbb{R}^n 上可积，假设在 $\mathbf{x} = \mathbf{x}_0$ 处 $\nabla g(\mathbf{x}_0) = \mathbf{0}$ 并且 $A = -\nabla^2 \ln g(\mathbf{x}_0)$ 为正定矩阵，则 $\ln g(\mathbf{x}) \approx \ln g(\mathbf{x}_0) - \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top A (\mathbf{x} - \mathbf{x}_0)$ 。进而近似地有

$$\int_{\mathbb{R}^n} g(\mathbf{x}) d\mathbf{x} \approx \frac{(2\pi)^{n/2} g(\mathbf{x}_0)}{\sqrt{|A|}}$$

^{*}物理学中常把 Laplace 近似积分法称为鞍点法 (saddle point method)。

事实上, 非负 n 元实值函数 $\sqrt{|A|}g(\mathbf{x})[(2\pi)^{n/2}g(\mathbf{x}_0)]^{-1}$ 近似地视作多元正态分布 $N_n(\mathbf{0}, A^{-1})$ 的密度函数。

定义 E.12. 已知 $f(A)$ 是关于实矩阵 $A = (a_{ij})_{m \times n}$ 的一个实值函数, 定义 $f(A)$ 相对于 A 的梯度矩阵为 $\partial f(A)/\partial A = (\partial f(A)/\partial a_{ij})_{m \times n}$, 常记作 $\nabla_A f$ 。

性质 E.5 (乘法法则和链式法则). 已知 A 为 $m \times n$ 矩阵, 若 $f(A)$ 和 $g(A)$ 都是有关矩阵 A 的实值函数, $h(y)$ 是单变量实值函数, 则

$$\begin{aligned}\frac{\partial f(A)g(A)}{\partial A} &= g(A)\frac{\partial f(A)}{\partial A} + f(A)\frac{\partial g(A)}{\partial A} \\ \frac{\partial h(f(A))}{\partial A} &= \frac{dh(y)}{dy}\frac{\partial f(A)}{\partial A}\end{aligned}$$

定理 E.14. 已知 A 为 $m \times n$ 阶矩阵, 对任意的 $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ 皆有 $\partial(\mathbf{x}^\top A \mathbf{y})/\partial A = \mathbf{x} \mathbf{y}^\top$ 。如果矩阵 A 是可逆方阵, 则 $\partial|A|/\partial A = |A|(A^{-1})^\top$ 。

例 E.4. $\ln \phi(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2} \ln |\Sigma^{-1}| + \text{常数}$ 是关于对称可逆矩阵 Σ^{-1} 的实值函数, 容易验证

$$\frac{\partial \ln \phi(\mathbf{x}|\boldsymbol{\mu}, \Sigma)}{\partial \Sigma^{-1}} = \Sigma - (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top$$

练习 E.1. 请验证下面的相对于 X 的梯度矩阵的结果:

$$\begin{aligned}\frac{\partial \text{tr}(AXB)}{\partial X} &= A^\top B^\top \\ \frac{\partial \text{tr}(AX^\top B)}{\partial X} &= BA \\ \frac{\partial \text{tr}(AXX^\top A^\top)}{\partial X} &= 2A^\top AX \\ \frac{\partial \text{tr}(AX^\top XA^\top)}{\partial X} &= 2XA^\top A\end{aligned}$$

附录 F

凸性与 Jensen 不等式

丹 丹麦数学家兼工程师 Johan Ludwig Jensen (1859-1925) 于 1906 年证明了有关凸函数性质的著名的 Jensen 不等式，如今它已演变成若干非常实用的形式，有着广泛的应用。例如，证明推广了的 Rao-Blackwell 定理、Kullback-Leibler 散度非负等结论都用到了该不等式。

Jensen 不等式谈论的是定义在凸集上的凸函数的性质，它有离散和连续两个版本，其中连续版才是 Jensen 发现的。因为连续版的证明比离散版的要难些，人们干脆把功劳都记到 Jensen 的头上。Jensen 不等式旗下一大堆著名的不等式：Jensen 不等式 \Rightarrow Young 不等式 \Rightarrow Hölder 不等式 \Rightarrow Cauchy-Schwarz 不等式。



定义 F.1 (凸集). 如果集合 $S \subset \mathbb{R}^d$ 上任意两点的连线段仍在 S 上，则称 S 为凸集 (convex set)。所谓“任意两点的连线段仍在 S 上”即 $\forall \mathbf{x}, \mathbf{y} \in S$,

$$\alpha \mathbf{x} + (1 - \alpha) \mathbf{y} \in S, \forall \alpha \in [0, 1]$$

定义 F.2 (凸函数). 凸集 S 上的实值函数 $g : S \rightarrow \mathbb{R}$ 称为 S 上的凸函数 (convex function)，如果它满足 $\forall \mathbf{x}, \mathbf{y} \in S, \forall \alpha \in [0, 1]$ 皆有

$$g[\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}] \leq \alpha g(\mathbf{x}) + (1 - \alpha) g(\mathbf{y})$$

凸函数直观的几何解释见图 F.1。

性质 F.1. 凸函数具有如下的一些性质，证明详见 [129]。

- ① 若 $d = 1$, $g(x)$ 是凸函数当且仅当 $\forall x \in S, g''(x) > 0$ 。例如, $g(x) = -\ln x$ 是凸函数, 还有 $e^x, |x|^t, t \geq 1$ 等等。

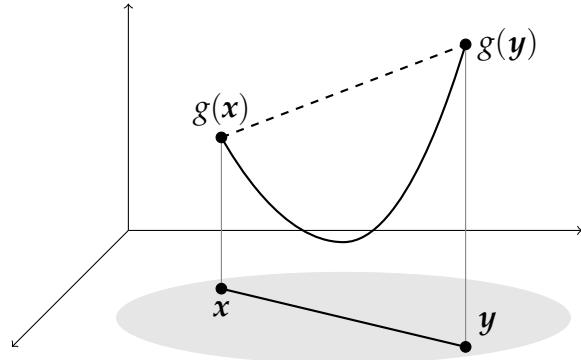
② 若 $d > 1$, $g(\mathbf{x})$ 是凸函数当且仅当 $\forall \mathbf{u} \in \mathbb{R}^d, \forall \mathbf{x} \in S$ 皆有

$$\mathbf{u}^\top \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x}^2} \mathbf{u} > 0$$

③ 若 $f(x), g(x)$ 是凸函数, 则 $\max\{f(x), g(x)\}$ 和 $f(x) + g(x)$ 也是凸函数。若 g 还是非减的, 则 $g[f(x)]$ 也是凸函数。

④ 凸性在仿射变换下不变: 若 $f(\mathbf{x})$ 是凸函数, 其中 $\mathbf{x} \in \mathbb{R}^m$, 则 $g(\mathbf{y}) = f(A_{m \times n} \mathbf{y} + \mathbf{c})$ 也是凸函数, 其中 $\mathbf{y} \in \mathbb{R}^n$ 且 $\mathbf{c} \in \mathbb{R}^m$ 。

图 F.1: 凸集 S 的几何含义是其上任意两点 \mathbf{x}, \mathbf{y} 的连线段仍在 S 上。 S 上的凸函数 g 的几何含义是线段 $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ 在 g 下的像在线段 $\alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y})$ (即图中的虚线) 的下方。



定理 F.1 (Jensen 不等式的离散版). 已知 $g(x)$ 是凸函数且 $\alpha_j > 0$ 满足 $\sum_{j=1}^n \alpha_j = 1$, 则下面的不等式成立。

$$g\left(\sum_{j=1}^n \alpha_j x_j\right) \leq \sum_{j=1}^n \alpha_j g(x_j) \quad (\text{F.1})$$

等号成立当且仅当 $x_1 = x_2 = \dots = x_n$ 或 $g(x)$ 是线性函数。

证明. 显然 $n = 2$ 时成立。设 $n \leq k$ 时都成立, 不妨设 $\alpha_1 \neq 1$,

$$\begin{aligned} g\left(\sum_{j=1}^{k+1} \alpha_j x_j\right) &= g\left[\alpha_1 x_1 + (1 - \alpha_1) \sum_{j=2}^{k+1} \frac{\alpha_j}{1 - \alpha_1} x_j\right] \\ &\leq \alpha_1 g(x_1) + (1 - \alpha_1) g\left[\sum_{j=2}^{k+1} \frac{\alpha_j}{1 - \alpha_1} x_j\right] \leq \sum_{j=1}^{k+1} \alpha_j g(x_j) \end{aligned} \quad \square$$

练习 F.1 (算术与几何平均不等式). 已知 z_1, z_2, \dots, z_n 是非负实数, 则

$$\sqrt[n]{\prod_{j=1}^n z_j} \leq \frac{1}{n} \sum_{j=1}^n z_j$$

著名的 Young 不等式和 Hölder 不等式^{*}可由 Jensen 不等式导出。

推论 F.1 (Young 不等式). 对任意实数 x, y 皆有

$$|xy| \leq \frac{|x|^r}{r} + \frac{|y|^s}{s}, \text{ 其中 } r > 1 \text{ 且 } \frac{1}{r} + \frac{1}{s} = 1 \quad (\text{F.2})$$

证明. 因为 $g(x) = \exp(x)$ 是凸函数, 由 Jensen 不等式, 我们得到

$$|xy| = \exp \left\{ \frac{\ln |x|^r}{r} + \frac{\ln |y|^s}{s} \right\} \leq \frac{|x|^r}{r} + \frac{|y|^s}{s} \quad \square$$

推论 F.2 (Hölder 不等式, 1889). 若实数 $r > 1$ 且 $1/r + 1/s = 1$, 则有

$$\begin{aligned} \left| \sum_{k=1}^n x_k y_k \right| &\leq \left(\sum_{k=1}^n |x_k|^r \right)^{1/r} \left(\sum_{k=1}^n |y_k|^s \right)^{1/s} \text{ 或者 } \langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_r \|\mathbf{y}\|_s \\ \left| \int_a^b f g du \right| &\leq \left(\int_a^b |f|^r du \right)^{1/r} \left(\int_a^b |g|^s du \right)^{1/s} \text{ 或者 } \langle f, g \rangle \leq \|f\|_r \|g\|_s \end{aligned} \quad (\text{F.3})$$

特别地, 当 $r = s = 2$ 时, Hölder 不等式也称为 Cauchy-Schwarz 不等式, 由法国数学家 Cauchy 于 1821 年证得。

证明. 由 Young 不等式 (F.2) 可得

$$\frac{|x_k|}{\|\mathbf{x}\|_r} \cdot \frac{|y_k|}{\|\mathbf{y}\|_s} \leq \frac{1}{r} \cdot \frac{|x_k|^r}{\sum_{k=1}^n |x_k|^r} + \frac{1}{s} \cdot \frac{|y_k|^s}{\sum_{k=1}^n |y_k|^s}, \text{ 其中 } k = 1, 2, \dots, n$$

将上述 n 个不等式相加, 便证得结果 (F.3)。 \square

其实, 公式 (F.1) 首先由德国数学家 O. Hölder (1859-1937) 于 1889 年提出 [70], 下面的结果才是 Jensen 于 1906 年提出并证明了的。

定理 F.2 (Jensen 不等式的连续版, 1906). 若 $g(x)$ 是 $S \subset \mathbb{R}$ 上凸函数, 则

$$g \left[\int_D \lambda(t)x(t)dt \right] \leq \int_D \lambda(t)g[x(t)]dt \quad (\text{F.4})$$

其中 $x(D) \subseteq S$ 且 $\forall t \in D, \lambda(t) \geq 0$ 满足 $\int_D \lambda(t)dt = 1$ 。等号成立当且仅当 $x(t)$ 在 D 上为常数或 g 在 $x(D)$ 上是线性函数。

*W. H. Young (1863-1942) 和 O. Hölder (1859-1937) 分别是英国和德国数学家。

附录 G

软件 R、Maxima 和 GnuPlot 简介

朝看花开满树红，暮看花落树还空。若将花比人间事，花与人间事一同。

龙牙《居遁》

古语道“工欲善其事，必先利其器”。开源的数学软件不胜枚举，经得起实践和时间考验的佼佼者就屈指可数了，然而不论怎么数，R、Maxima 和 GnuPlot 必列其中。这三个软件都是跨平台的，既支持命令行交互模式，也支持脚本。作者推荐在类 UNIX 环境使用它们。

本书鼓励以“用”为驱动熟练掌握这些优秀的工具软件，但由于篇幅和主题所限，在正文中无法过多地介绍这三门编程语言的细节，本附录所给的也仅仅是浮光掠影式的简介，读者可通过软件自带的手册或在线帮助文档学习它们。

G.1 R：最好的统计软件

R 是一门用于统计分析、统计计算和数据可视化的面向对象编程语言，它与 Bell 实验室 John Chambers 等人研发的 S 语言兼容^{*}，有时也称为 GNU S。R 的特点是一少部分的统计功能在 R 的底层实现，绝大多数基于经典统计技术和许多现代的统计方法的功能都是以包 (package) 的形式提供。R 的一个显著优点是与其他编程语言/数据库之间有很好的接口，如 C、Python、Gibbs 抽样工具 BUGS 或 JAGS 等。

“众人拾柴火焰高”，开源为 R 的普及铺平了道路，并使得 R 在短时间内轻松领先于 S-plus、SAS、Stata 等诸多优秀的统计软件，成为“新老皆宜”的选择。实践证明 R 是统计学研究和应用的利器，也是机器学习、模式识别、生物信息学、自然语言处理、计量经济学等涉及数据处理学科的不可多得的工具。

^{*}1998 年，John Chambers 因对 S 语言的杰出贡献获得了 ACM 软件系统奖。

R 按照应用领域，将工具包加以分类，如社会科学、计量经济学、金融、医学图像分析、遗传学、自然语言处理、机器学习、高性能计算等。同时，也按照方法类别对工具包进行了整理，如贝叶斯推断、多元统计学、聚类、试验设计、生存分析、时间序列分析、稳健统计方法等。

读者可以从 <http://cran.r-project.org> 或其镜像网站获得源码和可用的程序包，其中标准包和推荐包都经过严格的测试。另外，更多的针对具体问题的包可以通过网络得到。

例 G.1 (交互式，> 是命令行提示符). 利用 summary 函数考察一维数据的分布情况；利用 hist 绘出直方图，并叠加上密度估计曲线。

```
> attach(faithful)
> summary(eruptions)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.600 2.163 4.000 3.488 4.454 5.100
> hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE)
> lines(density(eruptions, bw=0.1))
> rug(eruptions)
```

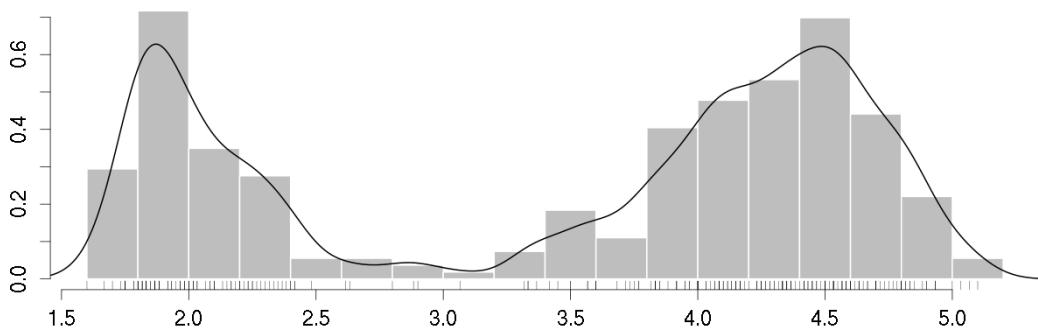


图 G.1: 直方图和密度函数估计。

例 G.2 (聚类). 从领域受限的大规模语料中提取关键词，计算词语之间的相似度，利用 hclust 函数画出它们的聚类树。

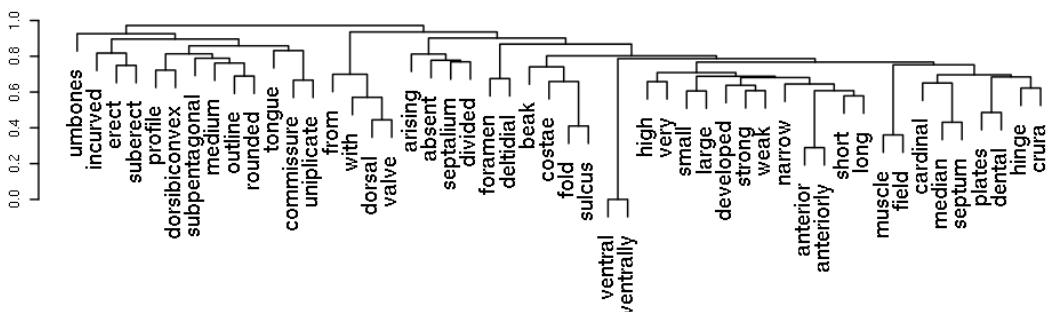


图 G.2: 从给定语料中提取关键词并计算它们的聚类。

例 G.3 (数据的可视化). 对于 Fisher 的 iris 数据 (四个特征, 三个类), 可以通过观察其指定分量来了解它们在空间中的分布情况。

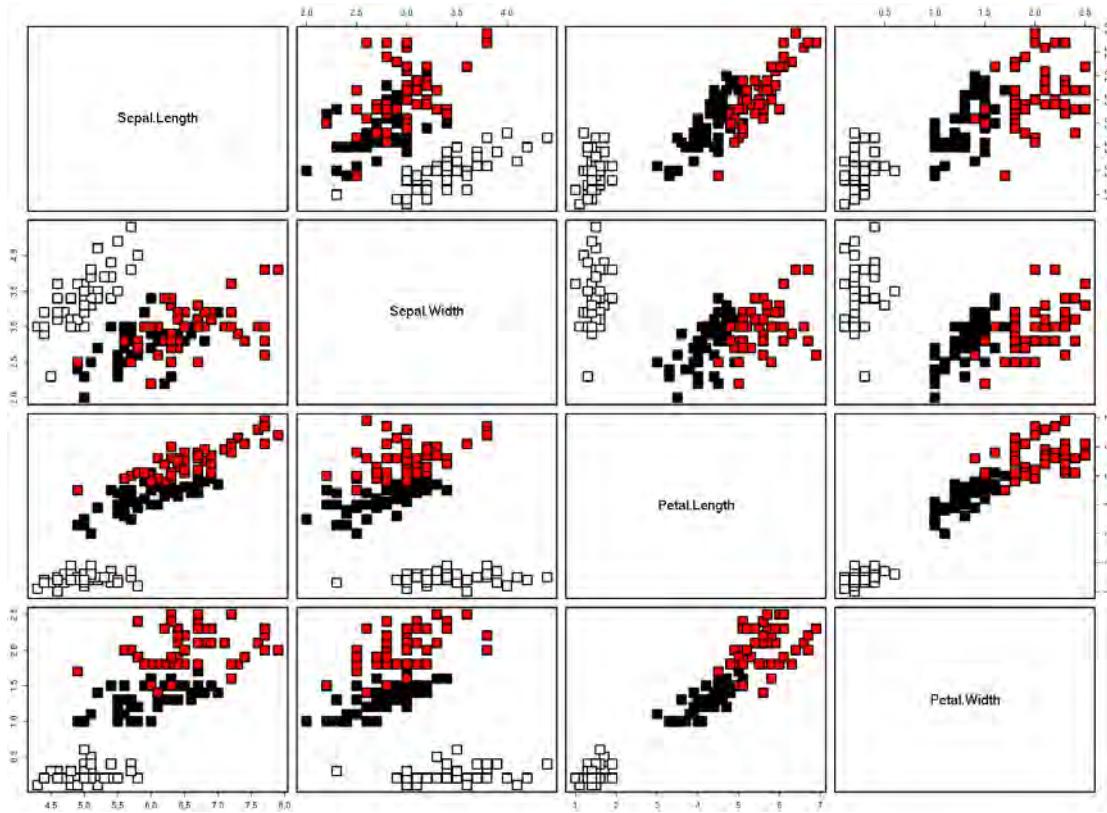


图 G.3: Fisher 的 iris 数据 (四个特征, 三个类) 的可视化。

G.2 Maxima: 符号计算的未来之路

Maxima 是 LISP 语言实现的用于公式推导和符号计算的计算机代数系统 (Computer Algebra System, CAS), 它的前身是 MIT 于 1968-1982 年间研发的计算机代数系统 Macsyma (CAS 的鼻祖之一), 更准确地说, 它是 Macsyma 的 GPL 衍生版本^{*}。

1982 年, MIT 将 Macsyma 源码拷贝移交给美国能源部, 该版本被称为 DOE Macsyma, 其中一份拷贝由 Texas 大学的 William F. Schelter 教授维护直至他 2001 年去世。1998 年, Schelter 从能源部获准以 GPL 方式发布 DOE Macsyma 源码。2000 年, Schelter 在 SourceForge 发起 Maxima 项目作为 DOE Macsyma 的延续。

Maxima 可与商用 CAS 软件 Maple 和 Mathematica 媲美, 甚至优于后者, 因为它有老当益壮的 LISP 语言做后盾而具有良好的可扩展性 (用户可以在 LISP 层定义函数, 在 Maxima 层调用它)。在开源盛世, Maxima 的生命力必将顽强。

^{*}GPL: GNU 通用公共许可证 (General Public License) 的简称, 是自由软件基金会发行的用于计算机软件的许可证。Macsyma 是 CAS 的鼻祖之一, 对后续的 CAS 产生过深远的影响, 也包括商用的 Maple 和 Mathematica。

例 G.4 (解线性递归式). 已知线性递归关系 $(n+4)T(n+2) = -T(n+1) + (n-1)T(n)$, 试求解 $T(n)$.

```
(%i1) load("solve_rec") $  
(%i2) solve_rec((n+4)*T[n+2] + T[n+1] - (n-1)*T[n], T[n]);
```

Maxima 给出的答案是

$$T_n = \frac{k_2(2n+1) - k_1(2n^2 + 2n - 1)(-1)^n}{(n-1)n(n+1)(n+2)}$$

例 G.5. 用组合数学的方法证明李善兰恒等式 (1.38) 并非易事, 可仅用 Maxima 的寥寥数行代码就能验证它。

```
load("simplify_sum") $  
assume(m > n) $  
sum ((binomial(n,k))^2 * binomial(m+2*n-k, 2*n), k, 0, n);  
simplify_sum(%);
```

例 G.6 (配方法). 给出偶数次多项式 p 关于变量 x 的配方结果。

```
/* 目的: 按照某指定的变量, 实现多项式的配方法, 得到“平方项 + 尾项” */  
/* 输出: 偶数次的多项式 p 关于它的某个变量 x 的配方结果 */
```

```
CompSq(p,x) := block([degree, coef, s, residual],  
    p : expand(p),  
    degree : hipow(p, x),  
    if oddp(degree) or degree = 0 then p else (  
        coef : coeff(p, x, n),  
        s : x^(n/2),  
        residual : ratsimp(p - coef * s^2),  
        while hipow(residual, x) > 0 do (  
            residual : ratsimp(first(divide(p - coef * s^2, 2 * coef * s, x))),  
            s : s + residual),  
        coef * s^2 + ratsimp(CompSq(p - coef * s^2, x))) $
```

请读者利用上面的 CompSq 函数解决第 66 页的例 1.50 中的配方问题。

例 G.7. 利用 Maxima 计算级数和不定积分, 例如,

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

$$\int \frac{1}{1+x^3} dx = -\frac{\ln(x^2 - x + 1)}{6} + \frac{\arctan \frac{2x-1}{\sqrt{3}}}{\sqrt{3}} + \frac{\ln(x+1)}{3} + C$$

```
sum (1/n^2, n, 1, inf), simpsum; integrate(1/(1+x^3), x);
```

例 G.8 (Chebyshev 问题). 接着第 92 页的例 1.72, 用 Maxima 计算式 (1.36) 定义的概率 $P(n)$, 并绘制折线图以探究 $P(n)$ 的极限。

```

/* 条件: 已知 a, b 是介于 1 和 n 之间的自然数 */
/* 目标: 计算 a, b 互素的概率, 即分数 a/b 不可约的概率 P(n) */
/* s 表示互素自然对 (a,b) 的个数 */
/* totient(j): 不超过 j 且与 j 互素的自然数个数 */
P(n) := (s:1,
  for j: 2 while j <= n do
    s: s + 2 * totient (j),
  float(s/n^2)) \$

/* 定义长度为 MAX 的数组 */
MAX: 10^2 \$

x: make_array (fixnum, MAX); y: make_array (fixnum, MAX) \$

/* 给数组赋值, 绘出 (n, P(n)) 折线图 */
for n:1 while n < MAX do (
  x[n]: n+1,
  y[n]: P(n+1)) \$

load(draw) \$

draw2d( xrange = [2, MAX], yrange = [0.60, 0.78],
  points_joined = true, point_type = 0,
  grid = true, color = black,
  line_width = 3, font_size = 18,
  points(x, y)) \$
```

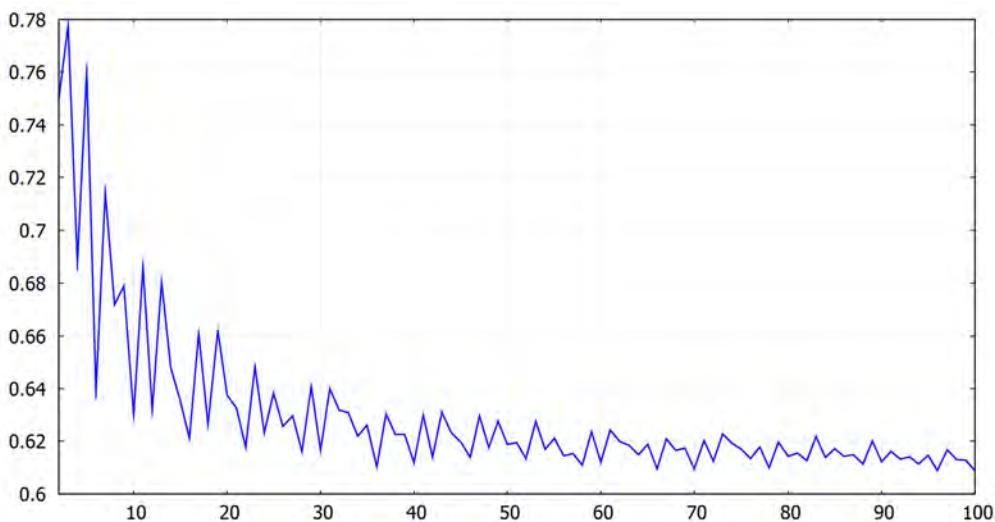


图 G.4: Chebyshev 问题: 令 $P(n)$ 为分子、分母随机地取自 $\{1, 2, \dots, n\}$ 的不可约分数的概率。通过 $P(2), \dots, P(100)$ 的折线图, 猜测 n 增大时, $P(n)$ 震荡着趋近某个值。事实上, $\lim_{n \rightarrow \infty} P(n) = 6/\pi^2 \approx 0.6079$, 具体推导见例 1.72。

例 G.9 (Goldbach 猜想). 1742 年, 德国数学家 Christian Goldbach (1690-1764) 提出一个猜想: 任意不小于 6 的偶数都可分解为两个奇素数之和。

在通往 Goldbach 猜想的途中, 中国著名数学家陈景润 (1933-1996) 于 1966 年取得了迄今为止最好的结果, “任何充分大的偶数都可表示为一个素数及一个不超过两个素数的乘积之和”, 简称为“1+2”。下面用 Maxima 给出了 100 的 Goldbach 分解, 并验证了 6 至 2×10^3 的偶数都满足 Goldbach 猜想。为简单起见, 算法未经优化。

```
(%i1) xprimep(x) := integerp(x) and (x > 1) and primep(x) $  
(%i2) BinaryDecomp : integer_partitions (100, 2) $  
(%i3) subset (BinaryDecomp, lambda ([x], every (xprimep, x))) ;  
(%o3) {[53, 47], [59, 41], [71, 29], [83, 17], [89, 11], [97, 3]}  
(%i4) GoldbachConjecture : true $  
(%i5) for n : 3 while n <= 10^3 do (  
    BinaryDecomp : integer_partitions (2*n, 2),  
    NoSolution : emptyp(subset (BinaryDecomp, lambda ([x], every (xprimep, x)))),  
    GoldbachConjecture : GoldbachConjecture and not(NoSolution)) $  
(%i6) GoldbachConjecture ;  
(%o6) true
```



目前, 利用计算机已经验证了 $n \leq 10^{18}$ 的偶数都满足 Goldbach 猜想。但面对无限个可能的情形, 有限的验证不等于证明, Goldbach 猜想依然是未解决的数学难题。

例 G.10. Maxima 可调用外部绘图程序 GnuPlot 实现绘图功能, 例如三维空间里可任意旋转的二维曲面。

```
(%i1) load(draw) $  
(%i2) draw(columns=2, gr3d(surface_hide = true,  
    explicit(x^2-y^2, x, -5, 5, y, -5, 5), explicit(6-x^2-y^2, x, -5, 5, y, -5, 5)),  
    gr3d(surface_hide = true, parametric_surface(cos(a) * (10 + b * cos(a/2)),  
        sin(a) * (10 + b * cos(a/2)), b * sin(a/2), a, -%pi, %pi, b, -1, 1))) $
```

例 G.11. 不断调用 draw 函数以产生模拟效果, 例如三维空间中粒子的布朗运动。

```
load(draw) $  
block([history:[[0,0,0]], lst, pos],  
for k:1 thru 10000 do  
    (lst: copylist(last(history)),  
     pos: random(3) + 1,  
     lst[pos]: lst[pos] + random(2)*2-1,  
     history: endcons(lst, history)),  
    draw3d(point_type = 0, points_joined = true, points(history))) $
```

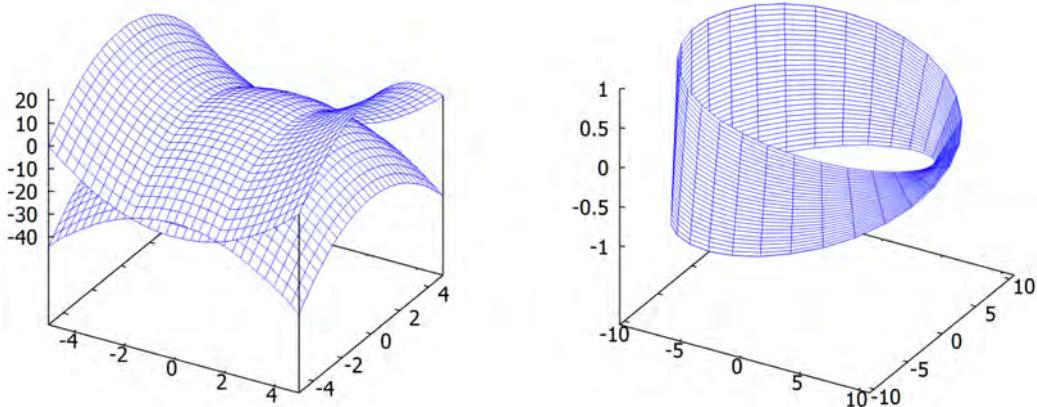


图 G.5: Maxima 的 draw 函数调用外部绘图程序 GnuPlot 完成绘图: 左图是曲面 $x^2 - y^2$ 与 $6 - x^2 - y^2$ 之交, 右图是不可定向曲面 Möbius 带。

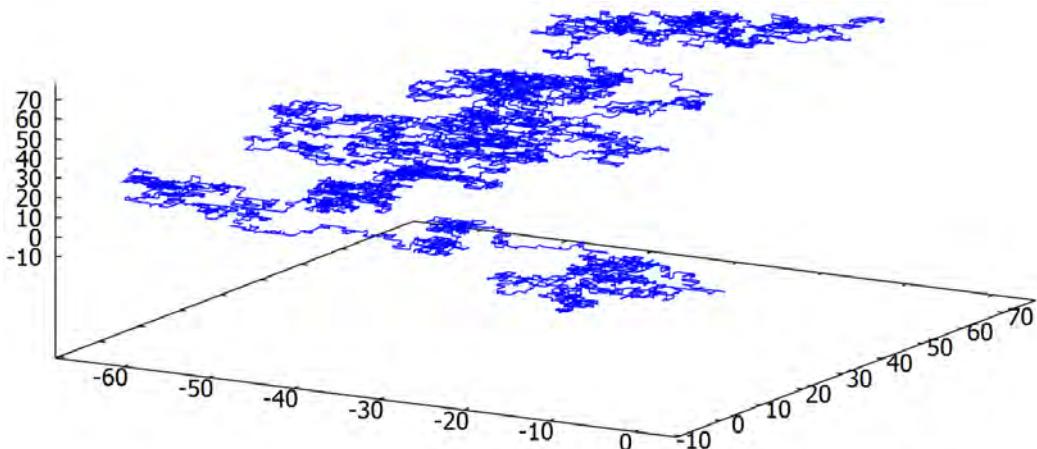


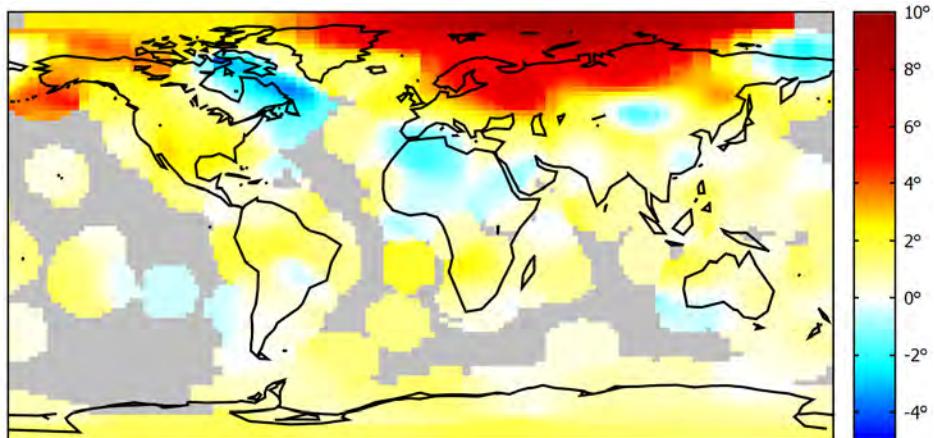
图 G.6: 三维空间中, 粒子 Brown 运动的随机模拟。

通过上面的例子, 读者能够感受到, 计算机为数学提供了宝贵的直观。然而, 从“有限”到“无限”是机器智能的鸿沟, 符号计算处理无穷集合时, 不能完全替代人类的思维。计算机能在多大程度上“做”数学, 这是个值得思考的问题 [38]。

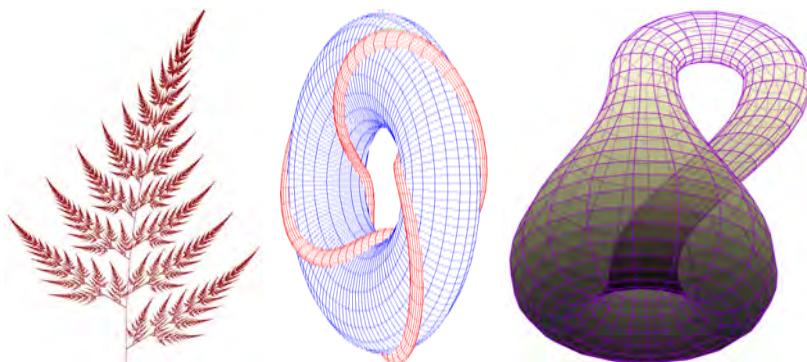
G.3 GnuPlot: 强大的函数绘图工具

GnuPlot 是一款轻便的科学绘图工具软件*, 是计算机代数系统 Maxima、数值计算工具 GNU Octave、计量经济分析软件 GRETL (Gnu Regression, Econometrics and Time-series Library) 等的绘图引擎。在函数绘图方面, GnuPlot 擅长绘制函数曲线、可三维旋转的二维曲面、向量场、等高线等, 也可用作数据的可视化。GnuPlot 几乎是无所不能的, 见图 G.7 所示的一些例子。

*GnuPlot 虽然名字中有“Gnu”, 但它尚不是 GNU 项目的一部分。从法律上, 可以免费使用 GnuPlot, 但不能免费分发 GnuPlot 的修改版本。



(a) 2005 年全球温度相对于 1951-1980 年温度的异常变化（灰色地区无数据）



(b) 分形、缠绕在环面上的三叶结、Klein 瓶

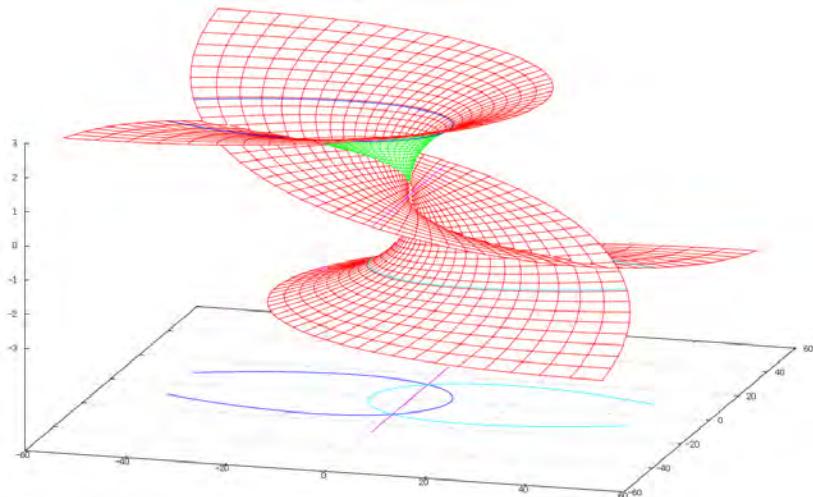
(c) 单复变函数 $f(z) = z^{1/3}$ 实部的 Riemann 曲面

图 G.7: 利用 GnuPlot 绘制的图形, 详见 <http://www.gnuplot.info>。

附录 H

习题答案或提示

1.1 $p_n = n!/n^n, p_{10} = 3.6288 \times 10^{-4}$, 即有空盒子的概率很大。

1.2 $C_n^2 n!/n^n$ 。提示：先选一个空盒子，有 n 种选法；恰有一个盒子空着意味着有一个盒子装着 2 个球。

1.3 由 $1 - (1 - p)^3 = 0.875$ 解得 $p = 0.5$ 。

1.4 $0.2^3 + 3 \times 0.8 \times 0.2^2 = 0.104$ 。

1.5 基本事件的总数为 n^k , 事件 A_m 所包含的基本事件个数为 $m^k - (m - 1)^k$, 故 $P(A_m) = [m^k - (m - 1)^k]/n^k$ 。

1.6 $1 - 2^4 C_{10}^4 / C_{20}^4 = 99/323$ 。

1.7 $P(A) = C_{10}^1 C_2^1 A_8^8 / A_{10}^{10} = 2/9$ 。

1.8 参照第 72 页的例 1.54, 所求概率为 $1 - e^{-1}$ 。

1.9 由例 1.14 的结果, $P(A_k) = C_m^k A_{Np}^k A_{Nq}^{m-k} / A_N^m$, 其中 $p = n/N, q = 1 - p$ 。

1.10 令 B 表示“第 k 次取出的是白球”。(1) $P(B) = w(w+b-1)/(w+b)! = w/(w+b)$;
(2) $P(B) = C_{w+b-1}^{w-1} / C_{w+b}^w = w/(w+b)$ 。

1.11 盒子装有 m 个可分辨的黑球和 n 个可分辨的白球, 下面两种取法等价: (1) 先从盒子里随机取出 k 个球, 再从剩下的白球中随机取出 r 个。(2) 先从白球中随机取出 r 个, 再从剩下的 $m+n-r$ 个黑球和白球中随机取出 k 个。

1.12 见第 781 页的例 G.5。

1.13 将 $(1+x+x^2+\cdots+x^k)^n$ 展开, 得到

$$\sum_{j=0}^n \binom{n}{j} (x+x^2+\cdots+x^k)^j = \sum_{j=0}^n \binom{n}{j} x^j \sum_{i=0}^{(k-1)j} \binom{j}{i}_k x^i = \sum_{j=0}^n \sum_{i=0}^{(k-1)j} \binom{n}{j} \binom{j}{i}_k x^{i+j}$$

令 $m = i + j$, 其变化范围是 $0, 1, \dots, kn$, 相应 i 的变化范围是 $0, \dots, \lfloor \frac{k-1}{k}m \rfloor$ 。于是, 上式右边是 $\sum_{m=0}^{kn} \sum_{i=0}^{\lfloor \frac{k-1}{k}m \rfloor} \binom{n}{m-i} \binom{m-i}{i}_k x^m$, 得证。

1.14 1/4。提示: 设其中两个边长为 x 和 y , 则第三个边长为 $1-x-y$, 由三角形三边的关系得到 x, y 的取值范围。

1.15 用 x 和 y 分别表示随机抽取的两个数, 则 “ $x+y \leq 6/5$ ” 在区域 $0 < x < 1, 0 < y < 1$ 中所占的比例是 $17/25$ 。

1.16 设至少要用 n 位情报员。令 A_i 表示“第 i 个情报员破译出密码”, 其中 $i = 1, 2, \dots, n$ 。由 $P(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - P(A_1^c A_2^c \dots A_n^c) = 1 - 0.4^n \geq 0.95$ 解得 $n \geq 3.27$, 故至少要使用 4 位情报员。

1.17 在每次抽取中, 某个球未被抽中的概率是 $1 - 1/n$ 。由抽取的独立性知, 所求概率为 $(1 - 1/n)^n \rightarrow 1/e \approx 36.8\%$ 。

1.18 (1) $P(ABC) = 0$; (2) $P(A+B+C) = 5/8$; (3) $P(A^c B^c C) = 1/8$; (4) $P(ABC + A^c BC + AB^c C + ABC^c) = 1/8$ 。

1.19 所求概率为 $P_k = C_{2n-k}^{n-k} p^{n+1} q^{n-k} + C_{2n-k}^{n-k} q^{n+1} p^{n-k}$ 。

1.20 假设两场比赛为一轮, 则甲在任意一轮比赛中获得一分和两分的概率分别为 $2\alpha\beta, \alpha^2$ 。令 A_k 表示“甲在第 k 轮获胜”, 则对应的概率为 $P(A_k) = \alpha^2(2\alpha\beta)^{k-1}$, 其中 $k = 1, 2, \dots$ 。所以甲获得奖牌的概率为 $\sum_{k=1}^{\infty} P(A_k) = \sum_{k=1}^{\infty} \alpha^2(2\alpha\beta)^{k-1} = \alpha^2/(1 - 2\alpha\beta)$ 。乙获得奖牌的概率为 $1 - \alpha^2/(1 - 2\alpha\beta) = \beta^2/(1 - 2\alpha\beta)$ 。

1.21 令 A = “第 k 次摸到黑球”, 则 A^c = “第 k 次摸到白球”, 且 $P(A) = 1 - P(A^c) = 1 - \frac{1}{n}(1 - \frac{1}{n})^{k-1}$ 。

1.22 在 n 次试验中, A 至少出现一次的概率为 $1 - [1 - P(A)]^n$, 随着 $n \rightarrow \infty$ 趋于 1。

1.23 $P(A_k) = 1/6 \cdot (5/6)^{k-1}$ 且 $A = \sum_{k=1}^{\infty} A_k$, 于是 $P(A) = 1$ 。

1.24 若 $A_j \in \mathcal{S}, j = 1, 2, \dots$ 两两互斥, 则 $\{(\bigcup_{j=1}^{\infty} A_j)^c, \bigcup_{j=1}^{\infty} A_j\}$ 是 Ω 的一个划分, $\{(\bigcup_{j=1}^{\infty} A_j)^c, A_1, A_2, \dots\}$ 也是。由已知条件, $(\bigcup_{j=1}^{\infty} A_j)^c + \bigcup_{j=1}^{\infty} A_j = \Omega = (\bigcup_{j=1}^{\infty} A_j)^c + A_1 + A_2 + \dots$, 进而 $P[(\bigcup_{j=1}^{\infty} A_j)^c] + P(\bigcup_{j=1}^{\infty} A_j) = P(\Omega) = 1 = P[(\bigcup_{j=1}^{\infty} A_j)^c] + P(A_1) + P(A_2) + \dots$, 于是 $P(\bigcup_{j=1}^{\infty} A_j) = P(A_1) + P(A_2) + \dots$ 。

1.25 大自然对独立试验的结果没有记忆，第 101 次抛出反面的概率还是 $1/2$ 。

1.26 当 $AB = \emptyset$ 时， $P(AB) - P(A)P(B)$ 最小，此时有 $P(AB) - P(A)P(B) = -P(A)P(B) \geq -[P(A) + P(B)]^2/4 = -[P(A + B)]^2/4 \geq -1/4$ 。

1.27 参考例 1.54，剩下的 $n - k$ 顶帽子无一物归原主的概率为 $P_{n-k,0} = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-k} \frac{1}{(n-k)!}$ ，问题所求的概率为 $P_{n,k} = P_{n-k,0}/k!$ 。

1.28 提示：利用 Borel-Cantelli 引理 1.1。

1.29 6/7。提示：令 B_1, B_2 分别表示“ A 至少出现 1 次”和“ A^c 至少出现一次”，计算 $P(B_2|B_1) = P(B_1B_2)/P(B_1)$ 。

1.30 由 $P(B|A) = P(B|A^c)$ 可得 $P(AB)/P(A) = P(A^cB)/P(A^c) = [P(B) - P(AB)]/[1 - P(A)]$ ，左右两端整理后即证得充分性。

1.31 $P(A) = P(B) = P(C) = 1/2, P(AC) = P(AB) = P(BC) = 1/4, P(ABC) = 0$ 。

1.32 令 A_1, A_2, A_3, A_4 分别表示乘火车、轮船、汽车、飞机，令 B = “朋友迟到”。

(1) 利用全概率公式， $P(B) = P(A_i)P(B|A_i) = 0.3 \times (1/4) + 0.2 \times (1/3) + 0.1 \times (1/12) + 0.4 \times 0 = 0.15$ 。(2) 利用 Bayes 公式， $P(A_1|B) = P(A_1)P(B|A_1)/P(B) = 0.3 \times 0.25/0.15 = 0.5$ 。

1.33 令 A 表示“期中考试及格”， B 表示“期末考试及格”，则 $P(B) = P(A)P(B|A) + P(A^c)P(B|A^c) = p^2 + (1-p)p/2 = p/2 + p^2/2$ 。(1) $P(A \cup B) = P(A) + P(B) - P(AB) = p + p/2 + p^2/2 - p^2 = 3p/2 - p^2/2$ ；(2) $P(A|B) = P(AB)/P(B) = 2p/(1+p)$ 。

1.34 满足 $\mu^*(A) = 0$ 的集合 A 及其子集都是 Carathéodory μ^* -可测的。往证 \mathcal{S} 是 Ω 上的一个 σ -域，以及 μ^* 在 (Ω, \mathcal{S}) 上满足可列可加性。

1.35 由性质 1.10 得到非交分解 $\bigcup_{k=1}^n A_k = A_1 + (A_2 - A_1) + (A_3 - A_1 \cup A_2) + \cdots + (A_n - A_1 \cup A_2 \cup \cdots \cup A_{n-1})$ ，由推论 1.3 和 Boole 不等式 (1.23) 可得

$$\begin{aligned} P\left(\bigcup_{k=1}^n A_k\right) &= \sum_{k=1}^n P(A_k) - P(A_1A_2) - P(A_1A_3 \cup A_2A_3) - \cdots \\ &\quad - P(A_1A_n \cup \cdots \cup A_{n-1}A_n) \geq \sum_{k=1}^n P(A_k) - \sum_{1 \leq i < j \leq n} P(A_iA_j) \end{aligned}$$

1.36 令 $A_t^{t+\Delta t}$ 表示“在 $(t, t + \Delta t]$ 内不与其它分子碰撞”，则 $A_{t+\Delta t} = A_t A_t^{t+\Delta t}$ ，进而 $P(A_{t+\Delta t}) = P(A_t)P(A_t^{t+\Delta t}|A_t) = P(A_t)[1 - \lambda \Delta t - o(\Delta t)]$ ，即 $P(A_{t+\Delta t}) - P(A_t) = -P(A_t)[\lambda \Delta t + o(\Delta t)]$ 。将 $P(A_t)$ 作为变量 t 的函数，令 $\Delta t \rightarrow 0$ 可得到微分方程 $dP(A_t)/dt = -\lambda P(A_t)$ ，解此方程即得 $P(A_t) = e^{-\lambda t}$ 。

1.37 令 D = “放回后仍为 MAXIMA”，令 H_1 = “脱落的两字母相同”， H_2 = “脱落的两字母不同”。所求概率为 $P(D) = P(H_1)P(D|H_1) + P(H_2)P(D|H_2) = C_2^1/C_6^2 + \frac{1}{2}(1 - C_2^1/C_6^2) = 17/30$ 。

1.38 提示：由定理 1.7，在概率空间 $(\Omega, \mathcal{S}, P_C)$ 上应用全概率公式。

1.39 令 A = “盒子里原装的是白球”，由无差别原则 $P(A) = 1/2$ 。令 B_1 = “取出的是白球”， B_2 = “盒子里剩下的是白球”， $P(B_1) = P(A)P(B_1|A) + P(A^c)P(B_1|A^c) = 3/4$ ， $P(B_1B_2) = 1/2$ ，故 $P(B_2|B_1) = P(B_1B_2)/P(B_1) = 2/3$ 。

1.40 (1) 令 B_k = “两次故障之间共生产 k 件正品”， A_n = “两次故障之间共生产 n 件产品”，其中 $n = 0, 1, 2, \dots$ 。当 $n < k$ 时，显然 $P(B_k|A_n) = 0$ 。利用全概率公式， $P(B_k) = \sum_{n=k}^{\infty} P(B_k|A_n)P(A_n) = \sum_{n=k}^{\infty} C_n^k p^k (1-p)^{n-k} \lambda^n e^{-\lambda} / n! = (\lambda p)^k e^{-\lambda p} / k!$ (在求解过程中用到了 e^x 在 $x=0$ 处的幂级数展开 $e^x = 1+x+x^2/2!+\dots+x^n/n!+\dots$)。
(2) 当 $m < k$ 时， $P(A_m|B_k) = 0$ ；当 $m \geq k$ 时， $P(A_m|B_k) = P(B_k|A_m)P(A_m)/P(B_k) = (\lambda q)^{m-k} e^{-\lambda q} / (m-k)!$ ，其中 $q = 1 - p$ 。

1.41 令 A_k = “取到盒子 A_k ”， $k = 0, 1, \dots, N$ ，由题意知 $P(A_k) = 1/(N+1)$ ；令 B_n = “连续 n 次有放回的抽取均为黑球”。于是， $P(B_n|A_k) = (k/N)^n$ 。由全概率公式可得 $P(B_n) = \sum_{k=0}^N P(A_k)P(B_n|A_k) \approx 1/(n+1)$ ，这里用到了不等式

$$\int_1^{N+1} (x-1)^n dx \leq \sum_{k=0}^N k^n \leq \int_1^{N+1} x^n dx$$

同理， $P(B_{n+1}) \approx 1/(n+2)$ 。所求概率 $P(B_{n+1}|B_n) \approx (n+1)/(n+2)$ 。

1.42 根据性质 1.10， $\bigcup_{j=1}^{\infty} A_j B_j$ 有如下非交分解。

$$\begin{aligned} P\left(\bigcup_{j=1}^{\infty} A_j B_j\right) &= P(A_1 B_1) + P[(A_1 B_1)^c A_2 B_2] + P[(A_1 B_1)^c (A_2 B_2)^c A_3 B_3] + \dots \\ &\geq P(A_1 B_1) + P(A_1^c A_2 B_2) + P(A_1^c A_2^c A_3 B_3) + \dots \\ &\geq P(A_1)P(B_1) + P(A_1^c A_2)P(B_2) + P(A_1^c A_2^c A_3)P(B_3) + \dots \\ &\geq \alpha P\left(\bigcup_{j=1}^{\infty} A_j\right) \end{aligned}$$

1.43 $P(A_1|BA_2) = \frac{P(A_1 A_2|B)P(B)}{P(BA_2)} = \frac{P(A_1|B)P(A_2|B)P(B)}{P(A_2|B)P(B)} = P(A_1|B)$ ，得证。

1.44 $P(\bigcap_{j=1}^{\infty} A_j)^c = P(\bigcup_{j=1}^{\infty} A_j^c) = \lim_{n \rightarrow \infty} P(\bigcup_{j=1}^n A_j^c) \leq \lim_{n \rightarrow \infty} \sum_{j=1}^n P(A_j^c) = 0$ 。几乎必然事件之交还是几乎必然事件。

2.1 往证 “ \Rightarrow ”：首先 p_j 是 Borel 可测的，这是因为 $p_j^{-1}(-\infty, b] = \{x \in \mathbb{R}^n : x_j \leq b\} \in \mathfrak{B}_n$ 。由性质 2.3 知， $h_j = p_j \circ h$ 仍然是 Borel 可测的。往证 “ \Leftarrow ”： $h^{-1}(-\infty, x] = \bigcap_{j=1}^n \{\omega \in \Omega : h_j(\omega) \leq x_j\} \in \mathcal{S}$ 。

2.2 所求分布列为 $P\{X = k\} = (1-p)^{k-1}p$ ，其中 $k = 1, 2, \dots$ 。

2.3 所求分布函数为 $F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$

2.4 随机变量 X_1, X_2 都具有分布列 $\frac{1}{6}\langle 1 \rangle + \frac{1}{6}\langle 2 \rangle + \dots + \frac{1}{6}\langle 6 \rangle$ 。

2.5 所求分布列为 $X \sim \frac{1}{64}\langle 1 \rangle + \frac{7}{64}\langle 2 \rangle + \frac{19}{64}\langle 3 \rangle + \frac{37}{64}\langle 4 \rangle$ 。

2.6 $P\{X = k\} = A_4^{k-1}C_4^1/A_8^k$ ，其中 $k = 1, 2, 3, 4, 5$ 。

2.7 若 $(n+1)p$ 为整数，则当 $k = (n+1)p$ 或 $k = (n+1)p - 1$ 时， $P\{X = k\}$ 取到最大值；若 $(n+1)p$ 不是整数时，则当 $k = \lfloor (n+1)p \rfloor$ 时， $P\{X = k\}$ 取到最大值。

2.8 $P\{X = 0\} = (1-p)^2 = 1 - P\{X \geq 1\} = 1 - 5/9 = 4/9$ ，从而 $p = 1/3$ 。
 $P\{Y \geq 1\} = 1 - P\{Y = 0\} = 1 - (1 - 1/3)^3 = 19/27$ 。

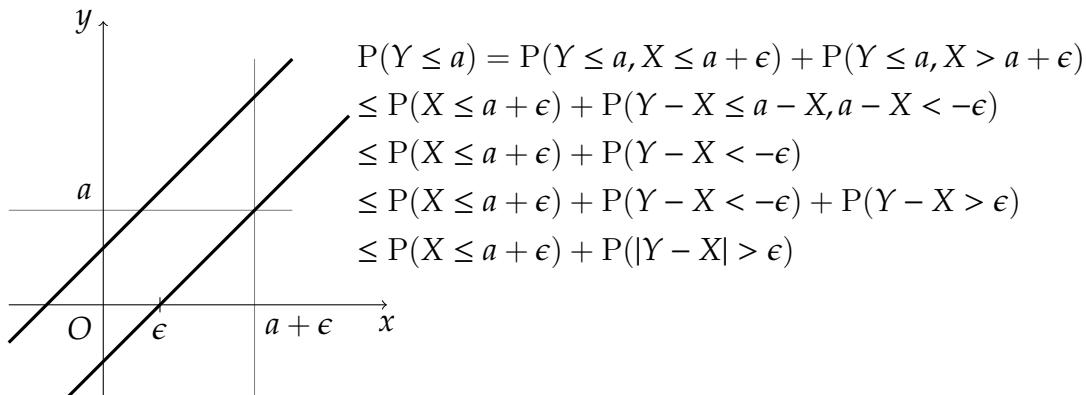
2.9 (1) $p = (1 + e^{-2\lambda})/2$ ；(2) $p = (1 + e^{-6})/2$ 。

2.10 分布函数为 $F(x) = \begin{cases} 0 & \text{当 } x < 0 \\ 1 + (x-1)\sqrt{1-x} & \text{当 } 0 \leq x < 1 \\ 1 & \text{当 } x \geq 1 \end{cases}$

2.11 验证 $G(x)$ 满足定理 2.3 所揭示的分布函数的充要条件。

2.12 均匀分布 $X \sim U[0, 1]$ 。

2.13 因为 $P(Y \leq a) = P(Y \leq a, X \leq a + \epsilon) + P(Y \leq a, X > a + \epsilon)$ ，进而



2.14 (1) 由 $\int_{-\infty}^{+\infty} f_X(x) = 1$ 求得 $a = 1/2$ 。 (2) 当 $x < 0$ 时, $F_X(x) = e^x/2$; 当 $x \geq 0$ 时, $F_X(x) = 1 - e^{-x}/2$ 。 (3) $1 - (e^{-2} + e^{-1})/2$ 。

2.15 由 $\sum_{k=1}^{+\infty} \frac{1}{ck!} = \frac{1}{c}(\sum_{k=0}^{+\infty} \frac{1}{k!} - 1) = \frac{1}{c}(e - 1) = 1$, 得 $c = e - 1$ 。

2.16 随机变量 Y 的概率密度函数为 $f_Y(y) = \begin{cases} |y - 1|/10 & \text{当 } -3 < y < 3 \\ 0 & \text{其它} \end{cases}$

2.17 (1) 当 $y > 0$ 时, $f_Y(y) = \exp\{-(\ln y)^2/2\}/(y\sqrt{2\pi})$; 当 $y \leq 0$ 时, $f_Y(y) = 0$ 。 (2) 当 $z > 0$ 时, $f_Z(z) = 4z \exp\{-z^4/2\}/\sqrt{2\pi}$; 当 $z \leq 0$ 时, $f_Z(z) = 0$ 。

2.18 当 $y \in (0, 1)$ 时, $F_Y(y) = P\{Y \leq y\} = P\{1 - e^{-2X} \leq y\} = P\{X \leq -\frac{1}{2}\ln(1-y)\} = F_X\{-\frac{1}{2}\ln(1-y)\} = y$ 。更一般的结果见定理 4.4。

2.19 由独立性假设, $F_Y(y) = F_{X_1}(y)F_{X_2}(y) \cdots F_{X_n}(y)$, 进而

$$F_Y(y) = \begin{cases} 0 & \text{当 } y \leq 0 \\ (y/a)^n & \text{当 } 0 < y < a \\ 1 & \text{当 } y \geq a \end{cases} \Rightarrow f_Y(y) = \begin{cases} ny^{n-1}/a^n & \text{当 } 0 < y < a \\ 0 & \text{其他} \end{cases}$$

2.20 $X \sim U[-2, 2]$ 。提示: 该方程有实根即事件 $\{X \leq -1\} \cup \{X \geq 2\}$, 为使此事件的概率等于 $1/4$, 唯有 $r > 1$ 。

2.21 Z 的取值范围是 $0, 1, \dots, 2k$, 则

$$\begin{aligned} P(Z = z) &= \sum_{j=0}^z P(X + Y = z, X = j) = \sum_{j=0}^z P(Y = z - j)P(X = j) \\ &= \begin{cases} \sum_{j=0}^z \frac{1}{(k+1)^2} = \frac{z+1}{(k+1)^2} & \text{若 } 0 \leq z \leq k \\ \sum_{j=z-k}^k \frac{1}{(k+1)^2} = \frac{2k-z+1}{(k+1)^2} & \text{若 } k+1 \leq z \leq 2k \end{cases} = \frac{k - |k - z| + 1}{(k+1)^2} \end{aligned}$$

2.22 所求密度函数为 $f_Z(z) = \begin{cases} 1 - z/2 & \text{当 } 0 < z < 2 \\ 0 & \text{其他} \end{cases}$

2.23 参考例 2.35 和例 2.17。

2.24 (4) $Z = \max(X, Y) \sim \frac{1}{10}\langle -1 \rangle + \frac{1}{5}\langle 1 \rangle + \frac{7}{10}\langle 2 \rangle$

2.25 $P\{X > 0, Y < 0\} = 1/3$ 。

2.26 (1) $a = 1/2, b = 1/\pi$; (2) $P\{X \geq 0, Y \geq 0\} = 9/32$ 。

2.27 $f_Z(z) = 12z(4z - z^2 - 2\ln z - 3)$, 其中 $0 < z < 1$ 。

2.28 仿照例 2.41, 求得 Z 的密度函数 $f_Z(z) = \begin{cases} \ln z - \ln 3 & \text{当 } 3 < z < 4 \\ \ln 4 - \ln 3 & \text{当 } 4 < z < 6 \\ 3 \ln 2 - \ln z & \text{当 } 6 < z < 8 \end{cases}$

2.29 令 $F_U(u)$ 是随机变量 U 的分布函数, 则

$$F_X(x) = \int_{-\infty}^{+\infty} F_Y\left(\frac{x}{u}\right) dF_U(u) = \int_0^1 F_Y\left(\frac{x}{u}\right) du$$

2.30 (1) 从 $f_X(x)f_Y(y) = f(x, y)$ 判定 X 与 Y 相互独立。

(2) 计算 $F_Z(z) = P\{Z \leq z\} = P\{X + Y \leq z\}$ 得到

$$F_Z(z) = \iint_{x+y \leq z} f(x, y) dx dy = \begin{cases} 0 & \text{当 } z < 0 \\ z - 1 + e^{-z} & \text{当 } 0 \leq z \leq 1 \\ 1 + (1 - e)e^{-z} & \text{当 } z > 1 \end{cases}$$

(3) $P\{Z > 3\} = 1 - P\{Z \leq 3\} = (e - 1)e^{-3}$ 。

2.31 由例 2.35 的结果, 随机变量 $Z = X/Y$ 的密度函数 $f_Z(z)$ 为

$$f_Z(z) = \frac{\sigma_X \sigma_Y \sqrt{1 - \rho^2}}{\pi(\sigma_Y^2 z^2 - 2\rho \sigma_X \sigma_Y z + \sigma_X^2)}$$

若 X, Y 独立, 则 X/Y 服从 Cauchy 分布 $\text{Cauchy}(0, \sigma_X/\sigma_Y)$ 。

2.32 设 $Z = \sqrt{Y/n}$ 的密度函数为 $f_Z(z)$, 从第 162 页的例 2.45 可知

$$f_Z(z) \propto \begin{cases} e^{-nz^2/2} z^{n-1} & \text{当 } z > 0 \\ 0 & \text{当 } z \leq 0 \end{cases}$$

再根据第 154 页的例 2.35 的结果, 随机变量 T 的密度函数

$$f_T(t) = \int_{-\infty}^{+\infty} \phi(zt) f_Z(z) |z| dz \propto \int_0^{+\infty} e^{-z^2(t^2+n)/2} z^n dz, \quad \text{令 } u = \frac{z^2(t^2+n)}{2}$$

$$\propto \left(\frac{t^2}{n} + 1\right)^{-(n+1)/2} \int_0^{+\infty} u^{(n-1)/2} e^{-u} du \propto \left(\frac{t^2}{n} + 1\right)^{-(n+1)/2}$$

$$\text{所以, } f_T(t) = C_n \left(\frac{t^2}{n} + 1\right)^{-(n+1)/2}, \quad \text{其中归一因子 } C_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)}$$

2.33 从例 2.45 可知随机变量 $U = X/m$ 和 $V = Y/n$ 的密度函数分别为

$$f_U(u) \propto \begin{cases} u^{m/2-1} e^{-mu/2} & \text{当 } u > 0 \\ 0 & \text{当 } u \leq 0 \end{cases} \quad \text{和} \quad f_V(v) \propto \begin{cases} v^{n/2-1} e^{-nv/2} & \text{当 } v > 0 \\ 0 & \text{当 } v \leq 0 \end{cases}$$

类似上一题的做法，随机变量 Z 的密度函数

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_U(uz) f_V(u) |u| du \\ &\propto z^{m/2-1} \int_0^{+\infty} e^{-u(mz+n)/2} u^{(m+n)/2-1} du, \quad \text{令 } w = \frac{u(mz+n)}{2} \\ &\propto \frac{z^{m/2-1}}{(mz+n)^{(m+n)/2}} \int_0^{+\infty} w^{(m+n)/2-1} e^{-w} dw \propto \frac{z^{m/2-1}}{(mz+n)^{(m+n)/2}} \\ \text{所以, } f_Z(z) &= \frac{C_{m,n} z^{m/2-1}}{(mz+n)^{(m+n)/2}}, \quad \text{其中 } z > 0 \text{ 且 } C_{m,n} = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{m/2} n^{n/2} \end{aligned}$$

2.34 (1) 在区域 D 上，随机向量 $(X, Y)^\top$ 的密度函数为 $f(x, y) = 2$ ，求得 X 的边缘密度为 $f_X(x) = \int_0^{1-x} 2 dy = 2(1-x)$ ，而 $f(x, y) \neq f_X(x)f_Y(y)$ ，所以 X 与 Y 不独立。 (2) $F_Z(z) = P(X + Y \leq z) = z^2, 0 \leq z \leq 1$ 。

$$\begin{aligned} 2.35 \quad (1) \quad f_X(x) &= \begin{cases} 2\sqrt{r^2 - x^2}/(\pi r^2) & \text{当 } |x| \leq r \\ 0 & \text{其他} \end{cases} \\ (2) \quad f_{X|Y}(x|y) &= \begin{cases} 1/(2\sqrt{r^2 - y^2}) & \text{当 } |x| \leq \sqrt{r^2 - y^2}, \text{ 其中 } |y| < r \\ 0 & \text{其他} \end{cases} \end{aligned}$$

2.36 由 $0 \leq E(X) \leq 1$ 和 $X^2 \leq X$ 得出 $V(X) = E(X^2) - [E(X)]^2 \leq E(X) - [E(X)]^2 \leq 1/4$ ，等号成立当且仅当 $P(X = 0) = P(X = 1) = 1/2$ 。

2.37 在 n' 次抽取中，令随机变量 X_i 服从 0-1 分布，表示标号 i 是否被抽中。随机变量 $Y = X_1 + X_2 + \dots + X_n$ 表示不同标号的个数。

$$EY = \sum_{i=1}^n EX_i = nEX_1 = nP(X_1 = 1) = n[1 - P(X_1 = 0)] = n\left[1 - \left(1 - \frac{1}{n}\right)^{n'}\right]$$

当 n 很大时，平均有 $n(1 - \exp{-n'/n})$ 个标号被抽中。

$$2.38 \quad P\{X \geq a\} = P\{Y \geq e^{\lambda a}\} = \int_{e^{\lambda a}}^{+\infty} dF_Y(y) \leq \int_{e^{\lambda a}}^{+\infty} ye^{-\lambda a} dF_Y(y) \leq e^{-\lambda a} E(Y)$$

2.39 令 $Y = X - \mu$, 利用结果 (2.13) 和 $\int_0^{+\infty} x \exp(-x^2/2) dx = 1$ 容易得到

$$E(|Y|) = 2 \int_0^{+\infty} y \phi(y|0, \sigma^2) dy = 2 \int_0^{+\infty} \frac{y}{\sigma} \phi\left(\frac{y}{\sigma}\right) dy = \sigma \sqrt{2/\pi}$$

2.40 由 $Z = \max(X, Y)$ 的分布函数 $F(z) = \int_{-\infty}^z \int_{-\infty}^z \phi(x, y|0, 0, 1, 1, \rho) dy dx$ 得到 Z 的密度函数 $f(z) = 2 \int_{-\infty}^z \phi(z, y|0, 0, 1, 1, \rho) dy$ 。参考第 141 页的例 2.25, 利用配方方法进一步将 $f(z)$ 简化为

$$f(z) = 2\phi(z)\Phi(z|\rho z, 1 - \rho^2) = 2\phi(z)\Phi(z|0, (1 + \rho)/(1 - \rho))$$

利用例 2.14 的结果, 不难求出 $E(Z) = \int_{-\infty}^{+\infty} zf(z) dz = \sqrt{(1 - \rho)/\pi}$ 。

2.41 提示: 利用 Jensen 不等式。

2.42 往证 “ \Leftarrow ”: 单点分布 $X \sim \langle 0 \rangle$ 的期望和方差都等于 0, 所以 $E(X^2) = V(X) + [E(X)]^2 = 0$ 。往证 “ \Rightarrow ”: 由 $E(X^2) = V(X) + [E(X)]^2$ 推出 $E(X) = 0, V(X) = 0$ 。对任意 $n \in \mathbb{N}$, 利用 Chebyshev 不等式 $P\{|X - E(X)| \geq 1/n\} \leq n^2 V(X)$ 进而得到 $P\{|X| < 1/n\} = 1$ 。而事件 $\{X = 0\}$ 即事件 $\bigcap_{n=1}^{\infty} \{|X| < 1/n\}$ 。由习题 1.44 的结果, $P(\bigcap_{n=1}^{\infty} \{|X| < 1/n\}) = 1$, 得证。

2.43 利用 Chebyshev 不等式 (2.75),

$$P\{0 < X < 2(m+1)\} = P(|X - (m+1)| < m+1) \geq 1 - \frac{V(X)}{m+1} = \frac{m}{m+1}$$

2.44 仿照第 193 页对 Markov 不等式的证明,

$$\begin{aligned} P(|X| \geq \epsilon) &= \int_{|x| \geq \epsilon} dF(x) \leq \frac{1}{\exp\{\epsilon^2\}} \int_{|x| \geq \epsilon} \exp\{x^2\} dF(x) \\ &\leq \frac{1}{\exp\{\epsilon^2\}} \int_{-\infty}^{+\infty} \exp\{x^2\} dF(x) = \frac{E(\exp\{X^2\})}{\epsilon^2} \end{aligned}$$

2.45 与上一题的证法类似, $\int_{|x| > t} dF(x) \leq \frac{1}{g(t)} \int_{|x| > t} g(x) dF(x)$ 。

2.46 定义随机变量 $X = ZI_{Z < \lambda EZ}$ 和 $Y = ZI_{Z \geq \lambda EZ}$, 其中 $I_{Z < \lambda EZ}$ 和 $I_{Z \geq \lambda EZ}$ 都是指示函数, 显然 $EZ = EX + EY$ 。一方面, $EX \leq \lambda EZ$ 。另一方面, 根据 Cauchy-Schwarz 不等式 (2.68),

$$EY \leq \sqrt{(EZ^2) \cdot (EI_{Z \geq \lambda EZ})} = \sqrt{(EZ^2) \cdot P(Z \geq \lambda EZ)}$$

2.47 若 $x < 0$, 则 $-x = \int_{-\infty}^{+\infty} (t - x) dF_X(t) \leq \int_x^{+\infty} (t - x) dF_X(t)$ 。进而,

$$x^2 \leq \left[\int_x^{+\infty} (t - x) dF_X(t) \right]^2 \leq \int_x^{+\infty} dF_X(t) \int_x^{+\infty} (t - x)^2 dF_X(t) \leq P(X \geq x)(\sigma^2 + x^2)$$

2.48 由独立性得到 $E(Z) = 2E(X) - E(Y) + 3 = 5, V(Z) = 2^2V(X) + V(Y) = 9$, 所以 $Z \sim N(5, 9)$ 。

2.49 (1) 由 $V(Z_1) = V(Z_2) = (a^2 + b^2)\sigma^2, E(Z_1) = E(Z_2) = 0, E(Z_1 Z_2) = E(a^2 X^2 - b^2 Y^2) = (a^2 - b^2)\sigma^2$ 得到 $\rho(Z_1 Z_2) = (a^2 - b^2)/(a^2 + b^2)$ 。(2) 当 $|a| = |b|$ 时, Z_1, Z_2 不相关; 否则 Z_1, Z_2 相关。对于正态分布的两个随机变量, 不相关与独立是等价的 (见例 2.85)。

2.50 令矩阵 A 的特征值为 $\lambda_1, \dots, \lambda_n$, 由 A 的正定性知它们都是整数。参考第 764 页的定义 E.1 对特征值的解释, 存在正交变换把向量 $(x_1 - \mu_1, \dots, x_n - \mu_n)^\top$ 变为 $(z_1, \dots, z_n)^\top$ 使得

$$\sum_{i,j=1}^n a_{ij}(x_i - \mu_i)(x_j - \mu_j) = \sum_{j=1}^n \lambda_j z_j^2$$

因为正交变换的雅可比行列式的绝对值等于 1, 于是

$$\begin{aligned} dx_1 \cdots dx_n &= dz_1 \cdots dz_n \\ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \dots, x_n) dx_1 \cdots dx_n &= \sqrt{\frac{|A|}{\pi^n}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp \left\{ -\sum_{j=1}^n \lambda_j z_j^2 \right\} \\ &= \sqrt{\frac{|A|}{\pi^n}} \prod_{j=1}^n \int_{-\infty}^{+\infty} \exp \left\{ -\lambda_j z_j^2 \right\} dz_j \\ &= \sqrt{\frac{|A|}{\pi^n}} \prod_{j=1}^n \sqrt{\frac{\pi}{\lambda_j}} \\ &= \sqrt{\frac{|A|}{\lambda_1 \cdots \lambda_n}} = 1, \text{ 根据第 765 页的性质 E.1} \end{aligned}$$

2.51 (1) $\rho = 0$; (2) X 与 Y 不独立。

2.52 求得 $(U, V)^\top$ 的联合密度函数

$$f(u, v) = \begin{cases} ue^{-u}/(1+v)^2 & \text{当 } u > 0, v > 0 \\ 0 & \text{其他} \end{cases}$$

再求 U 和 V 的边缘密度，得到

$$f_U(u) = \begin{cases} ue^{-u} & \text{当 } u > 0 \\ 0 & \text{当 } u \leq 0 \end{cases} \quad \text{且} \quad f_V(v) = \begin{cases} 1/(1+v)^2 & \text{当 } v > 0 \\ 0 & \text{当 } v \leq 0 \end{cases}$$

由 $f(u, v) = f_U(u)f_V(v)$ 推得 U 与 V 相互独立。

2.53 计算出 $E(X_n) = 0, V(X_n) = 2$ 。令 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ ，进而得出 $E(\bar{X}) = 0, V(\bar{X}) = 2/n$ ，由 Chebyshev 不等式推得 $1 \geq P(|\bar{X}| < \epsilon) \geq 1 - V(\bar{X})/\epsilon^2 = 1 - 2/(n\epsilon^2)$ 。

2.54 利用 Chebyshev 不等式， $P\{|X - E(X)| \geq \sqrt{2V(X)}\} \leq 1/2$ ，于是

$$E(X) - \sqrt{2V(X)} \leq M(X) \leq E(X) + \sqrt{2V(X)}$$

2.55 因为 $p/q > 1$ ，由 Jensen 不等式 (2.71)， $\{E(|X|^q)\}^{p/q} \leq E(|X|^{q(p/q)}) = E(|X|^p)$ ，即 $\{E(|X|^q)\}^{1/q} \leq \{E(|X|^p)\}^{1/p}$ 。

2.56 提示：利用 Markov 不等式 (2.72) 和 Lyapunov 不等式 (2.23) 可证。

2.57 $\gamma_{Y|X} = 1, \gamma_{X|Y} = 1/2, \rho(X, Y) = \sqrt{1/2}$ 。

2.58 令随机变量 $X_k \sim \frac{1}{n}\langle 1 \rangle + \frac{n-1}{n}\langle 0 \rangle$ 表示 k 是否为不动点， $k = 1, \dots, n$ ，则 $Y = \sum_{k=1}^n X_k$ 并且 $E(X_k) = \frac{1}{n}, V(X_k) = \frac{n-1}{n^2}$ ，另外 $Cov(X_j, X_k) = E(X_j X_k) - E(X_j)E(X_k) = \frac{1}{n^2(n-1)}$ 。于是 $E(Y) = \sum_{k=1}^n E(X_k) = 1$ ，进而有

$$V(Y) = \sum_{i=1}^n V(X_i) + 2 \sum_{1 \leq j < k \leq n} Cov(X_j, X_k) = \frac{n-1}{n} + \frac{2C_n^2}{n^2(n-1)} = 1$$

2.59 $E[\max(X^2, Y^2)] = E[\frac{1}{2}(X^2 + Y^2 + |X^2 - Y^2|)] = \frac{1}{2}[E(X^2) + E(Y^2) + E|X^2 - Y^2|] \leq \frac{1}{2}[V(X) + V(Y) + \sqrt{E(X+Y)^2 E(X-Y)^2}] = 1 + \sqrt{1 - \rho^2}$ 。

2.60 由第 174 页的例 2.61 的结果即得。

2.61 (1) $y = E(Y|X = x) = (1+x)/2$; (2) $x = E(X|Y = y) = y/2$ 。

2.62 因为 $\int_{-\infty}^{+\infty} f_\theta(x)dx = 1$, 所以 $\int_{-\infty}^{+\infty} \frac{\partial f_\theta(x)}{\partial \theta} dx = 0$ 。进而,

$$\begin{aligned} \int_{-\infty}^{+\infty} \frac{\partial f_\theta(x)}{\partial \theta} f_\theta(x) dx &= 0 \Rightarrow \int_{-\infty}^{+\infty} \frac{\partial \ln f_\theta(x)}{\partial \theta} f_\theta(x) dx = 0 \\ &\Rightarrow \int_{-\infty}^{+\infty} \left\{ \frac{\partial^2 \ln f_\theta(x)}{\partial \theta^2} f_\theta(x) + \frac{\partial \ln f_\theta(x)}{\partial \theta} \frac{\partial f_\theta(x)}{\partial \theta} \right\} dx = 0 \\ &\Rightarrow \int_{-\infty}^{+\infty} \left\{ \frac{\partial^2 \ln f_\theta(x)}{\partial \theta^2} + \left[\frac{\partial \ln f_\theta(x)}{\partial \theta} \right]^2 \right\} f_\theta(x) dx = 0 \end{aligned}$$

3.1 $\varphi(s, t) = \frac{1}{6}e^{i(-s-t)} + \frac{1}{6}e^{i(-s+t)} + \frac{1}{2}e^{i(s-t)} + \frac{1}{6}e^{i(s+t)} = \frac{1}{3}\cos(s+t) + \frac{2}{3}\cos(s-t) + \frac{i}{3}\sin(s-t)$

3.2 利用第 226 页的定理 3.3 来判定, 它们都不是特征函数。

3.3 因为 $\lim_{t \rightarrow \infty} \int_{-\infty}^{+\infty} \cos(tx)f(x)dx = 0$ 且 $\lim_{t \rightarrow \infty} \int_{-\infty}^{+\infty} \sin(tx)f(x)dx = 0$

3.4 利用 Markov 不等式, $P(X \geq x) = P(e^{tX} \geq e^{tx}) \leq E(e^{tX})/e^{tx}$ 。

3.5 由例 2.18 可知 Y 的密度函数, 其特征函数为

$$\varphi_Y(t) = \int_0^{+\infty} \frac{1}{\sqrt{2\pi y}} \exp\left\{ity - \frac{y}{2}\right\} dy \xrightarrow{y=z^2} (1-2it)^{-\frac{1}{2}}$$

3.6 如果 X 与 $-X$ 有相同的分布, 则它们有相同的特征函数。由式 (3.8), 有 $\varphi_X(t) = \varphi_{-X}(t) = \varphi_X(-t) = \overline{\varphi_X(t)}$, 故 $\varphi_X(t)$ 是实值函数。反之, 若 $\varphi_X(t)$ 是实值函数, 则 X 与 $-X$ 具有相同的特征函数, 进而由唯一性定理 3.14 知它们具有相同的分布。

3.7 (1) $\frac{1}{2}\langle 1 \rangle + \frac{1}{2}\langle -1 \rangle$; (2) $\frac{1}{4}\langle -2 \rangle + \frac{1}{2}\langle 0 \rangle + \frac{1}{4}\langle 2 \rangle$; (3) $U[-1, 1]$; (4) 随机变量 $Z = X + Y$ 的分布, 其中 $X, Y \stackrel{\text{iid}}{\sim} U[-1, 1]$, 称为区间 $[-2, 2]$ 上的三角分布, 其密度函数为

$$f(z) = \begin{cases} \frac{1}{4}(2+z) & \text{当 } -2 \leq z < 0 \\ \frac{1}{4}(2-z) & \text{当 } 0 \leq z \leq 2 \\ 0 & \text{其他} \end{cases}$$

3.8 利用 $\int_{-\infty}^{+\infty} f(x)dx = 1$ 求出 $c = a/2$, 再利用 $\int_0^{+\infty} \cos(tx)e^{-x}dx = 1/(1+t^2)$ 求出特征函数为 $a^2/(a^2 + t^2)$ 。

3.9 令 $X = \sum_{k=1}^n X_k$, 其中 $X_k = \begin{cases} 1 & \text{第 } k \text{ 次试验 } A \text{ 发生} \\ 0 & \text{第 } k \text{ 次试验 } A \text{ 不发生} \end{cases}$
于是 $\varphi_x(t) = \prod_{k=1}^n (q_k + p_k e^{it})$, 其中 $q_k = 1 - p_k$ 。

3.10 X_1 的特征函数是 $\varphi_X(t) = \sum_{m=0}^k \frac{1}{k+1} \exp(itm)$, 于是 Y 的特征函数是

$$\varphi_Y(t) = [\varphi_X(t)]^n = \frac{1}{(k+1)^n} \sum_{m=0}^{kn} \binom{n}{m}_{k+1} \exp(itm)$$

另外, $E(Y) = nE(X_1) = \frac{n}{k+1} \sum_{j=0}^k j = \frac{kn}{2}$ 。根据式 (3.11),

$$E(Y) = \frac{1}{i} \varphi'_Y(0) = \frac{1}{(k+1)^n} \sum_{m=0}^{kn} \binom{n}{m}_{k+1} m$$

3.11 X 和 Y 的特征函数分别为 $\varphi_X(t) = (pe^{it} + q)^m$ 和 $\varphi_Y(t) = (pe^{it} + q)^n$, 则 $Z = X + Y$ 的特征函数为 $\varphi_Z(t) = \varphi_X(t)\varphi_Y(t) = (pe^{it} + q)^{m+n}$, 即 $Z \sim B(m+n, p)$ 。

3.12 离散型随机变量 X_1 的特征函数为 $p(1 - qe^{it})^{-1}$, 进而 X 的特征函数为 $p^n(1 - qe^{it})^{-n}$, 再求出 X 的分布为 $P(X = k) = C_{n+k-1}^k p^n q^k$, 其中 $k = 0, 1, 2, \dots$, 即负二项分布 $\text{NegB}(n, p)$ 。

3.13 因为 $\varphi(t)$ 为实值的特征函数, 所以 $\varphi(t) = \int_{-\infty}^{+\infty} \cos(tx) dF(x)$ 。于是,

$$\begin{aligned} 1 - \varphi(2t) &= \int_{-\infty}^{+\infty} [1 - \cos(2tx)] dF(x) = 2 \int_{-\infty}^{+\infty} [1 - \cos(tx)][1 + \cos(tx)] dF(x) \\ &\leq 4 \int_{-\infty}^{+\infty} [1 - \cos(tx)] dF(x) = 4[1 - \varphi(t)] \end{aligned}$$

3.14 提示: 只需验证 $X \sim \langle 0 \rangle$ 或 $E(X^2) = 0$ 即可 (参见性质 2.30)。

3.15 提示: 利用第 241 页的推论 3.3。

3.16 该问题就是往证定理 3.3 中 Bochner 准则的必要性部分。

$$\begin{aligned} \text{左端} &= \sum_{k=1}^n \sum_{j=1}^n \left\{ \int_{-\infty}^{+\infty} e^{i(t_k - t_j)x} dF(x) \right\} z_k \bar{z}_j = \int_{-\infty}^{+\infty} \left\{ \sum_{k=1}^n \sum_{j=1}^n e^{i(t_k - t_j)x} z_k \bar{z}_j \right\} dF(x) \\ &= \int_{-\infty}^{+\infty} \left\{ \sum_{k=1}^n e^{it_k x} z_k \right\} \left\{ \sum_{j=1}^n e^{-it_j x} \bar{z}_j \right\} dF(x) = \int_{-\infty}^{+\infty} \left| \sum_{k=1}^n e^{it_k x} z_k \right|^2 dF(x) \geq 0 \end{aligned}$$

4.1 仿照例 4.12 的解法: 连续使用该程序两次, 如果产生 00 或 11, 则重新再来。利用对应关系 $10 \rightarrow 1, 01 \rightarrow 0$ 得到随机数。

4.2 $E(Y) = nE(X_1) = \frac{n}{k+1} \sum_{j=0}^k j = kn/2$, 另外 $V(Y) = nV(X_1) = k(k+2)n/12$ 。

4.3 $E(X) = (n+1)/2, V(X) = (n-1)(n+1)/12, c_s = 0, c_k = -6(n^2+1)/[5(n-1)(n+1)], c_v = \sqrt{(n-1)/[3(n+1)]}$ 。

4.4 利用特征函数的结果 (3.9), 求得 nX_n 的特征函数后进而得到 Y 的特征函数 $\exp\{\sum_{n=1}^{\infty} \lambda_n(e^{int} - 1)\}$, 其中 $\lambda_n = r^n/n$ 。再利用式 (3.13) 即得。

4.5 定义 Y 为 4 个随机数中不超过 a 的个数, 则由 $P(0 < X \leq a) = a$ 知 $Y \sim B(4, a)$ 。从 $P\{Y = 4\} = C_4^4 a^4 (1-a)^0 = 0.1$ 解得 $a \approx 0.5623$ 。

4.6 从左至右, 标记直方图中的小矩形为 R_1, R_2, \dots, R_n , 其面积分别是 a_1, a_2, \dots, a_n 。用 $\text{left}(R_i), \text{right}(R_i)$ 分别表示 R_i 底边区间的左右端点。首先抽取 $Y \sim a_1\langle 1 \rangle + a_2\langle 2 \rangle + \dots + a_n\langle n \rangle$ 的随机数 y^* , 然后抽取 $X \sim U[\text{left}(R_{y^*}), \text{right}(R_{y^*})]$ 的随机数即为所得。

4.7 先产生两点分布 $p\langle 1 \rangle + (1-p)\langle 2 \rangle$ 的随机数 j^* , 再产生 $N(\mu_{j^*}, \sigma_{j^*}^2)$ 的随机数 x^* 。

4.8 特征函数为 $\varphi(t) = \exp(-t^2/2)\Phi(it-r)/[1-\Phi(r)]$, 期望和方差分别为 $E(X) = \phi(r)/[1-\Phi(r)]$ 和 $V(X) = E(X^2) - [E(X)]^2 = 1 - E(X)[E(X) - r]$ 。

4.9 $E(Y) = -(1 + \ln 2)/2, V(Y) = \ln^2 2/4 + \ln 2/2 + 3/4$ 。

4.10 利用分布函数 $1 - \exp\{-\beta Y\} \sim U[0, 1]$ 可得 $h(x) = -\beta^{-1} \ln(1-x)$ 。

4.11 解法一: 利用第 297 页的练习 4.42 和练习 4.43。解法二: $2(X_1 + \dots + X_n)$ 与 χ_{2n}^2 的特征函数都为 $(1 - 2it)^{-n}$ 。

4.12 只需往证 $Y = -\ln X \sim \text{Expon}(1)$, 由上一题的结论便可得证。事实上, 由定理 2.9 知 Y 的密度函数为 $f(y) = \begin{cases} 0 & \text{当 } y \leq 0 \\ \exp(-y) & \text{当 } y > 0 \end{cases}$

4.13 提示: 参考例 3.18。当 n 为奇数时, $E(X^n) = 0, V(X^n) = (2n-1)!!$; 当 n 为偶数时, $E(X^n) = (n-1)!!$, $V(X^n) = (2n-1)!! - [(n-1)!!]^2$ 。

4.14 提示: $\max(X, Y) + \min(X, Y) = X + Y, \max(X, Y) - \min(X, Y) = |X - Y|$ 。根据 $X + Y \sim N(2\mu, 2\sigma^2)$ 及 $X - Y \sim N(0, 2\sigma^2)$ 可求出 $E(X+Y) = 2\mu$ 和 $E(|X-Y|) = 2\sigma/\sqrt{\pi}$, 于是 $E[\max(X, Y)] = \mu + \sigma/\sqrt{\pi}$ 且 $E[\min(X, Y)] = \mu - \sigma/\sqrt{\pi}$ 。

4.15 利用二元正态分布 $N(0, 0, 1, 1, \rho)$ 的密度函数的对称性,

$$E(\max(X, Y)) = \frac{1}{\pi \sqrt{1-\rho^2}} \iint_{x>y} x \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\} dx dy$$

做变量替换 $y - \rho x = u \sqrt{1 - \rho^2}$, 积分改写成

$$\begin{aligned} \frac{1}{\pi} \iint_{u < \frac{1-\rho}{\sqrt{1-\rho^2}}x} x \exp\left\{-\frac{x^2 + u^2}{2}\right\} dx du &= \frac{1}{\pi} \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} \left[\int_{\frac{\sqrt{1-\rho^2}}{1-\rho}u}^{+\infty} x e^{-\frac{x^2}{2}} dx \right] du \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} e^{-\frac{u^2}{1-\rho}} du = \sqrt{\frac{1-\rho}{\pi}} \end{aligned}$$

4.16 提示: $\exp[-(x^2 + y^2)/2] \sin x \sin y$ 关于 x, y 都是奇函数。

4.17 $E(|X|) = 1, V(|X|) = 1, \text{Cov}(X, |X|) = 0$ 。 X 与 $|X|$ 不独立。

4.18 由第 171 页的例 2.59, $X \sim \text{Cauchy}(0, 1)$ 的分布函数为 $F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$ 。
按照定义, 随机变量 Y 的分布函数为

$$\begin{aligned} F_Y(y) &= P\left(\frac{2X}{1-X^2} \leq y\right), \text{ 无论 } y \text{ 的正负, 皆可整理为} \\ &= \frac{1}{\pi} \arctan \frac{\sqrt{y^2+1}-1}{y} - \frac{1}{\pi} \arctan \frac{\sqrt{y^2+1}+1}{y} + \text{某常数} \\ \text{于是, } F'_Y(y) &= \frac{1}{y^2+1}, \text{ 即 } Y = \frac{2X}{1-X^2} \sim \text{Cauchy}(0, 1) \end{aligned}$$

4.19 $P(Y = y) = \beta^\alpha (\beta + 1)^{-\alpha} \Gamma(y + \alpha) [\Gamma(\alpha)y!(\beta + 1)^y]^{-1}, y = 0, 1, 2, \dots$

4.20 利用例 2.35 的结果可得 $f_Z(z) = \begin{cases} \frac{\Gamma(\alpha_1 + \alpha_2)z^{\alpha_1-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)(z+1)^{\alpha_1+\alpha_2}} & \text{当 } z > 0 \\ 0 & \text{当 } z \leq 0 \end{cases}$

4.21 (1) 分布函数 $F_U(u) = P\{\max(X_1, \dots, X_n) \leq u\} = P(X_1 \leq u, X_2 \leq u, \dots, X_n \leq u) = P(X_1 \leq u) \cdots P(X_n \leq u) = \begin{cases} 0 & \text{当 } u \leq 0 \\ (1 - e^{-\beta u})^n & \text{当 } u > 0 \end{cases}$
故 U 的密度函数为 $f_U(u) = F'_U(u) = \begin{cases} 0 & \text{当 } u \leq 0 \\ n\beta(1 - e^{-\beta u})^{n-1}e^{-\beta u} & \text{当 } u > 0 \end{cases}$
(2) $F_V(v) = P(\min(X_1, \dots, X_n) \leq v) = 1 - P(\min(X_1, \dots, X_n) > v) = 1 - P(X_1 > v) \cdots P(X_n > v) = 1 - [1 - F_{X_1}(v)] \cdots [1 - F_{X_n}(v)] = \begin{cases} 0 & \text{当 } v < 0 \\ 1 - e^{-n\beta v} & \text{当 } v \geq 0 \end{cases}$, 进而
 $f_V(v) = F'_V(v) = \begin{cases} n\beta e^{-n\beta v} & \text{当 } v > 0 \\ 0 & \text{当 } v \leq 0 \end{cases}$

4.22 由 $P\{X \geq 1\} = P\{X \leq 1\}$ 计算出 $\beta = \ln 2$, 所以 $P\{X \geq k\} = (1/2)^k$, 进而 $\sum_{k=1}^{\infty} P\{X \geq k\} = 1$ 。

4.23 分布 $\text{Pareto}(\alpha, \mu)$ 的随机数 $x^* = \mu u^{*-1/\alpha}$, 其中 u^* 是均匀分布 $U(0, 1)$ 的随机数。

4.24 利用定义 4.24。

4.25 $E(1/X) = \sqrt{2\pi}/2\sigma$ 。

4.26 由性质 4.39, 随机向量 $(X_1, \dots, X_{m_2})^\top$ 服从多项分布。设计 Bernoulli 试验: 要么事件 $\{X_1 + \dots + X_{m_1} = 1\}$ 发生, 要么其补事件 $\{X_{m_1+1} + \dots + X_{m_2} = 1\}$ 发生, 二者概率之比是 $(p_1 + \dots + p_{m_1}) : (p_{m_1+1} + \dots + p_{m_2})$ 。于是,

$$P(X_1 + \dots + X_{m_1} = 1) = \frac{p_1 + \dots + p_{m_1}}{p_1 + \dots + p_{m_2}} = p$$

观察到 $Y_2 = X_1 + \dots + X_{m_2} = y_2$ 意味着独立进行了 y_2 次这样的 Bernoulli 试验, 因此随机变量 $Y_1 = X_1 + \dots + X_{m_1}$ 服从二项分布 $B(y_2, p)$ 。

4.27 由性质 4.39, $(X_1, \dots, X_m, X'_{m+1})^\top \sim \text{Multin}(n; p_1, \dots, p_m, 1 - p_1 - \dots - p_m)$, 其中 $X'_{m+1} = n - X_1 - \dots - X_m$ 。再由性质 4.40,

$$(X_m, X'_{m+1})^\top | Y_m = y_m \sim \text{Multin}\left(n - y_m; \frac{p_m}{1 - p_1 - \dots - p_{m-1}}, \frac{1 - p_1 - \dots - p_m}{1 - p_1 - \dots - p_{m-1}}\right)$$

即, $X_m | Y_m = y_m \sim B(n - y_m, p_m / (1 - p_1 - \dots - p_{m-1}))$ 得证。

4.28 利用推论 4.4 和 Beta 分布的数字特征可证。

4.29 仿照性质 4.44 的证明, 并利用推论 4.5 可证。

4.30 $P(X_1 = x_1, \dots, X_n = x_n | \boldsymbol{p}) = \prod_{j=1}^k p_j^{n^{(j)}}$, 其中 $n^{(j)}$ 是 x_1, \dots, x_n 中 j 的个数。仿照第 330 页的性质 4.43, 我们得到欲求的概率函数如下:

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n | \boldsymbol{\alpha}) &= \int_{\Delta_{k-1}} P(X_1 = x_1, \dots, X_n = x_n | \boldsymbol{p}) \pi(\boldsymbol{p} | \boldsymbol{\alpha}) d\boldsymbol{p} \\ &= \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1 + \dots + \alpha_k + n)} \prod_{j=1}^k \frac{\Gamma(\alpha_j + n^{(j)})}{\Gamma(\alpha_j)} \\ &= \frac{B(\alpha_1 + n^{(1)}, \dots, \alpha_k + n^{(k)})}{B(\alpha_1, \dots, \alpha_k)} \end{aligned}$$

4.31 参考例 2.85 和例 2.44, $(W, V)^\top$ 服从正态分布且 $\text{Cov}(W, V) = 0$ 。

4.32 $X + Y \sim N(8, 18)$ 。

4.33 X_1, X_2, \dots, X_n 相互独立当且仅当 \mathbf{X} 的特征函数为

$$\varphi(\mathbf{t}) = \exp \left\{ i \sum_{j=1}^n t_j \mu_j - \frac{1}{2} \sum_{j=1}^n \sigma_j^2 t_j^2 \right\} = \varphi(t_1) \varphi(t_2) \cdots \varphi(t_n)$$

4.34

4.35 $E(W) = \sum_{j=1}^n E(\mathbf{X}_j \mathbf{X}_j^\top) = n\Sigma$, 因为 $E(\mathbf{X}_j \mathbf{X}_j^\top) = \Sigma$ 。

5.1 $E(X_k) = 0, V(X_k) = 2, k = 1, 2, \dots$, 由 Chebyshev 弱大数律可证。

5.2 算得 $E(X_k) = 0, V(X_k) = k^{2s}, k = 1, 2, \dots, n, \dots$ 并且 $\{X_k\}$ 相互独立, 有 $\frac{1}{n^2} V(\sum_{k=1}^n X_k) = \frac{1}{n^2} \sum_{k=1}^n V(X_k) = \frac{1}{n^2} \sum_{k=1}^n k^{2s} \leq n \cdot n^{2s}/n^2 = n^{2s-1}$, 所以当 $s < 1/2$ 时, 由 Markov 弱大数律可知 $\{X_k\}$ 满足弱大数律。

5.3 提示: 参看例 1.54。令 $X_j = \begin{cases} 1 & \text{第 } j \text{ 号球放入第 } j \text{ 号盒中} \\ 0 & \text{第 } j \text{ 号球未放入第 } j \text{ 号盒中} \end{cases}$

其中 $j = 1, 2, \dots, n$, 则 $S_n = \sum_{j=1}^n X_j$ 。由 $P\{X_j = 1\} = 1/n$ 求得 $V(X_j) = (n-1)/n^2, \text{Cov}(X_j, X_k) = 1/[n^2(n-1)]$, 利用式 (2.88) 算出 $V(S_n) = 1$ 。由 Markov 弱大数律可证。

5.4 提示: 利用式 (2.88) 往证 $\lim_{n \rightarrow \infty} \frac{1}{n^2} V(\sum_{i=1}^n X_i) = 0$, 再利用 Markov 弱大数律即可证得。这个结果被称为 Bernstein 定理。

$$\frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) + \frac{2}{n^2} \sum_{1 \leq j < k \leq n} \rho_{jk} \sqrt{V(X_j)} \sqrt{V(X_k)} \leq \frac{c}{n} + \frac{2c}{n^2} \sum_{1 \leq j < k \leq n} |\rho_{jk}|$$

因为当 $|k-j| \rightarrow \infty$ 时, $\rho_{kj} \rightarrow 0$, 故 $\forall \epsilon > 0$ 存在 $N > 0$ 使得当 $|k-j| > N$ 时, $|\rho_{jk}| < \epsilon/c$ 。对每一个暂时固定的 j , 满足条件 $k-j \leq N$ 的 ρ_{kj} 至多有 N 个, 从而满足 $0 < k-j \leq N$ 的 ρ_{kj} 至多有 Nn 个。同理, 对每一个暂时固定的 j , 满足 $k > j$ 的 ρ_{kj} 至多为 $n-j$ 个, 从而满足 $k-j > N$ 的 ρ_{kj} 至多为 $(n-1) + (n-2) + \cdots + 2 + 1 = n(n-1)/2$ 个。利用 $|\rho_{kj}| \leq 1$,

$$\begin{aligned} \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) &\leq \frac{c}{n} + \frac{2c}{n^2} \left(\sum_{0 < k-j \leq N} |\rho_{kj}| + \sum_{k-j > N} |\rho_{kj}| \right) \\ &\leq \frac{c}{n} + \frac{2c}{n^2} \left[Nn + \frac{n(n-1)}{2} \cdot \frac{\epsilon}{c} \right] = \frac{(2N+1)c}{n} + \left(1 - \frac{1}{n}\right)\epsilon \end{aligned}$$

由于 N 由 ϵ 确定, 故当 $n \rightarrow \infty$ 时, 有 $\lim_{n \rightarrow \infty} \frac{1}{n^2} V(\sum_{i=1}^n X_i) \leq \epsilon$ 。由 ϵ 的任意性即知随机变量序列 $\{X_i\}_{i=1}^\infty$ 满足 Markov 大数律条件。

5.5 因为 $\sum_{k=1}^{\infty} V(X_k)/k^2 < \infty$, 所以 $\forall \epsilon > 0$, 存在 N_1 使得 $m > N_1, n > m > N_1$ 时, 有 $\sum_{k=m+1}^n V(X_k)/k^2 < \epsilon/2$ 。又因为 $V(X_k)$ 有限, 存在 N_2 使得 $n > N_2$ 时, 有 $\frac{1}{n^2} \sum_{k=1}^m V(X_k) < \epsilon/2$ 。取 $N = \max(N_1, N_2)$, 当 $n > N$ 时, 有 $\frac{1}{n^2} V(\sum_{k=1}^n X_k) = \frac{1}{n^2} \sum_{k=1}^n V(X_k) \leq \frac{1}{n^2} \sum_{k=1}^m V(X_k) + \sum_{k=m+1}^n V(X_k)/k^2 < \epsilon/2 + \epsilon/2 = \epsilon$, 再利用 Markov 弱大数律可证。

5.6 提示: $\ln Y_n = \frac{1}{n} \sum_{j=1}^n \ln X_j$, 因为 $\ln X_i$ 也独立同分布, 可求得 $E(\ln X_j) = -1$ 。利用 Khinchin 弱大数律可证, 并得到 $c = e^{-1}$ 。

5.7 (1) 往证 “ \Rightarrow ”: 设 X_n 的分布函数为 $F_n(x)$, 则有

$$\begin{aligned} E[h(|X_n|)] &= \int_{|x|>\delta} h(|x|)dF_n(x) + \int_{|x|\leq\delta} h(|x|)dF_n(x) \\ &\leq \sup_{x \geq 0} h(x) \int_{|x|>\delta} dF_n(x) + h(\delta) \int_{|x|\leq\delta} dF_n(x) \leq cP(|X_n| > \delta) + h(\delta) \end{aligned}$$

对 $\forall \epsilon > 0, \exists \delta > 0$ 使 $h(\delta) < \epsilon/2$ 。对上述 ϵ , 存在 $N \in \mathbb{N}$ 使当 $n > N$ 时有 $P(|X_n| > \delta) < \epsilon/(2c)$, 于是 $E[h(|X_n|)] < \epsilon$ 。

(2) 往证 “ \Leftarrow ”: 说明 $E[h(|X_n|)] \geq h(\delta)P(|X_n| > \delta)$, 从而当 $n \rightarrow \infty$ 时 $P(|X_n| > \delta) \xrightarrow{P} 0$, 即 $X_n \xrightarrow{P} 0$ 。

5.8 提示: 令 $Y_n = \frac{2}{n(n+1)} \sum_{k=1}^n kX_k$, 则 $E(Y_n) = \mu, V(Y_n) \leq 4\sigma^2/(n+1)$ 。于是 $\forall \epsilon > 0$, 当 $n \rightarrow \infty$ 时有 $P\{|Y_n - \mu| \leq \epsilon\} \geq 1 - V(Y_n)/\epsilon^2 \rightarrow 1$ 。

5.9 满足中心极限定理。提示: 仿照例 5.21。

5.10 提示: 先证明 $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{L} N(0, 1)$, 且 $\frac{1}{n} \sum_{i=1}^n X_i$ 和 $\sqrt{\frac{1}{n} \sum_{i=0}^n X_i^2}$ 依概率收敛于 1, 然后利用定理 5.3。

5.11 由 Lindeberg-Lévy 中心极限定理,

$$P\left\{\sum_{i=1}^{100} X_i \geq 90\right\} \approx 1 - \Phi(-0.65) = \Phi(0.65) \approx 0.7422$$

5.12 令随机变量 X 表示事件 A 出现的次数, 利用定理 5.23,

$$P\{50 \leq X \leq 150\} \approx 2\Phi(10/\sqrt{3}) - 1 \approx 1$$

5.13 优等品个数 $X \sim B(100, 0.2)$, 利用 de Moivre-Laplace 中心极限定理 5.16, $P\{18 < X \leq 25\} \approx \Phi(1.25) + \Phi(0.5) - 1 \approx 0.5858127$ 。

5.14 所有的取整误差 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \stackrel{\text{iid}}{\sim} U(-0.5, 0.5)$, 由推论 5.2, 当 n 很大时近似地有 $\sum_{j=1}^n \varepsilon_j \sim N(0, n/12)$ 。于是,

$$P\{|\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_n| \leq 15\} \approx \Phi(15|0, n/12) - \Phi(-15|0, n/12) = 2\Phi\left(\frac{15}{\sqrt{n/12}}\right) - 1$$

将 $n = 1200$ 代入上式, 所求概率约为 $2\Phi(1.5) - 1 \approx 0.8663856$ 。

5.15 由题意知随机变量序列 $F(X_1), F(X_2), \dots, F(X_n), \dots \stackrel{\text{iid}}{\sim} U[0, 1]$, 由 Lindeberg-Lévy 中心极限定理 5.17 知

$$\lim_{n \rightarrow \infty} P\left\{ \frac{\sum_{i=1}^n F(X_i) - E[\sum_{i=1}^n F(X_i)]}{\sqrt{V[\sum_{i=1}^n F(X_i)]}} \leq x \right\} = \lim_{n \rightarrow \infty} P\left\{ \frac{\sum_{i=1}^n F(X_i) - n/2}{\sqrt{n/12}} \leq x \right\} = \Phi(x)$$

取 $x = 0$ 即可证得。

5.16 (1) 随机变量序列 $\{Y_n\}$ 独立同分布, 再利用大数律即可证。(2) $N(2\lambda^{-2}, 20n^{-1}\lambda^{-4})$ 。

5.17 $Y_n \sim N\left(m_2, \frac{m_4 - m_2^2}{n}\right)$ 。

5.18 $Y_j = X_{2j} - X_{2j-1}, j = 1, 2, \dots$ 独立同分布, 由中心极限定理求得 $c = 1/\sqrt{2}$ 。

5.19 令 $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} Poisson(1)$, 则 $E(X_n) = V(X_n) = 1$ 。由 Lindeberg-Lévy 中心极限定理, $P\{\sum_{k=1}^n X_k \leq n\} = P\{\sum_{k=1}^n (X_k - 1)/\sqrt{n} \leq 0\} \rightarrow \Phi(0) = 1/2$ 。并且, $\sum_{k=1}^n X_k \sim Poisson(n)$ 。于是, $P\{\sum_{k=1}^n X_k \leq n\} = e^{-n} \sum_{k=0}^n n^k/k! \rightarrow 1/2$ 得证。

6.1 Z_n 独立同分布于 0-1 分布, 参数是 $P(Z_n = 1) = 1 - P(Z_n = 0, Y_n = 0) = 1 - (1-p)(1-q) = p + q - pq$ 。Bernoulli 过程 U_n, V_n 的参数分别为 $pq, p(1-q)$ 。 $P(U_n = 1|V_n = 1) = 0 \neq pq = P(U_n = 1)$ 说明 U_n, V_n 并不独立。

6.2 即时最大值 M_t 和布朗运动 B_t 的联合密度函数是

$$f_{M_t, B_t}(m, x) = \frac{2(2m-x)}{t \sqrt{2\pi t}} \exp\left\{-\frac{(2m-x)^2}{2t}\right\}$$

于是, $f_{M_t}(m) = \int_{-\infty}^m f_{M_t, B_t}(m, x) dx = \sqrt{\frac{2}{\pi t}} \exp\left\{-\frac{m^2}{2t}\right\}$

$$E(M_t) = \int_0^{+\infty} m f_{M_t}(m) dm = \sqrt{\frac{2t}{\pi}}$$

6.3 空间对称性是显然的。

□ **时间平移性:** 显然, $W(0) = 0$ 。不妨设 $s < t$, 则 $W(t) - W(s) = B(t+h) - B(s+h) \sim N(0, t-s)$ 。不妨设 $[s, t] \cap [u, v] = \emptyset$, 则 $[s+h, t+h] \cap [u+h, v+h] = \emptyset$, 因此 $W(t) - W(s)$ 与 $W(v) - W(u)$ 是独立的。

□ **尺度变换性:** 不妨设 $s < t$, 则 $W_t - W_s$ 服从正态分布, 均值为零并且 $V(W_t - W_s) = V[(B_{\alpha t} - B_{\alpha s})/\sqrt{\alpha}] = (\alpha t - \alpha s)/\alpha = t - s$ 。

6.4

6.5 假设所有状态都是非常返的, 由**性质 6.7** 推出矛盾。

6.6

6.7 $f_{ii} = f_{ii}f_{ij}f_{ji} = 1 \Rightarrow f_{ij} = f_{ji} = 1$ 。

6.8 令 A 表示“迟早到达 j ”, 则

$$\begin{aligned} f_{ij} &= P(A|X_0 = i) = \sum_k P(A|X_0 = i, X_1 = k)P(X_1 = k|X_0 = i) \\ &= \sum_{k \in T} p_{ik}f_{kj} + \sum_{k \in K} p_{ik}f_{kj} + \sum_{k \notin T \cup K} p_{ik}f_{kj} = \sum_{k \in T} p_{ik}f_{kj} + \sum_{k \in K} p_{ik} \end{aligned}$$

其中, 最后一步是因为上一个习题和**性质 6.8**。

6.9 $0.4\langle \text{雨天} \rangle + 0.2\langle \text{晴天} \rangle + 0.4\langle \text{雪天} \rangle$ 。

6.10 解方程组 $x_j = \sum_i^n x_i p_{ij}$ 求得 $x_j = 1/n, j = 1, 2, \dots, n$ 。

6.11 将 $s = 1$ 代入 $H'_\infty(s) = G(H_\infty(s)) + sG(H_\infty(s))H'_\infty(s)$, 即得 $E(Z_\infty) = 1 + \mu E(Z_\infty)$ 。类似地, 将 $s = 1$ 代入

$$H''_\infty(s) = 2G'(H_\infty(s))H'_\infty(s) + sG''(H_\infty(s))[H'_\infty(s)]^2 + sG'(H_\infty(s))H''_\infty(s)$$

$$\text{进而, } H''_\infty(1) = \frac{\mu}{(1-\mu)^2} + \frac{\sigma^2}{(1-\mu)^3}$$

6.12 $EY = \frac{1}{n\alpha} + \dots + \frac{1}{\alpha}$

7.1 提示: 右边 $= \sum_{i=1}^n X_i^2 - n\bar{X}^2$, 利用 (7.3) 即得。

方法	优点	缺点
直方图	简单直观, 利于了解密度函数的形状。	需要划分区间, 这一人为因素有时严重影响结果。
ECDF	忠实地记录了观察样本, 数学定义严格。	略欠直观, 不易看出密度函数的形状。

7.3 提示: $\sum_{j=1}^n (X_j - c)^2 = \sum_{j=1}^n (X_j - \bar{X} + \bar{X} - c)^2 = \sum_{j=1}^n (X_j - \bar{X})^2 + n(\bar{X} - c)^2$, 类似于式 (2.57)。

7.4 $\bar{Y} = (\bar{X} - a)/b$ 且 $S_Y^2 = S_X^2/b^2$ 。

7.5 设 $E(X) = \mu, V(X) = \sigma^2$, 计算得 $V(X_i - \bar{X}) = \sigma^2(n-1)/n, E[(X_i - \bar{X})(X_j - \bar{X})] = -\sigma^2/n$, 因而 $\rho = \text{Cov}(X_i - \bar{X}, X_j - \bar{X})/V(X_i - \bar{X}) = E[(X_i - \bar{X})(X_j - \bar{X})]/V(X_i - \bar{X}) = -(n-1)^{-1}$ 。

7.6 利用性质 7.7: (1) $E(\bar{X}) = p$ 并且 $V(\bar{X}) = p(1-p)/n$ 。 (2) $E(S^2) = p(1-p)$ 。
(3) 当 $x < 0$ 时, $F_n^*(x) = 0$; 当 $0 \leq x < 1$ 时, $F_n^*(x) = 1 - m/n$; 当 $x \geq 1$ 时, $F_n^*(x) = 1$ 。

7.7 由性质 7.9 知 $9S^2/4^2 \sim \chi_9^2$, $P\{S^2 > a\} = P\{9S^2/4^2 > 9a/4^2\} = 0.1$, 所以 $9a/4^2 \approx \chi_9^2(0.9) \approx 14.684$, 进而 $a \approx 26.105$ 。利用 R 语言中的函数 qchisq(p,df) 可求得 χ_n^2 分布的 p -分位数。

7.8 参考例 7.14, $X_{(1)}$ 的分布函数为 $F_{X_{(1)}}(x) = \begin{cases} 1 - e^{-n\lambda x} & \text{当 } x \geq 0 \\ 0 & \text{当 } x < 0 \end{cases}$
于是, $E(X_{(1)}) = 1/(n\lambda), V(X_{(1)}) = 1/(n\lambda)^2$ 。

7.9 因为 $\bar{X}_1, \bar{X}_2 \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2/n)$, 所以 $\bar{X}_1 - \bar{X}_2 \sim N(0, 2\sigma^2/n)$ 。从 $P\{|\bar{X}_1 - \bar{X}_2| > \sigma\} = P\{|(\bar{X}_1 - \bar{X}_2)/\sqrt{2\sigma^2/n}| > \sigma/\sqrt{2\sigma^2/n}\} = 2[1 - \Phi(\sqrt{n/2})] = 0.01$ 得 $n = 14$ 。

7.10 提示: 由 $\text{Cov}(X_1 + X_2, X_1 - X_2) = 0$ 和例 2.85 先证明 $X_1 + X_2 \sim N(0, 2\sigma^2)$ 与 $X_1 - X_2 \sim N(0, 2\sigma^2)$ 相互独立, 再说明 $(X_1 + X_2)^2/(X_1 - X_2)^2 \sim F(1, 1)$, 故 $P\{(X_1 + X_2)^2/(X_1 - X_2)^2 < 4\} = \int_0^4 1/[\pi(1+y)y^{1/2}]dy = 2\arctan(2)/\pi \approx 0.70$ 。

7.11 提示: $\frac{Y_1 - Y_2}{\sigma/\sqrt{2}} \sim N(0, 1)$ 且 $2S^2/\sigma^2 \sim \chi_2^2$ 。

7.12 由 $E(\bar{X} - \mu)^2 = V(\bar{X}) = \frac{1}{n}V(X) = \frac{4}{n} \leq 0.1$ 得 $n \geq 40$ 。

7.13 (1) 由 Fisher 定理 3.16, $\bar{X} = (X_1 + X_2)/2$ 与 $S^2 = (X_1 - X_2)^2/2$ 相互独立, 于是 $(X_1 + X_2)^2$ 与 $(X_1 - X_2)^2$ 相互独立。(2) 因为 $X_1 + X_2, X_1 - X_2 \stackrel{\text{iid}}{\sim} N(0, 2\sigma^2)$, 所以 $\frac{(X_1+X_2)^2}{2\sigma^2}, \frac{(X_1-X_2)^2}{2\sigma^2} \stackrel{\text{iid}}{\sim} \chi_1^2$, 进而 $Y = \frac{(X_1+X_2)^2}{(X_1-X_2)^2} \sim F_{1,1}$ 。

7.14 由 $\frac{1}{\sqrt{n}}(X_1 + \dots + X_n) \sim N(0, 1)$ 和 $Y_1^2 + \dots + Y_n^2 \sim \chi_n^2$ 得到 $W \sim t_n$ 。

7.15 由 $V(X_1 - 2X_2) = 20, [(X_1 - 2X_2)/\sqrt{20}]^2 \sim \chi_1^2$ 得 $a = 1/20$ 。由 $V(3X_3 - 4X_4) = 100, [(3X_3 - 4X_4)/10]^2 \sim \chi_1^2$ 得 $b = 1/100$ 。

7.16 因 $\frac{1}{\sqrt{2}\sigma}(X_1 + X_2) \sim N(0, 1)$ 与 $\frac{1}{\sigma^2}(X_3^2 + X_4^2 + X_5^2) \sim \chi_3^2$ 独立, 故 $Y \sim t_3$, 得 $a = \sqrt{3/2}$ 。

7.17 由 $\sqrt{2}(\bar{X} - \bar{Y}) \sim N(0, 1)$, $P\{|\bar{X} - \bar{Y}| > 0.3\} = P\{\sqrt{2}|\bar{X} - \bar{Y}| > 0.3\sqrt{2}\} = 2[1 - \Phi(0.3\sqrt{2})] \approx 0.6714$ 。

7.18 $Y_1 \sim F(1, 1), Y_2 \sim F(2, 1), Y_3 \sim t_1$ 。

7.19 $Y_j \sim N((1+a)\mu, (n+2a+a^2)\sigma^2/n)$ 。

7.20 利用例 7.14 的结果, $X_{(k)}$ 的密度函数 $f_k(x)$ 为

$$f_k(x) = \begin{cases} 2kC_n^k x^{2k-1}(1-x^2)^{n-k} & \text{当 } x \in [0, 1] \\ 0 & \text{其他} \end{cases}$$

$$f_1(x) = \begin{cases} 2nx(1-x^2)^{n-1} & \text{当 } x \in [0, 1] \\ 0 & \text{其他} \end{cases} \quad f_n(x) = \begin{cases} 2nx^{2n-1} & \text{当 } x \in [0, 1] \\ 0 & \text{其他} \end{cases}$$

7.21 仿照例 7.18 的证法, 或利用性质 7.12。

7.22 二元正态分布的密度函数 $\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 见式 (2.22)。因为 $\sum_{j=1}^n (x_j - \mu_X)^2 = \sum_{j=1}^n (x_j - \bar{x})^2 + n\bar{x}^2 - 2n\mu_X\bar{x} + n\mu_X^2$ 可以表示为 $\bar{x}, \sum_{j=1}^n (x_j - \bar{x})^2$ 与 μ_X 定义的函数, $\sum_{j=1}^n (y_j - \mu_Y)^2$ 也有类似的结果。 $\sum_{j=1}^n (x_j - \mu_X)(y_j - \mu_Y) = \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y}) + \bar{x}n\bar{y} - n\mu_X\bar{y} - n\mu_Y\bar{x} + n\mu_X\mu_Y$ 可以表示为 $\bar{x}, \bar{y}, \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})$ 与 μ_X, μ_Y 的函数。由 Fisher 因子分解定理 7.7, 统计量 $(\bar{X}, \sum_{j=1}^n (X_j - \bar{X})^2, \bar{Y}, \sum_{j=1}^n (Y_j - \bar{Y})^2, \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}))^\top$ 是充分的。同样地, 统计量 $(\sum_{j=1}^n X_j, \sum_{j=1}^n X_j^2, \sum_{j=1}^n Y_j, \sum_{j=1}^n Y_j^2, \sum_{j=1}^n X_j Y_j)^\top$ 也是充分的。

8.1 (1) $c = 1/2(n-1)$; (2) $c = 1/n$ 。

8.2 $E(\hat{\theta}^2) = V(\hat{\theta}) + \theta^2 > \theta^2$ 。

8.3 $\hat{\mu}_1, \hat{\mu}_2$ 都是 μ 的无偏估计量且 $V(\hat{\mu}_1) \leq V(\hat{\mu}_2)$ 。

8.4 参考例 7.14, 求得 $X_{(1)}$ 的密度函数 $f_{X_{(1)}}(x) = \begin{cases} ne^{-n(x-\theta)} & \text{当 } x \geq \theta \\ 0 & \text{当 } x < \theta \end{cases}$ 于是, $E(X_{(1)}) = \theta + 1/n$, 而 $E(X) = \theta + 1, V(X) = 1$ 。计算得 $V(\hat{\theta}_2) = 1/n^2 \leq V(\hat{\theta}_1) = n^{-2} \sum_{j=1}^n V(X_j) = 1/n$ 。

8.5 提示: $E(X - \mu) = \sigma \sqrt{2/\pi}$ 。

8.6 $\partial \ln f_\beta(x)/\partial \beta = 1/\beta - x$, 进而 $I(\beta) = V(X) = 1/\beta^2$ (见第 297 页)。

8.7 $E(X) = \int_0^1 (\theta + 1)x^{\theta+1} dx = (\theta + 1)/(\theta + 2)$, 矩估计为 $\hat{\theta} = (2\bar{X} - 1)/(1 - \bar{X})$ 。似然函数为 $\mathcal{L}(\theta; x_1, x_2, \dots, x_n) = (\theta + 1)^n (\prod_{j=1}^n x_j)^\theta$, 解方程 $\partial \ell(\theta)/\partial \theta = n/(\theta + 1) + \sum_{j=1}^n \ln x_j = 0$ 得 θ 的最大似然估计 $\hat{\theta} = -n/\sum_{j=1}^n \ln X_j - 1$ 。

8.8 设盒中有 w 个白球, 则有 θw 个黑球, 且抽到白球和黑球的概率分别为 $1/(1+\theta)$, $\theta/(1+\theta)$ 。似然函数为 $\mathcal{L}(\theta) = [1/(1+\theta)]^k[\theta/(1+\theta)]^{n-k}$, 求得 θ 的最大似然估计为 $\hat{\theta} = n/k - 1$ 。

8.9 矩估计 $\hat{\theta} = (3 - \bar{X})/2$ 。最大似然估计 $\hat{\theta} = (2n_1 + n_2)/(2n)$, 其中 n_1, n_2 分别是样本中 1, 2 的个数。

8.10 由练习 4.28 知, $E(X) = \exp(\mu + \sigma^2/2)$, $V(X) = [E(X)]^2[\exp(\sigma^2) - 1]$ 。于是, $\hat{\theta}_1 = \exp(\hat{\mu} + \hat{\sigma}^2/2)$, $\hat{\theta}_1 = \hat{\theta}_1^2[\exp(\hat{\sigma}^2) - 1]$, 其中 $\hat{\mu}, \hat{\sigma}^2$ 分别是 μ, σ^2 的最大似然估计。即,

$$\begin{aligned}\hat{\theta}_1 &= \exp\left\{\frac{1}{n} \sum_{j=1}^n \ln X_j + \frac{1}{2n} \sum_{j=1}^n \left(\ln X_j - \frac{1}{n} \sum_{j=1}^n \ln X_j\right)^2\right\} \\ \hat{\theta}_2 &= \hat{\theta}_1^2 \left[\exp\left\{\frac{1}{n} \sum_{j=1}^n \left(\ln X_j - \frac{1}{n} \sum_{j=1}^n \ln X_j\right)^2\right\} - 1 \right]\end{aligned}$$

8.11 模仿例 8.27, $\mathcal{L}(\theta; x_1, \dots, x_n) = \begin{cases} \theta^{-n} & \text{当 } \theta \leq x_1, \dots, x_n \leq 2\theta \\ 0 & \text{其他} \end{cases}$
 因为 $\theta \leq \min_{1 \leq j \leq n} x_j \leq \max_{1 \leq j \leq n} x_j \leq 2\theta$, 所以 $\frac{\theta}{2} \leq \frac{1}{2} \min_{1 \leq j \leq n} x_j \leq \frac{1}{2} \max_{1 \leq j \leq n} x_j \leq \theta \leq \min_{1 \leq j \leq n} x_j$ 。
 又 \mathcal{L} 是 θ 的递减函数, 因此 $\hat{\theta} = \frac{1}{2} \max_{1 \leq j \leq n} X_j$ 。

8.12 未知参数 μ_X, σ_X^2, ρ 的 MLE 如下 (μ_Y, σ_Y^2 的 MLE 也是类似的)。

$$\begin{aligned}\hat{\mu}_X &= \frac{1}{n} \sum_{i=1}^n X_i & \hat{\sigma}_X^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_X)^2 \\ \hat{\rho} &= \frac{1}{n \hat{\sigma}_X \hat{\sigma}_Y} \sum_{i=1}^n (X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)\end{aligned}$$

仿照例 8.30, 先求得未知参数 $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)^\top$ 的 Fisher 信息矩阵及其

逆矩阵如下,

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_X^2} & -\frac{\rho}{\sigma_X \sigma_Y} & 0 & 0 & 0 \\ -\frac{\rho}{\sigma_X \sigma_Y} & \frac{1}{\sigma_Y^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{2-\rho^2}{4\sigma_X^4} & -\frac{\rho^2}{4\sigma_X^2 \sigma_Y^2} & -\frac{\rho}{2\sigma_X^2} \\ 0 & 0 & -\frac{\rho^2}{4\sigma_X^2 \sigma_Y^2} & \frac{2-\rho^2}{4\sigma_Y^4} & -\frac{\rho}{2\sigma_Y^2} \\ 0 & 0 & -\frac{\rho}{2\sigma_X^2} & -\frac{\rho}{2\sigma_Y^2} & \frac{1+\rho^2}{1-\rho^2} \end{pmatrix}$$

$$\mathcal{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y & 0 & 0 & 0 \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 & 0 & 0 & 0 \\ 0 & 0 & 2\sigma_X^4 & 2\rho^2 \sigma_X^2 \sigma_Y^2 & \rho(1-\rho^2) \sigma_X^2 \\ 0 & 0 & 2\rho^2 \sigma_X^2 \sigma_Y^2 & 2\sigma_Y^4 & \rho(1-\rho^2) \sigma_Y^2 \\ 0 & 0 & \rho(1-\rho^2) \sigma_X^2 & \rho(1-\rho^2) \sigma_Y^2 & (1-\rho^2)^2 \end{pmatrix}$$

利用定理 8.10, 当样本容量 n 很大时,

$$\text{MSE}(\mu_X, \hat{\mu}_X) = \frac{\sigma_X^2}{n} \quad \text{MSE}(\sigma_X^2, \hat{\sigma}_X^2) = \frac{2\sigma_X^4}{n} \quad \text{MSE}(\rho, \hat{\rho}) = \frac{(1-\rho^2)^2}{n}$$

独立性从 $\mathcal{I}^{-1}(\boldsymbol{\theta})$ 易得。直观上, 样本均值 $\hat{\mu}_X, \hat{\mu}_Y$ 的信息对于确定 $\hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\rho}$ 的联合分布没有增益。

8.13 要使得 $V(T) = p^2 V(\bar{X}) + (1-p)^2 V(\bar{Y}) = [p^2/n_1 + (1-p)^2/n_2] \sigma^2$ 达到最小, 解得 $p = n_1/(n_1 + n_2)$ 。

8.14 若 $\hat{\theta} = \sum_{j=1}^n k_j X_j$ 无偏, 则 $\sum_{j=1}^n k_j = 1$ 。要使目标函数 $V(\hat{\theta}) = \sum_{j=1}^n k_j^2 \sigma_j^2$ 达到最小, 利用 Lagrange 乘子法构造函数 $f(k_1^2, \dots, k_n^2) = \sum_{j=1}^n k_j^2 \sigma_j^2 - \lambda(\sum_{j=1}^n k_j - 1)$, 由 $\partial f / \partial k_j = 0$ 得到方程组 $2k_j \sigma_j^2 - \lambda = 0, j = 1, 2, \dots, n$, 连同 $\sum_{j=1}^n k_j = 1$ 解得 $k_j = \sigma_j^{-2} [\sum_{j=1}^n (1/\sigma_j^2)]^{-1}$ 时 $\hat{\theta}$ 的方差达到最小。

8.15 参见性质 8.5 第四种情况, 两边取对数即可。

8.16 根据结果 (7.18) 和性质 8.5 的结果, $E(L^2) = 4\sigma^2 t_{n-1, \alpha/2}^2 / n$ 。

8.17 参考例 7.14, $X_{(n)}$ 的密度函数为 $f_n(x) = \begin{cases} nx^{n-1} \theta^{-n} & \text{当 } 0 < x < \theta \\ 0 & \text{其他} \end{cases}$

故 $P\{X_{(n)} \leq \theta \leq c_n X_{(n)}\} = P(\theta/c_n \leq X_{(n)} \leq \theta) = \int_{\theta/c_n}^{\theta} nx^{n-1} \theta^{-n} dx = 1 - c_n^{-n}$, 要使此值等于 $1 - \alpha$ 只需取 $c_n = \alpha^{-1/n}$ 。

8.18 利用性质 7.11 得 $[F_{m-1,n-1,\alpha/2}S_Y^2/S_X^2, F_{m-1,n-1,1-\alpha/2}S_Y^2/S_X^2]$ 。

8.19 由第 296 页的定义 4.17 和性质 4.25, $2\lambda_1 \sum_{j=1}^m X_j = 2m\lambda_1 \bar{X} \sim \chi_{2m}^2$, 同理 $2n\lambda_2 \bar{Y} \sim \chi_{2n}^2$ 。取枢轴量 $\lambda_2 \bar{Y}/(\lambda_1 \bar{X}) \sim F_{2n,2m}$, 可求得 λ_2/λ_1 的置信度为 $1 - \alpha$ 的置信区间为 $[\bar{X}/(\bar{Y}F_{2m,2n,\alpha/2}), \bar{X}/(\bar{Y}F_{2m,2n,1-\alpha/2})]$ 。

9.1 取伪概率 $\gamma = P\{\sqrt{n}(\bar{X} - \mu_0)/\sigma < z_{1-\alpha}|H_1 \text{ 成立}\} = P\{\sqrt{n}(\bar{X} - \mu_1)/\sigma < z_{1-\alpha} - \sqrt{n}(\mu_1 - \mu_0)/\sigma|H_1 \text{ 成立}\} = \Phi[z_{1-\alpha} - \sqrt{n}(\mu_1 - \mu_0)/\sigma]$, 由分位数的性质, $z_\gamma = -z_{1-\gamma} = z_{1-\alpha} - \sqrt{n}(\mu_1 - \mu_0)/\sigma$, 进而得到 $n = [(z_{1-\alpha} + z_{1-\gamma})\sigma/(\mu_1 - \mu_0)]^2$ 。

9.2 参考例 9.10, 在水平 α 拒绝 H_0 的条件是 $\bar{x} \geq 1 + \frac{z_{1-\alpha}}{\sqrt{n}}$ 。取伪概率是 $\Phi(z_{1-\alpha} - \sqrt{n})$ 。

9.3 (1) 由功效函数的定义 9.8, $\beta_\delta(\theta) = P_\theta\{X_{(n)} < 2.9\} + P_\theta\{X_{(n)} > 4.2\}$, 而 $X_{(n)}$ 的分布函数见第 477 页的例 7.14。

$$\beta_\delta(\theta) = \begin{cases} 1 & \text{当 } \theta < 2.9 \\ (2.9/\theta)^n & \text{当 } 2.9 \leq \theta \leq 4.2 \\ 1 + (2.9^n - 4.2^n)/\theta^n & \text{当 } \theta > 4.2 \end{cases}$$

(2) 令 $\sup_{3 \leq \theta \leq 4} \beta_\delta(\theta) = 0.1$, 即 $(2.9/3)^n = 0.1$, 得到 $n = 68$ 。

9.4 对数似然比函数为 $\ln \lambda(x, y) = x + y - 1$, 拒绝零假设 $H_0 : (X, Y)^\top \sim N(0, 0, 1, 1, 0.6)$ 的充要条件是 $x + y > c$ 。若 H_0 成立, 由例 2.23 有 $X + Y \sim N(0, 3.2)$, 对于给定的水平 α , 由拒真概率 $P\{X + Y > c|H_0\} = \alpha$, 得到临界值 $c = 4\sqrt{5}z_{1-\alpha}/5$ 。

9.5 仿照例 9.17, 如果 $s_n^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 \leq \sigma_0^2$, 则广义似然比 $\lambda(\mathbf{x}) = 1$; 如果 $s_n^2 > \sigma_0^2$, 广义似然比 $\lambda(\mathbf{x})$ 与例 9.18 的相同。在水平 α 拒绝 H_0 的条件是 $\sum_{j=1}^n (x_j - \bar{x})^2 / \sigma_0^2 > \chi_{n-1,1-\alpha}^2$ 。

9.6 广义似然比为

$$\lambda(\mathbf{x}, \mathbf{y}) = \left[1 + \frac{mn(\bar{x} - \bar{y})^2}{(m+n)^2 \hat{\sigma}^2} \right]^{-(m+n)/2}, \text{ 其中 } \hat{\sigma}^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n}$$

显然, $\lambda(\mathbf{x}, \mathbf{y})$ 是 $|\bar{x} - \bar{y}|/\hat{\sigma}$ 的减函数, 所以拒绝 H_0 的条件是 $|\bar{x} - \bar{y}|/\hat{\sigma} > c$ 。根据性质 7.11,

$$T(\mathbf{X}, \mathbf{Y}) = \frac{(\bar{X} - \bar{Y}) \sqrt{mn(m+n-2)/(m+n)}}{\sqrt{(m-1)s_X^2 + (n-1)s_Y^2}} \sim t_{m+n-2}$$

因此, 在水平 α 拒绝 H_0 的条件是 $|T(\mathbf{x}, \mathbf{y})| \geq t_{m+n-2, 1-\alpha/2}$ 。

9.7 参考第 566 页的例 9.21, 零假设 $H_0 : \mu = 100$ 在显著水平 $\alpha = 0.05$ 被拒绝, 即这批零件长度不合格。

9.8 设零假设为 “ H_0 : Mendel 遗传规律成立”。利用 Pearson χ^2 检验, 得到 $\chi^2 = (30 - 25)^2/25 + (48 - 50)^2/50 + (22 - 25)^2/25 = 1.44 < \chi_{2,0.95}^2 = 5.991465$ 。故在水平 $\alpha = 0.05$ 数据无法拒绝 H_0 。

9.9 若零假设 “ H_0 : 硬币是均匀的” 成立, 令 $p_k = P(X = k) = 1/2^k, k = 1, 2, \dots, 6$ 且 $p_7 = P(X \geq 7) = 1 - P(X \leq 6)$ 。利用 Pearson χ^2 检验, 得到 $\chi^2 = 3.83 < \chi_{6,0.95}^2 = 12.59159$, 故在水平 $\alpha = 0.05$ 数据无法拒绝 H_0 , 即此硬币是均匀的。

9.10 当 $k = 2$ 时, $X_1 \sim B(n, p_1)$ 。计算 Pearson χ^2 统计量,

$$\begin{aligned}\chi^2 &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{[n - X_1 - n(1 - p_1)]^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} \xrightarrow{L} \chi_1^2, \text{ 当 } n \rightarrow \infty \text{ 时}\end{aligned}$$

10.1 (1) 由 $A = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 1 & 2 \end{pmatrix}$ 得 $A^T A = \begin{pmatrix} 6 & 0 \\ 0 & 5 \end{pmatrix}$, 因为 $A^T A$ 非奇异, 根据式 (10.11) 得到 $\hat{\beta}_1 = (X_1 + 2X_2 + X_3)/6$ 且 $\hat{\beta}_2 = (-X_2 + 2X_3)/5$; (2) 利用定理 10.3 知, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/6), \hat{\beta}_2 \sim N(\beta_2, \sigma^2/5)$ 。因为 $A^T A$ 对角阵, 所以 $\hat{\beta}_1, \hat{\beta}_2$ 相互独立。

10.2 根据定理 4.10 和 B 是幂等矩阵, $\mathbf{X} - A\hat{\beta} = B\mathbf{X} \sim N(\mathbf{0}, \sigma^2 B)$ 。

10.3 $T = \frac{\hat{\beta}_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{t_{xx}}}}, t_{n-2}, |T| \geq t_{n-2, \frac{\alpha}{2}}$ 。

10.4 百分比的均值有显著差异。

10.5 (1) $\left(\hat{a} - S_e \frac{t_{n-1,\alpha} \sqrt{\bar{x} + s_x^2}}{s_x \sqrt{n}}, \hat{a} + S_e \frac{t_{n-1,\alpha} \sqrt{\bar{x} + s_x^2}}{s_x \sqrt{n}} \right)$
 (2) $\left(\hat{b} - S_e \frac{t_{n-1,\alpha}}{s_x \sqrt{n}}, \hat{b} + S_e \frac{t_{n-1,\alpha}}{s_x \sqrt{n}} \right)$
 (3) $\left(\frac{(n-1)S_e^2}{\chi_{\alpha/2,n-1}^2}, \frac{(n-1)S_e^2}{\chi_{1-\alpha/2,n-1}^2} \right)$

其中 S_e 是标准残差, $t_{n-2,\alpha}$ 是自由度为 $n - 2$ 的 t 分布水平 α 双侧分位数, 而 $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j, s_x^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$ 。

10.6 (1) $\hat{y} = 6.28 + 0.18x$; (2) $(9.02, 9.30)$ 。

11.1 未完成

12.1 (1) 利用公理 ③, 令 $C = B$ 得到 $P\{A|B\}P\{B|B\} = P\{AB|B\}$, 再利用公理 ① 便得证。(2) 由公理 ②, $P\{AB|B\} + P\{A^c B|B\} = P\{B|B\} = 1$, 再利用第一个结果便证

得 $P\{A|B\} \leq 1$ 。(3) 从 $P\{\emptyset|B\} = P\{\emptyset + \emptyset|B\} = 2P\{\emptyset|B\}$ 立得。(4) 从第一个和第三个结果可证。(5) 由公理 ③ 可得欲证等式左右两端都为 $P\{AB|C\}$ 。(6) $P\{A|B'\} = P\{AA'B|B'\} = P\{AA'|BB'\}P\{B|B'\} \leq P\{AA'|B\} = P\{A'|B\} - P\{A^c A'|B\} \leq P\{A'|B\}$ 。

12.2 令 $Y = cX, \eta = c\sigma$, 其中 $c > 0$, 则 Y 的密度函数 $\eta^{-1}f(\eta^{-1}y)$ 为尺度密度函数。从结构上看, $X \sim \sigma^{-1}f(\sigma^{-1}x)$ 与 $Y \sim \eta^{-1}f(\eta^{-1}y)$ 的样本空间和参数空间都是一样的, 参数 σ 和 η 应该具有相同的无信息先验, 不妨设它为 π 。参照第 653 页的例 12.13 的做法, 对任意的 $A \subseteq (0, \infty)$ 有 $P(\sigma \in A) = P(\eta \in A) = P(\sigma \in c^{-1}A)$, 其中 $c^{-1}A = \{c^{-1}\sigma : \sigma \in A\}$ 。于是,

$$\int_A \pi(\sigma)d\sigma = \int_{c^{-1}A} \pi(\sigma)d\sigma = \int_A c^{-1}\pi(c^{-1}\sigma)d\sigma$$

由 A 的任意性可得 $\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma)$, 特别地 $\pi(c) = c^{-1}\pi(1)$ 。再由 c 的任意性, 所以选择 σ 的无信息先验为 $\pi(\sigma) = \frac{1}{\sigma}$ 。利用此结果可得例 12.14 的第二款。

12.3 对数似然函数 $\ell(\boldsymbol{\theta}; \mathbf{x})$ 为

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{x}) &= \sum_{j=1}^k x_j \ln \theta_j = \sum_{j=1}^{k-1} x_j \ln \theta_j + \left(n - \sum_{j=1}^{k-1} x_j\right) \ln \left(1 - \sum_{j=1}^{k-1} \theta_j\right) \\ -E\left\{\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{X})}{\partial \theta_j^2}\right\} &= -E\left\{-\frac{X_j}{\theta_j^2} - \frac{X_k}{\theta_k^2}\right\} = \frac{n}{\theta_j} + \frac{n}{\theta_k}, \text{ 其中 } j = 1, \dots, k-1 \\ -E\left\{\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{X})}{\partial \theta_i \partial \theta_j}\right\} &= -E\left\{-\frac{X_k}{\theta_k^2}\right\} = \frac{n}{\theta_k}, \text{ 其中 } i, j = 1, \dots, k-1, i \neq j \end{aligned}$$

计算得 $\det \mathcal{I}(\boldsymbol{\theta}) \propto (\theta_1 \theta_2 \cdots \theta_k)^{-1}$, 于是参数 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)^\top$ 的 Jeffreys 先验为 $\pi(\boldsymbol{\theta}) \propto (\theta_1 \theta_2 \cdots \theta_k)^{-1/2}$ 。例 12.15 是此问题的特例。

12.4 $p(\beta|\mathbf{X} = \mathbf{x}) \propto \beta^{n-1} \exp\{-\beta \sum_{j=1}^n x_j\}$, 即 $\beta|\mathbf{X} = \mathbf{x} \sim \text{Gamma}(n, \sum_{j=1}^n x_j)$ 。

13.1 仿照例 13.2, 验证当 $O = \text{MMMM}$ 时 $P(O|\boldsymbol{\theta}) = 0.09691563$ 取得最大值。

15.1 令 x_j 取值 1, 0 分别表示选择和不选择物品 j , 则 0-1 背包问题就是在带约束条件 $\sum_{j=1}^n w_j x_j \leq W$ 之下, 求函数 $f(x_1, \dots, x_n) = \sum_{j=1}^n c_j x_j$ 的最大值。

附录 I

参考文献

参考文献

- [1] 《中国大百科全书数学卷》. 中国大百科全书出版社, 1988.
- [2] 《英汉数学词汇》. 科学出版社, 1997.
- [3] 《现代数学手册》. 华中科技大学出版社, 1999.
- [4] A. D. Aleksandrov. Mathematics, Its Essence, Methods and Role, 《数学 它的内容, 方法和意义》. Publishers of the USSR Academy of Sciences, Moscow, 1956.
- [5] S. Amari. Differential-Geometrical Methods in Statistics, volume 28 of Lecture Notes in Statistics. Springer-Verlag, Berlin, 1985.
- [6] R. B. Ash and C. A. Doléans-Dade. Probability & Measure Theory. Elsevier, second edition, 2000.
- [7] R. F. Bass. Stochastic Processes. Cambridge University Press, 2011.
- [8] D. R. Bellhouse. The Reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth. Statistical Science, 19(1):3–43, 2004.
- [9] J. O. Berger. Statistical Decision Theory and Bayesian Analysis, 《统计决策论及贝叶斯分析》. Springer-Verlag New York, Inc., second edition, 1985.
- [10] J. M. Bernardo and A. F. M. Smith. Bayesian Theory. John Wiley & Sons, Inc., 1994.
- [11] P. J. Bickel and K. A. Doksum. Mathematical Statistics: Basic Ideas and Selected Topics, 《数理统计 基本概念及专题》. Holden-Day, Inc., 1977.
- [12] P. J. Bickel and K. A. Doksum. Mathematical Statistics: Basic Ideas and Selected Topics, volume 1. Prentice-Hall, Inc., second edition, 2001.
- [13] P. Billingsley. Probability and Measure. John Wiley & Sons, Inc., 1995.

- [14] P. Billingsley. Convergence of Probability Measures. John Wiley & Sons, Inc., second edition, 1999.
- [15] C. M. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, 1995.
- [16] C. M. Bishop. Pattern Recognition and Machine Learning. Spring Science +Business Media, LLC, 2006.
- [17] E. A. Bishop. Foundations of Constructive Analysis. McGraw-Hill, Inc., 1967.
- [18] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.
- [19] G. E. P. Box and M. E. Muller. A note on the generation of random normal deviates. The Annals of Mathematical Statistics, 29(2):610–611, 1958.
- [20] G. E. P. Box and G. C. Tiao. Bayesian Inference in Statistical Analysis. Addison-Wesley, 1973.
- [21] Leo Breiman. Statistics: with a View toward Applications. Houghton Mifflin Company, 1973.
- [22] Leo Breiman. Probability. Society for Industrial and Applied Mathematics (SIAM), 1992.
- [23] Leo Breiman. Statistical modeling: The two cultures. Statistical Science, 16(3): 199–231, 2001.
- [24] G. Casella and R. L. Berger. Statistical Inference. Duxbury Press, second edition, 2002.
- [25] Y. S. Chow, H. Robbins, and D. Siegmund. The Theory of Optimal Stopping. Dover Publications, 1991.
- [26] K. L. Chung. A Course in Probability Theory, 《概率论教程》. Academic Press, third edition, 2001.
- [27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Introduction to Algorithms. McGraw-Hill Companies, second edition, 2001.
- [28] R. Courant and F. John. Introduction to Calculus and Analysis, 《微积分和数学分析引论》. Springer-Verlag New York, Inc., 1989.

- [29] H. Cramér. Mathematical Methods of Statistics, 《统计学数学方法》. Princeton University Press, 1946.
- [30] A. C. Davison and D. V. Hinkley. Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.
- [31] A. P. Dawid. Conditional independence in statistical theory. Journal of the Royal Statistical Society, Series B, 41:1–31, 1979.
- [32] M. H. DeGroot. Optimal Statistical Decisions. McGraw-Hill, New York, 1970.
- [33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B, 39:1–38, 1977.
- [34] Luc Devroye. Non-Uniform Random Variate Generation. Springer Science + Business Media New York, 1986.
- [35] D. K. Dey and C. R. Rao, editors. Bayesian Thinking: Modeling and Computation, volume 25 of Handbook of Statistics. Elsevier, 2005.
- [36] J. L. Doob. Stochastic Processes. John Wiley & Sons, New York, 1953.
- [37] N. R. Draper and H. Smith. Applied Regression Analysis. John Wiley & Sons, 1998.
- [38] H. L Dreyfus. What Computers Can't Do: The Limits of Artificial Intelligence. Harper & Row New York, 1979.
- [39] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. John Wiley & Sons, Inc., 2001.
- [40] R. Durrett. Probability: Theory and Examples. Duxbury Press, third edition, 2005.
- [41] B. Efron. Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1):1–26, 1979.
- [42] B. Efron and R. Tibshirani. An Introduction to the Bootstrap. Chapman & Hall, 1993.
- [43] Bradley Efron and Trevor Hastie. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press, 2016.

- [44] I. Ekeland. *Le Calcul, l'imprévu*, 《计算出人意料 开普勒到托姆的时间图景》. Le Seuil, 1984.
- [45] W. Feller. *An Introduction to Probability Theory and Its Applications*, 《概率论及其应用》第一卷, 胡迪鹤译, volume 1. John Wiley & Sons, Inc., 1968.
- [46] W. Feller. *An Introduction to Probability Theory and Its Applications*, 《概率论及其应用》第二卷, 李志阐、郑元禄译, volume 2. John Wiley & Sons, Inc., 1971.
- [47] R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics*. Basic Books, 2011.
- [48] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society London, Series A*, 222A:309–368, 1922.
- [49] R. A. Fisher. *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press Inc., 1990.
- [50] J. L. Folks. *Ideas of Statistics*, 《统计思想》. John Wiley & Sons, Inc., 1981.
- [51] D. Freedman, R. Pisani, R. Purves, and A. Adhikari. *Statistics*. W. W. Norton & Company, Inc., 1991.
- [52] King Sun Fu. *Syntactic Methods in Pattern Recognition*. Elsevier, 1974.
- [53] R.C. Geary. The distribution of the student's ratio for the non-normal samples. *Supplement to the Journal of the Royal Statistical Society*, 3:178–184, 1936.
- [54] B. R. Gelbaum and J. M. H. Olmsted. *Counterexamples in Analysis*. Dover Publications, 2003.
- [55] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition, 2004.
- [56] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- [57] J. E. Gentle. *Elements of Computational Statistics*. Springer Science+Business Media, Inc., 2002.

- [58] B. V. Gnedenko. The Theory of Probability, 《概率论教程》. Mir Publishers, Moscow, third edition, 1978.
- [59] G. H. Golub and C. F. van Loan. Matrix Computations, 《矩阵计算》. John Hopkins University Press, 1996.
- [60] I. J. Good. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. The MIT Press, 1965.
- [61] W. S. Gosset. The probable error of a mean. Biometrika, 6:1–25, 1908.
- [62] Z. Govindarajulu. Elements of Sampling Theory and Methods. Prentice Hall, 1999.
- [63] R. L. Graham, D. E. Knuth, and O. Patashnik. Concrete Mathematics: A Foundation for Computer Science. Addison Wesley Publishing Company, Inc., second edition, 2002.
- [64] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82:711–732, 1995.
- [65] Ulf Grenander and Michael I Miller. Pattern Theory: From Representation to Inference. Oxford University Press, 2007.
- [66] A. Hald. A History of Probability and Statistics and Their Applications before 1750. John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.
- [67] A. Hald. A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713–1935. Springer Science+Business Media, LLC, 2007.
- [68] P. R. Halmos. Measure Theory, 《测度论》. Springer Verlag, 1974.
- [69] Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
- [70] G. H. Hardy, J. E. Littlewood, and G. Pólya. Inequalities. Cambridge University Press, second edition, 1952.
- [71] J. Havil. Gamma: Exploring Euler’s Constant. Princeton University Press, 2003.
- [72] T. P. Hettmansperger. Statistical Inference Based on Ranks. John Wiley & Sons, 1984.
- [73] A. Heyting. Intuitionism: An Introduction. North-Holland, 1971.

- [74] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- [75] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57. ACM, 1999.
- [76] R. A. Horn and C. R. Johnson. Matrix Analysis. Cambridge University Press, 1985.
- [77] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.
- [78] Wikimedia Foundation Inc. Wikipedia: The free encyclopedia. <http://en.wikipedia.org>.
- [79] Kiyosi Itô. Introduction to Probability Theory. Cambridge University Press, 1986.
- [80] E. T. Jaynes. The well-posed problem. *Foundations of Physics*, 3:477–493, 1973.
- [81] H. S. Jeffreys. Theory of Probability. Oxford University Press, third edition, 1961.
- [82] N. L. Johnson and S. Kotz. Urn Models and Their Application: An Approach to Modern Discrete Probability Theory. John Wiley & Sons Inc., 1977.
- [83] R. A. Johnson and D. W. Wichern. Applied Multivariate Statistical Analysis. Pearson Education, Inc., fifth edition, 2003.
- [84] O. Kallenberg. Foundations of Modern Probability. Springer, New York, second edition, 2002.
- [85] E. P. C. Kao. An Introduction to Random Processes. Wadsworth Publishing Company, 1997.
- [86] J. G. Kemeny and J. L. Snell. Finite Markov Chains. Nostrand, Princeton, 1960.
- [87] W. J. Kennedy and J. E. Gentle. Statistical Computing. New York, 1980.
- [88] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

- [89] M. Kline. Mathematical Thought From Ancient to Modern Times, 《古今数学思想》. Oxford University Press, 1972.
- [90] D. E. Knuth. The T_EXbook. Addison-Wesley Professional, 1984.
- [91] D. E. Knuth. The Art of Computer Programming, volume 2. Addison-Wesley Publishing Company, Inc., third edition, 1998.
- [92] D. Koller and N. Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.
- [93] A. N. Kolmogorov. Foundations of The Theory of Probability. Chelsea Publishing Company, 1956.
- [94] A. N. Kolmogorov and S. V. Fomin, Halmos. Elements of the Theory of Functions and Functional Analysis. Graylock Press, 1961.
- [95] A. N. Kolmogorov and A. P. Yushkevich, editors. Mathematics of the 19th Century: Mathematical Logic, Algebra, Number Theory, Probability Theory, volume 1. Springer Basel AG, 1992.
- [96] S. Kotz and N. L. Johnson, editors. Breakthroughs in Statistics: Foundations and Basic Theory, volume 1. Spring-Verlag New York, Inc., 1992.
- [97] S. Kotz and N. L. Johnson, editors. Breakthroughs in Statistics: Methodology and Distribution, volume 2. Spring-Verlag New York, Inc., 1992.
- [98] S. Kotz and N. L. Johnson, editors. Breakthroughs in Statistics: Methodology and Distribution, volume 3. Spring-Verlag New York, Inc., 1997.
- [99] Samuel Kotz and Saralees Nadarajah. Multivariate t-Distributions and Their Applications. Cambridge University Press, 2004.
- [100] E. L. Lehmann. Nonparametrics: Statistical Methods based on Ranks. Hoden-Day, San Francisco, 1975.
- [101] E. L. Lehmann. Testing Statistical Hypotheses. Spring-Verlag New York, Inc., second edition, 1997.
- [102] E. L. Lehmann. Elements of Large-Sample Theory. Spring-Verlag New York, Inc., 1999.
- [103] E. L. Lehmann and G. Casella. Theory of Point Estimation, 《点估计理论》. Spring-Verlag New York, Inc., second edition, 1998.

- [104] T. Leonard and J. S. J. Hsu. Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers. Cambridge University Press, 1999.
- [105] D. V. Lindley. The philosophy of statistics. *The Statistician*, 49(3):293–337, 2000.
- [106] Jun S Liu, Faming Liang, and Wing Hung Wong. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, 2000.
- [107] M Loèeve. Probability Theory. Springer-Verlag Inc., fourth edition, 1977.
- [108] P. McCullagh and J. Nelder. Generalized Linear Models. Chapman and Hall, London, 1989.
- [109] G. J. McLachlan and T Krishnan. The EM Algorithm and Extensions. Wiley-Interscience, second edition, 2008.
- [110] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.
- [111] T. M. Mitchell. Machine Learning. The McGraw-Hill Companies, Inc., 1997.
- [112] David Mumford and Agnès Desolneux. Pattern Theory: The Stochastic Analysis of Real-World Signals. CRC Press, 2010.
- [113] R. M. Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.
- [114] J. Neyman. On the problem of confidence intervals. *Annals of Mathematical Statistics*, 6(3):111–116, 1935.
- [115] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London*, (236):333–380, 1937.
- [116] J. Neyman. Fiducial argument and the theory of confidence intervals. *Biometrika*, 32(2):128–150, 1941.
- [117] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, (231):289–337, 1933.

- [118] J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000.
- [119] K. Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.
- [120] V. V. Petrov. Sums of Independent Random Variables. Springer-Verlag, 1975.
- [121] V. V. Petrov. Limit Theorems for Sums of Independent Random Variables, 《独立随机变量之和的极限定理》. Nauka, Moscow, 1987.
- [122] Henri Poincaré. Science and Hypothesis, 《科学与假设》, 叶蕴理译. Science Press, 1905.
- [123] S. J. Press. Subjective and Objective Bayesian Statistics: Principles, Models, and Applications. John Wiley & Sons, Inc., 2003.
- [124] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes: The Art of Scientific Computing. New York: Cambridge University Press, third edition, 2007.
- [125] I. Prigogine. The End of Certainty: Time, Chaos, and the New Laws of Nature. The Free Press, 1997.
- [126] M. H. Quenouille. Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11:18–44, 1949.
- [127] M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 61:353–360, 1956.
- [128] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [129] C. R. Rao. Linear Statistical Inference and Its Applications. John Wiley & Sons, Inc., second edition, 1973.
- [130] C. R. Rao. R. A. Fisher: The founder of modern statistics. *Statistical Science*, 7(1):34–48, 1992.
- [131] A. Rényi. On a new axiomatic theory of probability. *Acta Mathematica Hungarica*, 6(3-4):285–335, 1955.
- [132] A. Rényi. Probability Theory. American Elsevier Publishing Company, Inc., New York, 1970.
- [133] A. Rényi. A Diary on Information Theory. Wiley and Sons, 1984.

- [134] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, 59:731–792, 1997.
- [135] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
- [136] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Spring Science +Business Media, LLC., second edition, 2004.
- [137] V. K. Rohatgi. *An introduction to Probability Theory and Mathematical Statistics*. John Wiley & Sons, Inc., 1976.
- [138] M. Rosenblatt. *Random Processes*. Spring-Verlag New York Inc., 1974.
- [139] S. M. Ross. *Random Processes, 《随机过程》*. John Wiley & Sons, Inc., 1983.
- [140] W. Rudin. *Principles of Mathematical Analysis*. The McGraw-Hill Companies, Inc., third edition, 1976.
- [141] W. Rudin. *Real and Complex Analysis*. The McGraw-Hill Companies, Inc., third edition, 1987.
- [142] W. Rudin. *Functional Analysis*. The McGraw-Hill Companies, Inc., second edition, 1991.
- [143] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [144] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Spring-Verlag New York, Inc., 1995.
- [145] A. N. Shirayev. *Probability*. Spring-Verlag New York Inc., 1984.
- [146] B. W. Silverman. *Density Estimation*. London: Chapman and Hall, 1986.
- [147] M. Stephens. Bayesian analysis of mixture models with an unknown number of components: An alternative to reversible jump methods. *Annals of Statistics*, 28:40–74, 2000.
- [148] S. M. Stigler. Thomas Bayes' Bayesian inference. *Journal of the Royal Statistical Society, Series A*, 145:250–258, 1982.

- [149] M. A. Tanner. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Spring-Verlag New York, Inc., 1996.
- [150] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.
- [151] R. A. Thisted. Elements of Statistical Computing: Numerical Computation. CRC Press, 1988.
- [152] J. W. Tukey. Bias and confidence in not-quite large samples. *The Annals of Statistics*, 29(2):614–623, 1958.
- [153] A. W. van der Vaart. Asymptotic statistics. Cambridge University Press, New York, 1998.
- [154] W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. The McGraw-Hill Companies, Inc., 1997.
- [155] A. Wald. Sequential Analysis. John Wiley & Sons, New York, 1947.
- [156] A. Wald. Statistical Decision Functions. Wiley, New York, 1950.
- [157] L. Wasserman. All of Nonparametric Statistics, 《现代非参数统计》. Springer-Verlag New York, Inc., 2005.
- [158] S. Weisberg. Applied Linear Regression, 《应用线性回归》. John Wiley & Sons, Inc., second edition, 1985.
- [159] K. M. Wolter. Introduction to Variance Estimation, 《方差估计引论》. Springer-Verlag New York, Inc., 1985.
- [160] R. Wong. Asymptotic Approximations of Integrals. Academic Press, San Diego, 1989.
- [161] 伊藤清. 《確率論の基礎》, 中译本《伊藤清概率论》, 闫理坦译. 人民邮电出版社, 2011.
- [162] 华罗庚. 《华罗庚科普著作选集》. 上海教育出版社, 1984.
- [163] 夏道行, 严绍宗. 《实变函数与应用泛函分析基础》. 上海科学技术出版社, 1982.

- [164] 张贤达. 《矩阵分析与应用》. 清华大学出版社, 2004.
- [165] 李文林. 《数学珍宝: 历史文献精选》. 科学出版社, 1998.
- [166] 王梓坤. 《概率论基础及其应用》. 北京师范大学出版社, 1996.
- [167] 陈家鼎, 孙山泽, 李东风. 《数理统计学讲义》. 高等教育出版社, 1993.
- [168] 陈希孺. 《高等数理统计学》. 中国科技大学出版社, 1999.
- [169] 陈希孺. 《数理统计学简史》. 湖南教育出版社, 2000.
- [170] 陈希孺, 陈桂景等. 《线性模型参数的估计理论》. 科学出版社, 1985.

附录 J

符号表

符号表

2^A 或 $\mathcal{P}(A)$	集合 A 的幂集合	$H(X)$	随机变量 X 的熵
β_k	k 阶绝对矩	$H(X, Y)$	随机变量 X, Y 的联合熵
$\beta_\delta(\theta)$	功效函数	$K(f/g)$	Kullback-Leibler 信息散度
\mathbf{x}^\top	列向量 \mathbf{x} 转置	$\lambda(\mathbf{x}; \theta_0, \theta_1)$	似然比
χ^2_η	χ^2 分布	$\langle \mathbf{x}, \mathbf{y} \rangle$	或者 $\mathbf{x}^\top \mathbf{y}$ 向量 \mathbf{x}, \mathbf{y} 的内积
χ^{-2}_η	逆 χ^2 分布	$M(X)$	或 m_X 随机变量 X 的中位数
$\text{Cov}(X, Y)$	随机变量 X, Y 的协方差	$\mathcal{F}(g)$	函数 g 的 Fourier 变换
$\text{Cov}_X(t_1, t_2)$ 或 $\gamma_X(t_1, t_2)$	协方差函数	$I(\theta)$	参数 θ 的 Fisher 信息量
$\Delta[a, b; c]$	三角形分布	\mathfrak{B}_n	\mathbb{R}^n 上的 Borel σ 域
$\det(A)$	方阵 A 的行列式	$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$	似然函数
$E(X)$ 或 μ_X	随机变量 X 的期望	\mathcal{S}	σ 域
$E_X(t)$	均值函数	$\text{Beta}(\alpha, \beta)$	Beta 分布
$\ell(\boldsymbol{\theta}; \mathbf{x})$	对数似然函数	$\text{Expon}(\beta)$	指数分布
$\text{erf}(x)$	误差函数	$\text{Inv-Gamma}(\alpha, \beta)$	逆 Gamma 分布
$\Gamma(x)$	Gamma 函数	$I(X, Y)$	随机变量 X, Y 的互信息
c_s	偏度系数	$\mu(A)$	A 的测度
c_k	峰度系数	μ_k	k 阶中心矩
$\gamma_X(h)$	自协方差函数	$\nabla_{\mathbf{x}}^2 f$ 或 $H(f)$	函数 $f(\mathbf{x})$ 的海森矩阵

$\nabla_{\mathbf{x}} f$	函数 $f(\mathbf{x})$ 的梯度	C_n^k	组合数
$\nabla_A f$	矩阵函数 $f(A)$ 的梯度	c_v	变异系数
Ω	基本事件集合	$d(F, G)$	分布 F, G 的 Kolmogorov 距离
\bar{X}	样本均值	$f * g$	f, g 的卷积
$\perp \!\! \perp \{X_1, X_2, \dots, X_n\}$	独立的随机变量	$F_n^*(x)$	经验分布函数
$\perp \!\! \perp_Y \{X_1, X_2, \dots, X_n\}$	关于 \mathbf{Y} 条件独立	$F_X(x)$	随机变量 X 的分布函数
$\Phi(x)$	标准正态分布的分布函数	$F_{m,n}$	自由度为 (m, n) 的 F 分布
$\phi(x)$	标准正态分布的密度函数	H_0	零假设
$P(A)$	事件 A 的概率	H_1	备择假设
$\text{proj}_{\mathbf{x}} \mathbf{y}$	向量 \mathbf{y} 在 \mathbf{x} 上的投影	$i \leftrightarrow j$	状态 i, j 是连通的
$\rho(X, Y)$ 或 $\rho_{X,Y}$	X, Y 的相关系数	$I_A(x)$	集合 A 的指示函数
$\sigma(\mathcal{A})$	由 \mathcal{A} 生成的 σ 域	I_n	n 阶单位矩阵
σ_X	随机变量 X 的标准差	$J(x)$	非负判定函数
$V(X)$ 或 σ_X^2	随机变量 X 的方差	J_g	g 的雅可比矩阵
$V_X(t)$	方差函数	$K(z)$	Kolmogorov 分布函数
$\varphi_X(t)$	随机变量 X 的特征函数	$m(A)$	A 的 Lebesgue 测度
$\{X_t\}$	随机过程	m_k	k 阶原点矩
A^c	A 的补集	O	零矩阵
A_n^k	排列数	$p\langle 1 \rangle + (1-p)\langle 0 \rangle$	0-1 分布
A^\top	矩阵 A 的转置	P_n	n -步转移矩阵
A_k	样本 k 阶矩	P_∞ 或 P^∞	∞ -步转移矩阵
$A_n \downarrow A$	集合降序至 $A = \bigcap_{n=1}^{\infty} A_n$	q_α	α -分位数
$A_n \uparrow A$	集合升序至 $A = \bigcap_{n=1}^{\infty} A_n$	$R(z)$	Rényi 分布函数
B_k	样本 k 阶中心矩	S^2	样本方差
		t_n	自由度为 n 的 t 分布
		$X_n \xrightarrow{a.s.} X$	几乎必然收敛

$X_n \xrightarrow{L} X$ 依分布收敛	$\text{MSE}(\theta, T)$ 统计量 T 对 θ 的均方误差
$X_n \xrightarrow{P} X$ 依概率收敛	$\text{Multin}(n; p_1, p_2, \dots, p_k)$ 多项分布
$X_{(j)}$ 第 j 个次序统计量	$\text{NegB}(n, p)$ 负二项分布
x_+ 正截尾函数	$\text{N}_n(\boldsymbol{\mu}, \Sigma)$ n 元正态分布
$\text{BIAS}(\theta, T)$ 统计量 T 对 θ 的偏倚	$\text{P\'olya}(n, p, a)$ P\'olya 分布
$\text{B}(n, p)$ 二项分布	$\text{Pareto}(\alpha, \mu)$ Pareto 分布
$\text{Cauchy}(\mu, \lambda)$ Cauchy 分布	$\text{pa}(x)$ x 的父辈节点
$\text{Dirichlet}(\boldsymbol{\alpha})$ Dirichlet 分布	$\text{Poisson}(\lambda)$ Poisson 分布
$\text{Erlang}(n, \beta)$ Erlang 分布	$\text{Rayleigh}(\sigma)$ Rayleigh 分布
$\text{Gamma}(\alpha, \beta)$ Gamma 分布	$\text{SN}(\mu, \sigma^2; \beta)$ 偏正态分布
$\text{Geom}(p)$ 几何分布	$\text{tr}(A)$ 方阵 A 的迹
$\text{Hyper}(b, w, n)$ 超几何分布	$\text{U}[a, b]$ 区间 $[a, b]$ 上的均匀分布
$\text{Laplace}(\mu, \sigma)$ Laplace 分布	$\text{Weibull}(\lambda, \alpha)$ Weibull 分布
$\text{logN}(\mu, \sigma^2)$ 对数正态分布	$\text{Wigner}(r)$ Wigner 半圆分布
$\text{Maxwell}(\sigma)$ Maxwell 分布	$\text{Wishart}_n(\Sigma, d)$ Wishart 分布

附录 K

术语索引

术语索引

- F -比, 605
 F -范数, 755
 L_r 范数, 185
 $M/M/1$ 排队模型, 419
 U 统计量, 452
 χ^2 分布, 125, 290
 μ -可测的, 56
 σ -有限的, 56
 k -分类试验, 317
 k -参数指数族, 478
 n 维 Borel 可测空间, 52
 n -步 Fibonacci 序列, 57
 n -步转移矩阵, 388
 p -值, 541
 p -范数, 754
 t 检验, 557
 z 检验, 557
 $ARMA(p, q)$ 过程, 616
 $AR(p)$ 过程, 616
 $MA(q)$ 过程, 616
0-1 分布, 254
Dirichlet 函数, 750
Arnold 变换, 11
Baum-Welch 变量, 676
Baum-Welch 算法, 676
Bayes 公式, 78, 639
Bayes 规则, 632
Bayes 风险, 632
Bell 数, 47
Bernoulli 分布, 254
Bernoulli 弱大数律, 3
Bernoulli 试验, 60
Beta-二项分布, 323
Bienaymé 公式, 173
Boltzmann 分布, 304
Borel σ 域, 52
Borel 0-1 准则, 85
Borel 函数, 103
Borel 可测函数, 103
Borel 可测集, 52
Borel 强大数律, 353
Borel 集, 52
Carathéodory 扩张定理, 56
Cauchy-Schwarz 不等式, 185, 766
Chapman-Kolmogorov 方程, 389, 408
Cholesky 分解, 756
CR 界, 496
Cramér-Lévy 定理, 278
Cramér-Rao 下界, 496
Cramér-Rao 不等式, 496
DA 算法, 736
de Moivre-Laplace 中心极限定理, 3
delta 函数, 213
Dirac delta 函数, 213
Dirichlet 分布, 321
Dirichlet-多项分布, 323

- Erlang 分布, 290
 Euler 公式, 212
 Fisher 信息度量, 487
 Fisher 信息矩阵, 486
 Fisher 信息量, 486
 Fourier 变换, 211
 Fourier 逆变换, 212
 Frobenius 范数, 754
 Galton-Watson 过程, 403
 Gamma 分布, 288
 Gauss-Laplace 分布, 119
 Gibbs 分布, 304
 Gibbs 抽样, 722
 Gibbs 抽样器, 722
 Gini 混乱度, 164
 GLR 检验, 549
 GnuPlot, 773
 Gram 矩阵, 756
 Gram-Schmidt 正交化, 755
 Hölder 不等式, 185
 Jaynes 熵, 648
 Jeffreys 先验, 645
 K-S 检验, 568, 569
 Kolmogorov 分布函数, 459
 Kolmogorov 检验, 458, 567, 568
 Kolmogorov 距离, 127
 Kullback-Leibler 信息散度, 180
 Kullback-Leibler 散度, 180
 Lévy 不等式, 191
 Lévy 连续性定理, 239
 Lagrange 乘子, 648
 Laplace 分布, 283
 Laplace-Gauss 分布, 63
 Lebesgue 可测集, 55
 Lebesgue 测度, 55
 Lebesgue 积分, 750
 Lebesgue-Stieltjes 积分, 751
 Lindeberg 条件, 365
 Lindeberg-Lévy 中心极限定理, 363
 logistic 函数, 274
 logit 函数, 655
 Lyapunov 条件, 367
 Markov 性, 386, 408
 Markov 毯, 681
 Markov 网络, 679
 Markov 边界, 681
 Markov 链, 386
 Markov 随机场, 679
 Maxima, 769
 Maxwell 分布, 157, 307
 Metropolis 比, 715, 716
 Metropolis 规则, 716
 Metropolis-Hastings 算法, 719
 ML-II 先验, 649
 Monte Carlo 方法, 35
 Neyman-Pearson 原则, 534
 Pólya 准则, 220
 Pólya 分布, 260
 Pólya 近似, 121
 Pareto 分布, 301
 Pearson χ^2 统计量, 565
 Perron 根, 759
 Poisson 过程, 413
 R, 767
 Rényi 分布函数, 459
 Rényi 定理, 459
 Rényi 熵, 181
 Rayleigh 分布, 157, 306
 Riemann ζ 函数, 89

- Riemann-Stieltjes 积分, 745
Schur 补, 759
Shannon 熵, 176
Smirnov 检验, 458, 569
Smirnov 统计量, 569
Tribonacci 序列, 19
UMVU 估计, 496
Viterbi 变量, 674
Viterbi 路径, 674
von Neumann 舍选法, 705
Weibull 分布, 305
Wiener 过程, 423
Wiener-Bachelier 过程, 379, 423
Wigner 分布, 308
Wigner 矩阵, 333
Wishart 分布, 333
一致收敛, 748
一致最优检验, 543
一致最大功效, 543
一致最小方差无偏估计, 496
一致有界, 348, 366
三角分布, 785
三角形分布, 277
上下文无关语言, 90
上侧 α -分位数, 158
上极限, 47
上鞅, 381
下侧 α -分位数, 158
下极限, 47
下鞅, 381
不充分理由原则, 80
不可数样本空间, 50
不可约的, 391
不可能事件, 8
不次于, 631
不相关的, 195
不相容, 25
不连续点, 64
两两独立, 142
两因素方差分析, 601
两样本的 Kolmogorov-Smirnov 检验, 569
两点分布, 114
中位数, 158
中心极限定理, 360, 361
中心矩, 182
临界值, 538
主成分, 204
主成分分析, 204, 758
主观概率, 3, 45
主题模型, 660
乘法法则, 75
事件, 50
二元正态分布, 130
二分类, 81
二项分布, 116
二项过程, 377
互信息, 179
互斥, 25
互达的, 391
交事件, 51
产生器, 247
代数, 49
伊藤积分, 431
伊藤过程, 431
众数, 159
优于, 543
伪逆, 585
伪随机数产生器, 245
估计量, 482
似然, 82, 639
似然函数, 505, 639

- 似然方程组, 506
 似然比, 531
 似然比检验, 544
 位置参数, 63, 643
 位置密度函数, 643
 余事件, 51
 依分布收敛, 229
 依概率收敛, 339
 信任分布, 525
 信任区间, 525
 信任系数, 525
 信念度, 45, 634
 信念网络, 144
 假设检验, 529
 假阳性, 81
 假阴性, 81
 偏倚, 485
 偏差平方和, 606
 偏度系数, 182
 偏正态分布, 281
 儿辈节点, 681
 充分统计量, 475
 先验概率, 72
 全概率公式, 78
 全面试验, 602
 公理化方法, 44
 共现, 561
 共轭先验, 645
 关于 m 奇异连续, 65
 关于 m 绝对连续, 65
 内积矩阵, 756
 冗余参数, 517
 决策规则, 631
 几乎处处, 66
 几乎必然, 67
 几乎必然收敛, 353
 几何分布, 262
 几何布朗运动, 432
 几何概率, 28
 凸函数, 764
 凸集, 764
 函数, 274, 286
 刀切估计量, 501
 分位数, 158
 分布, 111, 157
 分布函数, 111
 分布列, 116
 分布族, 443
 分支过程, 406
 切片抽样, 724
 划分, 46
 列正交矩阵, 757
 列联表, 571
 列联表检验, 571
 初始分布, 408
 删失数据, 696
 到达率, 412, 413
 功效, 533
 功效函数, 538
 加法法则, 68
 势, 533
 势函数, 538
 区间估计, 483, 512
 升序, 69
 半不变量, 225
 半圆分布, 308
 半正交矩阵, 757
 半正定, 756
 半正定函数, 756
 协方差, 194
 协方差函数, 382
 协方差矩阵, 196
 单位脉冲函数, 213
 单侧 Kolmogorov 统计量, 568

- 单侧 Smirnov 统计量, 569
单侧检验, 531
单值映射, 102
单参数指数族, 546
单因素方差分析, 601
单峰分布, 159
单样本的 Kolmogorov-Smirnov 检验, 568
单点分布, 114
单调似然比, 546
即时最大值, 426
卷积, 149, 150, 744
原假设, 529
原点矩, 182
参数为 p 的 Bernoulli 过程, 376
参数假设, 529
参数总体, 443
参数空间, 443
参数统计推断, 482
双侧检验, 531
双侧滑动平均, 618
反射公式, 287
发射矩阵, 668
取伪概率, 533
取伪错误, 532
变差, 746
变异系数, 183
古典概率模型, 17
句法模式识别, 91
可交换性, 142, 657
可交换的, 142, 657
可容决策规则, 631
可测函数, 103
可测映射, 103
可测空间, 49
可测集, 49
可积的, 750
可逆的, 401, 617
可逆跳 MCMC 方法, 721
右偏, 281
合并样本方差, 558
后辈节点, 681
后验分布, 639
后验期望损失, 629
后验概率, 72
后验预测分布, 640
向前变量, 671
向后变量, 672
向后算法, 672
吸收壁, 387
周期, 391
和事件, 51
响应变量, 575
回归, 200
回归值, 576
回归函数, 575
回归分析, 576
回归平方和, 590
回归曲线, 200
回归直线, 202
回归系数, 203, 578
因果的, 616
团, 678
图模型, 667
在线学习, 447
均值, 163, 164
均值函数, 382
均匀分布, 118, 130
均匀收敛, 748
均方根误差, 170
均方误差, 170, 485
域, 49
基本事件, 9
基本事件集合, 9
填补, 734

- 增量学习, 586
 备择假设, 529
 复合事件, 10
 复合假设, 530
 外测度, 55
 多元总体, 442
 多元统计学, 442
 多峰分布, 159
 多指标随机过程, 376
 多试 Metropolis 算法, 720
 多重填补, 734
 多项分布, 317
 大数律, 339
 大样本问题, 446
 奇异值, 756
 奇异值分解, 757
 奇异型分布, 115
 奇异概率测度, 65
 奇异连续, 65
 字节, 176
 季节分量, 618
 完全似然函数, 688
 完全数据, 688
 完全最大似然估计, 688
 实现, 378
 客观概率, 3, 60
 密度函数, 117, 129
 对数, 506
 对数似然函数, 505
 对称差事件, 51
 对立事件, 51
 小样本问题, 446
 尺度参数, 63, 662
 尾事件, 85
 层级先验, 647, 649
 峰度系数, 182
 左偏, 281
 左逆, 585
 差事件, 51
 布朗桥, 428
 布朗运动, 379
 带漂移的布朗运动, 428
 常返的, 393
 幂等矩阵, 758
 平均差异, 171
 平稳, 612
 平稳分布, 399
 平稳增量, 410
 平稳增量过程, 410
 平稳的, 383
 平稳转移概率, 408
 并事件, 51
 广义 EM 算法, 689
 广义似然比, 549
 广义似然比检验, 549
 广义线性模型, 582
 建议分布, 716
 开集, 49
 弃一, 501
 弱大数律, 339
 弱平稳, 612
 弱平稳的, 383
 弱收敛, 229
 弱相合估计, 490
 强大数律, 353
 强平稳, 612
 强平稳的, 383
 强度, 413
 强相合估计, 490
 归一因子, 157, 224
 归一性, 408
 微分熵, 177
 必然事件, 8
 总体, 442

- 总体分布, 442
总偏差平方和, 604
总平方和, 590
成列删除, 734
成对删除, 734
截尾正态分布, 335
扩充数据, 688
扩张, 56
抽样分布, 464
拉普拉斯矩阵, 756
拒真概率, 533
拒真错误, 532
拒绝域, 536
拓扑, 49
拓扑空间, 49
拟合优度检验, 565
拟合优度的 Pearson χ^2 检验, 566
拟合值, 576
指数分布, 290
指示函数, 108
损失函数, 628
排队论, 419
接受函数, 716, 719, 720
接受域, 536
接受率, 715
接续规则, 651, 652
效用, 163
效能, 246
数学建模, 440
数学规划, 583
数据增扩, 736
数据矩阵, 576
数据转换似然, 644
整合性, 325
方差, 171
方差-协方差矩阵, 195
方差函数, 382
方差分析, 601
方差齐性, 581
无信息先验, 643
无偏估计, 494
无差别原则, 80
无限总体, 442
时间序列, 376
时齐 Makrov 链, 408
时齐的, 386
显著概率, 541
显著水平, 536
更新函数, 414
更新方程, 418
更新过程, 410, 414
最优化, 583
最佳线性预测, 612
最佳预测, 612
最大似然估计, 506
最大熵先验, 647, 648
最小二乘估计, 584
最小二乘原则, 200
最小二乘法, 200, 582
有偏估计, 494
有效估计, 497
有界变差, 746
有限总体, 442
有限样本空间, 50
有限维分布族, 381
期望, 163, 164
期望损失, 162, 629
期望损失原则, 629
李善兰恒等式, 96
条件分布函数, 137, 138
条件分布列, 136
条件方差, 172
条件期望, 168
条件概率, 73, 634

- 条件概率函数, 136
 条件概率空间, 73, 634
 条件熵, 178
 条件独立, 93, 143
 条件随机场, 680
 极值, 451
 极大团, 678
 极小极大决策规则, 632
 极小极大行为, 632
 极差, 451
 枢轴量, 516
 标准 Cauchy 分布, 285
 标准偏正态分布, 281
 标准化, 120
 标准化的, 175
 标准差, 172
 标准布朗运动, 423
 标准正态分布, 119
 样本, 446
 样本 k 阶中心矩, 447
 样本 k 阶矩, 447
 样本中位数, 452
 样本值, 446
 样本偏度系数, 463
 样本分布, 446
 样本协方差矩阵, 469
 样本变异系数, 463
 样本均值, 447
 样本容量, 446
 样本峰度系数, 463
 样本方差, 447
 样本方差-协方差矩阵, 469
 样本点, 9, 446
 样本相关系数, 469
 样本空间, 50
 样本统计量, 447
 样本路径, 378
 样本量, 446
 核, 453
 根节点, 144
 梯度, 759, 760
 梯度矩阵, 763
 检验函数, 536
 检验统计量, 538
 概率, 57, 58
 概率函数, 116
 概率分布, 111
 概率图模型, 144, 667
 概率密度函数, 66, 117
 概率母函数, 252
 概率测度, 57, 58
 概率测度空间, 57
 概率潜在语义标引, 659
 概率空间, 57, 58
 概率质量函数, 116
 模式理论, 91
 模拟退火, 727
 次序统计量, 451
 正交多项式, 595
 正交多项式回归, 596
 正则方程, 584
 正定矩阵, 756
 正常返的, 396
 正态分布, 119, 742
 正态过程, 381
 正截尾函数, 275
 正比关系, 62
 残差, 577
 残差向量, 577
 残差平方和, 582, 590
 比特, 176
 水平, 536, 600
 水平 α 检验, 538
 水平 α -UMP 检验, 543

- 水库抽样, 76
法则, 301
测度, 55
测度空间, 55
海森矩阵, 761
混沌, 10
混淆矩阵, 81
渐近无偏估计, 495
渐近正态性, 492
渐近正态的, 492
滑动平均, 150
滑过的, 393
滞后算子, 616
潜在 Dirichlet 分配, 659
点估计, 483
熵, 176
父辈节点, 144
特征函数, 214
状态, 376
状态空间, 376
独立, 83
独立同分布, 141
独立同分布样本, 448
独立增量, 410
独立增量过程, 410
独立的, 140, 227
理论值, 576
理论频次, 565, 566
生存分析, 696
生灭过程, 419
白噪声, 614
目标变量, 575
直方图, 61
相关关系, 575
相关函数, 382
相关系数, 198
相合估计, 490
相对熵, 180
真阳性, 81
真阴性, 81
矩, 182
矩估计, 503
矩母函数, 214, 218
离散均匀分布, 251
离散时间 Markov 链, 386
离散时间过程, 376
离散样本空间, 50
离散熵, 176
离散的, 116
离散高斯过程, 381
种子, 245
秩, 452
秩为 k 的 Eckart 近似, 758
秩统计量, 452
积事件, 51
积分, 750
积测度, 56
窗宽, 150
等价关系, 46
等腰三角形分布, 152, 277
简单 Monte Carlo 方法, 703
简单假设, 530
简单函数, 750
简单滑动平均, 150
简单随机变量, 108
简单随机样本, 448
简单随机游动, 377
类别分布, 116
类型 II 最大似然先验, 647, 649
系统函数, 744
系统误差, 485
紧凑 SVD, 757
累积分布函数, 111
纯不连续的概率测度, 65

- 线性假设, 592
 线性同余产生器, 247
 线性同余序列, 245
 线性回归模型, 581
 线性时间序列, 614
 线性模型, 581
 线性过程, 615
 组内偏差平方和, 604
 组合产生器, 247
 组间偏差平方和, 604
 细致均衡条件, 401
 经典分解模型, 618
 经验 Bayes 方法, 72
 经验分布函数, 454
 经验累积分布函数, 454
 经验频次, 565, 566
 绝对矩, 182
 绝对连续, 65
 统计假设, 529
 统计决策理论, 625
 统计量, 447
 维单纯形, 321
 维随机变量, 127
 维随机向量, 127
 罩函数, 706
 置信上限, 513
 置信下限, 513
 置信区间, 512
 置信度, 512
 置信水平, 512
 置信系数, 513
 联合, 129
 联合分布函数, 127
 联合概率, 75
 联合熵, 177
 自助法, 471
 自协方差函数, 612
 自相关函数, 384, 612
 茎叶图, 449
 行为空间, 628
 行正交矩阵, 757
 补事件, 51
 观测的 Fisher 信息矩阵, 654
 观测矩阵, 576
 规则, 631
 解释变量, 575
 计数测度, 55
 计数过程, 377
 试验设计, 602
 误差函数, 119
 误差平方和, 583
 谱半径, 759
 谱密度, 615
 贝叶斯推断, 80, 639
 贝叶斯期望损失, 629
 贝叶斯检验, 630
 贝叶斯网络, 144
 贝叶斯行为, 629
 负二项分布, 264
 贡献率, 205
 超先验, 649
 超几何分布, 260
 超参数, 649, 658
 趋势分量, 618
 轨道, 381
 转移矩阵, 386, 668
 边缘分布, 134
 边缘密度, 134
 过程, 375
 连接函数, 582
 连续性修正, 259
 连续时间 Markov 链, 408
 连续时间过程, 376
 连续样本空间, 50

- 连续概率测度, 64
连续熵, 177
连续的, 117
连续高斯过程, 381
连通的, 391
迭对数律, 357
迹, 755
逆 χ^2 分布, 291
逆 CDF 法, 273
逆 Gamma 分布, 289
逆像, 102
遍历的, 398
邻接矩阵, 756
配方法, 64, 770
重 Bernoulli 试验, 60
重对数律, 357
重尾分布, 301
长尾分布, 303
阶后移算子, 616
降序, 69
随机上下文无关文法, 91
随机事件, 8, 50
随机变量, 106
随机变量序列, 227
随机序列, 376
随机微分方程, 433
随机数, 244
随机服务系统理论, 419
随机模拟, 35
随机游动, 614
随机现象, 8
随机试验, 8
随机误差, 575
随机过程, 375
隐 Markov 模型, 668
雅可比矩阵, 760
集成学习, 94
零假设, 529
零常返的, 396
零测集, 55
非主观先验, 643
非参数假设, 529, 530
非参数总体, 443
非可容决策规则, 631
非周期的, 392
非季节模型, 618
非常返的, 393
非正常先验, 643
非确定下推自动机, 90
非负判定函数, 117
面板, 638
鞅, 381
鞍点法, 762
预测值, 576
预测分布, 638
预烧, 315, 715
频次表, 448
频率表, 448
风险函数, 631
验前概率, 72
验后概率, 72
高斯函数, 62
高斯分布, 119
高斯过程, 381
齐性 Markov 链, 408
齐性 Poisson 过程, 413
齐性的, 386