应用概率统计基础及算法

于江生

前言：概率论 (probability theory) 源于十七世纪对赌博的研究，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，探索的是随机现象的数量规律。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、工程科学、社会科学、人文科学、军事科学等诸多应用领域，凡是涉及数据的收集、整理、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实践工具。随着计算机科学的发展，概率统计的应用价值也越来越得到凸显。

既然概率统计有这么广泛的应用背景，建模、算法设计与实现就显得很重要。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，为此推荐使用开源的统计计算软件 R（与 S 语言兼容，可阅读 S-Plus 的文档 [89]）、计算机代数系统 Maxima 和科学绘图工具 GnuPlot 来完成概率统计实践，旨在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。除此之外，本书还将介绍用于贝叶斯建模和随机模拟计算的软件 BUGS (Bayesian inference Using Gibbs Sampler, [86])，以及与它兼容的开源 Gibbs 抽样工具 JAGS (Just Another Gibbs Sampler)，R 为它们提供了很好的接口。读者在互联网上可找这些开源软件的使用手册、帮助文档等。另一个革新的地方是增加了对概率统计历史和现状的简介，包括近些年取得的一些成果和相关数学家的学术功绩。

为什么学习概率统计需要计算机辅助或实践？这是因为计算机科学是成功地运用了数学的典范，一方面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，也就是计算机科学大师 Donald Ervin Knuth (1938-) 称之为“具体数学”的东西 [44]。另一方面计算机辅助使得抽象的数学概念和理论变得更容易理解，甚至能帮助人们深入研究。更不用说计算

*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”。

机实践能把“纸上谈兵”的数学模型变成可行的算法并加以实现，理论在显示强大威力的同时也露出有趣好玩的一面。美国数学家 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。”在信息科学的时代里，概率统计与计算机实践难分难舍，谁也无法忽略计算机的作用，它是帮助我们走向应用和直观的工具，本书正想阐明这一点。

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，如定理的证明。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 (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。没人知道计算机在这条路上到底能走多远，有笑话说 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。

---

*美国数学家、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够了。目前找到的正例越来越多，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，在 1995 年 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”。如何利用计算机证明 Fermat 大定理依然吸引着一批数学家和计算机科学家。

†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。感兴趣的读者可参阅荷兰数学家 Ared Heyting (1898-1980) 的著作《直觉主义导论》 [49]。

‡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。1967 年，美国数学家 Errett Albert Bishop (1928-1983) 出版了专著《构造性分析基础》[22]，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。

除了实用性，数学和谐的美也是值得追求的。德国数学家 Carl Gustav Jacob Jacobi (1804-1851) 在给友人的信中说道，“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……。”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。概率统计方法最终都落实到可行的算法上，无论是为了“心智的荣耀”还是实际的应用，皆需要基础理论之可靠和算法设计之精巧。

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，但一些带有词根变化用作形容词的外国人名还是保留了中文译名，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》 [7] 和《现代数学手册》 [8] 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 LaTeX 系统进行排版*，所有科学绘图都由作者通过 R、Maxima、GnuPlot、MetaPost 完成。数学家肖像的图片取自互联网，恕不一一标明其出处。书中特殊符号说明如下：

	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
□	证明结束		条目、特款
§1.2.3	第一章第二节的第三小节		临时小注解
☆, ✯, ★	标注在习题前，分别表示“有点难”、“难”、“很难”		

*有关 TeX 排版系统的介绍详见 D. Knuth 的力作 The TeXbook，它的 TeX 源码也是很好的“入门读物”。另外，绘图语言 MetaPost 也值得推荐，它可以通过对图形进行数学描述来制作矢量图，然后嵌入 TeX 文档生成精美插图。

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。而在每一节的开始也有“本节内容”和“学习目标”，为便于读者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，多是为了强调某种方法或让读者“照葫芦画瓢”，一般有答案或提示。而“问题”和“注记”则是为了引发思考和做更深入细致的解释，它们在正文中起到点缀的作用，如果读不过去可以先放下来，在掌握了正文主要内容后再来阅读比较好些。正文里某些新概念的解释被放置在脚注中，如 Tribonacci 数列等。课后习题中有难度的习题标有难度级别，为便于自学，附录 I 提供了“习题答案或提示”。正文中穿插了一些关键概念、重要方法或经典结果的数学史，也粗略地介绍了相关数学家的相关成就以及他们提出的思想，因为它们也是数学文化不可缺少的组成部分。另外，还有一些课外阅读的建议，它们有助于扩展知识面和增加趣味性，读者可依自己的目标和实际情况而定。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。

本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，要写一本涵盖所有重要结果的基础教科书也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。本书共分四个部分：概率论基础（包括随机事件、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等内容），数理统计学初步（包括参数估计、假设检验、线性模型的回归分析与方差分析、非参数统计推断、贝叶斯统计学概要等），概率统计中的一些实用算法（包括隐 Markov 模型的 Viterbi 算法、参数估计的期望最大化算法、最大熵算法、随机模拟算法等），附录（一些重要的补充内容放在了附录里，如对 R、Maxima 和 GnuPlot 的简介，一些模拟试验的 R 演示源码、正态分布的由来、Riemann-Stieltjes 积分、可测函数与 Lebesgue 积分、矩阵计算、凸性与 Jensen 不等式等）。该书正文几乎不

涉及抽象测度论，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。对某些较深入的内容和较复杂的证明，都以特殊符号标出，或者指明在哪些参考文献中可以找到详解，读者在阅读时可以有选择地略过或者按图索骥。虽然正文不要求掌握这些结论的严格证明，但结论本身还是需要了解的，我们把它们当作本书的“边缘”知识并不是因为它们不重要，而是篇幅和主题所限。

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，他们对初稿提出了许多宝贵的意见。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。感激导师程民德先生，他引导作者由数学转入信息科学领域。另外，作者怀着羞愧的心面对父母和妻子多年来无私的爱护与支持，他们一直鼓励作者完成此书的写作，但愿这本书能算作感恩的回报。

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。由于作者能力所限，书中难免有各种错误或不妥之处，诚恳地欢迎读者指正，以便在后续版本中不断提高它的质量。希望这本新书对读者们有所裨益，并能带来阅读的快乐。

于江生
北京大学信息科学技术学院计算机系
邮件地址：yujs@pku.edu.cn

目录

第一部分

概率论基础

概率论的起源：十七世纪初至中叶法国流行一种赌博游戏，连续掷一个骰子 4 次，赌是否出现 1 点。热衷于赌博的法国显贵 Chevalier de Méré (1607-1648) 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：连续掷两个骰子 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（这是对的），所以掷 24 次双骰子出现一对 1 点的机会等同于掷 4 次单骰子出现 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，提出了期望值的概念（源自赌资分配问题，见 §2.3）。

荷兰数学家 Christiaan Huygens (1629-1695) 受 Pascal 和 Fermat 工作的启发，于 1657 年发表了第一部概率论的著作《论赌博中的推断》(De Ratiociniis in Ludo Aleae)。对概率论的研究也引发了人们对知识、推理等哲学问题的思考，Huygens 在给友人的信中说到，“我坚信我们对凡事都不确定，而是或然地了解万物。”一百多年后，法国数学家 Pierre-Simon Laplace (1749-1827) 在其著作《概率的分析理论》(Théorie Analytique des Probabilités, 1812) 中表达了同样的思想，“自然界绝大多数重要问题都是概率问题。严格地讲，我们的一切知识几乎都是或然性的，只有很少的事物对我们来说是知其所以然的。即使在数学中，归纳和类比这些发现真理的基本方法也是建基于概率的。因此人类知识的整个系统都和概率论息息相关。”

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

*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。

的规律。Pascal 和 Fermat 对概率论开创性的工作在瑞士数学家 Jacob Bernoulli (1654-1705) 的推动下得到了更深入的发展，他的遗著《推测术》(Ars Conjectandi) 于 1713 年发表，其内容涉及了排列组合的一般理论、当时概率论的经典结果及应用和后来以他的名字命名的*弱大数律*等。该书奠定了概率论的基础，是该领域早期的集大成之作。

此后，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 $\mathrm{B}(n, 1/2)$，相继在他的论著《机遇论》(Théorie du Hasard, 1718) 和《级数和求积的分析杂论》(Miscellanea Analytica de Seriebus et Quadraturis, 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 $\mathrm{B}(n, 1/2)$。这一成果的思想超越了那个年代，当时无人能理解，直到 Laplace 在《概率的分析理论》中再度提起，并将之推广到二项分布的一般形式，这就是著名的 de Moivre-Laplace *中心极限定理* (central limit theorem)*，堪称古典概率论的无冕之王。十九世纪末至二十世纪初，俄国圣彼得堡学派对极限定理进行了深入的研究，定理本身和证明方法的重要性都逐渐被世人所知，所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。

十九世纪古典概率论经过法国数学家 Laplace 和 Siméon-Deni Poisson (1781-1840)、德国数学家 Carl Friedrich Gauss (1777-1855)、英国理论物理学家兼数学家 James Clerk Maxwell (1831-1879)、美国理论物理学家兼数学家 Josiah Willard Gibbs (1839-1903)、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) 等人的进一步研究，积累出很多漂亮的结果，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，比

*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。

较著名的是法国数学家 Joseph Bertrand (1822-1900) 在其著作《概率计算》(Calcul des Probabilités, 1889) 中给出几何概率的悖论（详见 §1.1.2 和 §1.2.2）。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。

1900 年，德国数学家 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。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。直到三十年代随着大数律的深入研究，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来临。

1933 年，苏联数学家 Andrey Nikolaevich Kolmogorov (1903-1987) 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 [57] 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。现代概率论经过 A. N. Kolmogorov、Aleksandr Yakovlevich Khinchin (1894-1959)、Jarl Waldemar Lindeberg (1876-1932)、Paul Pierre Lévy (1886-1971)、William Feller (1906-1970)、Norbert Wiener (1894-1964)、Joseph Leo Doob (1910-2004)、伊藤清 (Kiyoshi Itô, 1915-2008) 等数学家的大力推动，目前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。

*测度论 (measure theory) 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 [47]。二十世纪初建立的 Lebesgue 测度和 Lebesgue 积分理论以及随后创立的抽象测度和积分理论为概率的公理化奠定了基础。

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。

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科学领域，如物理学、化学、生物学、医学、心理学、经济学、社会学、教育学、政治学等，以及所有的工程科学领域都能找到广泛的应用，有关概率论的数学知识是不可或缺的，它的实用价值也是毋庸置疑的。为了更好地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。

- W. Feller 的《概率论及其应用》上卷 [35]（此书深入浅出，对预备知识要求较少。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）。
- B. V. Gnedenko (1912-1995) 的《概率论教程》 [40]（内容浓缩，书末提供了《统计学要领》和《概率论简史》。作者 Gnedenko 是 Kolmogorov 的学生）。
- 王梓坤院士的《概率论基础及其应用》 [6]（书中有丰富的应用实例，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）。
- A. N. Shiryayev 的《概率论》 [84]（莫斯科大学的经典教材，此书被列为美国研究生用数学丛书 GTM 第 95 号，作者 Shiryayev 也是 Kolmogorov 的学生）。
- 钟开莱 (1917-2009) 的《概率论教程》 [25]（钟开莱是著名的美籍华裔数学家、概率论专家）。

前三部著作无需测度论基础，非常适合初学者；后两部面向数学专业，需要一定的数学基础。这些推荐书籍的共同特点是：(1) 强调了概率论的思想方法，(2) 语言优美例子丰富。

第一章

随机事件与概率论的公理化

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是否发生：若是则称之为*必然事件* (certain event)，否则称之为*不可能事件* (impossible event)。必然事件有如，(1) 一个标准大气压下，纯水在100ºC 沸腾；(2) 物体在无外力作用下速度保持不变；(3) 光线通过引力场将发生偏移。不可能事件有如，(4) 太阳从西方升起；(5) 欧式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，在得知发生与否之前，我们称之为*随机事件*。如 (1) 抛一枚均匀的硬币出现正面；(2) 一个均匀的骰子掷出奇数点；(3) 明年三月份交货的黄金期货价格为每盎司 765.50 美元；(4) 未来十年全球温度将持续上升；(5) 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，哪里有什么随机性可言？事实上，“在相同条件下”这一要求并非是绝对的，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，历史上曾有多位学者，像英国数学家 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，下面是他们的一些试验结果。

表 1.1: 历史上一些抛硬币试验的结果。

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

对随机现象数量规律的研究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，抛硬币的基本事件集合 Ω = {正面, 反面}，其中 {正面} 和 {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该特效药的其他胃癌患者也作为观察对象，只要试验结果对修改了的条件不太敏感，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“骰子掷出奇数点”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“骰子掷出奇数点”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“骰子掷出奇数点”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“骰子掷出奇数点”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被称为**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，反之亦然，后文对二者不再区分。于是，集合论理所当然地成为概率论的数学基础。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，所以它只能表示不可能事件。

例 1.1. 一枚均匀的硬币连续抛两次，用 H 表示正面 (head)，T 表示反面 (tail)，基本事件集合是 $\Omega = \{(T, T), (T, H), (H, T), (H, H)\}$。

为方便起见，Ω 的元素依次简写为 $0, 1, 2, 3$。显然，集合 Ω 的元素个数（也称为 Ω 的势）为 4，记作 $|\Omega| = 4$。幂集合 2^Ω 共有 $2^{|\Omega|} = 2^4 = 16$ 个元素，我们利用开源软件*Maxima 列出它们并验证其个数。

*本着数学理论联系计算机实践的原则，有三个开源 (open source) 软件贯穿全书：(1) 用于统计计算的 R；(2) 用于符号计算的 Maxima；(3) 用于函数绘图的 GnuPlot（有关这三个软件的简介见附录 A）。读者也可以选择使用某些商业软件来完成本门课程的计算机实践，如 MatLab、S-PLUS、Maple 或 Mathematica。

```
(%i1) powerset({0,1,2,3});
(%o1) {{}, {0}, {0, 1}, {0, 1, 2}, {0, 1, 2, 3}, {0, 1, 3}, {0, 2}, {0, 2, 3},
       {0, 3}, {1}, {1, 2}, {1, 2, 3}, {1, 3}, {2}, {2, 3}, {3}}
(%i2) cardinality(%);
(%o2) 16
```

其中，$\{0,3\}$ 或 $\{(T,T),(H,H)\}$ 表示事件“两次抛出的结果相同”，$\{1,2,3\}$ 或 $\{(T,H),(H,T),(H,H)\}$ 表示事件“至少抛出一个正面”。

确定性和随机性之间并无绝对界限，确定的动力学系统有时也可以“表现出”随机性，即混沌 (chaos)*，其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”。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，还有所谓的“蝴蝶效应”——北京一只蝴蝶偶尔扇动了几下翅膀将引发数月后美国德克萨斯州的一场飓风。

例 1.2. 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 $f(x_0), f[f(x_0)], \cdots$ 的折线图。初始值的小扰动 0.0001 在长期迭代后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：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。

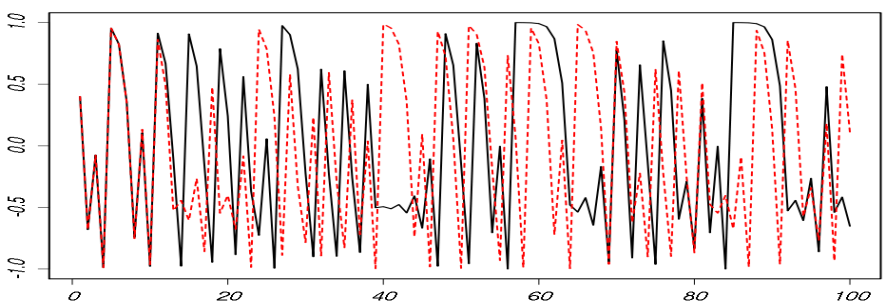


图 1.1: 初值的小扰动在经过一定时间的演化后引发了混沌系统的随机变化。

*对混沌理论和复杂系统感兴趣的读者可参阅科普著作 [34,71]。

例 1.3 (Arnold 变换)**.** 可逆的周期映射 $\Gamma : (x, y)^{\mathsf{T}} \mapsto (2x + y, x + y)^{\mathsf{T}} \mod 1$ 被称为 Arnold 变换。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。下图是俄国数学家 Vladimir Igorevich Arnold (1937-) 的头像经过 Arnold 变换 Γ 的数次迭代后的结果。

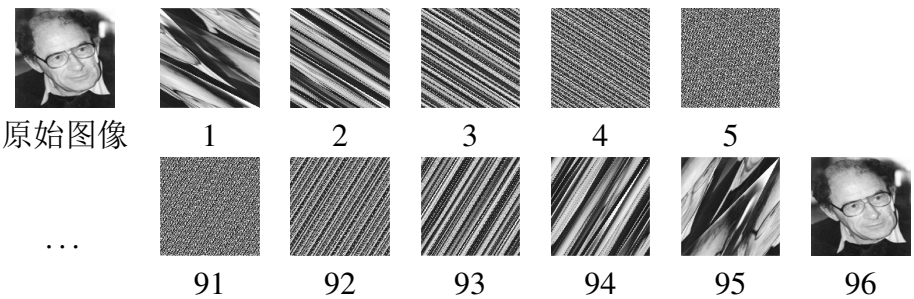
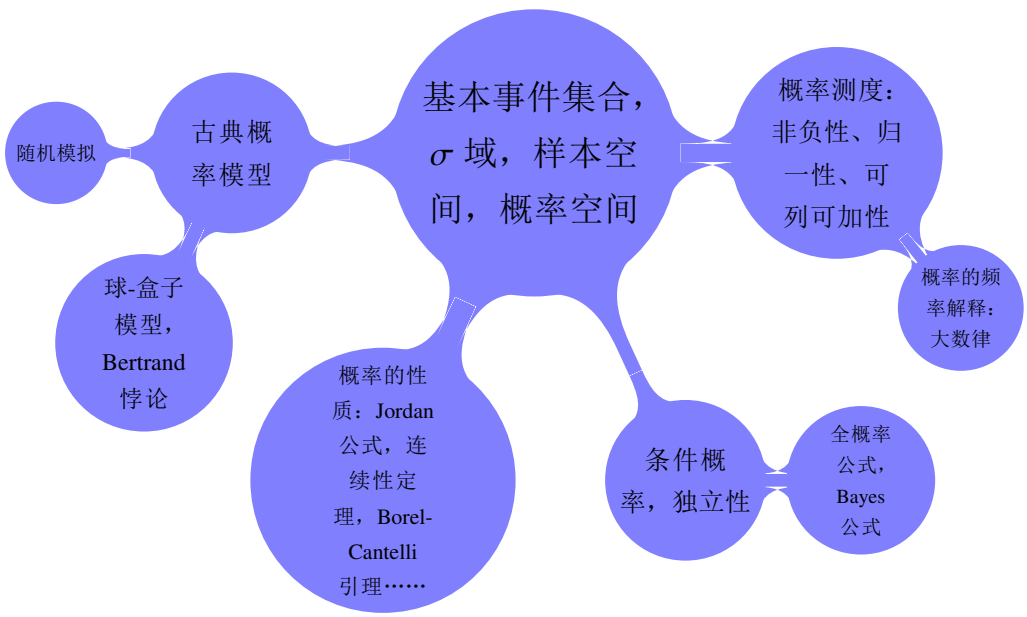


图 1.2: 原始图像大小为 128×128（像素），经过 96 次迭代后又重新得到原始图像，在这一过程中，该动力系统的似乎表现出随机性，其实不然。第 n 次迭代的结果总是第 $n-1$ 次迭代的结果按照固定的模式分割拼装而成。

练习 1.1. 举更多有关随机现象的例子，阅读 [6] 的附录《论随机性》。



1.1　古典概率模型

赌博的历史与人类文明史一样久远，可为何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对复杂的组合计算，我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和医生，也是个狂热赌徒。Cardano 在他的著作《论赌博游戏》(Liber de Ludo Aleae, 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 $i, j = 1, 2, \cdots, 6$，而不是 21 个无序对，并且每个有序对出现的机会都是等同的，即 1/36。Cardano 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著名的数学家 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，并用构成随机事件 A 的基本事件占

基本事件集合 Ω 几成来刻画 A 出现的机会，即 A 的概率 $\mathsf{P}(A)$。显然，

$$0 \le \mathsf{P}(A) \le 1，其中 A \subseteq \Omega \tag{1.1}$$

$$\mathsf{P}(\emptyset) = 0 并且 \mathsf{P}(\Omega) = 1 \tag{1.2}$$

例 1.5. 把标号为 a, b 的两个球随机放入两个盒子的所有可能的结果是 $\Omega = \{(ab|-), (a|b), (b|a), (-|ab)\}$。假设每个基本事件发生的机会等同，则出现空盒子的概率是 $1/2$。如果这两个球无标号或不可分辨 (indistinguishable)，则基本事件集合 $\Omega = \{(**|-), (*|*), (-|**)\}$，出现空盒子的概率是 $2/3$。

练习 1.2. 把三个球随机地放入三个盒子，球可分辨和不可分辨时各有多少个可能的结果？请读者仿照上例不厌其烦地列出具体结果。
答案：分别有 27 个和 10 个可能的结果。

例 1.6. 已知随机试验 $\mathcal{E}$ 由两个步骤组成：抛一次硬币，然后再掷一次骰子，则基本事件集合为 $\Omega = \Omega_1 \times \Omega_2$，其中 $\Omega_1 = \{H, T\}$ 和 $\Omega_2 = \{1, 2, \cdots, 6\}$ 分别是抛硬币和掷骰子的基本事件集合。推而广之，如果随机试验 $\mathcal{E}$ 的 n 个步骤依次为随机试验 $\mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_n$，所对应的基本事件集合分别为 $\Omega_1, \Omega_2, \cdots, \Omega_n$，则 $\mathcal{E}$ 的基本事件集合为

$$\begin{aligned}\Omega &= \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n \\ &= \left\{\left(\omega^{(1)}, \omega^{(2)}, \cdots, \omega^{(n)}\right) : \omega^{(k)} \in \Omega_k, k = 1, 2, \cdots, n\right\}\end{aligned} \tag{1.3}$$

如果基本事件集合 Ω_k 的势皆有限，不妨设 $|\Omega_k| = m_k < \infty$，则由式 (1.3) 得知基本事件集合 Ω 的势为

$$|\Omega| = \prod_{k=1}^{n} |\Omega_k| = \prod_{k=1}^{n} m_k \tag{1.4}$$

例如，连续抛一枚硬币 n 次，任一基本事件 $\omega \in \Omega$ 都是 H, T 组成的长度为 n 的序列。下面用 Maxima 具体描述一下 $n = 3$ 的情形，

```
(%i1) Omega: cartesian_product ({H, T}, {H, T}, {H, T});
(%o1) {[H, H, H], [H, H, T], [H, T, H], [H, T, T], [T, H, H], [T, H, T], [T, T,
       H], [T, T, T]}
(%i2) cardinality(Omega);
(%o2) 8
```

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*。

解决古典概率问题有时需要很高的技巧，本书并不强调这些技巧，而更愿把古典概率模型视为通向概率公理化的必经之路，一方面为了展示历史的原貌，另一方面也为了使一些基本概念的引入显得更自然些。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。*”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 [45,46]，也可参阅数学科普名著《数学——它的内容，方法和意义》 [12] 第十一章（由 Kolmogorov 撰写）。

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†，并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 随机模拟方法，借助此方法帮助读者粗略认识概率的频率解释。

学习目标 (1) 能给出随机试验的基本事件集合，并会用集合表示随机事件；(2) 构建古典概率模型解决实际问题，包括利用“球-盒子”模型描述问题、计算几何概率等；(3) 了解 Bertrand 悖论；(4) 理解 Monte Carlo 随机模拟的思想；(5) 尝试 R 语言的简单编程。

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

†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思考。

1.1.1 排列与组合

排列和组合的方法在中学数学里就有介绍，譬如 k 个球放入 n 个盒子 ($n \geq k$)，每个盒子至多装一个球。如果球是可分辨的，则共有 $A_n^k = n(n-1)\cdots(n-k+1)$ 种放法；如果球是不可分辨的，则共有 $C_n^k = A_n^k/k! = n!/[k!(n-k)!]$ 种放法。最常用的组合公式有

$$C_n^k = C_n^{n-k} \quad \text{和} \quad C_n^k = C_{n-1}^k + C_{n-1}^{k-1} \tag{1.5}$$

以及著名的 Stirling 公式（它的真正发现者是 A. de Moivre，见附录 C）

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}} = 1 \quad \text{或} \quad n! \sim \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n} \tag{1.6}$$

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 $\mathsf{P}(\{\omega\}) = 1/n$ 且 $\sum_{\omega\in\Omega}\mathsf{P}(\{\omega\}) = 1$。随机事件 $A \subseteq \Omega$ 的概率是

$$\mathsf{P}(A) = \frac{|A|}{|\Omega|} = \frac{n_A}{n} = \frac{A \text{ 中基本事件的个数}}{\text{所有基本事件的个数}} \tag{1.7}$$

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

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 $|A|$ 和 $|\Omega|$，所需的技巧也仅限于此。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，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。

*假定这些球的大小、质地等物理特性都一样。如果要求球是可分辨的，可以通过球的颜色或标号对它们进行区分。有时候颜色和标号都要用到，譬如约定“n 个可辨的黑球”意味着 n 个黑球的标号分别为 $1, 2, \cdots, n$。

†当提到“n 个盒子”，读者可以缺省地认为盒子的标号分别 $1, 2, \cdots, n$。

例 1.7 (生日问题). 假设一年有 365 天，随机选取 n 个人 ($n \le 365$)，问至少两人生日相同的概率是多少？

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 $1, 2, \cdots, n$ 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。而基本事件就是把 n 个球放入 N 个盒子，每个球都有 N 种选择，故共有 $|\Omega| = N^n$ 种放法。所以 $\mathsf{P}(A^c) = A_N^n/N^n$，进而原问题的解是 $\mathsf{P}(A) = 1 - \mathsf{P}(A^c) = 1 - A_N^n/N^n$。下面利用 GNU 统计软件 R 计算人数取 $n = 1, 2, \cdots, 50$ 所对应的 $\mathsf{P}(A)$ 并画出竖线图。

```
## 生日问题: n <= 365 个人中至少两人生日相同的概率?
## 输出: n 个人当中至少两人生日相同的概率 P(A)
N <- 365                        # 一年的天数
n <- 50                         # 选取的人数
InitProb  <- matrix(1,n,1)      # 一个 n 维的列向量的初始化

## 计算 n 个人当中没有人生日相同的概率
for (i in 2:n){
  InitProb[i] <- InitProb[i-1] * (N-i+1)/N
}
Prob <- 1 - InitProb            # 生日问题的解, 输出一个 n 维列向量
idx  <- n - sum(Prob>0.5) + 1 # 概率大于 50% 所需最少人数

## 绘图参数设定
op  <- par(lwd = "2",                          # 线的宽度
           font = "3", font.axis = "3",        # 选择字体
           cex = 1.2, cex.axis = 1.2,          # 字体大小
           mar = par("mar")+c(0,0,0,0))        # 输出图形边白宽度的设定

## 对生日问题的结果进行绘图
plot(Prob, xlab = "人 数 n", ylab = "P(A)", type = "h")
points(Prob, type = "l")                       # n 个结果之间连线

## 标记出临界值: 至少两人生日相同的概率超过 50% 所需的最少人数
points(idx, Prob[idx], type = "p")             # 标记出临界值的位置
points(1:idx, rep(Prob[idx], idx), type = "l")
legend(idx, Prob[idx], xjust = 1, yjust = 0,  # 打印出临界值及其对应的概率值
       paste("n=", idx, ", P(A)=", floor(1000 * Prob[idx])/1000),
       bg = "NA", box.col = "NA")
```

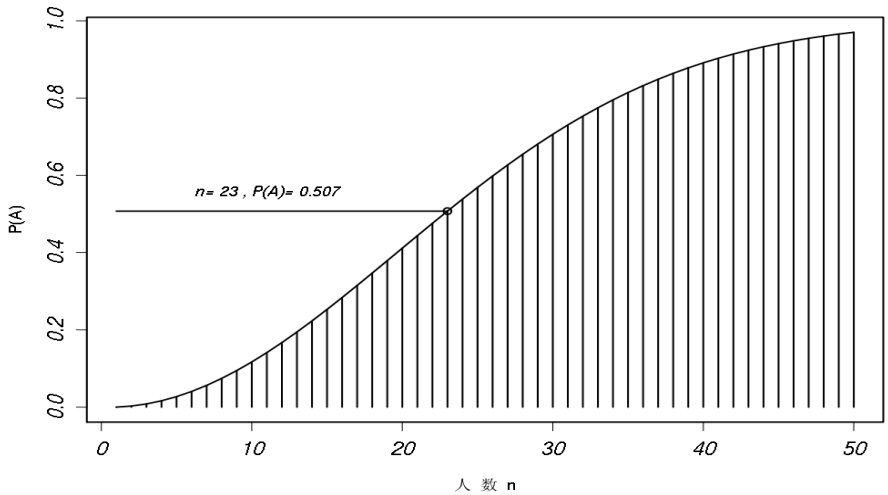


图 1.3: 随机选取的 n 个人中至少两人生日相同的概率。程序求得临界值 $n = 23$ 是使得 $\mathsf{P}(A) > 50\%$ 所需最少的人数。对于 $n = 50$，概率 $\mathsf{P}(A)$ 飙升至 97%！即随机挑选的 50 人中至少两人生日相同的概率为 97%。联想到组合数学里的抽屉原则，这个结果多多少少有点出乎意料。

例 1.8. 试证明：对于任一自然数 $n \in \mathbb{N}$，皆有

$$\sum_{k=0}^{n} \left(C_n^k\right)^2 = C_{2n}^n \tag{1.8}$$

证明. 假设盒子里有 n 个可辨的黑球和 n 个可辨的白球，从中选出 n 个球共有 C_{2n}^n 种不同的选法，等价于“先从 n 个黑球中选 k 个黑球出来，再从 n 个白球中选 $n-k$ 个白球出来”，其中 $k = 0, 1, \cdots, n$。 □

例 1.9. 盒子里有 m 个球，包括 mp 个黑球和 mq 个白球，其中 $p, q \in (0, 1)$ 满足 $p + q = 1$。一次随机抽取一个球，有放回地抽取 n 次，恰有 k 次抽到黑球的概率为 $\mathsf{P}(k) = C_n^k p^k q^{n-k}$。

证明. 抽取一次的基本事件集合是 $\Omega = \{黑_1, \cdots, 黑_{mp}, 白_1, \cdots, 白_{mq}\}$，抽取 n 次的基本事件集合是 $\Omega^n = \underbrace{\Omega \times \cdots \times \Omega}_{n\text{ 个}}$，所以 $|\Omega^n| = m^n$。

记集合 $\Omega_{黑} = \{黑_1, \cdots, 黑_{mp}\}$ 且 $\Omega_{白} = \{白_1, \cdots, 白_{mq}\}$。有序 n 元组 $\boldsymbol{t} = (t_1, \cdots, t_n)$ 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 $\boldsymbol{t}$。“恰有 k 次黑球”的事件是 $A_k = \bigcup_{\boldsymbol{t}} \Omega_{t_1} \times \cdots \times \Omega_{t_n}$，

$$\mathsf{P}(k) = \mathsf{P}(A_k) = \frac{|A_k|}{|\Omega^n|} = \frac{C_n^k (mp)^k (mq)^{n-k}}{m^n} = C_n^k p^k q^{n-k} \tag{1.9}$$

显然，$\mathsf{P}(k)$ 就是 $\sum_{k=0}^{n} \mathsf{P}(k) = 1 = (p+q)^n$ 二项式展开的第 $k+1$ 项。 □

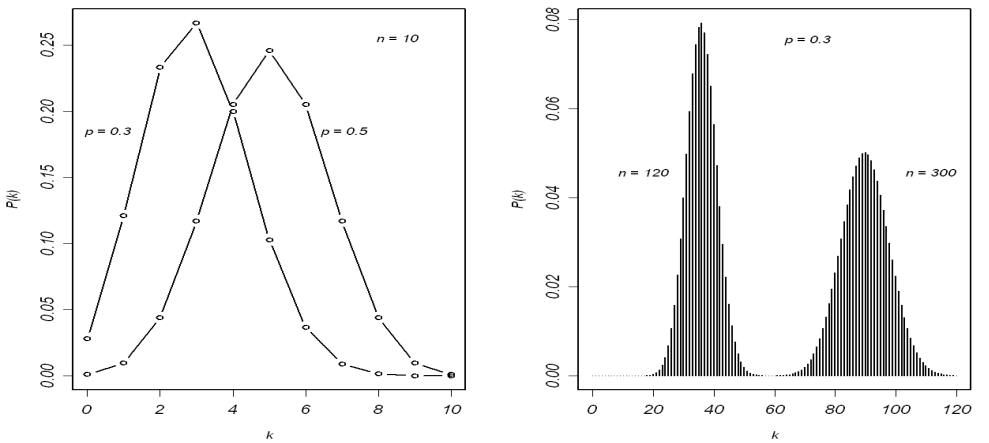


图 1.4: 对 $\mathsf{P}(k) = C_n^k p^k (1-p)^{n-k}$ 的直观认识：若 $p = 0.5$，概率 $\mathsf{P}(k)$ 关于 np 是对称的；否则，$\mathsf{P}(k)$ 是非对称的。如左图所示，$n = 10, p = 0.3$ 和 0.5 所对应的 $\mathsf{P}(k)$ 的折线图。但当 n 很大时，即便 $p \neq 0.5$，关于 np，概率 $\mathsf{P}(k)$ 也呈现出很好的对称性，见右边的竖线图。这是为什么呢？等读者学完第五章的 de Moivre-Laplace 中心极限定理就明白其中的道理了。

例 1.10. 已知 N 件产品中有 n 件次品（其余为正品），令 A_k 表示事件“在随机抽取的 $m \leq n$ 件产品恰有 k 件次品”，试求概率 $\mathsf{P}(A_k)$ =?

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 Ω 的势。从 n 个黑球中选出 k 个有 C_n^k 种选法，剩下的 $N-n$ 个白球中选出 $m-k$ 个有 C_{N-n}^{m-k} 种选法，因此所要求的概率为

$$\mathsf{P}(A_k) = \frac{C_n^k C_{N-n}^{m-k}}{C_N^m} \tag{1.10}$$

由 $\sum_{k=0}^{m} \mathsf{P}(A_k) = \sum_{k=0}^{m} \sharp(A_k)/\sharp(\Omega) = 1$ 这一事实可“顺手牵羊”证得

$$\sum_{k=0}^{m} C_n^k C_{N-n}^{m-k} = C_N^m \tag{1.11}$$

利用 Maxima 验证一下上式，毫无疑问结果是正确的。

```
(%i1) load("simplify_sum") $
(%i2) assume(n>m, N>n+m) $
(%i3) simplify_sum( sum (binomial(n,k) * binomial(N-n,m-k), k, 0, m));
                                  N (N - 1)!
(%o3)                             -----------
                                  m! (N - m)!
```

例 1.11 (统计物理模型)**.** 有 n 个粒子处于 N 个不同的能级上 ($n \le N$)，请计算指定的 n 个能级上各有一个粒子的概率 p_1，和恰好有 n 个能级各有一个粒子的概率 p_2。这个问题相当于将 n 个球放入 N 个盒子中，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，即可以对它们进行编号，则 $p_1 = n!/N^n, p_2 = A_N^n/N^n$（与例 1.7 相同）。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。

- 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 $p_1 = 1/C_N^n, p_2 = 1$。该模型适用于费米子，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，可以做这样的随机试验：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，然后两个盒子之间的球都将放入左边的盒子。这样，共有 $b = C_{N+n-1}^{N-1} = C_{N+n-1}^{n}$ 种放法。于是，$p_1 = 1/b = 1/C_{N+n-1}^{n}, p_2 = C_N^n/b = C_N^n/C_{N+n-1}^{n}$。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。

练习 1.3. 用 R 语言实现例 1.11 描述的三个物理模型。

二十世纪二十年代，量子力学发展迅速。Albert Einstein (1879-1955) 不满意量子力学的 Copenhagen 解释，与 Niels Bohr (1885-1962) 展开多年的论战。1926 年，Einstein 在给 Max Born (1882-1970) 的信中说道，“无论如何，我确信上帝不掷骰子”，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* (uncertainty principle) 只是权宜之计。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大多数物理学家接受并成为量子力学的基石之一。英国物理学家 Stephen Hawking (1942-) 说，“上帝不但掷骰子，有时还把骰子掷到它们无法被看到的地方。”

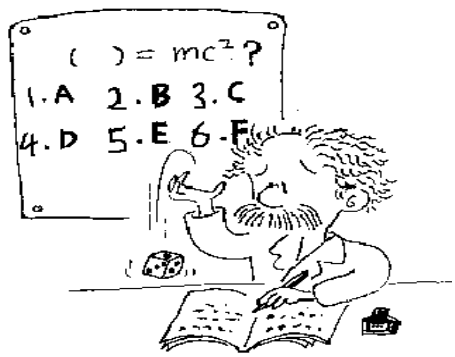


图 1.5: 上帝不掷骰子，但我们掷。

注记 1.1. Hawking 的这番话意味着大自然并没有把所有的随机现象都展示在人类面前。既便如此，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，下面举一个例子来说明：有 n 个盒子，标号分别为 $1,2,\cdots,n$，每个盒子里都有 m 种不同颜色的球。甲、乙二人玩一个游戏，游戏规则和已知条件如下。

游戏规则： 甲随机地选取一个“初始”盒子，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；按照转移概率再选取下一个盒子，重复刚才的过程……。游戏要求甲所选盒子的序列对乙来说是不可观察的，乙能得到的观察数据就是甲汇报的球的颜色序列。试问：乙如何猜出颜色序列所对应的盒子序列？

*在很多文献中也译为“测不准原理”。

已知条件: (1) 各种颜色的球在每个盒子中所占的比例；(2) 盒子 k 第一次被选中的概率为 p_k，满足 $\sum_{k=1}^{n} p_k = 1$；(3) 当前从盒子 i 里摸球，下次从盒子 j 里摸球的转移概率为 p_{ij}，满足 $\sum_{j=1}^{n} p_{ij} = 1$。

解决之道: 可以利用隐 Markov 模型的 Viterbi 算法（细节见 §12.2）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、词性标注、蛋白质结构预测等实际问题。

例 1.12. 一个盒子里有 10 个球，编号分别为 $1, 2, \cdots, 10$。一次抽取一球，有放回地抽取 6 次，试求下列事件的概率：

- A：至少出现两次 5 号球。
- A_n：所抽编号之和为 $n \in \mathbb{N}$。

解. 参考例 1.6，共有 10^6 个基本事件。$\mathsf{P}(A) = \sum_{k=2}^{6} C_6^k \cdot 9^{6-k}/10^6 = 0.114265$。事件 A_n 所含基本事件的个数等于不定方程 $\sum_{j=1}^{6} x_j = n$ 满足约束条件 $1 \le x_j \le 10$ 的解的个数（其中 x_j 表示第 j 次所抽取编号），即多项式 $(x + x^2 + \cdots + x^{10})^6$ 中项 x^n 的系数，可用 Maxima 求之。

```
(%i1) polynomial: expand((sum(x^k,k,1,10))^6) $
(%i2) coeff(polynomial, x, 25) ;
(%o2) 30492
```

故 $\mathsf{P}(A_{25}) = 0.030492$。当 n 取 $6, 7, \cdots, 60$ 之外的自然数时 $\mathsf{P}(A_n) = 0$。

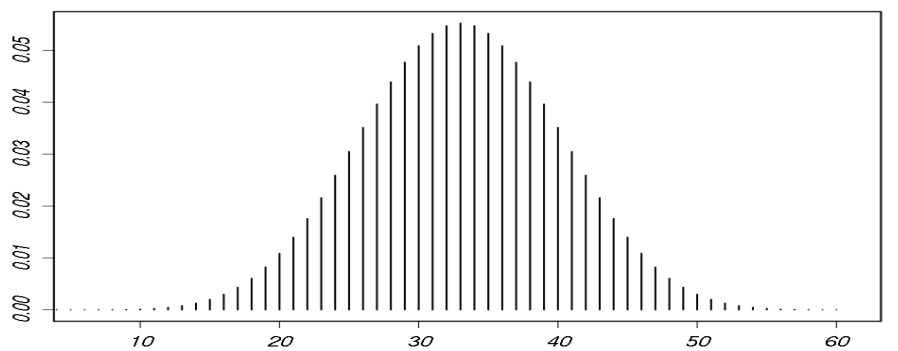


图 1.6: 所抽 6 个球的编号之和为 $6, 7, \cdots, 60$ 的概率呈现出很好的对称性。

1.1.2　几何概率

古典概率并不限于有限的基本事件集合，譬如下面的投钉问题：往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 $A \subseteq \Omega$（右图中阴影部分）的概率？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。为求得钉落于子区域 A 的概率 $\mathsf{P}(A)$，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 $A_{内}(k) \subseteq A$，其中小正方形的个数记为 $N_{内}(k)$；把那些并集恰好覆盖住 A 的小正方形组成的区域记为 $A_{外}(k) \supseteq A$，其中小正方形的个数记为 $N_{外}(k)$，显然有

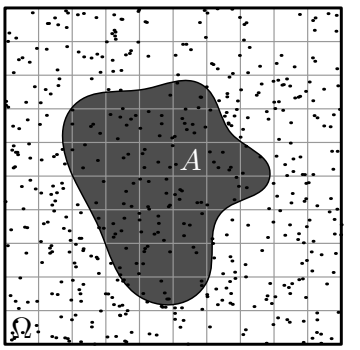


图 1.7: 投钉问题：均匀投钉于正方形 Ω 上，钉落于有面积的子区域 A 内的概率是 A 与 Ω 的面积之比。

$$\frac{N_{内}(k)}{N(k)} = \mathsf{P}\left[A_{内}(k)\right] \leq \mathsf{P}(A) \leq \mathsf{P}\left[A_{外}(k)\right] = \frac{N_{外}(k)}{N(k)} \tag{1.12}$$

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

$$\lim_{k\to\infty} \frac{N_{内}(k)}{N(k)} = \lim_{k\to\infty} \frac{N_{外}(k)}{N(k)} = \frac{\mu(A)}{\mu(\Omega)} \tag{1.13}$$

所以钉落于 A 的概率*为

$$\mathsf{P}(A) = \frac{\mu(A)}{\mu(\Omega)} \tag{1.14}$$

*§5.1 将介绍 Borel 强大数律，该结果保证通过大量的投钉试验，落于区域 A 的钉数 m 与落于 Ω 上的总钉数 n 之比 m/n 可作为 $\mathsf{P}(A)$ 的估计值。现在考虑投钉问题的反问题：区域 Ω 的面积 $\mu(\Omega)$ 已知，子区域 $A \subset \Omega$（阴影部分）的面积 $\mu(A)$ 未知，如何近似求解？由式 (1.14) 显然可以用 $m/n \cdot \mu(\Omega)$ 来逼进 $\mu(A)$。这种通过大量随机试验给出数值近似解的方法称为 Monte Carlo 方法，将在下一小节继续介绍。

依据具体情况 $\mu(\cdot)$ 有时也可以表示长度、体积等度量。几何概率问题的解决步骤一般为：(1) 确定基本事件集合 $\Omega \subset \mathbb{R}^n$；(2) 刻画出随机事件 $A \subseteq \Omega$；(3) 通过式 (1.14) 计算 A 的概率。请看下例，

例 1.13 (见面问题)**.** 甲乙二人约定 11 点至 12 点在某地见面，等候 20 分钟对方若不来就离开。如果两人可在 11 点至 12 点的任何时刻到达该地，并且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 $|x-y| \le 20$，所以用 $A = \{(x,y) \in \Omega : |x-y| \le 20\}$（右图中的阴影部分）来表示随机事件“两人见面”。由式 (1.14) 计算得到 $\mathsf{P}(A) = \mu(A)/\mu(\Omega) = 1-(2/3)^2 = 5/9$，其中 $\mu(A)$ 和 $\mu(\Omega)$ 分别为区域 A 和 Ω 的面积。

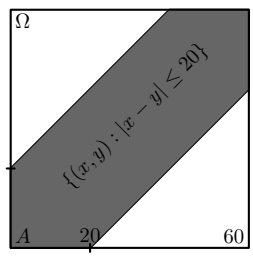


图 1.8: 阴影部分表示两人能见面。

练习 1.4. $0 < k < 1$ 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 1.14. 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 (1.14) 得 $\mathsf{P}(A) = 1/4$。请读者顺着上述结果继续论证，三个点的连线“构成直角三角形”的概率为 0，“构成钝角三角形”的概率为 3/4。

例 1.15. 已知二次多项式 $f(x) = x^2 + 2ax + b$，其中系数满足 $|a| \le A, |b| \le B$。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 $b \le a^2$，分下面两种情况讨论。

1. 如果 $B \geq A^2$，则 $p = (2AB + 2\int_0^A a^2 da)/\mu(\Omega) = 1/2 + A^2/(6B)$。

2. 如果 $B \leq A^2$，则 $p = (4AB - 2\int_0^B \sqrt{b}db)/\mu(\Omega) = 1 - \sqrt{B}/(3A)$。

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子。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 1.16** (Bertrand 悖论)**.** 在单位圆内随机取一条弦，试求：该弦长度超过单位圆内接正三角形边长 $\sqrt{3}$ 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。

1. 由于对称性，不妨预设了弦的方向，则有唯一直径垂直于该方向。只有交直径于 1/4 分点 α 与 3/4 分点 β 之间，才能使弦长超过内接正三角形的边长，见图 1.9 之 (1)。所以，$p = 1/2$。

2. 由于对称性，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 1.9 之 (2) 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. 弦的中点唯一决定弦的位置，想让弦长满足条件，它的中点 M 必须落于半径为 1/2 的同心圆内，见图 1.9 之 (3)。所以，$p = 1/4$。

---

*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。

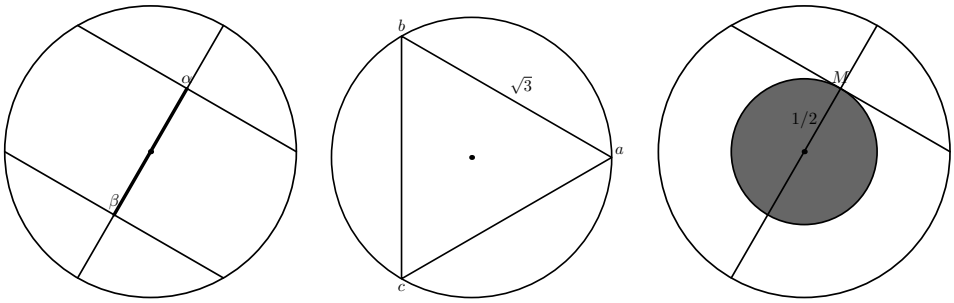


图 1.9: Bertrand 问题三种不同解法的直观图示。所有解法听起来都蛮有道理，但答案只能有一个，到底哪个解法是正确的呢？

Bertrand 悖论是几何概率的隐患，它的“症结”出在哪里呢？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，它们都是对的。上述三种解法的取弦方式分别是：

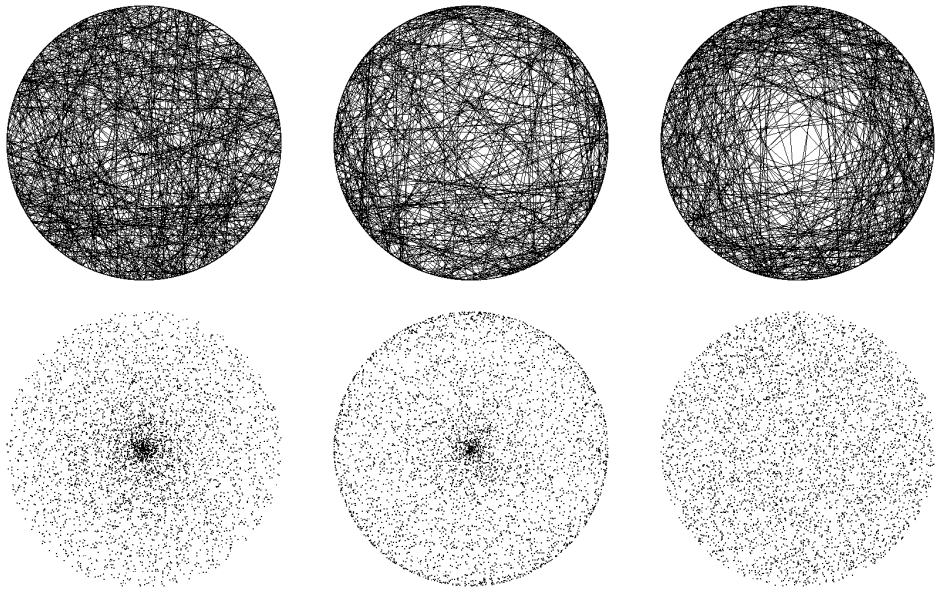
1. 按照方向区间 $[0, 2\pi)$ 上的均匀分布*随机选定一个角度 θ，再按照 $(0, 1)$ 上的均匀分布随机选定一个极径长度 ρ，构造过点 $(\rho\cos\theta, \rho\sin\theta)$ 的弦。

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（即两个点），构造以此二点为端点的弦。

3. 按照圆盘 $\{(x, y) : x^2 + y^2 < 1\}$ 上的均匀分布选定一点，构造以此点为中点的弦。

图 1.10 是上述三种不同取弦方式所随机生成的 $n = 5000$ 条弦及相应的中点（为了能清楚显示，图中只画出了前 $m = 500$ 条弦），其 R 源码见附录 B。请读者观察这些随机弦的中点的分布有什么不同，选出你认为最合理的选弦方式。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，再次深入讨论它的成因。

*读者可将区域 $A = [0, 2\pi)$ 上的均匀分布理解为 A 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。



(1)　(2)　(3)

图 1.10: Bertrand 问题三种不同的取弦方式所随机生成的 $n = 5000$ 条弦及其中点（得到此结果的 R 源码见附录 B），这些中点在圆心周围的疏密程度显然是不同的，为什么呢？

考察半径为 1/2 的同心圆内随机弦中点所占的比例，以第二种选弦方式为例，应为 1/3，理由见右图。请读者验证对于第一和第三种选弦方式，该比例分别为 1/2 和 1/4。这解释了图 1.10 中不同选弦方式下随机弦的中点在圆心周围的稀疏程度为何是不同的。有趣的是，这些比例恰为满足 Bertrand 问题条件的随机弦的概率。

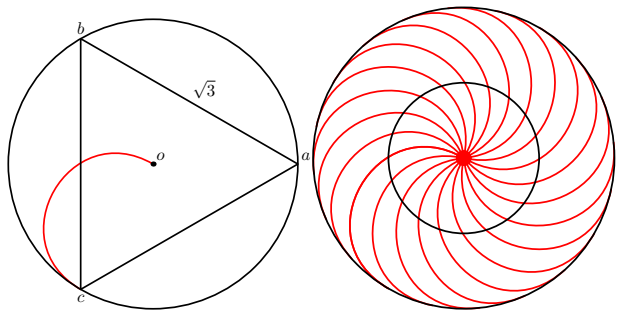


图 1.11: 第二种选弦方式下，以 oc 为直径的半圆周 $\widehat{oc}$ 被线段 bc 分割为两段弧，其上随机弦中点的比例是 1 : 2。以点 o 为中心旋转弧 $\widehat{oc}$ 一周，除了圆心 o，圆内任一点均被“扫描”过一次。

下面，对这两个小节所讨论的古典概率的主要性质做一个概要总结，这些性质对概率的公理化将起到一定的启发作用。

性质 1.2. 对于某个（离散的或连续的）古典概率问题，Ω 是它的基本事件集合，利用式 (1.7) 或式 (1.14) 不难验证概率满足下面的性质。

1. 非负性：对于事件 $A \subseteq \Omega$，总有 $\mathsf{P}(A) \geq 0$。

2. 归一性：必然事件 Ω 的概率等于 1，即 $\mathsf{P}(\Omega) = 1$。

3. 可加性：如果随机事件 $A, B \subseteq \Omega$ 满足 $A \cap B = \emptyset$，则称它们是*互斥的或不相容的*，一定有 $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B)$。如果随机事件 $A_1, A_2, \cdots, A_n$ 两两互斥，则概率满足下面的“有限可加性”。

$$\mathsf{P}\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mathsf{P}(A_k) \tag{1.15}$$

推而广之，如果可数（或称可列）个随机事件 $A_1, A_2, \cdots, A_k, \cdots$ 两两互斥，则概率满足下面的“可列可加性”。

$$\mathsf{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathsf{P}(A_k) \tag{1.16}$$

譬如在例 1.13 中，可以把区域 $A = \{(x, y) \in \Omega : |x - y| \leq 20\}$ 划分为可数个子块，$A_k = \{(x, y) \in A : 60/(k+1) < x \leq 60/k, 60/(k+1) < y \leq 60/k\}, k = 1, 2, \cdots$，显然 $A_1, A_2, \cdots$ 两两互斥且 $\bigcup_{k=1}^{\infty} A_k = A$，

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

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。数学上必须以测度论为基础，才能建立起概率论的大厦。对测度论感兴趣的读者可以参阅 Halmos 的《测度论》 [47]。

1.1.3* Monte Carlo 方法

法国博物学家、数学家、作家 Comte de Buffon (1707-1788) 是一位百科全书式的学者，代表作是 44 卷的巨著《自然史》(Histoire naturelle, générale et particulière)。1733 年 Buffon 曾考虑过下面的投针试验，并于 1777 年在《自然史》的增刊中纂文《或然算术试验》(Essai d'arithmétique morale) 给出了解，这是最早对 Monte Carlo 方法和几何概率的研究。

例 1.17 (Buffon 投针试验)**.** 随机地往一个宽度为 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？

解. 令针的中点 M 到两平行线的最短距离为 y，夹角为 φ，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。针与平行线相交当且仅当 $y \le L/2 \sin\varphi$。而针的位置 (φ, y) 的变化范围是 $\Omega = [0,\pi]\times[0, D/2]$，其面积为 $D\pi/2$。于是，针与平行线相交的概率是

$$p = \frac{1}{D\pi/2}\int_0^{\pi}\frac{L}{2}\sin\varphi d\varphi = \frac{2L}{D\pi} \tag{1.17}$$

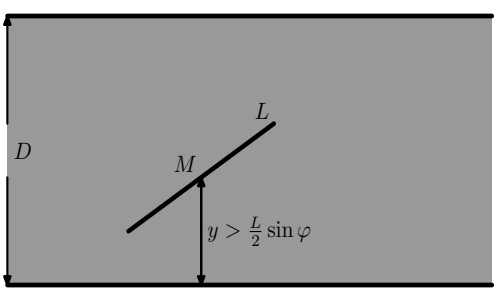


图 1.12: Buffon 投针试验中针的状态。

✎我们找到了一个近似计算 圆周率 π 的方法：通过足够多次的投针试验，得到针与平行线相交的频率（譬如在 n 次投针中有 m 次交于平行线)，并把该频率 m/n 用作 p 的估计值，根据式 (1.17) 有 $\pi \approx \hat{\pi} = 2Ln/(Dm)$。1901 年，意大利数学家 Mario Lazzarini 做了

Buffon 投针试验 ($D = 1, L = 5/6$)，看是否能得到 π 的估计值 $\hat{\pi} = 355/113 \approx 3.1415929$，即祖冲之的密率*。把问题逆过来想，如果能得到密率，相交次数 m 与投次 n 显然有 $m = 113n/213$。于是，这位“事后诸葛亮”以 213 次投针为一批，共投了 $3408 = 213 \times 16$ 次针，终于以 $1808 = 113 \times 16$ 次相交“幸运地”得到了密率。

例 1.18. 实质上，Buffon 投针试验相当于往区域 $\Omega = [0, \pi] \times [0, D/2]$ 内投钉，考察落于子区域 $0 \le y \le L/2 \sin\varphi$ 的概率的随机试验。

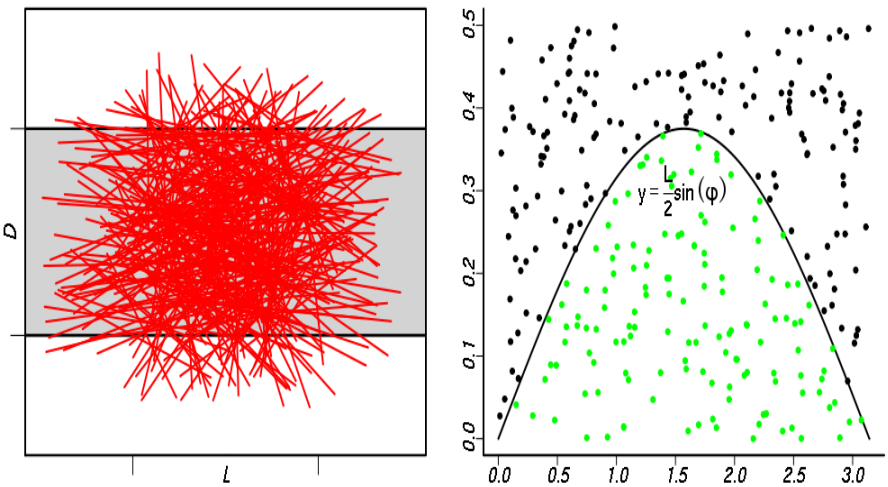


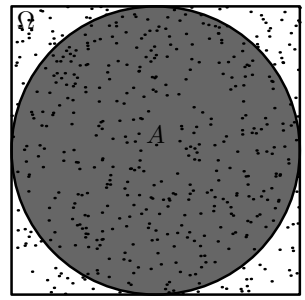
图 1.13: Buffon 投针试验：设定 $D = 1, L = 0.75$，投针 $n = 300$ 次。

例 1.19. 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。

解. 模拟试验产生 n 个“钉”均匀地分布于区域 $\Omega = [-1/2, 1/2] \times [-1/2, 1/2]$ 上，判断它们中哪些落到内接圆盘 A 上，其个数设为 m 个，π 的估计值即为 $\hat{\pi} = 4m/n$。

*祖冲之 (429-500)，字文远，我国南北朝时期著名的数学家和天文学家。据《隋书 · 律历志》记载，祖冲之得到圆周率的近似值“圆径一百一十三，圆周三百五十五”，这是最早的圆周率渐近分数表示，精确到第六位小数，该发现领先世界一千年。祖冲之在其专著《缀术》中描述了如何得到这一结果，但因当时“学官莫能究其深奥，是故废而不理”而失传，是数学史中的一件憾事。数学家华罗庚 (1910-1985) 曾撰文《从祖冲之的圆周率谈起》 [2] 来纪念这位伟大的古代学者。

共进行了三次模拟试验，每次模拟试验的规模都是 10000 次投钉，试验所得圆周率的估计值 $\hat{\pi}$ 如下图所示。有时 $\hat{\pi}$ 很接近真实值，但由于没有判定何时停止模拟的算法，投钉次数并非多多益善。在不知道真实值的前提下，需要用最优停止理论来指导模拟试验的规模 [24]，该话题不在本书范围之内。



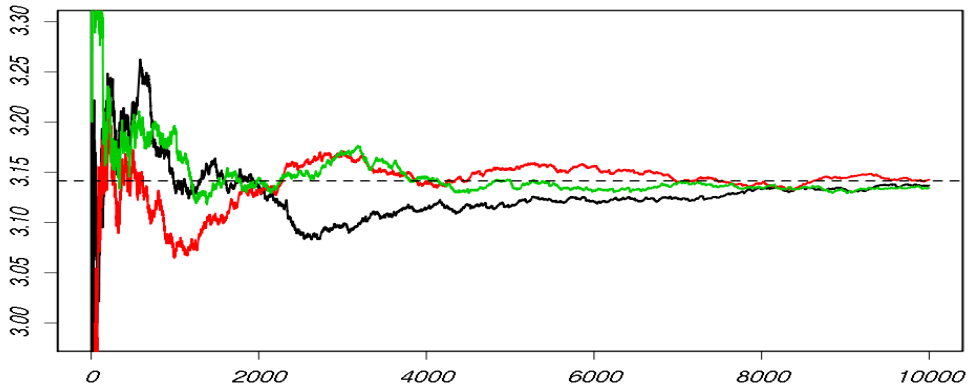


图 1.14: 利用投钉试验估算圆周率（R 源码见附录 B）。此方法贵在启发性，其实用性因收敛速度慢、稳健性欠佳而远不如用级数法估算圆周率。

问题 1.1. 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？请说明理由。

例 1.20. 通过实例人们了解到，利用 Buffon 投针试验所得的圆周率估计值围绕在真实值周围，它们的误差有没有规律可言？为此，我们独立做了三批 Buffon 投针试验，每批试验投针 $n = 10000$ 次。估计值 $\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_n$ 围绕着真实值 π 上下振荡，R 源码见附录 B。

为了直观地考察误差 $\hat{\pi}_k - \pi$，计算 $\hat{\pi}_k$ 落于开区间 $(\pi - \epsilon, \pi + \epsilon)$ 内的频率，即比例 $\sharp\{\hat{\pi}_k : |\hat{\pi}_k - \pi| < \epsilon, k = 1, 2, \cdots, n\}/n$。随着投针次数 n 的增加，该比例越来越大，说明越来越多的估计值集中在真实值的“不远处”。该模拟试验取 $\epsilon = 0.08$，试验结果如图 1.15 所示。

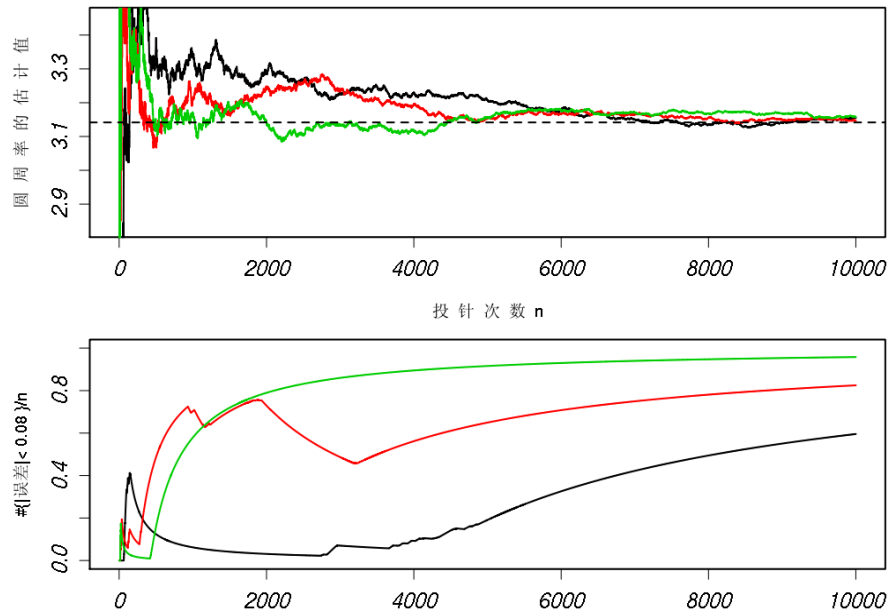


图 1.15: 利用 Buffon 投针试验估算圆周率（上图），所得估计值 $\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_n$ 落于开区间 $(\pi - \epsilon, \pi + \epsilon)$ 内的频率随着 n 的增大也越来越大（下图）。

定义 1.2. 像本小节中的几个例子，利用大量的随机抽样来解决实际问题的计算方法统称 Monte Carlo *方法*，也称*随机模拟方法*。

出于计算上的需要人们对随机模拟方法进行了系统的研究，最初的一些算法是由 Stanislaw 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 Monaco 最有名的赌场 Monte Carlo 来命名它，据说 Ulam 的叔叔经常在那里出没。有关随机模拟技术及其算法的详细介绍见第十四章。

*几乎所有的编程语言都自带有随机数产生器，特别是 [0, 1] 上的均匀分布，见第四章。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，以及对伪随机数的统计检验 [55]。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。

†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)。

1.2　概率论的公理化

起源于古希腊，公理化方法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）

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式几何公理体系。“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。

1933 年，著名的苏联数学家 Andrey Nikolaevich Kolmogorov (1903-1987) 出版了专著《概率论基础》[56]，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，由此蓬勃发展起来的概率论已成为经典数理统计学的基础。与此同时，也有一些学者致力于非传统概率论的研究，如贝叶斯学派的概率论。本书的绝大多数内容都是基于 Kolmogorov 公理体系的，对贝叶斯概率论将在第十一章予以介绍。

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。我们约定

- 在不引起歧义的情况下，$A\cap B, A\backslash B$ 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 $\mathscr{S}, \mathfrak{B}$ 等。

性质 1.3 (de Morgan 律)**.** 令 $\mathfrak{A}$ 是某些集合组成的类，则

$$\left(\bigcup_{A\in\mathfrak{A}} A\right)^c = \bigcap_{A\in\mathfrak{A}} A^c \quad \text{且} \quad \left(\bigcap_{A\in\mathfrak{A}} A\right)^c = \bigcup_{A\in\mathfrak{A}} A^c \tag{1.18}$$

性质 1.4. 已知 $A_1, A_2, \cdots, A_n, \cdots$ 是一个集合的序列，则

$$\text{元素 } a \text{ 属于无穷多个 } A_n \quad \Leftrightarrow \quad a \in \bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty} A_n \tag{1.19}$$

$$\text{元素 } a \text{ 仅不属于有限多个 } A_n \quad \Leftrightarrow \quad a \in \bigcup_{k=1}^{\infty}\bigcap_{n=k}^{\infty} A_n \tag{1.20}$$

证明. 往证式 (1.20)：$\exists k\in\mathbb{N}$ 使得 $n\geq k$ 时 $a\in A_n$，所以 $a\in\bigcup_{k=1}^{\infty}\bigcap_{n=k}^{\infty} A_n$；

反之亦然。式 (1.19) 留作练习或参考引理 1.1 的证明。 □

有时将 $\bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty} A_n$ 和 $\bigcup_{k=1}^{\infty}\bigcap_{n=k}^{\infty} A_n$ 分别简记为 $\overline{\lim_{n\to\infty}} A_n$ 和 $\underline{\lim_{n\to\infty}} A_n$，显然

$$\underline{\lim_{n\to\infty}} A_n \subseteq \overline{\lim_{n\to\infty}} A_n \tag{1.21}$$

定义 1.3 (划分)**.** 给定集合 A，$\{A_j : A_j \subseteq A, j = 1, 2, \cdots\}$ 称为 A 的一个划分当且仅当 A_j 两两不交且 $\bigcup_{j=1}^{\infty} A_j = A$。

性质 1.5. 集合 A 上的划分与 A 上的等价关系是一一对应的。例如，

```
(%i1) equiv_classes ({1, 2, 3, 4, 5, 6, 7},     /* 等价关系: 模 3 同余 */
                     lambda ([x, y], remainder (x - y, 3) = 0));
(%o1) {{1, 4, 7}, {2, 5}, {3, 6}}
```

问题 1.2. 如果集合 A 的势为 $n < \infty$，它有多少个不同的划分？

解. 势为 n 的集合所有不同划分的个数为第 n 个 Bell 数，记作 B_n，满足递归关系 $B_{n+1} = \sum_{k=0}^{n} C_n^k B_k$。下面用 Maxima 列出 $B_0, B_1, \cdots, B_{18}$。

```
(%i1) makelist (belln (i), i, 0, 18);
(%o1) [1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, 678570, 4213597,
       27644437, 190899322, 1382958545, 10480142147, 82864869804, 682076806159]
```

本节内容 出于集合运算的封闭性要求，第一小节利用 σ 域定义了样本空间，明确说明了什么是随机事件。在 Kolmogorov 公理体系里，概率测度是定义在样本空间上的实值集函数 (set function)，满足非负性、归一性和可列可加性，“样本空间+概率测度=概率空间”是第二小节的主题，我们从概率空间角度回顾了 Bertrand 悖论。例 1.34 帮助读者加深理解概率的频率解释。第三小节描述了概率的一些重要的性质，特别是概率的连续性定理。

学习目标 (1) 理解 σ 域、样本空间、概率测度等关键概念；(2) 熟练掌握概率的性质，如 Jordan 公式、概率的连续性定理等。(3) 初步了解 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 $\mathscr{P}(\Omega)$。已知 $\mathscr{S} \subseteq 2^\Omega$ 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、差等集合运算。如果 $\mathscr{S}$ 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 $\mathscr{S}$ 上的集合运算就没有多大意义。

定义 1.4. 设 $\mathscr{F}$ 是非空集合 Ω 的某些子集构成的类，如果 $\mathscr{F}$ 中任意两个元素的交集、并集、差集都仍在 $\mathscr{F}$ 中，则称 $\mathscr{F}$ 是 Ω 上的一个域。

定义 1.5 (σ 域)**.** 已知 $\mathscr{S}$ 是非空集合 Ω 上的一个域，它是 Ω 上的一个 σ 域 (σ-field) 或 σ 代数 (σ-algebra) 当且仅当 $\mathscr{S}$ 满足：
❶ 对补运算封闭，即对任意的 $A \in \mathscr{S}$，有 $A^c \in \mathscr{S}$。
❷ 对可数并运算封闭，即 $\forall A_1, A_2, \cdots \in \mathscr{S}$，有 $\bigcup_{j=1}^{\infty} A_j \in \mathscr{S}$。
二元组 $(\Omega, \mathscr{S})$ 称作可测空间*，$\mathscr{S}$ 中的元素称作可测集。

性质 1.6. 显然，2^Ω 是一个 σ 域。如果 $\mathscr{S}$ 是 Ω 上的一个 σ 域，从上述定义容易证明（请读者练习）$\mathscr{S}$ 满足以下性质。

- ❑ 含空集和全集，即 $\emptyset, \Omega \in \mathscr{S}$。该性质可替换定义 1.5 中的 ❶。
- ❑ 对可数交运算封闭，即 $\forall A_1, A_2, \cdots \in \mathscr{S}$，有 $\bigcap_{j=1}^{\infty} A_j \in \mathscr{S}$。
- ❑ 对差运算封闭，即对于任意的 $A, B \in \mathscr{S}$，有 $A \backslash B \in \mathscr{S}$。

如果 Ω 的某些子集组成的非空类 $\mathscr{A}$ 对集合运算不封闭怎么办呢？我们可以对 $\mathscr{A}$ 稍加扩充使之变成一个 σ 域，见下面的定义。

定义 1.6 (生成的 σ 域)**.** 已知类 $\mathscr{A} \subseteq 2^\Omega$ 非空，显然，所有包含 $\mathscr{A}$ 的 σ 域之交 $\mathscr{S}_0$ 仍是 σ 域，它是包含 $\mathscr{A}$ 的唯一最小的 σ 域，称之为由 $\mathscr{A}$ 生成的 σ 域，记作 $\mathscr{S}_0 = \sigma(\mathscr{A})$。

*所谓“空间”就是一个非空集合上赋予了某种结构。例如，非空集合 X 的某些子集构成的类 $\mathscr{T}$ 如果满足以下条件，则称 $(X, \mathscr{T})$ 是一个拓扑空间，称 $\mathscr{T}$ 为 X 上的一个拓扑 (topology)，称 $\mathscr{T}$ 中的元素为 $(X, \mathscr{T})$ 的开集。(1) $X, \emptyset \in \mathscr{T}$；(2) 若 $A, B \in \mathscr{T}$，则 $A \cap B \in \mathscr{T}$；(3) $\mathscr{T}$ 中任意多元素之并集合仍在 $\mathscr{T}$ 中。

例 1.21. 已知 $A \subset \Omega$ 非空，则 $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$。

例 1.22 (Borel σ 域)**.** 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 域被称为 Borel σ 域，简记作 $\mathfrak{B}_1$，它包含所有单点集合和所有的区间。这是因为，

$$\{x\} = \bigcap_{n=1}^{\infty}(x - 1/n, x], \quad (x, y) = (x, y] - \{y\}, \quad [x, y] = (x, y] + \{x\}$$

$$[x, y) = \{x\} + (x, y] - \{y\}, \quad (x, \infty) = (-\infty, x]^c$$

$(\mathbb{R}, \mathfrak{B}_1)$ 称为一维 Borel（可测）空间。实数集合 $\mathbb{R}$ 的许多子集不在 $\mathfrak{B}_1$ 中，我们把 $\mathfrak{B}_1$ 中的元素称为 Borel 可测集或 Borel 集，此概念是测度论的奠基者之一、法国数学家 Émile Borel (1871-1956) 于 1898 年引入的。Borel 集的概念非常重要，将用于 Borel 可测函数和随机变量的定义。类似地，直角坐标平面 $\mathbb{R}^2$ 上的 Borel σ 域 $\mathfrak{B}_2$ 是由所有形如 $(a, b] \times (c, d]$ 的矩形组成的类所生成的 σ 域，试想 $\mathfrak{B}_2$ 都包含 $\mathbb{R}^2$ 怎样的子集合？

练习 1.5. 请读者验证：Borel σ 域 $\mathfrak{B}_1$ 等同于实数轴 $\mathbb{R}$ 上所有开区间组成的类所生成的 σ 域，也等同于形如 $(-\infty, b]$ 的左开右闭区间组成的类所生成的 σ 域。

定义 1.7. 已知 $g : \Omega \to \mathbb{R}$ 是定义在可测空间 $(\Omega, \mathscr{S})$ 上的实值函数，如果对任意实数 $r \in \mathbb{R}$ 皆有 $\{\omega : g(\omega) \le r\} \in \mathscr{S}$，即 $(\mathbb{R}, \mathfrak{B}_1)$ 的可测集 $(-\infty, r]$ 的逆像仍是可测集*，则称 g 为可测函数。特别地，当 $(\Omega, \mathscr{S})$ 是 Borel 可测空间时，这样的函数 g 被称为 Borel 可测函数或 Borel 函数。

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 F 列举可测函数的一些常用性质，读者也可以参阅 W. Rudin 的名著《数

*有点儿类似于拓扑空间之间的连续映射，即任意开集的逆像仍是开集。

学分析原理》 [79] 的第十一章或测度论 [47]、函数论 [57] 教材。我们平常用到的实函数多是 Borel 函数，§2.1 将用可测函数来定义随机变量，§2.1.3 将论证 Borel 函数把随机变量依然映为随机变量。

定义 1.8 (样本空间，此概念由 R. von Mises 引进)**.** 已知 $\mathscr{S}$ 是随机试验基本事件集合 Ω 上的一个 σ 域，特称可测空间 $(\Omega, \mathscr{S})$ 为一个样本空间 (sample space)，称 $\mathscr{S}$ 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 $(\Omega, \mathscr{S})$ 为有限样本空间。如果 Ω 至多可数，则称 $(\Omega, \mathscr{S})$ 为离散样本空间。如果 Ω 不可数，则称 $(\Omega, \mathscr{S})$ 为不可数样本空间，特别地，当 $\Omega = \mathbb{R}^k$，我们称 $(\Omega, \mathscr{S})$ 为连续样本空间。

例 1.23. 考虑 Buffon 投针试验，在例 1.17 中，$\Omega = [0, \pi] \times [0, D/2]$，定义 $\mathscr{S}$ 是由形如以下的 Ω 的子集生成的 σ 域：$(a, b] \times (c, d]$，其中 $0 \le a < b \le \pi, 0 \le c < d \le D/2$。我们称 $\{A \cap \Omega : A \in \mathfrak{B}_2\}$ 为 Ω 的 Borel σ 域，常记作 $\Omega \cap \mathfrak{B}_2$。请读者验证上面定义的 $\mathscr{S} = \{A \cap \Omega : A \in \mathfrak{B}_2\}$。

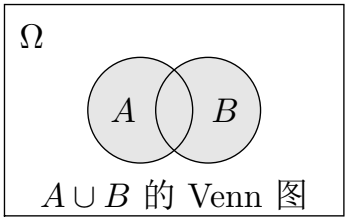
定义 1.9. 若事件 A 发生则事件 B 也发生，则“称 A 包含于 B”或“B 包含 A”，记作 $A \subseteq B$ 或 $B \supseteq A$。

例 1.24. 令 Ω 是抛两次硬币随机试验的基本事件集合，$\mathscr{S} = 2^\Omega$ 使得 $(\Omega, \mathscr{S})$ 构成一个样本空间。集合 $A = \{(T, H), (H, T)\} \in \mathscr{S}$ 表示事件“抛得一正一反”；$B = \{(T, H), (H, T), (H, H)\} \in \mathscr{S}$ 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。

定义 1.10 (并事件)**.** 由样本空间的定义知，若 $A, B \in \mathscr{S}$，则 $A \cup B \in \mathscr{S}$，即 $A \cup B$ 也是一个随机事件，我们称之为 A, B 的并事件或和事件*。若事件 A 发生，则事件 $A \cup B$ 也发生。如在例 1.24 中，$\{(T, T)\} \cup \{(H, H)\} = \{(T, T), (H, H)\}$ 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”。

*如果事件 $A_1, A_2, \cdots, A_n$ 两两互斥，我们常用 $A_1 + A_2$ 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嫌冗余地提醒读者它们是“非交并”。

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 $A\backslash B$（也记作 $A-B$）、对称差事件 $A\triangle B=(A\backslash B)\cup(B\backslash A)$、$A$ 的补事件（或称余事件、对立事件）A^c，请读者一一画出它们的 Venn 图。



$A\cup B$ 的 Venn 图

练习 1.6. 解释随机事件间的关系 $A\subseteq A\cup B$ 和 $AB\subseteq A$。

例 1.25. 已知 $A,B,C\in\mathscr{S}$ 是三个随机事件，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) $A^cB^cC^c$；(2) $A\cup B\cup C$；(3) $AB\cup BC\cup CA$；(4) $A(BC)^c\cup B(AC)^c\cup C(AB)^c$；(5) $(AB\cup BC\cup CA)^c$；(6) $(ABC)^c$ 或 $A^c\cup B^c\cup C^c$。

定义 1.11. 若事件 A,B 的交集为空集，则称它们互斥，意味着 A 与 B 不能同时发生。显然，事件 A,A^c 是互斥的，而且必有一个发生。

性质 1.7. 已知事件 $A_1,A_2,\cdots,A_n,\cdots\in\mathscr{S}$，则对于任意 $n\in\mathbb{N}$，事件 $\bigcup_{j=1}^n A_j$ 和 $\bigcup_{j=1}^\infty A_j$ 分别具有非交分解

$$\bigcup_{j=1}^{n}A_j=A_1+(A_1^cA_2)+(A_1^cA_2^cA_3)+\cdots+(A_1^c\cdots A_{n-1}^cA_n)\tag{1.22}$$

$$\bigcup_{j=1}^{\infty}A_j=\sum_{n=1}^{\infty}A_1^c\cdots A_{n-1}^cA_n\tag{1.23}$$

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……。

1.2.2 Kolmogorov 公理体系

定义 1.12 (概率)**.** 给定可测空间 $(\Omega, \mathscr{S})$，定义在 $\mathscr{S}$ 上的集函数 μ 若满足以下条件，则称 $(\Omega, \mathscr{S}, \mu)$ 为测度空间 (measure space)，称 μ 为测度。

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

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

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

❸ 空集测度为零：$\mu(\emptyset) = 0$。测度为零的可测集简称为零测集，在函数论的许多问题中都不起主要作用，可以乎略不计。

特别地，当测度空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 中 $(\Omega, \mathscr{S})$ 是一个样本空间并且 $\mathsf{P}(\Omega) = 1$ 时，特称 P 为概率测度，简称概率 (probability)，称 $(\Omega, \mathscr{S}, \mathsf{P})$ 为一个概率测度空间或概率空间。

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。”

也可以避而不谈测度空间，类比古典概率的性质 1.2，现代概率论奠基人 A. N. Kolmogorov 在《概率论基础》中最初是这样给出概率的 Kolmogorov 公理体系的 [56]。

定义 1.13 (Kolmogorov, 1933)**.** 已知随机试验的样本空间 $(\Omega, \mathscr{S})$，如果 $\mathscr{S}$ 上的实值集函数 P 满足以下条件，则称之为概率。

❶ 非负公理：对于任一随机事件 $A \in \mathscr{S}$，总有 $\mathsf{P}(A) \geq 0$，其中 $\mathsf{P}(A)$ 称为事件 A 的概率。

❷ 归一公理：$\mathsf{P}(\Omega) = 1$，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，则 $\mathsf{P}(A + B) = \mathsf{P}(A) + \mathsf{P}(B)$。

Kolmogorov 论证这样给出的概率公理体系是和谐的。应研究的需求，“有限可加公理”被补充加强为

❸ 可列可加公理：若事件 $A_j \in \mathscr{S}, j = 1, 2, \cdots$ 两两互斥，则

$$\mathsf{P}\left(\sum_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathsf{P}(A_j) \tag{1.24}$$

问题 1.3. 表面上如何理解 Kolmogorov 的这三条公理？

- 归一性约定必然事件 Ω 的概率是 1，由可列可加性可得 $\mathsf{P}(\emptyset) = 0$，即不可能事件 $\emptyset$ 的概率是 0。再由概率 $\mathsf{P}(\cdot)$ 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 $\mathsf{P}(\cdot)$ 的非负性保证了级数 $\sum_{j=1}^{\infty} \mathsf{P}(A_j)$ 绝对收敛，求和与次序无关。式 (1.24) 给出了求非交并事件 $\sum_{j=1}^{\infty} A_j \in \mathscr{S}$ 的概率的合理方法。

注记 1.2. “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。作为类比，这里“概率”的含义就是上述三条公理及其蕴含的性质，整个经典概率论的大厦就是建基于此的，除非特别说明，第一部分所介绍的缺省地都是经典概率论的内容。

注记 1.3. 公理化一方面明确了基本概念、规范了研究基础，另一方面又允许百花齐放、百家争鸣。非 Kolmogorov 体系的概率论最著名的就

*奥地利哲学家 Ludwig Wittgenstein (1889-1951) 是语言哲学的奠基人，是二十世纪最有影响力的哲学家之一，他认为“意义即使用”(Meaning is use)。

是贝叶斯理论*，1970 年匈牙利数学家 Alfréd Rényi (1921-1970) 在其遗作《概率论》中描述的贝叶斯概率公理体系 [76] 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。本书第十一章将对 Rényi 公理体系给予简介，并详述不同学派对“概率”的理解和哲学之争。

例 1.26. 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 $\{H,(T,H),(T,T,H),(T,T,T,H),\cdots\}$ 可以简记作 $\{1,2,3,4,\cdots\}$ 或自然数集合 $\mathbb{N}$。显然，$(\mathbb{N},2^{\mathbb{N}})$ 构成一个样本空间。定义 $(\mathbb{N},2^{\mathbb{N}})$ 上的概率测度如下：

$$\mathsf{P}(\{n\})=\frac{1}{2^n},\qquad n=1,2,\cdots$$

集合 $A_{2n-1}=\{2n-1\}$ 表示事件“第 $2n-1$ 次抛出第一个正面”，其概率为 $\mathsf{P}(A_{2n-1})=1/2^{2n-1}$，此处 $n=1,2,\cdots$。集合 $A=\{2n-1:n=1,2,\cdots\}$ 表示事件“第一个正面出现之前已经抛过偶数次”，其概率为

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

例 1.27 (n 重 Bernoulli 试验)**.** 假设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 $\mathsf{P}(k)$?

解. 该问题翻译成球-盒子模型即例 1.9 所述，请读者通过比较确认之。所以 $\mathsf{P}(k)=C_n^k p^k(1-p)^{n-k}$，其中 $k=0,1,\cdots,n$。

例 1.28 (连续正面问题)**.** 随机试验：连续抛一枚均匀的硬币 n 次，不妨设 $n\geq 3$。请构建概率空间 $(\Omega_n,\mathscr{S}_n,\mathsf{P}_n)$，并计算事件 $H_3=$“出现至少 3 个相连正面”的概率 $\mathsf{P}_n(H_3)$。

---

*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。

†像抛硬币这种只出现两个非此即彼结果的随机试验被称为 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。

解法 1. 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，定义 $\mathscr{S}_n = 2^{\Omega_n}$ 且 $\mathsf{P}(\omega) = 2^{-n}$，则 $(\Omega_n, \mathscr{S}_n, \mathsf{P}_n)$ 构成一个概率空间。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}。事件 H_3 可用下面的剪枝二叉树递归地构造出来：

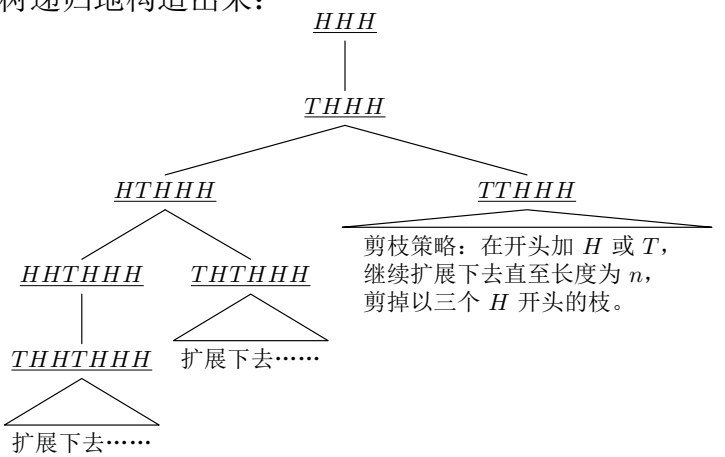


图 1.16: 树的高度为 $n-2$，根节点（即第 1 层节点）表示一开始就连抛三个正面，第 k 层节点表示事件“从第 k 抛开始出现至少三个相连正面”。

树中的每个节点代表一个事件，这些事件之间两两互斥，其并集合就是 H_3。各层节点个数 $T_k, k = 1, 2, \cdots$ 构成 Tribonacci 数列*。如果把图 1.16 中的树扩展到 $n+2$ 层，所有的叶子节点去掉末尾的 $THHH$ 就是 H_3^c 所含之基本事件，所以 $\mathsf{P}_n(H_3) = 1 - 2^{-n}T_{n+2}$，也等于 $\sum_{k=0}^{n-2} 2^{-2-k}T_k$。当 $n \to \infty$ 时，由 Tribonacci 数列的性质知，$\mathsf{P}_n(H_3) \to 1$。它的概率含义是什么？读者还能顺便得出什么结果呢？

解法 2. 事件 H_3 可分解为两个互斥事件：(1) 前 $n-1$ 抛中 H_3 发生，概率为 $p_{n-1} = \mathsf{P}_{n-1}(H_3)$；(2) 在第 n 抛后 H_3 才发生，即前 $n-4$ 抛中 H_3 不发生，第 $n-3$ 次抛出反面，最后 3 次抛出正面。于是，得到下面的线性递归关系：$p_n = p_{n-1} + 2^{-4}(1 - p_{n-4})$，满足初始条件 $p_0 = p_1 = p_2 = 0, p_3 = 2^{-3}$。列举几个结果：$p_{30} \approx 0.9078, p_{40} \approx 0.9601, p_{50} \approx 0.9827, p_{60} \approx 0.9925$。另外，利用 Maxima 还可以得到 p_n

*即满足线性递归关系 $T_k = T_{k-1} + T_{k-2} + T_{k-3}, k \geq 3$ 和初始条件 $T_0 = 0, T_1 = T_2 = 1, T_3 = 2$ 的数列，具有性质 $\sum_{k=0}^{\infty} x^k T_k = x/(1 - x - x^2 - x^3)$。

的解析表达式，代码如下。

```
load("solve_rec") $
rec: p[n] - p[n-1] - (1-p[n-4])/2^4 $
solve_rec (rec, p[n], p[0]=0, p[1]=0, p[2]=0, p[3]=1/2^3) $
ratsimp(minfactorial(factcomb(%)));
```

✂ **例 1.29** (连续正面问题的极限版)**.** 假设某硬币出现正面的概率是 $\mathsf{P}(H) = h > 0$，连续抛该枚硬币无穷次，则事件 $H_t =$ “出现至少 $t < \infty$ 个相连正面”的概率是 1。

证明. 该随机试验的基本事件都是 H 和 T 构成的无穷长度的字符串，把它们当作输入，事件 H_t 的概率可视为下面的确定有限状态自动机 $\mathbb{A}_t$ 停机的概率。

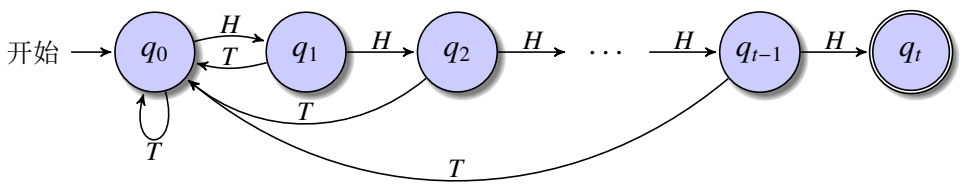


图 1.17: 有限状态自动机 $\mathbb{A}_t$：输入字母表是 $\Sigma = \{H, T\}$，q_0 是初始状态，q_t 是接受状态（亦称终止状态）。如果当前状态是 q_k，意味着当前输入字符是 H 且之前恰有 $\max(0, k-1)$ 连续的 H。

设某输入串的当前状态 q_k，该输入串能否导致 $\mathbb{A}_t$ 停机与到达 q_k 之前的输入无关。把 $\mathbb{A}_t$ 的初始状态改为 q_k，设其停机的概率为 p_k，则 $p_t = 1, p_k = hp_{k+1} + (1-h)p_0$，其中 $k = 0, 1, \cdots, t-1$，解之得 $p_0 = p_1 = \cdots = p_t = 1$，即 $\mathsf{P}(H_t) = 1$。 □

例 1.30. 为了记述的方便，定义 $x \in \mathbb{R}$ 的实值函数 $\phi(x|\mu, \sigma^2)$ 如下：

$$\phi(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \text{ 其中参数 } \mu \in \mathbb{R}, \sigma > 0 \quad (1.25)$$

特别地，我们简记 $\phi(x|0,1)$ 为 $\phi(x)$。令 $\Omega = \mathbb{R}, \mathscr{S} = \mathfrak{B}_1$，验证如下定义

的 $\mathscr{S}$ 上的实值集函数 P 是一个概率：在任意区间 $A \subseteq \Omega$ 上

$$\mathsf{P}(A) = \int_A \phi(x)dx \quad 或 \quad \mathsf{P}(A) = \int_A \phi(x|\mu,\sigma^2)dx \tag{1.26}$$

证明. 下面逐一验证实值集函数 $\mathsf{P}(A) = \int_A \phi(x)dx$ 满足概率测度的三条公理。显然 $\mathsf{P}(A) \geq 0$ 且 P 满足可列可加性（积分的性质）。接下来验证 $\mathsf{P}(\Omega) = 1$，即练习 1.2.2 中阴影部分的面积等于 1。令 $\mathsf{P}(\Omega) = m$，则

$$m^2 = \int_{-\infty}^{+\infty} \phi(x)dx \int_{-\infty}^{+\infty} \phi(y)dy = \frac{1}{2\pi} \iint_{\mathbb{R}^2} \exp\left\{-\frac{x^2+y^2}{2}\right\} dxdy$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{+\infty} \rho \exp\left\{-\frac{\rho^2}{2}\right\} d\rho d\theta = 1 \text{（令 } x = \rho\cos\theta, y = \rho\sin\theta\text{）} \ \square$$

练习 1.7. 请读者自行验证式 (1.25) 的情形。提示：通过变量替换 $y = (x-\mu)/\sigma$ 把 $\phi(x|\mu,\sigma^2)$ 变换为 $\phi(y)$。

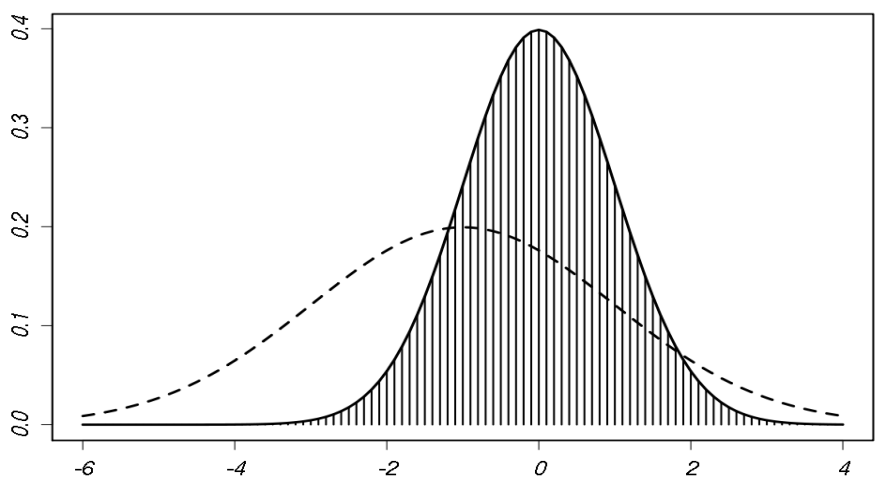


图 1.18: 实线是 $\phi(x)$，虚线是 $\phi(x|-1,4)$。函数 $\phi(x|\mu,\sigma^2)$ 的曲线呈钟型，关于 $x = \mu$ 对称，我们称 μ 为位置参数。参数 σ^2 刻画的是 $\phi(x|\mu,\sigma^2)$ 的“体形”：σ^2 越小曲线越“高瘦”，被称为尺度参数。

✂ **例 1.31.** 已知参数 τ^2, μ, σ^2，利用 $\phi(\theta|\cdot,\cdot)$ 的性质计算下面的积分

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

解. 首先发现 $\int_{-\infty}^{+\infty} m(x)dx = 1$，所以函数 $\phi(x|\theta,\tau^2)\phi(\theta|\mu,\sigma^2)$ 中影响积分的只有 x,θ，而 τ^2,μ,σ^2 只是起到归一化的作用。先利用“正比于”关系（记作 $\propto$）简化一下被积函数，

$$\phi(x|\theta,\tau^2)\phi(\theta|\mu,\sigma^2) \propto \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\sigma^2} + \frac{(x-\theta)^2}{\tau^2}\right]\right\}$$

利用配方法我们“凑”得

$$\frac{(\theta-\mu)^2}{\sigma^2} + \frac{(x-\theta)^2}{\tau^2} = \left\{\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta - \left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\right]^2 + \frac{(x-\mu)^2}{\sigma^2+\tau^2}\right\} + g(\tau^2,\mu,\sigma^2)$$

其中函数 $g(\tau^2,\mu,\sigma^2)$ 不含变量 θ 和 x，所以

$$\phi(x|\theta,\tau^2)\phi(\theta|\mu,\sigma^2) \propto \phi\left(\theta\left|\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right.\right)\phi(x|\mu,\sigma^2+\tau^2)$$

由于 $\int_{-\infty}^{+\infty} m(x)dx = 1$，所以 $m(x) = \phi(x|\mu,\sigma^2+\tau^2)$，即

$$\int_{-\infty}^{+\infty} \phi(x|\theta,\tau^2)\phi(\theta|\mu,\sigma^2)d\theta = \phi(x|\mu,\sigma^2+\tau^2) \tag{1.27}$$

以上方法仅凭借 $\phi(x|\cdot,\cdot)$ 在 $\mathbb{R}$ 上积分为 1 这一事实来定性地求解积分。读者可可以利用 Maxima 辅助印证上面给出的结果是正确的。

```
load (distrib) $
assume (sigma>0 and tau>0) $
integrate (pdf_normal(x,theta,tau) * pdf_normal(theta,mu,sigma), theta, minf,
       inf);
```

练习 1.8. 令尺度参数 $\lambda > 0$，实值函数 $p(x)$ 定义为

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{当 } x \geq 0 \\ 0 & \text{当 } x < 0 \end{cases} \tag{1.28}$$

试证明：$\mathsf{P}(A) = \int_A p(x)dx$ 是样本空间 $(\mathbb{R}, \mathfrak{B}_1)$ 的概率测度，其中 A 是 $\mathbb{R}$ 上任意区间。

例 1.32. 1957 年，美国语言学家 Noam Chomsky (1928-) 发表了《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)。其中，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应用于自然语言处理、生物信息学等研究领域。随机上下文无关文法 (stochastic context-free grammar, SCFG) 就是每个产生式被赋予了概率的上下文无关文法，如同隐 Markov 模型扩展了正则文法一样（详见第十二章）。下面给出随机上下文无关文法的一个样例。

$$\begin{aligned}
S &\to NP + VP \quad [1.00] \\
NP &\to Pronoun \quad [0.10] \\
&\mid Name \quad [0.10] \\
&\mid Noun \quad [0.15] \\
&\mid Article + Noun \quad [0.6] \\
&\mid NP + PP \quad [0.05] \\
VP &\to Verb \quad [0.60] \\
&\mid VP + NP \quad [0.20] \\
&\mid VP + PP \quad [0.20] \\
PP &\to Prep + NP \quad [1.00] \\
Noun &\to \text{telescope[0.001]|microscope[0.001]|boy[0.01]|girl[0.01]|}\cdots \\
Verb &\to \text{saw[0.02]|study[0.01]|}\cdots \\
Pronoun &\to \text{I[0.20]|you[0.10]|it[0.30]|}\cdots \\
Article &\to \text{the[0.30]|a[0.35]|every[0.02]|}\cdots \\
Prep &\to \text{in[0.20]|to[0.30]|on[0.04]|with[0.10]|}\cdots
\end{aligned}$$

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）。另外，还要假设这些重写规则是独立的。

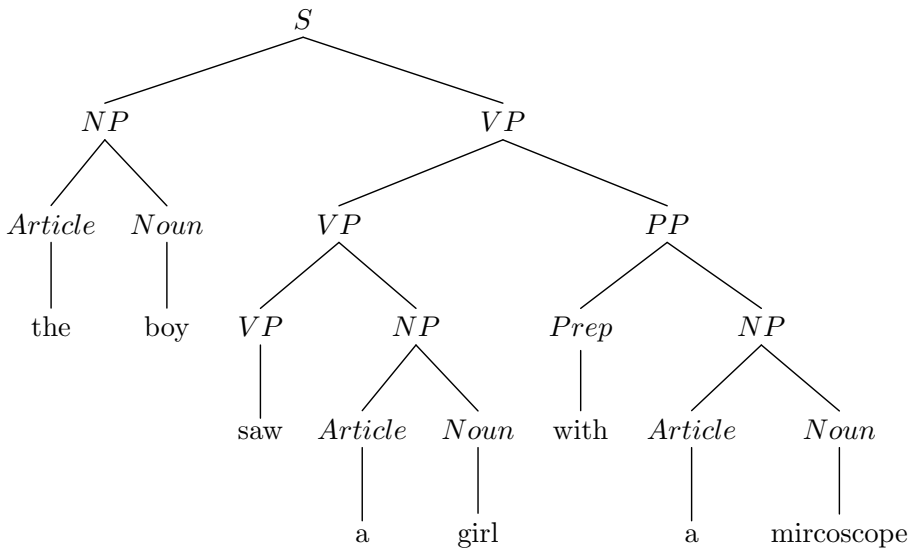


图 1.19: 按照给定的重写规则，句子 the boy saw a girl with a microscope 有两个不同的编译结果。由图示的句法树导出的逻辑表示在语义上是错误的，虽然该句法树的产生概率 p 相对另外一个还稍大些，其中 $p = 1.00 \times (0.6 \times 0.20) \times (0.3 \times 0.01 \times 0.2 \times 1) \times (0.02 \times 0.6 \times 0.1 \times 0.6) \times (0.35 \times 0.01 \times 0.35 \times 0.001) = 6.3504 \times 10^{-14}$。请读者给出另外一棵句法树，并计算它的产生概率。

✂ **例 1.33** (回顾 Bertrand 悖论)**.** 在例 1.16 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，样本空间 $(\Omega, \mathscr{S})$ 依次为

1. $\Omega = [0, 2\pi) \times (0, 1)$，选 $\mathscr{S} = \{A \cap \Omega : A \in \mathfrak{B}_2\}$，即 Ω 的 Borel σ 域，则 $E = [0, 2\pi) \times (0, 1/2)$ 且 $\mathsf{P}(E) = 1/2$。

2. $\Omega = [0, 2\pi) \times [0, 2\pi) - \{(x, x) : x \in [0, 2\pi)\}$，选 $\mathscr{S}$ 为 Ω 的 Borel σ 域，则 $E = \{(x, y) : 2\pi/3 < |x - y| < 4\pi/3\}$ 且 $\mathsf{P}(E) = 1/3$（请读者仿照例 1.13 验证之）。

3. $\Omega = \{(x, y) : x^2 + y^2 < 1\}$，选 $\mathscr{S}$ 为 Ω 的 Borel σ 域，则 $E = \{(x, y) : x^2 + y^2 < 1/2\}$ 且 $\mathsf{P}(E) = 1/4$。

三种不同的随机取弦方法对应着三个不同的样本空间，进而对随机事件的定义也不相同，得出不同的结论就不令人诧异了。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，贝叶斯学派知名学者 Edwin Thompson Jaynes (1922-1998) 撰文 [52] 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 [16,52]。

注记 1.4. 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 $(\Omega, \mathscr{S}, \mathsf{P})$ 上进行的，任意随机事件 $E \in \mathscr{S}$ 都明确定义好了的。概率的公理化并没有指明如何构造 $(\Omega, \mathscr{S}, \mathsf{P})$ 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。

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。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 J. Bernoulli (1654-1705) 于十七世纪末发现的弱大数律，它揭示了随机现象中蕴藏的客观性。例 1.34 通过试验了解频率与概率的关系，有助于理解弱大数律。

定理 1.1 (Bernoulli 弱大数律)**.** 已知随机事件 A 的概率 $\mathsf{P}(A) = p$，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

$$\lim_{n\to\infty} \mathsf{P}\left\{\left|\frac{m}{n} - p\right| \le \epsilon\right\} = 1 \tag{1.29}$$

证明. 详见 §5.1，这里先给出一个直观的论证。利用 §2.3.2 即将介绍的

Hoeffding 不等式*(2.88) 得到如下推论：当 $n \to \infty$ 时，

$$\mathsf{P}\left\{\left|\frac{m}{n} - p\right| \le \epsilon\right\} = \sum_{m=\lceil np-n\epsilon \rceil}^{\lfloor np+n\epsilon \rfloor} C_n^m p^m (1-p)^{n-m} \ge 1 - 2\exp\left\{-2n\epsilon^2\right\} \to 1 \quad \square$$

例 1.34. 抛一枚不均匀的硬币，假设正面出现的概率为 $\mathsf{P}(H) = 0.6$。连续抛该硬币 n 次，在这 n 重 Bernoulli 试验中出现了 m 次正面，其频率 m/n 与概率 $\mathsf{P}(H)$ 之间是怎样的关系？

解. 出现正面的概率 $\mathsf{P}(H)$ 并不是指试验次数 $n \to \infty$ 时，随机事件频率 m/n 的极限。事实上，根本无法谈论 m/n 的极限，因为 m 是不确定的。但是，读者可以通过直方图†考察 m/n 散落在 $\mathsf{P}(H)$ 周围的情况。

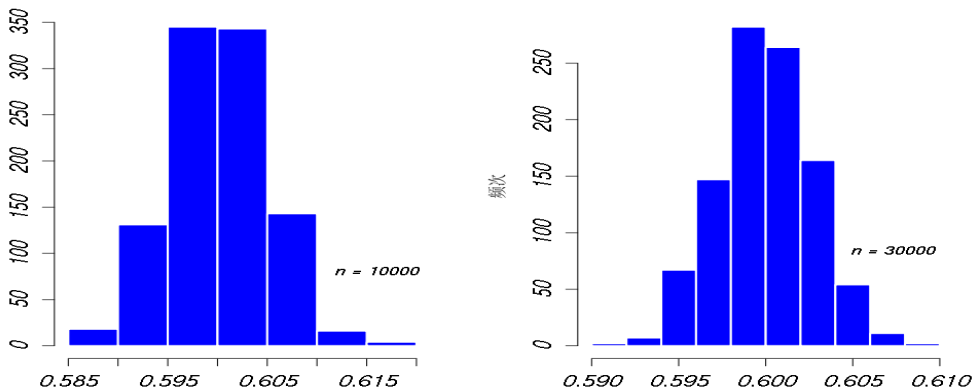


图 1.20: 抛硬币 $n = 10000$ 次或 $n = 30000$ 次，把出现正面的频率 m/n 记录下来。重复此过程 $k = 1000$ 次，得到了 $k = 1000$ 个频率结果，由直方图不难发现：(1) 频率基本围绕在 $\mathsf{P}(H)$ 两侧且偏离 $\mathsf{P}(H)$ 越远越少有发生。(2) 右边的直方图比左边的“瘦”些，这说明试验次数 n 越多，试验所得的频率越紧密“团结”在 $\mathsf{P}(H)$ 周围，误差 $|m/n - \mathsf{P}(H)|$ 大于给定正数 ϵ 的机会也越小。

*Wassily Hoeffding (1914-1991)，美国统计学家，非参数统计学的奠基者之一。

†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。直方图一般用来探索分析数据的分布情况，即显示观察值在何处聚集以及疏散程度。

1.2.3　概率的一些基本性质

已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$，根据 Kolmogorov 的三条公理，不难得到概率 $\mathsf{P}(\cdot)$ 的一些基本性质如下：

定理 1.2. 对于任意事件 $A \in \mathscr{S}$，皆有 $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$。

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

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

但由 $\mathsf{P}(A) = 0$ 推导不出 $A = \emptyset$。回顾会面问题，在例 1.13 中，样本空间 $(\Omega, \mathscr{S})$ 为：$\Omega = [0, 60] \times [0, 60]$，$\mathscr{S}$ 是 Ω 上的 Borel σ 域。集合 $A = \{(x, x) : 0 \le x \le 60\} \in \mathscr{S}$ 非空，表示事件“两人同时到达”，但 $\mathsf{P}(A) = 0$。类似地，由 $\mathsf{P}(A) = 1$ 也不能得出 $A = \Omega$，请读者说明理由。

定理 1.3. 如果 $A, B \in \mathscr{S}$ 且 $A \subseteq B$，则 $\mathsf{P}(A) \le \mathsf{P}(B)$。

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

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

定理 1.4. 已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$，若 $\{A_j \in \mathscr{S} : j = 1, 2, \cdots\}$ 是 Ω 的一个划分，则 $\sum_{j=1}^{\infty} \mathsf{P}(A_j) = 1$ 且对于任意事件 $B \in \mathscr{S}$ 皆有

$$\mathsf{P}(B) = \sum_{j=1}^{\infty} \mathsf{P}(BA_j) \tag{1.30}$$

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

定理 1.5 (和事件的概率)**.** 如果 $A, B \in \mathscr{S}$，则有如下的*加法法则*。

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(AB) \tag{1.31}$$

证明. 事件 $A\cup B$ 有非交分解 $A\cup B=(A-B)+(B-A)+AB$，所以 $\mathsf{P}(A\cup B)=\mathsf{P}(A-B)+\mathsf{P}(B-A)+\mathsf{P}(AB)$。同理，$\mathsf{P}(A)=\mathsf{P}(A-B)+\mathsf{P}(AB),\mathsf{P}(B)=\mathsf{P}(B-A)+\mathsf{P}(AB)$。三式联立即得证。 □

推论 1.3. 对任意事件 $A,B\in\mathscr{S}$，总有 $\mathsf{P}(A\cup B)\leq\mathsf{P}(A)+\mathsf{P}(B)$ 且 $\mathsf{P}(A-B)=\mathsf{P}(A)-\mathsf{P}(AB)$。

练习 1.9 (Boole 不等式)**.** 对于随机事件 $A_1,A_2,\cdots\in\mathscr{S}$，总有

$$\mathsf{P}\left(\bigcup_{k=1}^{\infty}A_k\right)\leq\sum_{k=1}^{\infty}\mathsf{P}(A_k)\quad\text{或者}\quad\mathsf{P}\left(\bigcap_{k=1}^{\infty}A_k\right)\geq 1-\sum_{k=1}^{\infty}\mathsf{P}(A_k^c)\tag{1.32}$$

定理 1.6 (Jordan 公式)**.** 作为定理 1.5 的一般化有如下结果：对于任意事件 $A_1,\cdots,A_n\in\mathscr{S}$，下面的等式成立。

$$\begin{aligned}\mathsf{P}\left(\bigcup_{k=1}^{n}A_k\right)=&\sum_{k=1}^{n}\mathsf{P}(A_k)-\sum_{k_1<k_2}^{n}\mathsf{P}(A_{k_1}A_{k_2})+\\&\sum_{k_1<k_2<k_3}^{n}\mathsf{P}(A_{k_1}A_{k_2}A_{k_3})+\cdots+(-1)^{n+1}\mathsf{P}\left(\prod_{k=1}^{n}A_k\right)\end{aligned}\tag{1.33}$$

证明. 对 n 进行数学归纳法，请读者给出证明的具体细节。 □

例 1.35. 任取有限集合 $A=\{a_1,a_2,\cdots,a_n\}$ 的一个置换，即 A 到自身的一个一一映射，试问：该置换没有不动点的概率？

解. 任取 A 的一个置换相当于把标号为 $1,2,\cdots,n$ 的 n 个球放入 n 个盒子，每个盒子只能放一个球，总共有 $n!$ 中不同的放法。令 A_k 表示第 k 个盒子里装着第 k 个球，则

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

第 k_1 个盒子装着第 k_1 个球，且第 k_2 个盒子装着第 k_2 个球的概率 $\mathsf{P}(A_{k_1}A_{k_2})=(n-2)!/n!$，于是 $\sum_{k_1<k_2}^{n}\mathsf{P}(A_{k_1}A_{k_2})=C_n^2(n-2)!/n!=1/2!$。依次

类推，$\mathsf{P}(A_{k_1}A_{k_2}A_{k_3}) = (n-3)!/n!$ 且 $\sum_{k_1<k_2<k_3}^{n} \mathsf{P}(A_{k_1}A_{k_2}A_{k_3}) = C_n^3(n-3)!/n! = 1/3!, \cdots\cdots$。式 (1.33) 右侧的每一个求和项都能算出，于是

$$\mathsf{P}\Big(\text{至少有一个不动点}\Big) = \mathsf{P}\left(\bigcup_{k=1}^{n} A_k\right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!}$$

$$\mathsf{P}\Big(\text{没有不动点}\Big) = 1 - \mathsf{P}\left(\bigcup_{k=1}^{n} A_k\right) = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n}\frac{1}{n!}$$

当 n 很大时，随机选取的置换没有不动点的概率约为 e^{-1}。

定理 1.7 (概率的连续性定理)**.** 已知 $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$ 是随机事件的非升序列，则

$$\mathsf{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathsf{P}(A_n) \tag{1.34}$$

这个定理将在 §2.1 用于证明分布函数右连续。

证明. 见右图，$A = \bigcap_{n=1}^{\infty} A_n$ 是 $A_1, A_2, \cdots$ 共有的“核”。由非交分解 $A_n = \sum_{k=n}^{\infty} A_k A_{k+1}^c + A$，

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

因为正项级数 $\sum_{k=1}^{\infty} \mathsf{P}(A_k A_{k+1}^c)$ 收敛，所以余项极限为零，即 $\lim\limits_{n\to\infty} \sum_{k=n}^{\infty} \mathsf{P}(A_k A_{k+1}^c) = 0$，进而得证 $\mathsf{P}(A) = \lim\limits_{n\to\infty} \mathsf{P}(A_n)$。 □

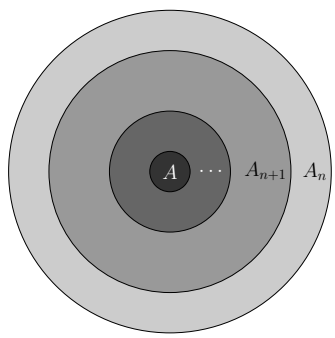


图 1.21: $A_k A_{k+1}^c$ 就像一个套一个的环，两两不交。

推论 1.4. 已知随机事件的非减序列 $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$，则

$$\mathsf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathsf{P}(A_n) \tag{1.35}$$

一般地，我们把 $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots \supseteq A = \bigcap_{n=1}^{\infty} A_n$ 简记作 $A_n \downarrow A$，把 $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots \subseteq A = \bigcup_{n=1}^{\infty} A_n$ 简记作 $A_n \uparrow A$。

推论 1.5 (连续公理)**.** 若随机事件的序列 $A_1, A_2, \cdots, A_n, \cdots$ 满足 $A_n \downarrow \emptyset$，则 $\lim\limits_{n\to\infty} \mathsf{P}(A_n) = 0$。

引理 1.1 (Borel-Cantelli*, 1909, 1917)**.** 已知随机事件的序列 $A_1, A_2, \cdots, A_n, \cdots$ 满足 $\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty$，则 P(无穷多个 A_n 发生) $= 0$。

证明. 如果无穷多个 A_n 发生，则对于任意有限的 k 都有 $\bigcup_{n=k}^{\infty} A_n$ 发生，于是如下定义的事件 A 发生：$A = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$，即 $\overline{\lim\limits_{n\to\infty}} A_n$。反之，若事件 A 发生，则有无穷多个 A_n 发生，请读者参考结果 (1.19) 验证。明显地，$B_k = \bigcup_{n=k}^{\infty} A_n$ 是一个递减的序列，其中 $k = 1, 2, \cdots$。利用定理 1.7 和式 (1.32)，可以得出“无穷多个 A_n 发生”的概率为零，即

$$\mathsf{P}(A) = \mathsf{P}\left(\bigcap_{k=1}^{\infty} B_k\right) = \lim_{k\to\infty} \mathsf{P}(B_k) \leq \lim_{k\to\infty} \sum_{n=k}^{\infty} \mathsf{P}(A_n) = 0 \qquad \square$$

定理 1.8. 连续公理与可列可加公理等价。

证明. 只需往证“连续公理”$\Rightarrow$“可列可加公理”：已知事件 $A_n, n = 1, 2, \cdots$ 两两互斥，令 $A = \sum_{n=1}^{\infty} A_n$ 且 $B_k = \sum_{n=k}^{\infty} A_k$。显然，$B_k \supseteq B_{k+1}, k = 1, 2, \cdots$。而且 $B_k \downarrow \emptyset$，如若不然，则有无穷多个 A_n 发生，与 $A_1, A_2, \cdots$ 两两互斥矛盾。所以，$\lim\limits_{k\to\infty} \mathsf{P}(B_k) = 0$。按照加法公理，

$$\mathsf{P}(A) = \sum_{n=1}^{k} \mathsf{P}(A_n) + \mathsf{P}(B_{k+1}) = \lim_{k\to\infty} \sum_{n=1}^{k} \mathsf{P}(A_n) = \sum_{n=1}^{\infty} \mathsf{P}(A_n) \qquad \square$$

*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。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。

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》(Essay Towards Solving a Problem in the Doctrine of Chances, 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。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)。如今我们只能从 Bayes 的这篇重要论文中探究他的思想，多数研究者将之归为主观贝叶斯主义 [15,87]。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：(1) 经过多年的概率哲学基础之争 [60]，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。(2) 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，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。第十一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，于是用它来做推断的手法也不相同：Bayes 公式是贝叶斯推断的核心，它通过事件的验后概率（或称后验概率，posterior probability）或它与验前概率（或称先验概率，prior probability）的比较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推断。

条件概率引发了对随机事件独立性和条件独立性的思考：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 [67]。

本节内容 第一小节在 Kolmogorov 概率公理的框架之下给出了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，通过实例介绍了如何利用 Bayes 公式进行推断。第三、四小节深入讨论了随机事件的独立性和条件独立性。最后通过两个有趣的例子，一个是“兼听则明，偏听则暗”的多专家系统，另一个是贝叶斯垃圾邮件过滤，来说明可应用条件独立性假设简化验后概率的计算。

学习目标 (1) 掌握条件概率的定义及其性质；(2) 熟练运用全概率公式和 Bayes 公式计算条件概率；(3) 会利用 Bayes 公式做推断；(4) 充分理解随机事件的独立性和条件独立性。

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 1.36. 掷两个均匀的骰子，基本事件集合是 $\Omega = \{(i,j) : i,j = 1,2,\cdots,6\}$，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 $\mathsf{P}(A) = 1/6, \mathsf{P}(B) = 5/18, \mathsf{P}(AB) = 1/18$。现在已知事件 B 发生了，问 A 发生的概率？

解. 已知事件 B 发生了，所以掷双骰子的随机试验的结果只可能是集合 $\{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)\}$ 中的某一个，进而得出 A 发生的概率是 $1/5$，数值上等于 $\mathsf{P}(AB)/\mathsf{P}(B)$。

定理 1.9 (条件概率)**.** 已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$，且事件 $B \in \mathscr{S}$ 满足 $\mathsf{P}(B) > 0$。对任意事件 $A \in \mathscr{S}$，定义 B 发生情况下，A 发生的条件概率 (conditional probability) 为

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(AB)}{\mathsf{P}(B)} \tag{1.36}$$

有时也记作 $\mathsf{P}_B(A)$，它也是 $\mathscr{S}$ 上的一个概率测度，即 $(\Omega, \mathscr{S}, \mathsf{P}_B)$ 也是一个概率空间，称为条件概率空间。

证明. 下面依次验证 P_B 满足 Kolmogorov 的三条公理。

1. $\forall A \in \mathscr{S}$，显然有 $\mathsf{P}_B(A) = \mathsf{P}(AB)/\mathsf{P}(B) \geq 0$，非负性成立。

2. $\mathsf{P}_B(\Omega) = \mathsf{P}(\Omega B)/\mathsf{P}(B) = 1$，归一性成立。

3. 如果 $A_1, A_2, \cdots \in \mathscr{S}$ 两两不交，则

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

综上所述，$(\Omega, \mathscr{S}, \mathsf{P}_B)$ 构成一个概率空间。 □


读者可以把 $(\Omega, \mathscr{S}, \mathsf{P}_B)$ 看作是由 $(\Omega, \mathscr{S}, \mathsf{P})$ 和事件 B 诱导出来的概率空间，用来考察 B 发生的情况下 $A \in \mathscr{S}$ 的概率。

练习 1.10. 已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$，且事件 $B \in \mathscr{S}$ 满足 $\mathsf{P}(B) > 0$。请读者验证：(1) $\mathsf{P}(A|B) = 1$，其中 $B \subseteq A \in \mathscr{S}$，并请读者给出直观的解释。(2) 对于定理 1.9 定义的条件概率空间 $(\Omega, \mathscr{S}, \mathsf{P}_B)$，如果事件 $C \in \mathscr{S}$ 满足 $\mathsf{P}_B(C) > 0$，则 $\mathsf{P}_B(A|C) = \mathsf{P}(A|BC)$。(3) $\mathscr{S}_B = \mathscr{S} \cap B = \{E \cap B : E \in \mathscr{S}\}$ 也是 B 上的 σ 域；对于事件 $C \in \mathscr{S}_B$，定义 $\mathsf{P}_B(C) = \mathsf{P}(C)/\mathsf{P}(B)$，则 $(B, \mathscr{S}_B, \mathsf{P}_B)$ 构成一个概率空间。

练习 1.11. 考虑恰有两个孩子的家庭：基本事件集合 $\{bb, bg, gb, gg\}$，其中 b 表示男孩，g 表示女孩，假定每个基本事件的概率都是 1/4。若已知某家庭有一个女孩，问该家庭有两个女孩的概率？（答案：1/3）

例 1.37. 已知 $\mathsf{P}(B) > 0$，则

$$\mathsf{P}(A|B) \geq \frac{\mathsf{P}(A) + \mathsf{P}(B) - 1}{\mathsf{P}(B)} \tag{1.37}$$

证明. 由 $\mathsf{P}(A \cup B) \leq 1$ 和定理 1.5 得，$\mathsf{P}(B) - \mathsf{P}(AB) \leq \mathsf{P}(A^c)$ 或者 $\mathsf{P}(A|B) \geq 1 - \mathsf{P}(A^c)/\mathsf{P}(B) = [\mathsf{P}(A) + \mathsf{P}(B) - 1]/\mathsf{P}(B)$，得证。 □

定理 1.10. 如果 $\mathsf{P}(AB) > 0$，则有下面的结果成立。

1. 积事件 AB 的概率*有如下的分解：

$$\mathsf{P}(AB) = \mathsf{P}(A|B)\mathsf{P}(B) = \mathsf{P}(B|A)\mathsf{P}(A) \tag{1.38}$$

2. 如果 $\mathsf{P}(A|B) > \mathsf{P}(A)$，则 $\mathsf{P}(B|A) > \mathsf{P}(B)$。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 $\mathsf{P}(AB) > 0$，所以 $\mathsf{P}(A) > 0$，同理 $\mathsf{P}(B) > 0$。由条件概率的定义式 (1.36) 可得式 (1.38)。 □

*积事件 AB 的概率也称为 A, B 的联合概率，记作 $\mathsf{P}(A, B)$。

推论 1.6 (积事件的概率)**.** 把式 (1.38) 推广到 $A_1A_2\cdots A_n$ 的情形，就有了下面的*乘法法则*：已知概率空间 $(\Omega,\mathscr{S},\mathsf{P})$，且事件 $A_1,A_2,\cdots,A_n\in\mathscr{S}$ 满足 $\mathsf{P}(A_1A_2\cdots A_{n-1})>0$，则

$$\mathsf{P}(A_1A_2\cdots A_n)=\mathsf{P}(A_1)\mathsf{P}(A_2|A_1)\mathsf{P}(A_3|A_1A_2)\cdots\mathsf{P}(A_n|A_1A_2\cdots A_{n-1}) \quad (1.39)$$

证明. 由已知可得 $\mathsf{P}(A_1A_2\cdots A_n)=\mathsf{P}(A_1A_2\cdots A_{n-1})\mathsf{P}(A_n|A_1A_2\cdots A_{n-1})$。因为 $A_1A_2\cdots A_{n-1}\subseteq\cdots\subseteq A_1A_2\subseteq A_1$ 且 $\mathsf{P}(A_1A_2\cdots A_{n-1})>0$，所以 $\mathsf{P}(A_1A_2\cdots A_{n-2})>0,\cdots,\mathsf{P}(A_1A_2A_3)>0,\mathsf{P}(A_1A_2)>0,\mathsf{P}(A_1)>0$，据此对 $\mathsf{P}(A_1A_2\cdots A_{n-1})$ 可以继续分解下去……。 □

例 1.38. 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？

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 $\mathsf{P}(A_1A_2A_3^c)$，利用乘法法则，即式 (1.39) 可得

$$\mathsf{P}(A_1A_2A_3^c)=\mathsf{P}(A_1)\mathsf{P}(A_2|A_1)\mathsf{P}(A_3^c|A_1A_2)=\frac{5}{100}\times\frac{4}{99}\times\frac{95}{98}\approx 0.002$$

例 1.39. 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”，$k=1,2,\cdots,n$。

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

例 1.40. 盒子里有一黑一白两个球，一次抽取一个球，直至抽到黑球。如果抽到白球，除了放回还再要补充两个白球回盒子，问前 n 次抽球中黑球不出现的概率 $\mathsf{P}(n)$？

解. 前 n 次抽球中每次都是白球，所以

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

利用 Stirling 公式 (1.6)，当 n 充分大时，$\mathsf{P}(n)\approx 1/\sqrt{\pi n}$。

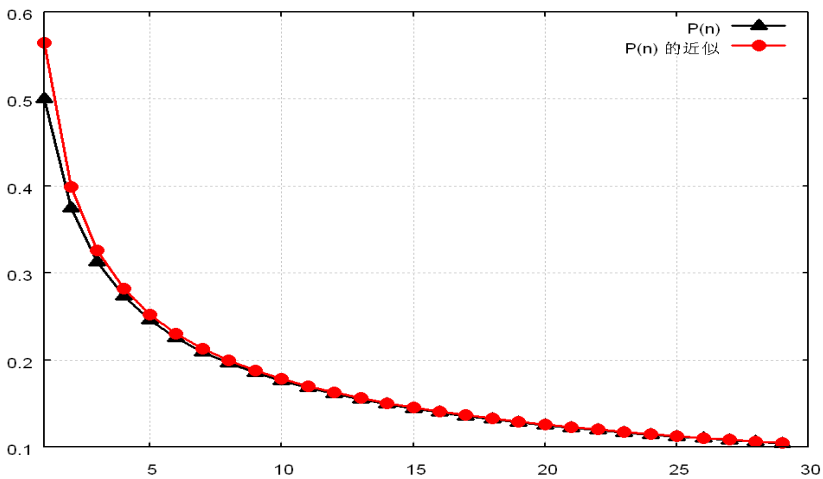


图 1.22: 例 1.40 中 $\mathsf{P}(n)$ 及其近似的折线图。当 $n\geq 10$ 时，$\mathsf{P}(n)$ 与其近似 $1/\sqrt{\pi n}$ 就已经非常接近了。

练习 1.12. 把式 (1.39) 做一个推广：已知概率空间 $(\Omega,\mathscr{S},\mathsf{P})$，事件 $B,A_1,A_2,\cdots,A_n\in\mathscr{S}$ 满足 $\mathsf{P}(BA_1A_2\cdots A_{n-1})>0$，则

$$\mathsf{P}(A_1A_2\cdots A_n|B)=\mathsf{P}(A_1|B)\prod_{k=2}^{n}\mathsf{P}(A_k|BA_1A_2\cdots A_{k-1}) \tag{1.40}$$

特别地，我们得到了式 (1.38) 的条件概率情形：

$$\mathsf{P}(A_1A_2|B)=\mathsf{P}(A_2|BA_1)\mathsf{P}(A_1|B)\quad\text{或}\quad\mathsf{P}(A_2|BA_1)=\frac{\mathsf{P}(A_1A_2|B)}{\mathsf{P}(A_1|B)} \tag{1.41}$$

1.3.2　全概率公式与 Bayes 公式

问题 1.4. 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 $\mathsf{P}(A_1) = \mathsf{P}(A_2) = 1/2$，问摸到白球的概率？提示：利用下面的全概率公式。

定理 1.11 (全概率公式)**.** 已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 且 $\{A_j \in \mathscr{S} : \mathsf{P}(A_j) \neq 0, j = 1, 2, \cdots\}$ 是 Ω 的一个划分，则对任一事件 B 皆有

$$\mathsf{P}(B) = \sum_{j=1}^{\infty} \mathsf{P}(A_j)\mathsf{P}(B|A_j) \tag{1.42}$$

证明. 由式 (1.30) 和式 (1.38) 可证。 □

对上面的问题 1.4，摸到白球的概率是

$$\begin{aligned}\mathsf{P}(\text{白球}) &= \mathsf{P}(A_1)\mathsf{P}(\text{白球}|A_1) \\ &\quad + \mathsf{P}(A_2)\mathsf{P}(\text{白球}|A_2) \\ &= \frac{1}{2}\left(\frac{3}{5} + \frac{1}{5}\right) = \frac{2}{5}\end{aligned}$$

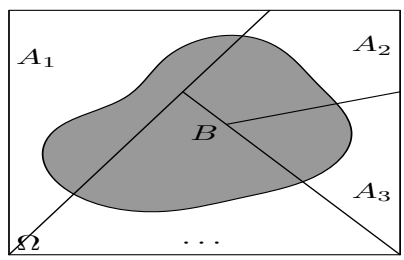


图 1.23: 直观解释全概率公式。

问题 1.5. 还是问题 1.4 的条件，如果摸到白球，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。

定理 1.12 (Bayes 公式或逆概率公式)**.** 还是全概率公式的条件，则对任一事件 B，如果 $\mathsf{P}(B) > 0$，我们有

$$\mathsf{P}(A_j|B) = \frac{\mathsf{P}(A_j)\mathsf{P}(B|A_j)}{\sum_{j=1}^{\infty} \mathsf{P}(A_j)\mathsf{P}(B|A_j)} \tag{1.43}$$

证明. 由 $\mathsf{P}(A_j|B) = \mathsf{P}(A_j)\mathsf{P}(B|A_j)/\mathsf{P}(B)$ 和全概率公式可证。 □

推论 1.7. 由 Bayes 公式，总有 $\sum_{j=1}^{\infty} \mathsf{P}(A_j|B) = 1$。

例 1.41. 在问题 1.4 中，把有关 $\mathsf{P}(A_1), \mathsf{P}(A_2)$ 的条件放宽为 $\mathsf{P}(A_1)+\mathsf{P}(A_2) = 1$，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 $\mathsf{P}(A_1) = p, \mathsf{P}(A_2) = 1 - p$，其中 $0 < p < 1$。

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

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）。而在我们的理念中，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，$\mathsf{P}(A_1|\text{白})$ 和 $\mathsf{P}(A_2|\text{白})$ 才具有客观意义，进而利用 Bayes 公式做推断才是合理的。按照不充分理由原则 (principle of insufficient reason)*，我们假定 $\mathsf{P}(A_1) = \mathsf{P}(A_2) = 1/2$（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 $\mathsf{P}(A_1|\text{白球}) = 3/4 > \mathsf{P}(A_2|\text{白球}) = 1/4$，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“相信某随机事件会发生”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 $\mathsf{P}(A_1|\text{白球})$ 与试验之前的信念度 $\mathsf{P}(A_1)$ 的差别 $\mathsf{P}(A_1|\text{白球}) - \mathsf{P}(A_1)$，不正表示了观察数据“白球”给信念度带来的变化吗？

*经济学家 John Maynard Keynes (1883-1946) 也称之为无差别原则 (principle of indifference)，它约定：当我们对基本事件的概率一无所知的时候，每个基本事件都假定是等概率的。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。

$\mathsf{P}(A_1|\text{白球}) - \mathsf{P}(A_1) = 3p/(2p+1) - p > 0$ 和 $\mathsf{P}(A_2|\text{白球}) - \mathsf{P}(A_2) = (1-p)/(2p+1) - (1-p) < 0$ 说明观察数据支持答案“A_1”，不管验前认为 $\mathsf{P}(A_1)$ 有多大。这种考察验前、验后信念度的改变来评判观察数据支持哪个论断的方法便是贝叶斯推断 (Bayesian inference) 的方法。

例 1.42. 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“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 $\mathsf{P}(A_1) = 1/10, \mathsf{P}(A_2) = 9/10$。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，即患有 Z 病者被诊断为“有 Z 病”或“阳性”的概率 $\mathsf{P}(\text{白球}|A_1)$，此值越大诊断越灵敏。此例中，$\mathsf{P}(\text{白球}|A_1) = 3/5$ 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，即未患 Z 病者被诊断为“没 Z 病”或“阴性”的概率 $\mathsf{P}(\text{黑球}|A_2)$，此值越大诊断越精确。此例中，$\mathsf{P}(\text{黑球}|A_2) = 4/5$ 说明医院诊断没 Z 病的正确率为 4/5。

应用上面刚介绍过的两个推断方法，我们再讨论上例的问题。

- 频率派：因为 $\mathsf{P}(A_1|\text{白球}) = 1/4 < \mathsf{P}(A_2|\text{白球}) = 3/4$，所以选“$A_2$”。这意味着，医院诊断有 Z 病，实际未患 Z 病的可能性更大些。使用这种推断方法的患者会乐观地想，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，我为什么那么倒霉就是那些少数人之一呢？一定是医院诊断有误！

- 贝叶斯学派：如果医院已诊断有 Z 病，则应该更相信患有 Z 病。这似乎更符合常人的心态——医院的诊断具有一定的说服力，看到这样的诊断书谁会高兴起来呢？

1.3.3 随机事件的独立性

如果 $\mathsf{P}(B)>0$，$\mathsf{P}(A|B)=\mathsf{P}(A)$ 意味着事件 B 发生与否丝毫不影响 A 的概率，由此引出了独立事件的定义。

定义 1.14 (独立性)**.** 已知概率空间 $(\Omega,\mathscr{S},\mathsf{P})$，事件 $A,B\in\mathscr{S}$ 相互独立 (independent) 当且仅当 $\mathsf{P}(AB)=\mathsf{P}(A)\mathsf{P}(B)$，记作 $A\perp\!\!\!\perp B$。更一般地，事件 $A_1,A_2,\cdots,A_n\in\mathscr{S}$ 独立，记作 $\perp\!\!\!\perp\{A_1,A_2,\cdots,A_n\}$，当且仅当对于 $1,2,\cdots,n$ 的任意子序列 $k_1<k_2<\cdots<k_s$，

$$\mathsf{P}(A_{k_1}A_{k_2}\cdots A_{k_s})=\mathsf{P}(A_{k_1})\mathsf{P}(A_{k_2})\cdots\mathsf{P}(A_{k_s}) \tag{1.44}$$

性质 1.8 (独立与互斥)**.** 如果事件 A 与 B 独立，且 $\mathsf{P}(A)>0,\mathsf{P}(B)>0$，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 $\mathsf{P}(A)>0,\mathsf{P}(B)>0$，则 A 与 B 不独立。

定理 1.13. 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 $\mathsf{P}(AB^c)=\mathsf{P}(A-AB)=\mathsf{P}(A)-\mathsf{P}(AB)=\mathsf{P}(A)[1-\mathsf{P}(B)]=\mathsf{P}(A)\mathsf{P}(B^c)$，于是 $\perp\!\!\!\perp\{A,B^c\}$ 得证。其他的证明类似。 □

练习 1.13. 如果事件 $A_1,A_2,\cdots,A_n$ 独立，则 $A_1^c,A_2^c,\cdots,A_n^c$ 独立。

练习 1.14. 已知概率空间 $(\Omega,\mathscr{S},\mathsf{P})$，且事件 $A,B,C\in\mathscr{S}$ 满足 $\mathsf{P}(A)>0,\mathsf{P}(B)>0$。试证明：

- 如果 A,B 独立，则 $\mathsf{P}(C|A)=\mathsf{P}(B)\mathsf{P}(C|AB)+\mathsf{P}(B^c)\mathsf{P}(C|AB^c)$。
  提示：利用全概率公式 $\mathsf{P}_A(C)=\mathsf{P}_A(B)\mathsf{P}_A(C|B)+\mathsf{P}_A(B^c)\mathsf{P}_A(C|B^c)$。
- 若上式成立并满足 $\mathsf{P}(C|AB)\neq\mathsf{P}(C|A)$ 且 $\mathsf{P}(C)>0$，则 A,B 独立。
  提示：由条件 $\mathsf{P}(C|AB)\neq\mathsf{P}(C|A)$ 可证得 $\mathsf{P}_A(C|B)\neq\mathsf{P}_A(C|B^c)$，从 $[\mathsf{P}(B)-\mathsf{P}_A(B)][\mathsf{P}_A(C|B)-\mathsf{P}_A(C|B^c)]=0$ 得出 $\mathsf{P}(B)=\mathsf{P}_A(B)$。

定理 1.14. 已知随机事件 $A_1, A_2, \cdots, A_n, \cdots$ 相互独立且 $\sum_{n=1}^{\infty} \mathsf{P}(A_n) = \infty$，则 P(无穷多个 A_n 发生) = 1。

证明. 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

$$\begin{aligned}
\mathsf{P}\left\{\text{无穷多个 } A_n \text{ 发生}\right\} &= \lim_{k\to\infty} \mathsf{P}\left(\bigcup_{n=k}^{\infty} A_n\right) \\
&= \lim_{k\to\infty}\left[1 - \mathsf{P}\left(\bigcap_{n=k}^{\infty} A_n^c\right)\right] = 1 - \lim_{k\to\infty} \mathsf{P}\left(\bigcap_{n=k}^{\infty} A_n^c\right) = 1，\text{这是因为} \\
\mathsf{P}\left(\bigcap_{n=k}^{\infty} A_n^c\right) &= \prod_{n=k}^{\infty} \mathsf{P}(A_n^c) = \prod_{n=k}^{\infty}[1 - \mathsf{P}(A_n)]，\text{由不等式 } 1 - x \le e^{-x} \text{ 得出} \\
&\le \prod_{n=k}^{\infty} \exp\{-\mathsf{P}(A_n)\} = \exp\left\{-\sum_{n=k}^{\infty} \mathsf{P}(A_n)\right\} = 0
\end{aligned}$$

最后一步是因为对于任意有限的 k，皆有 $\sum_{n=k}^{\infty} \mathsf{P}(A_n) = \infty$，得证。 □

由 Borel-Cantelli 引理 1.1 及定理 1.14，对于一列相互独立的随机事件 $A_1, A_2, \cdots, A_n, \cdots$，总有下面的结论成立。

$$\mathsf{P}\left(\overline{\lim_{n\to\infty}} A_n\right) = \begin{cases} 0 & \text{当且仅当级数 } \sum_{n=1}^{\infty} \mathsf{P}(A_n) \text{ 收敛} \\ 1 & \text{当且仅当级数 } \sum_{n=1}^{\infty} \mathsf{P}(A_n) \text{ 发散} \end{cases} \tag{1.45}$$

即按照级数数 $\sum_{n=1}^{\infty} \mathsf{P}(A_n)$ 是否发散来判定 $\{A_n : n = 1, 2, \cdots\}$ 中无穷多个事件同时发生的概率为 1 还是为 0。判定准则式 (1.45) 被称为 Borel 0-1 律或者 Borel 0-1 准则，其中判定事件 $\overline{\lim_{n\to\infty}} A_n$ 的概率为 0 时不需要 $\{A_n\}$ 的独立性假设。这一结果将在 §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，例如下面的数论问题。

问题 1.6. 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，随机选取一个小数，问它的小数点后面数字 0 出现的概率？

解. 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 发现了下述性质*：

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

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 $1,2,\cdots,k-1$ 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 5.10）：

$$\mathsf{P}\left\{\overline{\lim_{n\to\infty}}\frac{\left|S_n(x)-\frac{n}{10}\right|}{\sqrt{n\ln\ln n}}=\frac{3\sqrt{2}}{10}\right\}=1$$

问题 1.7. (1) 事件 A_1,A_2,A_3 两两独立，问 A_1,A_2,A_3 是否独立？(2) 如果 $\mathsf{P}(A_1A_2A_3)=\mathsf{P}(A_1)\mathsf{P}(A_2)\mathsf{P}(A_3)$，问 A_1,A_2,A_3 是否独立？

例 1.43 (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“球标号的第 k 个位置是 1”，其中 $k=1,2,3$，则 $\mathsf{P}(A_1)=\mathsf{P}(A_2)=\mathsf{P}(A_3)=1/2,\mathsf{P}(A_1A_2A_3)=0$。

- $\mathsf{P}(A_1|A_2)=\mathsf{P}(A_1)=1/2$，即 A_1 与 A_2 独立。类似地，A_1 与 A_3 独立，A_2 与 A_3 独立。
- $\mathsf{P}(A_1)\mathsf{P}(A_2)\mathsf{P}(A_3)=1/8$，故 A_1,A_2,A_3 不独立。

例 1.44 (Kac 反例)**.** 已知基本事件集合 $\Omega=\{1,2,3,4\}$，以及 $\mathsf{P}(\{1\})=\sqrt{2}/2-1/4,\mathsf{P}(\{2\})=1/4,\mathsf{P}(\{3\})=3/4-\sqrt{2}/2,\mathsf{P}(\{4\})=1/4$。令 $A_1=\{1,3\},A_2=\{2,3\},A_3=\{3,4\}$，则 $\mathsf{P}(A_1)=\mathsf{P}(\{1\})+\mathsf{P}(\{3\})=1/2,\mathsf{P}(A_2)=\mathsf{P}(A_3)=1-\sqrt{2}/2$。

- 满足 $\mathsf{P}(A_1A_2A_3)=\mathsf{P}(\{3\})=3/4-\sqrt{2}/2=\mathsf{P}(A_1)\mathsf{P}(A_2)\mathsf{P}(A_3)$。
- $\mathsf{P}(A_1A_2)\neq\mathsf{P}(A_1)\mathsf{P}(A_2)$ 说明 A_1,A_2,A_3 不独立。

*说某事件“以概率 1”发生，在概率论里可算作很强烈的“语气”了，常用“几乎必然”(almost surely, *a.s.*) 来作它的同义语。

✂ **例 1.45** (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 $(\Omega_n, \mathscr{S}_n, \mathsf{P}_n)$ 定义为 $\Omega_n = \{1,2,\cdots,n\}\times\{1,2,\cdots,n\}, \mathscr{S}_n = 2^{\Omega_n}$ 且 $\forall\omega\in\Omega_n, \mathsf{P}_n(\omega) = 1/n^2$。定义概率 $\mathsf{P}(n)$ 如下：

$$\mathsf{P}(n) = \mathsf{P}_n\left\{\text{自然数对 } (a,b)\in\Omega_n \text{ 互素}\right\},\ \text{其中 } n\geq 2$$

如果原问题有解，则意味着 $n\to\infty$ 时 $\mathsf{P}(n)$ 的极限存在。为了直观地了解 $\mathsf{P}(n)$，构造 Maxima 的函数 `P(n)` 如下，并绘出它的曲线以探究 $\mathsf{P}(n)$ 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) */
P(n) := (s:1,              /* s 表示互素自然对 (a,b) 的个数 */
  for j: 2 while j <= n do
  s: s + 2*totient (j), /* totient(j): 不超过 j 且与 j 互素的自然数个数 */
  float(s/n^2)) $

/* 定义长度为 MAX 的数组 */
MAX: 10^2 $               /* MAX 为用户给定的上限 */
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,           /* 自变量数组的赋值 */
     y[n]: P(n+1)) $      /* 因变量数组的赋值 */
load(draw) $              /* 导入绘图包 */
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)) $
```

❑ 设 $\{2,\cdots,n\}$ 里所有的素数为 $p_1 < \cdots < p_{r_n}$，若分数 a/b 可约，则存在素数 $p\in\{p_1,\cdots,p_{r_n}\}$ 是 a,b 公因子，即 a,b 选自

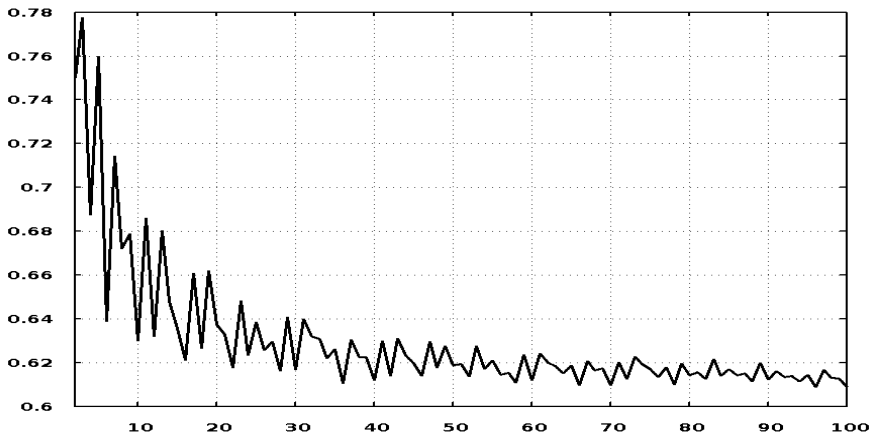


图 1.24: Chebyshev 问题：令 $\mathsf{P}(n)$ 为分子、分母取自 $\{1,2,\cdots,n\}$ 的不可约分数的概率。通过 $\mathsf{P}(2),\cdots,\mathsf{P}(100)$ 的折线图，猜测 n 增大时，$\mathsf{P}(n)$ 震荡着趋近某个值。事实上，$\lim\limits_{n\to\infty}\mathsf{P}(n)=6/\pi^2\approx 0.60792710185403$。

$p,2p,\cdots,kp$，其中 $k=\lfloor n/p\rfloor$。

所以，$\dfrac{(k-1)^2}{n^2}<\mathsf{P}_n\left\{\text{素数 } p \text{ 是 } a,b \text{ 的公因子}\right\}=\dfrac{k^2}{n^2}$，进而

$$1-\frac{1}{p^2}\le \mathsf{P}_n\left\{\text{素数 } p \text{ 不是 } a,b \text{ 的公因子}\right\}=1-\frac{k^2}{n^2}<1-\frac{1}{p^2}+\frac{4}{np}$$

显然，当 $n\to\infty$ 时，$\mathsf{P}_n\left\{\text{素数 } p \text{ 不是 } a,b \text{ 的公因子}\right\}\to 1-1/p^2$。

❑ 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 $\mathsf{P}_n(A_iA_j)$，其中 $i\neq j$。

$$\begin{aligned}\mathsf{P}_n(A_iA_j)&=1-\mathsf{P}_n(A_i^c)-\mathsf{P}_n(A_j^c)+\mathsf{P}_n(A_i^cA_j^c)\\&=1-\mathsf{P}_n(A_i^c)-\mathsf{P}_n(A_j^c)+\frac{k_{ij}^2}{n^2},\ \text{其中 } k_{ij}=\left\lfloor\frac{n}{p_ip_j}\right\rfloor\\&\to 1-\frac{1}{p_i^2}-\frac{1}{p_j^2}+\frac{1}{p_i^2p_j^2}=\mathsf{P}_n(A_i)\mathsf{P}_n(A_j),\ \text{当 } n\to\infty \text{ 时}\end{aligned}$$

❑ 类似地，利用 Jordan 式 (1.33) 可证：当 n 充分大时，近似地有

$A_1,\cdots,A_{r_n}$ 独立。于是，任选一个分数不可约的概率是

$$\mathsf{P}\left\{\text{分数不可约}\right\}=\lim_{n\to\infty}\mathsf{P}(n)=\lim_{n\to\infty}\prod_{j=1}^{r_n}\mathsf{P}_n(A_j)=\prod_{p\text{ 是素数}}\left(1-\frac{1}{p^2}\right)$$
$$=\left[\prod_{p\text{ 是素数}}\left(1+\frac{1}{p^2}+\frac{1}{p^4}+\cdots\right)\right]^{-1}=\left[\sum_{n=1}^{\infty}\frac{1}{n^2}\right]^{-1}=\frac{6}{\pi^2}\qquad(1.46)$$

表 1.2: Chebyshev 问题：计算 $\mathsf{P}(10^k)$，发现 $\mathsf{P}(10^4)$ 已很接近极限值 $6/\pi^2$。

k	$\mathsf{P}(10^k)$	$\mathsf{P}(10^k)-6/\pi^2$
1	0.63	2.2073e-2
2	0.6087	7.7290e-4
3	0.608383	4.5590e-4
4	0.60794971	2.2608e-5
5	0.6079301507	3.0488e-6
6	0.607927104783	2.9290e-9
7	0.60792712854483	2.6691e-8
8	0.6079271032731814	1.4192e-9

注记 1.5. 例 1.45 非常有趣，它牵扯到数论中的一些经典结果。在式 (1.46) 的推导中用到了 $1-x=(1+x+x^2+x^3+\cdots)^{-1}$，也可以直接推得结果为 $1/\zeta(2)=6/\pi^2$，其中单复变函数 $\zeta(s)=\sum_{n=1}^{\infty}1/n^s$ 是定义在区域 $\{s=\sigma+it\in\mathbb{C}:s\text{ 的实部 }\Re(s)=\sigma>1\}$ 上的 ζ 函数*，这是因为瑞士数学家 Leonhard Paul Euler (1707-1783) 证得：当实数 $x>1$ 时，

$$\zeta(x)\equiv\sum_{n=1}^{\infty}\frac{1}{n^x}=\prod_{p\text{ 是素数}}\left(1-\frac{1}{p^x}\right)^{-1}$$

*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 $s=-2,-4,-6,\cdots$ 等负偶数是 ζ 函数的平凡零点。Riemann 在该论文中提出了猜想：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，至今尚未被证明或证伪。1896 年，法国数学家 Jacques Salomon Hadamard (1865-1963) 和比利时数学家 Charles Jean de la Vallée-Poussin (1866-1962) 通过证明 ζ 函数没有形如 $1+it$ 的零点证明了素数定理：令 $\pi(n)$ 表示 $\{2,3,\cdots,n\}$ 里素数的个数，则 $\pi(n)\sim n/\ln n$。

例 1.46. 已知电话在时间段 t 内被呼叫 k 次的概率是

$$\mathsf{P}_t(k) = \frac{(at)^k}{k!}e^{-at}，\text{其中 } a \text{ 为常数，} k = 0, 1, 2, \cdots \tag{1.47}$$

如果相邻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 $\mathsf{P}_{2t}(s)$？其中 $s = 0, 1, 2, \cdots$。

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 $A_{2t}^s = A_t^0A_t^s + A_t^1A_t^{s-1} + \cdots + A_t^sA_t^0$。

$$\begin{aligned}
\mathsf{P}_{2t}(s) &= \mathsf{P}\left(A_{2t}^s\right) \\
&= \sum_{j=0}^{s} \mathsf{P}\left(A_t^jA_t^{s-j}\right) \\
&= \sum_{j=0}^{s} \mathsf{P}_t(j)\mathsf{P}_t(s-j) \\
&= \sum_{j=0}^{s} \frac{(at)^s e^{-2at}}{j!(s-j)!} \\
&= \frac{(2at)^s}{s!}e^{-2at}
\end{aligned}$$

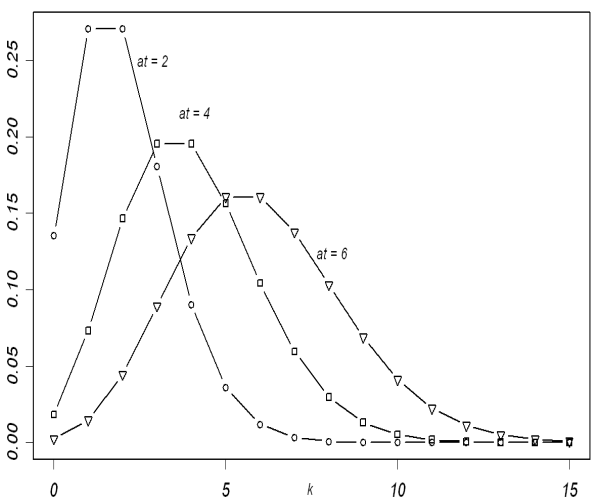


图 1.25: 在例 1.46 中，at 分别取 $\lambda = 2, 4, 6$ 时 $\mathsf{P}(k) = \lambda^k e^{-\lambda}/k!$ 的折线图：λ 越大折线图越呈现中间高两头低的对称性。

问题 1.8. 连续抛一枚硬币，每次抛都是独立的且 $\mathsf{P}(H) = \mathsf{P}(T) = 1/2$。若头 10 次都是正面，有人觉得第 11 次抛该硬币出现反面的机会更大些，因为 Bernoulli 弱大数律说，抛的次数趋向无穷，正面频率接近 $\mathsf{P}(H) = 1/2$ 的概率趋向 1。你如何认为？

解. 机会还是 1/2，因为大自然对独立试验的结果没有记忆。如果不知道正面出现的概率，则需要利用统计方法通过样本对之进行估计，这种情况下推断自然要受到观察结果的影响（见本书的第二部分）。

1.3.4* 条件独立性及其性质

定义 1.15 (条件独立性)**.** 已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$，且事件 $B \in \mathscr{S}$ 满足 $\mathsf{P}(B) > 0$。由定理 1.9 定义的条件概率空间 $(\Omega, \mathscr{S}, \mathsf{P}_B)$ 中，如果 $A_1, A_2 \in \mathscr{S}$ 使得 $\mathsf{P}_B(A_1A_2) = \mathsf{P}_B(A_1)\mathsf{P}_B(A_2)$，则称 A_1, A_2 关于 B 条件独立 [30,67]，记作 $A_1 \perp\!\!\!\perp A_2|B$ 或 $A_2 \perp\!\!\!\perp A_1|B$。更一般地，事件 $A_1, A_2, \cdots, A_n \in \mathscr{S}$ 关于 B 条件独立，记作 $\perp\!\!\!\perp \{A_1, A_2, \cdots, A_n\}|B$，当且仅当对于 $1, 2, \cdots, n$ 的任意子序列 $k_1 < k_2 < \cdots < k_s$，

$$\mathsf{P}_B(A_{k_1}A_{k_2}\cdots A_{k_s}) = \mathsf{P}_B(A_{k_1})\mathsf{P}_B(A_{k_2})\cdots\mathsf{P}_B(A_{k_s}) \tag{1.48}$$

性质 1.9. 与性质 1.8 类似，如果 A, B 互斥，即 $AB = \emptyset$，且 $\mathsf{P}(A_1|B) > 0, \mathsf{P}(A_2|B) > 0$，则 A_1, A_2 关于 B 不条件独立。

例 1.47. 由 $\mathsf{P}(A_1A_2|B) = \mathsf{P}(A_1|B)\mathsf{P}(A_2|B)$ 能推导出 $\mathsf{P}(A_1A_2) = \mathsf{P}(A_1)\mathsf{P}(A_2)$ 或 $\mathsf{P}(A_1A_2|B^c) = \mathsf{P}(A_1|B^c)\mathsf{P}(A_2|B^c)$ 吗？

解. 不能！下面构造一个反例。左下图是一个 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 $\mathsf{P}(A_1A_2|B) = 2/12 = 1/6$ 且 $\mathsf{P}(A_1|B)\mathsf{P}(A_2|B) = 6/12 \cdot 4/12 = 1/6$。显然 $\mathsf{P}(A_1A_2) = 1/6 < \mathsf{P}(A_1)\mathsf{P}(A_2)$。事件 A_1, A_2 关于 B^c 不是条件独立的，这是因为 $\mathsf{P}(A_1A_2|B^c) = 1/6$，但 $\mathsf{P}(A_1|B^c) = 13/24$，$\mathsf{P}(A_2|B^c) = 9/24$。

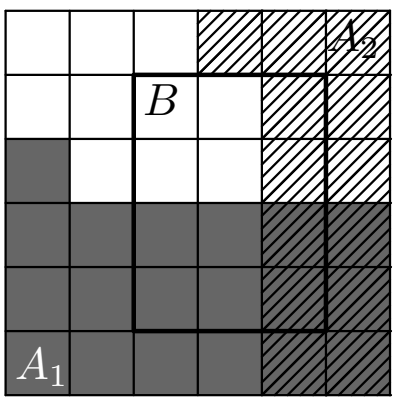


图 1.26: 条件独立 $\not\Rightarrow$ 独立。

性质 1.10. 若 Ω 是基本事件集合且 $A_1 \perp\!\!\!\perp A_2|\Omega$，则 $A_1 \perp\!\!\!\perp A_2$。即独立性是条件独立性的一个特例，请读者验证之。

定理 1.15. 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。

证明. 与定理 1.13 的证明类似，请读者补全。 □

例 1.48 (兼听则明，偏听则暗)**.** 某投资人总是听取三个理财顾问 A_1, A_2, A_3 中的多数意见。令 G 表示市场利于投资，已知 $\mathsf{P}(G) = 0.7$。$\mathsf{P}(A_1|G) = 0.95$ 表示市场利于投资的条件下顾问 A_1 建议投资的概率，$\mathsf{P}(A_1|G^c) = 0.2$ 则表示市场不利于投资的条件下顾问 A_1 建议投资的概率，显然它们刻画了顾问 A_1 的理财水平。其他两个理财顾问的水平情况是：$\mathsf{P}(A_2|G) = 0.75, \mathsf{P}(A_2|G^c) = 0.1, \mathsf{P}(A_3|G) = 0.8, \mathsf{P}(A_3|G^c) = 0.25$。假设三个理财顾问独立工作、互不影响，即 $\perp\!\!\!\perp \{A_1, A_2, A_3\}|G$ 且 $\perp\!\!\!\perp \{A_1, A_2, A_3\}|G^c$。试问：投资人决策正确的概率 $\mathsf{P}(D)$?

解. 由全概率公式得 $\mathsf{P}(D) = \mathsf{P}(D|G)\mathsf{P}(G) + \mathsf{P}(D|G^c)\mathsf{P}(G^c)$，其中，

$$\mathsf{P}(D|G) = \mathsf{P}(A_1A_2A_3|G) + \mathsf{P}(A_1A_2A_3^c|G) + \mathsf{P}(A_1A_2^cA_3|G) + \mathsf{P}(A_1^cA_2A_3|G)$$
$$\mathsf{P}(D|G^c) = \mathsf{P}(A_1^cA_2^cA_3^c|G^c) + \mathsf{P}(A_1A_2^cA_3^c|G^c) + \mathsf{P}(A_1^cA_2^cA_3|G^c) + \mathsf{P}(A_1^cA_2A_3^c|G^c)$$

由条件独立性假设 $\perp\!\!\!\perp \{A_1, A_2, A_3\}|G$ 和 $\perp\!\!\!\perp \{A_1, A_2, A_3\}|G^c$ 可得

$$\mathsf{P}(A_1A_2A_3|G) = \mathsf{P}(A_1|G)\mathsf{P}(A_2|G)\mathsf{P}(A_3|G) = 0.95 \cdot 0.75 \cdot 0.8 = 0.57$$
$$\mathsf{P}(A_1^cA_2^cA_3^c|G^c) = \mathsf{P}(A_1^c|G^c)\mathsf{P}(A_2^c|G^c)\mathsf{P}(A_3^c|G^c) = 0.8 \cdot 0.9 \cdot 0.75 = 0.54$$

其他项可类似计算。于是，$\mathsf{P}(D|G) = 0.9325, \mathsf{P}(D|G^c) = 0.915$，最终结果为 $\mathsf{P}(D) = 0.92725$。事实上，无论市场如何风云变化，即无论 $\mathsf{P}(G)$ 如何取值，$\mathsf{P}(D)$ 总是高于任何一个理财顾问的水平。

注记 1.6. 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

*学习机 (learner) 就是实现某一特定任务的算法，它能够通过经验修改自身以求达到更好的效果。如，人工神经网络 (artificial neural network)、支持向量机 (support vector machine, SVM) 等 [20,21,77]。

下其效果比依靠单个专家的要好些。但这并不意味着“三个臭皮匠凑个诸葛亮”——如果皮匠水平太臭，再多也没用。如何为决策问题选择“专家团”是机器学习关注的话题，一般来说“专家”之间要相互独立、各有所长等等。

练习 1.15. 一群学生中男女各半，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玩电脑游戏。令 M 表示事件“所选学生为男生”，B 表示事件“所选学生喜欢打篮球”，C 表示事件“所选学生喜欢玩电脑游戏”。已知条件可形式地表示为

$$\mathsf{P}(B|M) = 3/4, \qquad \mathsf{P}(B|M^c) = 1/4$$
$$\mathsf{P}(C|M) = \mathsf{P}(C|M^c) = 1/2, \qquad \mathsf{P}(M) = \mathsf{P}(M^c) = 1/2$$

假设男生喜欢打篮球和喜欢玩电脑游戏是独立的，即 $\perp\!\!\!\perp \{B, C\}|M$，女生也是如此，即 $\perp\!\!\!\perp \{B, C\}|M^c$。试证明：$B, C$ 独立，但 B, C, M 不独立。

问题 1.9. 一般情况下能从 $\perp\!\!\!\perp \{B, C\}|M$ 且 $\perp\!\!\!\perp \{B, C\}|M^c$ 推导出 $\perp\!\!\!\perp \{B, C\}$ 吗？若是，请证明；若否，请构造反例。

例 1.49 (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：

1. 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 $L = (w_1, w_2, \cdots, w_n)$。

2. 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 $\mathsf{P}(S|w_j) = p_j$，进而 $\mathsf{P}(S^c|w_j) = 1 - p_j$。

*也可以仅考虑名词，或者用户感兴趣的词集。

3. 为了降低计算复杂度，一般假设 $w_1, w_2, \cdots, w_n$ 关于 S 和关于 S^c 都是条件独立的，即

$$\mathsf{P}(L|S) = \mathsf{P}(w_1|S)\cdots\mathsf{P}(w_n|S) = \prod_{j=1}^{n} p_j \prod_{j=1}^{n} \mathsf{P}(w_j)/[\mathsf{P}(S)]^n$$

$$\mathsf{P}(L|S^c) = \mathsf{P}(w_1|S^c)\cdots\mathsf{P}(w_n|S^c) = \prod_{j=1}^{n}(1-p_j) \prod_{j=1}^{n} \mathsf{P}(w_j)/[\mathsf{P}(S^c)]^n$$

显然这个假设不符合语言学事实，但为了使算法可行，模型粗糙一点儿也是迫不得已。

4. 由无差别原则，假定垃圾邮件和非垃圾邮件的验前概率为 $\mathsf{P}(S) = \mathsf{P}(S^c) = 1/2$，并利用 Bayes 公式计算 S 的验后概率

$$\mathsf{P}(S|L) = \frac{\mathsf{P}(L|S)}{\mathsf{P}(L|S) + \mathsf{P}(L|S^c)} = \frac{\prod_{j=1}^{n} p_j}{\prod_{j=1}^{n} p_j + \prod_{j=1}^{n}(1-p_j)} \tag{1.49}$$

再利用例 1.41 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。

性质 1.11 (条件独立的性质)**.** 已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$。在此性质中，我们约定“$A_1 \perp\!\!\!\perp A_2|B$”暗含着 $A_1, A_2, B \in \mathscr{S}$ 且 $\mathsf{P}(BA_1A_2) > 0$。

1. 如果 $A_1 \perp\!\!\!\perp A_2|B$，则 $\mathsf{P}(A_1|B) = \mathsf{P}(A_1|BA_2)$。这意味着，即使 A_2 发生了，也不影响 $\mathsf{P}(A_1|B)$。同理，$\mathsf{P}(A_2|B) = \mathsf{P}(A_2|BA_1)$。

2. 如果 $(A_1A_2) \perp\!\!\!\perp A_3|B$，则 $A_1 \perp\!\!\!\perp A_3|B$ 且 $A_2 \perp\!\!\!\perp A_3|B$。

3. 如果 $A_1 \perp\!\!\!\perp (A_2A_3)|B$，则 $A_1 \perp\!\!\!\perp A_2|BA_3$。

4. 如果 $A_1 \perp\!\!\!\perp A_2|BA_3, A_1 \perp\!\!\!\perp A_3|B$，则 $A_1 \perp\!\!\!\perp (A_2A_3)|B$。

5. 如果 $A_1 \perp\!\!\!\perp A_2|BA_3, A_1 \perp\!\!\!\perp A_3|BA_2$，则 $A_1 \perp\!\!\!\perp (A_2A_3)|B$。

证明. 性质 (2) 之 $(A_1A_2) \perp\!\!\!\perp A_3|B \Rightarrow A_2 \perp\!\!\!\perp A_3|B$ 留给读者练习。

$$
\begin{aligned}
\mathsf{P}(A_1|BA_2) &= \frac{\mathsf{P}(A_1A_2|B)\mathsf{P}(B)}{\mathsf{P}(BA_2)} = \frac{\mathsf{P}(A_1|B)\mathsf{P}(A_2|B)\mathsf{P}(B)}{\mathsf{P}(A_2|B)\mathsf{P}(B)} \\
&= \mathsf{P}(A_1|B)，\text{性质 (1) 得证} \\
\mathsf{P}(A_1A_3|B) &= \frac{\mathsf{P}(A_1|B)\mathsf{P}(A_3|BA_1)}{\mathsf{P}(A_3|B)}\mathsf{P}(A_3|B)，\text{由式 (1.41)} \\
&= \mathsf{P}(A_1|B)\mathsf{P}(A_3|B)\frac{\mathsf{P}(A_3|BA_1A_1A_2)}{\mathsf{P}(A_3|BA_1A_2)}，\text{由已知条件} \\
&= \mathsf{P}(A_1|B)\mathsf{P}(A_3|B)，\text{性质 (2) 得证} \\
\mathsf{P}(A_1A_2|BA_3) &= \mathsf{P}(A_1|BA_3)\frac{\mathsf{P}(A_2A_3|BA_1)}{\mathsf{P}(A_3|BA_1)}，\text{由式 (1.41)} \\
&= \mathsf{P}(A_1|BA_3)\frac{\mathsf{P}(A_2A_3|B)}{\mathsf{P}(A_3|B)}，\text{由性质 (1)、(2) 和已知条件} \\
&= \mathsf{P}(A_1|BA_3)\mathsf{P}(A_2|BA_3)，\text{性质 (3) 得证} \\
\mathsf{P}(A_1A_2A_3|B) &= \mathsf{P}(A_1A_2|BA_3)\mathsf{P}(A_3|B)，\text{由式 (1.41)} \\
&= \mathsf{P}(A_1|BA_3)[\mathsf{P}(A_2|BA_3)\mathsf{P}(A_3|B)]，\text{由已知条件} \\
&= \mathsf{P}(A_1|B)\mathsf{P}(A_2A_3|B)，\text{性质 (4) 得证} \\
\mathsf{P}(A_1A_2A_3|B) &= \mathsf{P}(A_1A_3|BA_2)\mathsf{P}(A_2|B) \\
&= \mathsf{P}(A_1|BA_2)[\mathsf{P}(A_3|BA_2)\mathsf{P}(A_2|B)] \\
&= \mathsf{P}(A_1|BA_2A_3)\mathsf{P}(A_2A_3|B)，\text{由性质 (1) 和已知条件} \\
&= \mathsf{P}(A_1|B)\mathsf{P}(A_2A_3|B)，\text{性质 (5) 得证}
\end{aligned}
$$

□

1.4 习题

1.1. 掷 3 粒骰子，问至少有 1 粒出现 6 点的概率？

1.2. 投掷一对均匀的骰子两次，求在两次中均得不到点数之和为 7 点或 11 点的概率。

1.3. 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？

1.4. k 个盒子各装 n 个球，编号为 $1,2,\cdots,n$，从每个盒子中各取一个球，计算所得到的 k 个球中最大编号是 m 的概率。

1.5. 某射手在 3 次射击中至少命中 1 次的概率为 0.875，则该射手在 1 次射击中命中的概率为多少？

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. 10 人中有一对夫妇，他们随机地围坐在一张圆桌周围聊天，求这对夫妇正好坐在一起的概率？

1.8. 将 n 个球放入 n 个盒子，问恰有一个盒子空着的概率？

☆ 1.9. 盒子里有 n 球，编号为 $1,2,\cdots,n$。现有编号为 $1,2,\cdots,n$ 的 n 个人分别随机地从盒子中取走一个球，求至少有一个人拿到相同编号的概率？

1.10. 例 1.10 中的抽取是无放回的，当 N 充分大而 m 不大时，$\mathsf{P}(A_k) \approx C_m^k p^k (1-p)^{m-k}$，其中 $p = n/N$，即有放回的抽取和无放回的抽取相差无几。

1.11. 盒子里装有 w 个白球，b 个黑球。不放回地一次一个摸取，试求：(1) 同色球可辨；(2) 同色球不可辨这两种情况下，第 k 次摸出白球的概率，其中 $1 \le k \le w+b$。

1.12. 利用球-盒子模型证明：已知非负整数 m,n,k,r 满足 $k \le \min(m,n)$ 且 $r+k \le n$，则 $\sum_{j=0}^{k} C_m^{k-j} C_n^j C_{n-j}^r = C_{m+n-r}^k C_n^r$。

☆ 1.13. 利用 Maxima “证明”李善兰恒等式*：已知非负整数 m,n 满足 $n \le m$，则

$$\sum_{j=0}^{n} \left(C_n^j\right)^2 C_{m+2n-j}^{2n} = \left(C_{m+n}^n\right)^2 \tag{1.50}$$

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，则事件“两个数之和小于 6/5”的概率是多少？

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. 已知随机事件 A,B,C 的概率为 $\mathsf{P}(A) = \mathsf{P}(B) = \mathsf{P}(C) = 1/4$ 且 $\mathsf{P}(AB) = \mathsf{P}(AC) = 0, \mathsf{P}(BC) = 1/8$，试用集合运算表示下述事件并求出它们相应的概率。

- ❑ A,B,C 同时发生。
- ❑ A,B,C 中至少有一个事件发生。
- ❑ A,B 不发生，C 发生。
- ❑ A,B,C 中至少有两个事件发生。

1.18. 不论随机事件 A 的概率 $\mathsf{P}(A) > 0$ 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。

---

*李善兰 (1810-1882)，字竟芳，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。

1.19. 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。

1.20. A, B 是两个随机事件，试证明：$|\mathsf{P}(AB) - \mathsf{P}(A)\mathsf{P}(B)| \le 1/4$。

1.21. 已知随机事件序列 $A_1, A_2, \cdots$ 满足 $\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty$，试证明：$\mathsf{P}(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c) = 1$ 并说明其含义。

1.22. 设 Bernoulli 试验中事件 A 的概率 $\mathsf{P}(A) = 1/2$。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。

1.23. 设 A, B 是任意二事件，其中 A 的概率不等于 0 和 1，证明：$\mathsf{P}(B|A) = \mathsf{P}(B|A^c)$ 是事件 A 与 B 独立的充分必要条件。

1.24. Kolmogorov 公理体系中的归一性和可列可加性等价于：若 $\{B_k \in \mathscr{S} : k = 1, 2, \cdots\}$ 是 Ω 的一个划分，则 $\sum_{k=1}^{\infty} \mathsf{P}(B_k) = 1$。

1.25. 假设 $(\Omega, \mathscr{S}, \mathsf{P})$ 是一个给定的概率空间，对任意事件 $A_1, \cdots, A_n \in \mathscr{S}$，我们有

$$\mathsf{P}\left(\bigcup_{k=1}^{n} A_k\right) \ge \sum_{k=1}^{n} \mathsf{P}(A_k) - \sum_{k_1<k_2}^{n} \mathsf{P}(A_{k_1}A_{k_2}) \tag{1.51}$$

1.26. 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 $(\Omega, \mathscr{S}, \mathsf{P})$。用 A_k 表示“第 k 次掷出首个 6 点”，$k = 1, 2, \cdots$；用 A 表示“将掷出 6 点”。试求：$\mathsf{P}(A_k)$ 和 $\mathsf{P}(A)$。

☆ 1.27. Banach 问题：某个喜好抽烟的数学家左右口袋各带着一盒火柴，每盒皆有 n 根火柴。他以概率 p 选左口袋的火柴，以概率 $q = 1-p$ 选右口袋的火柴。问一盒空而另一盒剩 k 根火柴 $(k = 0, 1, \cdots, n)$ 的概率 P_k？

☆ 1.28. 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 2 获得奖牌为止，分别求出甲、乙获得奖牌的概率。

☆ 1.29. 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 $\mathsf{P}(A_t)$。

1.30. 盒子里有 $n - 1$ 只黑球和 1 只白球，每次从盒中随机摸出一球，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。

1.31. 盒子里装有一球，此球可能是白球也可能是黑球，现在放一个白球到盒子中，然后再从中随机地取出一球，若取出的球是白球，问盒子里剩下的球也是白球的概率。

1.32. 袋中装有 m 枚正品硬币、n 枚次品硬币（次品硬币的两面均印有国徽）。在袋中任取一枚硬币抛掷 r 次，已知每次都得到国徽，问这枚硬币是正品的概率？

1.33. 商标“MAXIMA”中有 2 个字母脱落，有人捡起随意放回，问放回后仍为“MAXIMA”的概率？

☆ 1.34. 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 $n = 0, 1, 2, \cdots$。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。试计算：

- ❑ 两次故障间共生产 k 件正品的概率，其中 $k = 0, 1, 2, \cdots$。
- ❑ 若已知在某两次故障间生产了 k 件正品，求生产线共生产 m 件产品的概率。

1.35. 盒子里有 6 个白球，4 个黑球。不放回地抽取两次，每次任取一球，问：(1) 第二次才取出白球的概率；(2) 发现其中之一是白球，另一个也是白球的概率。

☆ 1.36. 有 $N+1$ 个盒子 $A_0, A_1, \cdots, A_N$，假设 N 非常之大。盒子 A_k 有 k 个黑球，$N-k$ 个白球，$k = 0, 1, \cdots, N$。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。

☆ 1.37. 令 $A_0 = \emptyset$，若对于 $j = 1, 2, \cdots$ 皆有 $A_j A_{j-1}^c \cdots A_0^c$ 与 B_j 独立，则

$$\mathsf{P}\left\{\bigcup_{j=1}^{\infty} A_j B_j\right\} \geq \alpha \mathsf{P}\left\{\bigcup_{j=1}^{\infty} A_j\right\}，\text{其中 } \alpha = \inf_j \mathsf{P}(B_j) \tag{1.52}$$

更一般地，如果 $(A_j + A_j')(A_{j-1} + A_{j-1}')^c \cdots (A_0 + A_0')^c$ 与 B_j 独立，也与 B_j' 独立，则

$$\mathsf{P}\left\{\bigcup_{j=1}^{\infty} A_j B_j + A_j' B_j'\right\} \geq \alpha \mathsf{P}\left\{\bigcup_{j=1}^{\infty} A_j + A_j'\right\}，\text{其中 } \alpha = \inf_j \left\{\mathsf{P}(B_j), \mathsf{P}(B_j')\right\}$$

☆ 1.38. 设事件 $A_j, j = 1, 2, \cdots$ 满足 $\mathsf{P}(A_j) = 1$，试证明：$\mathsf{P}(\bigcap_{j=1}^{\infty} A_j) = 1$。

第二章

随机变量及其数字特征

在数学发展史中，对变量的认识曾带来观念和方法的革命*。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。

同以前大家了解的变量一样，随机变量也可以取不同的值，它的取值范围就是随机试验所有可能的结果。与传统变量不一样的是，随机变量的取值是不确定的，总是以某一概率来取值。例如，

例 2.1. 已知随机事件 A 发生的概率是 p，在 n 重 Bernoulli 试验中 A 出现的次数 X 是一个随机变量，它的取值范围是 $0, 1, \cdots, n$，其中取 k 的概率是 $\mathsf{P}(X = k) = C_n^k p^k (1-p)^{n-k}$，见例 1.9 和例 1.27。自此可以方便地探讨 X 的性质，而不必计较它的具体取值，譬如，$\mathsf{P}(X \le x)$ 关于 x 是非减的——这就是变量数学的好处。

当随机试验的结果不是实数时，如抛硬币，我们要想办法把将随机事件与实数域 $\mathbb{R}$ 建立起联系以使得对随机事件的描述可抽象为随机

*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》(La Géométrie, 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》 [54]。

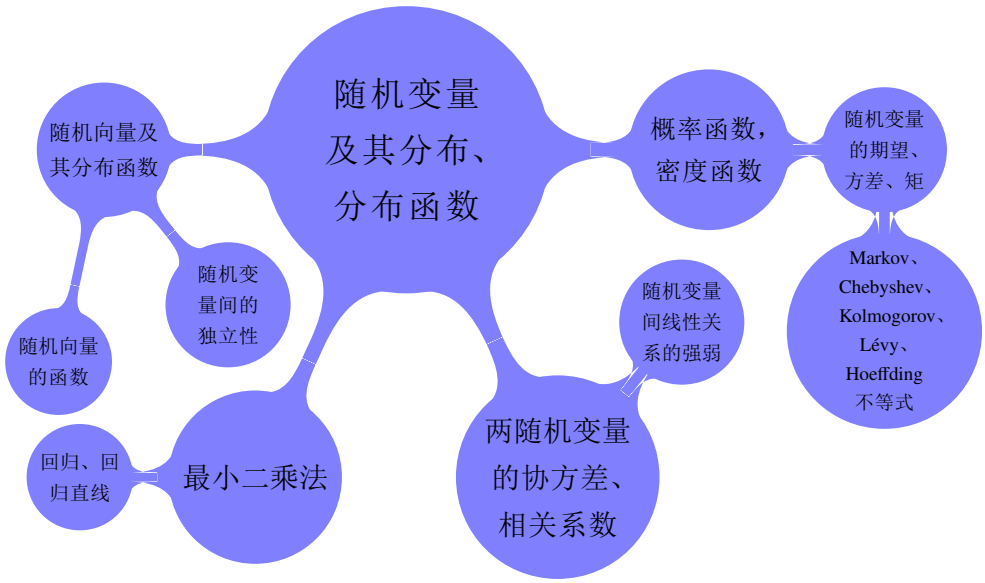
变量 X 在 $\mathbb{R}$ 上取值，这就是本章第一节给出的随机变量的严格定义，它依赖于集合论的如下结果。

性质 2.1. 已知单值映射 $g:\Omega\to\Lambda$，定义 $B\subseteq\Lambda$ 的逆像为 $g^{-1}(B)=\{\omega\in\Omega: f(\omega)\in B\}$。对于 $B, B_k\subseteq\Lambda$，其中 k 属于某个指标集 K（可以是不可数的），可以证明 $g^{-1}(\cdot)$ 具有如下性质：

$$g^{-1}(B^c)=[g^{-1}(B)]^c \tag{2.1}$$

$$g^{-1}\left(\bigcap_{k\in K}B_k\right)=\bigcap_{k\in K}g^{-1}(B_k)\ \text{且}\ g^{-1}\left(\bigcup_{k\in K}B_k\right)=\bigcup_{k\in K}g^{-1}(B_k) \tag{2.2}$$

随机变量分为离散型和连续型，不管哪种类型，每个随机变量 X 都唯一对应着一个分布函数 $F(x)=\mathrm{P}(X\le x)$，它承载着 X 的所有信息就像随机变量的“基因”。有时随机变量 X 的一些数字特征，如期望、方差、矩等，就足以描绘 X 的“形象”，这些数字特征都是分布函数的“深加工产品”，它们导出了几个应用广泛的不等式：Markov 不等式、Chebyshev 不等式、Kolmogorov 不等式、Hoeffding 不等式等。而基于最小二乘法的回归则揭示了两个随机变量之间潜在的函数关系。



2.1　随机变量及其基本性质

为了在样本空间 $(\Omega, \mathscr{S})$ 和 $(\mathbb{R}, \mathfrak{B}_1)$ 之间搭建桥梁，自然联想到令单值函数 $X : \Omega \to \mathbb{R}$ 满足条件：任一 $B \in \mathfrak{B}_1$ 的逆像 $X^{-1}(B) \in \mathscr{S}$，这样就使得 $(\mathbb{R}, \mathfrak{B}_1)$ 的一个随机事件通过 X^{-1} 对应到 $(\Omega, \mathscr{S})$ 的某一随机事件。由性质 2.1 知，对集合的并运算（或交运算）而言，逆映射 X^{-1} 是 $\mathfrak{B}_1$ 与 $\mathscr{S}$ 间的同态映射。

定义 2.1 (随机变量)**.** 已知样本空间 $(\Omega, \mathscr{S})$，随机变量 (random variable, rv)*是一个单值函数 $X : \Omega \to \mathbb{R}$，使得任一 Borel 集 $B \in \mathfrak{B}_1$（见例 1.22）的逆像 $X^{-1}(B)$ 是一个随机事件，即 $\forall B \in \mathfrak{B}_1$

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathscr{S} \tag{2.3}$$

即，随机变量 X 就是样本空间 $(\Omega, \mathscr{S})$ 到一维 Borel 空间 $(\mathbb{R}, \mathfrak{B}_1)$ 的可测函数（见定义 1.7）。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。

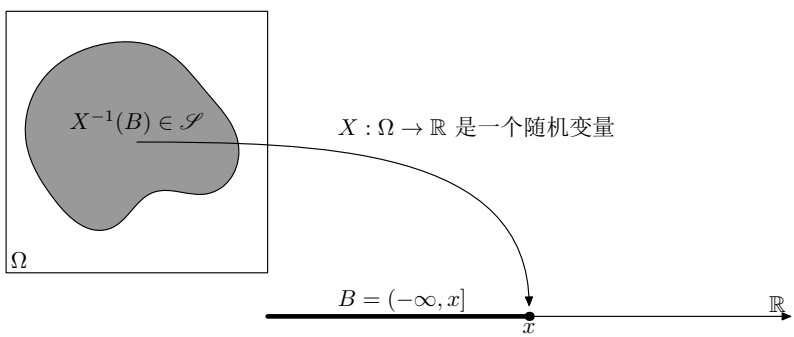


图 2.1: 单值函数 $\Omega \xrightarrow{X} \mathbb{R}$ 是一个随机变量当且仅当逆映射 X^{-1} 把任意一个左开右闭区间 $(-\infty, x]$ 映为某一随机事件 $\{X \leq x\} \in \mathscr{S}$。

*本书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)**.** 已知样本空间 $(\Omega, \mathscr{S})$，单值函数 $X : \Omega \to \mathbb{R}$ 是一个随机变量当且仅当 $\forall x \in \mathbb{R}$，皆有 $\{\omega : X(\omega) \le x\} \in \mathscr{S}$。

证明. “⇒” 是显然的，因为 $(-\infty, x] \in \mathfrak{B}_1$。下面往证 “⇐”：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 “⇐” 得证。 □

例 2.2. 掷骰子的基本事件集合是 $\Omega = \{1, 2, 3, 4, 5, 6\}$，定义 $\mathscr{S} = 2^{\Omega}$。考虑如下定义的单值函数 $X : \Omega \to \mathbb{R}$，

$$X(k) = k，\text{其中 } k = 1, 2, \cdots, 6$$

容易验证

$$\{\omega : X(\omega) \le x\} = \{X \in (-\infty, x]\} = \begin{cases} \emptyset \in \mathscr{S} & \text{当 } x < 1 \\ \{1\} \in \mathscr{S} & \text{当 } 1 \le x < 2 \\ \quad \vdots & \\ \Omega \in \mathscr{S} & \text{当 } x \ge 6 \end{cases}$$

所以，X 是样本空间 $(\Omega, \mathscr{S})$ 上定义的随机变量。

例 2.3 (指示函数)**.** 集合 A 的指示函数 (indicator function) I_A 定义为

$$I_A(x) = \begin{cases} 1 & \text{当 } x \in A \\ 0 & \text{当 } x \notin A \end{cases} \tag{2.4}$$

I_A 是样本空间 $(\Omega, \mathscr{S})$ 上定义的随机变量当且仅当 $A \in \mathscr{S}$，这是因为

$$\{\omega : I_A(\omega) \le x\} = \begin{cases} \emptyset & \text{当 } x < 0 \\ A^c & \text{当 } 0 \le x < 1 \\ \Omega & \text{当 } x \ge 1 \end{cases}$$

令 $A_1, A_2, \cdots, A_n$ 是 Ω 的一个划分，我们称 $X = \sum_{j=1}^{n} x_j I_{A_j}$ 为一个简单随机变量，其中 $x_j \in \mathbb{R}$ 有限。

练习 2.1. 请读者验证指示函数的以下性质：

- $A \subseteq B$ 当且仅当 $I_A \leq I_B$。特别地，$A = B$ 当且仅当 $I_A = I_B$。
- $I_{A^c} = 1 - I_A$，$I_{AB} = I_A I_B$ 且 $I_{A\cup B} = I_A + I_B - I_{AB}$。

定理 2.2. 已知 X 是定义在样本空间 $(\Omega, \mathscr{S})$ 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 $(\Omega, \mathscr{S})$ 上的随机变量。

证明. 往证 $\{\omega : aX(\omega) + b \leq x\} = \{aX + b \leq x\} \in \mathscr{S}$，事实上

$$\{aX + b \leq x\} = \begin{cases} \{X \leq (x-b)/a\} \in \mathscr{S} & \text{当 } a > 0 \\ \{X \geq (x-b)/a\} \in \mathscr{S} & \text{当 } a < 0 \\ \Omega \in \mathscr{S} & \text{当 } a = 0 \text{ 且 } x \geq b \\ \emptyset \in \mathscr{S} & \text{当 } a = 0 \text{ 且 } x < b \end{cases}$$

所以，$aX + b$ 也是一个随机变量。 □

本节内容 第一小节讨论随机变量的分布函数及其性质。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。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 $g(X)$。

学习目标 (1) 掌握分布函数的充要条件；(2) 学会利用分布函数或分布列、密度函数描述随机变量及其函数；(3) 初步了解均匀分布、正态分布。

2.1.1 随机变量的分布与分布函数

已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 上定义的随机变量 $X : \Omega \to \mathbb{R}$，如何构造样本空间 $(\mathbb{R}, \mathfrak{B}_1)$ 上相应的概率测度？我们需要“提炼”出下面的概念。

定义 2.2 (分布)**.** 对于任意 $A \in \mathfrak{B}_1$，集函数 $F_X(A) = \mathsf{P}(\{X \in A\})$ 称为随机变量 X 在概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 上的*概率分布* (probability distribution)，简称*分布*。谈论一个随机变量，必先交待清楚它的分布。

定理 2.3. 集函数 $F_X(\cdot)$ 是 $(\mathbb{R}, \mathfrak{B}_1)$ 上的概率测度。

证明. 若 $A_j \in \mathfrak{B}_1, j = 1, 2, \cdots$ 两两不交，则

$$F_X\left(\bigcup_{j=1}^{\infty} A_j\right) = \mathsf{P}\left(\left\{X \in \bigcup_{j=1}^{\infty} A_j\right\}\right) = \mathsf{P}\left(\bigcup_{j=1}^{\infty} \{X \in A_j\}\right)$$
$$= \sum_{j=1}^{\infty} \mathsf{P}\{X \in A_j\} = \sum_{j=1}^{\infty} F_X(A_j)$$

显然 $\forall B \in \mathfrak{B}_1, 0 \le F_X(B) \le 1$ 且 $F_X(\mathbb{R}) = 1$。 □

特别地，随机变量 X 在左开右闭区间 $(-\infty, x]$ 上的分布显得尤为重要，这是因为：(i) 所有左开右闭区间生成了 Borel σ 域 $\mathfrak{B}_1$，见例 1.22。(ii) 直接讨论分布这一集函数不很方便。

定义 2.3 (分布函数)**.** 如下定义的实值函数 $F : \mathbb{R} \to [0, 1]$ 被称为随机变量 X 的*分布函数* (distribution function)*，它与分布之间是一一对应的，承载了随机变量的所有的信息。

$$F(x) = \mathsf{P}\{\omega : X(\omega) \le x\} = \mathsf{P}\{X^{-1}(-\infty, x]\} \tag{2.5}$$

在不引起歧义的情况下，也写作 $F(x) = \mathsf{P}(X \le x)$ 或 $\mathsf{P}\{X \in (-\infty, x]\}$。

*一些概率论的经典教材中对分布函数的定义是 $\mathsf{P}(X < x)$，这导致分布函数是左连续的 [40,61]，而按照现在流行的定义分布函数 $\mathsf{P}(X \le x)$ 是右连续的 [36,78]。差别由约定俗成引起，绝大多数结果都不受其影响。

定理 2.4 (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 的分布函数当且仅当 $F(x)$ 是 (1) 非减的，(2) 右连续的*，(3) 满足 $F(-\infty)=0$ 和 $F(+\infty)=1$。

证明. 往证“$\Rightarrow$”：(1) 如果 $x_1>x_2$，则 $\{\omega: X(\omega)\le x_1\}\supseteq\{\omega: X(\omega)\le x_2\}$，所以 $F(x)$ 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 $F(x_n)\to F(x)$。令 $A_n=\{X\in(x,x_n]\}$，则 $A_n\downarrow\emptyset$。由定理 1.7 有 $\lim\limits_{n\to\infty}\mathsf{P}(A_n)=0$，即 $\lim\limits_{n\to\infty}F(x_n)-F(x)=0$。(3) 令序列 $\{x_n\}$ 单调升趋于 $+\infty$，则 $\{X\le x_n\}\uparrow\Omega$。由推论 1.4，有 $F(+\infty)=\lim\limits_{x_n\to\infty}\mathsf{P}(\{X\le x_n\})=1$。类似可证 $F(-\infty)=0$，留给读者。

往证“$\Leftarrow$”：在样本空间 $(\mathbb{R},\mathfrak{B}_1)$ 上，定义 $\mathsf{P}\{(-\infty,x]\}=F(x)$，易证 P 是 $(\mathbb{R},\mathfrak{B}_1)$ 上的概率测度。$F(x)$ 是定义在概率空间 $(\mathbb{R},\mathfrak{B}_1,\mathsf{P})$ 上的随机变量 $\mathbb{1}:\mathbb{R}\to\mathbb{R}$（即 $\mathbb{R}$ 到自身的恒等映射）的分布函数。 □

例 2.4. 已知随机变量 X 的分布函数为 $F(x)$，设 $a<b$，则 $X^{-1}(-\infty,a]\subseteq X^{-1}(-\infty,b]$。进而，

$$\begin{aligned}\mathsf{P}\{X\in(a,b]\}&=\mathsf{P}\{X^{-1}(a,b]\}=\mathsf{P}\{X^{-1}(-\infty,b]\cap(X^{-1}(-\infty,a])^c\}\\&=\mathsf{P}\{X^{-1}(-\infty,b]\}-\mathsf{P}\{X^{-1}(-\infty,a]\}=F(b)-F(a)\end{aligned}$$

同理，$\mathsf{P}\{X\in(a,b)\}=F(b-)-F(a)$，其中 $F(b-)$ 表示 $F(x)$ 在 $x=b$ 点的左极限。还有，$\mathsf{P}\{X\in[a,b)\}=F(b-)-F(a-)$, $\mathsf{P}\{X\in[a,b]\}=F(b)-F(a-)$，留作练习。

例 2.5 (单点分布)**.** 概率空间 $(\Omega,\mathscr{S},\mathsf{P})$ 上有如下定义的随机变量 X，

$$X(\omega)=c,\qquad\forall\omega\in\Omega$$

即 c 的逆像是 Ω，于是 $\{X=c\}$ 的概率为 1，记作 $\mathsf{P}(X=c)=1$。

*即 $F(x)=F(x+)$，其中 $F(x+)$ 表示 F 在 x 点的右极限。

X 的分布函数为

$$F(x)=\begin{cases}0 & \text{当 } x<c\\ 1 & \text{当 } x\geq c\end{cases} \tag{2.6}$$

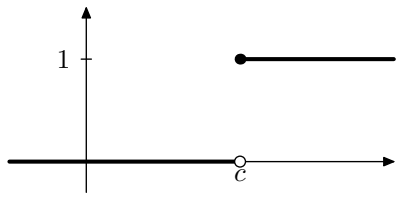


图 2.2: 单点分布：分布函数是有一个跳跃点的阶梯函数。

并称 X 服从**单点分布** (one-point distribution)，记作 $X\sim\langle c\rangle$。分布函数曲线见右图，空心点表示“抠掉”此点，实心点强调包含此点。

例 2.6 (两点分布)**.** 定义在概率空间 $(\Omega,\mathscr{S},\mathsf{P})$ 上的随机变量 X 如果满足 $X(\Omega)=\{a,b\}$（不妨设 $b<a$），且 $\mathsf{P}(X=a)=p,\mathsf{P}(X=b)=1-p$，其中 $0<p<1$，则称 X 服从**两点分布** (two-point distribution)，记作 $X\sim p\langle a\rangle+(1-p)\langle b\rangle$。两点分布 $X\sim p\langle 1\rangle+(1-p)\langle 0\rangle$ 特称为 0-1 分布。

两点分布 $X\sim p\langle a\rangle+(1-p)\langle b\rangle$ 的分布函数为阶梯函数

$$F(x)=\begin{cases}0 & \text{当 } x<b\\ 1-p & \text{当 } b\leq x<a\\ 1 & \text{当 } x\geq a\end{cases} \tag{2.7}$$

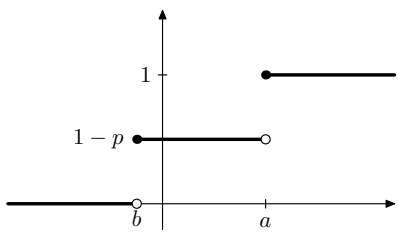


图 2.3: 两点分布的分布函数是有两个跳跃点的阶梯函数。

单点分布和两点分布的随机变量都是简单随机变量（见例 2.3）。

定理 2.5. 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。 □

以奥地利数学家 Eduard Helly (1884-1943) 命名的下述有关存在性的结果在研究分布函数的极限时非常有用。

定理 2.6 (Helly 选择定理)**.** 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 $F(x)=\lim\limits_{k\to\infty}F_{n_k}(x)$。

证明. 详见 Gnedenko 的《概率论教程》 [40] 第七章。此定理也称为第一 Helly 定理。 □

在 Helly 选择定理中，$F(x)$ 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 $G(x)$ 是一个分布函数且非负常数 a, b, c 满足 $a + b + c = 1$。显然 $F_n(x) \to F(x) = b + cG(x)$ 使得 $F(x)$ 不是一个分布函数，原因是 $F(-\infty) = b, F(+\infty) = 1 - a$。

2.1.2 离散型与连续型随机变量

定义 2.4 (分布列)**.** 已知概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 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 $X(\Omega) = \{x_1, x_2, \cdots\}$ 且 $\mathsf{P}(X = x_j) = p_j$，显然 $\sum_{j=1}^{\infty} p_j = 1$。我们称 $\mathsf{P}(X = x_j) = p_j$ 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。

表 2.1: 为了描述离散型随机变量，我们常用下面的分布列。有时候，也用 $X \sim p_1\langle x_1\rangle + p_2\langle x_2\rangle + \cdots + p_j\langle x_j\rangle + \cdots$ 来表示。

X	x_1	x_2	$\cdots$	x_j	$\cdots$
$\mathsf{P}(X = x_j)$	p_1	p_2	$\cdots$	p_j	$\cdots$

离散型随机变量 X 也可用指示函数来表示，

$$X(\omega) = \sum_{j=1}^{\infty} x_j I_{\{X=x_j\}}(\omega) \tag{2.8}$$

X 的分布函数为简单函数（见附录 F）

$$F(x) = \sum_{x_j \le x} [F(x_j) - F(x_j - 0)] = \sum_{x_j \le x} \mathsf{P}(X = x_j) = \sum_{j=1}^{\infty} p_j J(x - x_j) \tag{2.9}$$

其中，$J(\cdot)$ 称为非负判定函数，定义为

$$J(x) = \begin{cases} 0 & \text{当 } x < 0 \\ 1 & \text{当 } x \ge 0 \end{cases} \tag{2.10}$$

定义 2.5 (密度函数)**.** 概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 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

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{2.11}$$

其中，$f(x)$ 称为 X 的*概率密度函数* (probability density function, pdf) 或简称为*密度函数*，有时为区别其他随机变量的密度函数，也记作 $f_X(x)$。

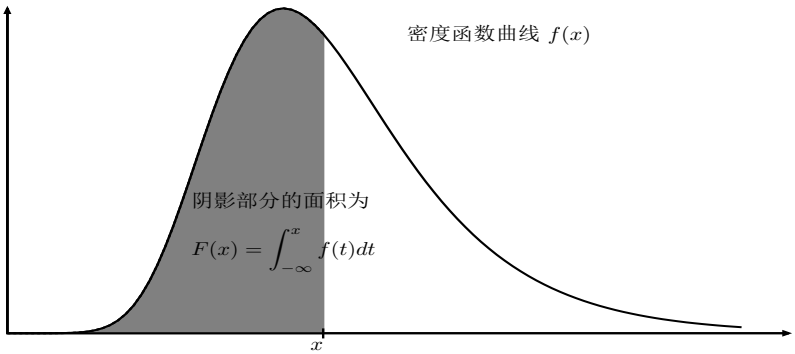


图 2.4: 密度函数曲线 $f(x)$ 与 x 轴围成的面积为 1。随机变量 X 落于区间 $(-\infty, x]$ 的概率为 $F(x)=\int_{-\infty}^{x} f(t)dt$，即阴影部分的面积。

例 2.7 (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*，记作 $X \sim \mathrm{U}[a,b]$。

$$F(x)=\begin{cases} 0 & \text{当 } x<a \\ (x-a)/(b-a) & \text{当 } a\le x<b \\ 1 & \text{当 } x\ge b \end{cases} \tag{2.12}$$

容易验证下面的函数是 X 的密度函数，

$$f(x)=\begin{cases} 1/(b-a) & \text{如果 } x\in[a,b] \\ 0 & \text{如果 } x\notin[a,b] \end{cases} \tag{2.13}$$

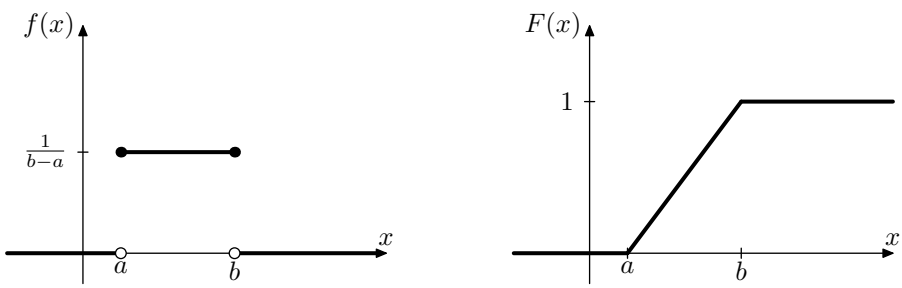


图 2.5: 随机变量 $X \sim \mathrm{U}[a,b]$ 的密度函数 $f(x)$ 和分布函数 $F(x)$。

例 2.8 (正态分布)**.** 若连续型随机变量 X 的密度函数为例 1.30 所描述的式 (1.25)，即

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

则称 X 服从参数为 (μ,σ^2) 的正态分布 (normal distribution)，也称高斯分布 (Gaussian distribution)*或 Gauss-Laplace 分布，记作 $X\sim \mathrm{N}(\mu,\sigma^2)$，其分布函数 $\Phi(x|\mu,\sigma^2)=\int_{-\infty}^{x}\phi(z|\mu,\sigma^2)dz$ 具有以下性质（请读者证明）。

$$\int_{-\infty}^{\infty}\left[\Phi(x|\mu_1,\sigma_1^2)-\Phi(x-t|\mu_2,\sigma_2^2)\right]dx=t+\mu_2-\mu_1 \tag{2.14}$$

正态分布在概率论中充当着十分重要的角色，特别地，分布 $X\sim\mathrm{N}(0,1)$ 称为**标准正态分布**，它所对应的密度函数为 $\phi(x)$，分布函数记作 $\Phi(x)$。函数 $\Phi(x)$ 没有显式表达，人们制作了它的数值表，通过查表来完成计算，现在这些琐事都可以交给计算机来做了（R 语言的 `pnorm` 函数）。直观上，$\Phi(x)$ 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 $\Phi(-x)=1-\Phi(x)$。

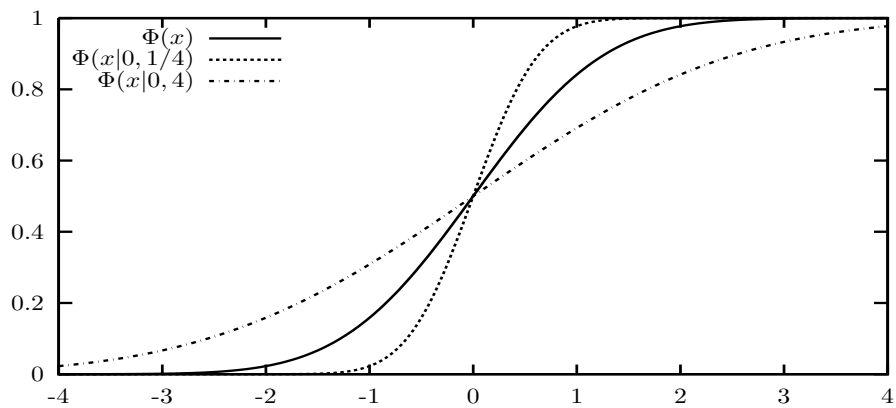


图 2.6: 正态分布 $X\sim\mathrm{N}(0,\sigma^2)$ 的分布函数曲线 $\Phi(x|0,\sigma^2)$，其中实线是 $\Phi(x)$。不难发现尺度参数 σ^2 越小，曲线越“陡”。

*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 C 和 §4.2.2。

用 $\Phi(x)$ 定义的*误差函数* (error function) $\mathrm{erf}(x)$ 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。

$$\mathrm{erf}(x) = 2\Phi(\sqrt{2}x) - 1 \tag{2.15}$$

性质 2.2. 正态分布 $X \sim \mathrm{N}(\mu, \sigma^2)$ 与标准正态分布之间有下面的关系。

$$\frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1) \tag{2.16}$$

变换 $(X-\mu)/\sigma$ 被称为 X 的*标准化*，它把对 $F_X(x) = \mathsf{P}(X \le x)$ 的计算“转嫁”到 $\Phi(\cdot)$ 上，不难验证下面的公式。

$$\mathsf{P}(X \le x) = \mathsf{P}\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{2.17}$$

$$\text{特别地，}\mathsf{P}(|X-\mu| \le r\sigma) = \mathsf{P}\left(\left|\frac{X-\mu}{\sigma}\right| \le r\right) = 2\Phi(r) - 1 \tag{2.18}$$

例如，$\mathsf{P}(|X-\mu| \le 3\sigma) = 2\Phi(3) - 1 > 99.7\%$，这一事实被称为“$3\sigma$ 原则”，它说明随机变量 $X \sim \mathrm{N}(\mu, \sigma^2)$ “几乎”都落于 $(\mu - 3\sigma, \mu + 3\sigma)$ 内。

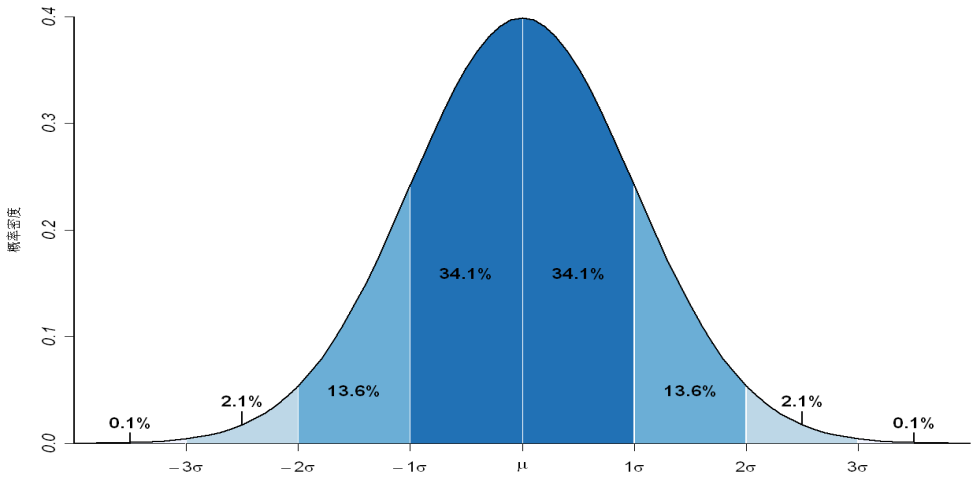


图 2.7: 计算得 $\mathsf{P}(|X-\mu| > 3\sigma) \approx 0.0027$。直观含义是：随机变量 X 距离 μ 超过 3σ 的概率小于 0.3%，换言之，X 落于 $(\mu - 3\sigma, \mu + 3\sigma)$ 的概率大于 99.7%。

例 2.9. 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 $T \sim \mathrm{N}(\mu, \sigma^2)$，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

$$X = \begin{cases} t_{\min} & \text{当 } T \le t_{\min} \\ T & \text{当 } t_{\min} < T < t_{\max} \\ t_{\max} & \text{当 } T \ge t_{\max} \end{cases}$$

这个随机变量既不是离散型的，也不是连续型的。它的分布函数为

$$F_X(x) = \begin{cases} 0 & \text{当 } x < t_{\min} \\ \Phi[(x-\mu)/\sigma] & \text{当 } t_{\min} \le T < t_{\max} \\ 1 & \text{当 } x \ge t_{\max} \end{cases}$$

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，读者很容易鉴别之。

性质 2.3. 连续型随机变量 X 的密度函数 $f(x)$ 和分布函数 $F(x)$ 具有以下性质，请读者给出证明。

$$\int_{-\infty}^{+\infty} f(x)dx = F(+\infty) = 1 \tag{2.19}$$

$$\mathsf{P}(a < X \le b) = F(b) - F(a) = \int_a^b f(x)dx \tag{2.20}$$

$$F'(x) = f(x)，\text{其中 } x \text{ 是 } f(x) \text{ 的连续点} \tag{2.21}$$

定理 2.7. 已知 $f(x)$ 是一个在 $\mathbb{R}$ 上非负的、可积的实函数，且满足 $\int_{-\infty}^{+\infty} f(x)dx = 1$，则 $f(x)$ 一定是某个随机变量的密度函数。

证明. 令 $F(x) = \int_{-\infty}^{x} f(t)dt$，显然 $F(x)$ 非减且 $F(-\infty) = 0, F(+\infty) = 1$。请读者验证 $F(x)$ 右连续。 □

2.1.3　随机变量的函数

已知在概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 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 $(\Omega, \mathscr{S}, \mathsf{P})$ 上定义的随机变量？

定理 2.8. 若 X 是概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 上定义的随机变量，Borel 可测函数（见定义 1.7） g 使得 $Y = g(X)$ 依然是 $(\Omega, \mathscr{S}, \mathsf{P})$ 上定义的随机变量。

证明. 因为 g 是 Borel 可测函数，所以 $g^{-1}(-\infty, y] \in \mathfrak{B}_1$，进而 $\{Y \leq y\} = \{g(X) \leq y\} = \{X \in g^{-1}(-\infty, y]\} \in \mathscr{S}$，由定义 2.1 得证。 □

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 F 列举了可测函数的性质，对 Borel 可测函数也是适用的。

定理 2.9. 设离散型随机变量 X 具有如表 2.4 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 $x_{j_1}, x_{j_2}, \cdots$ 是所有 $g(x_j)$ 的逆像，则 $\mathsf{P}\left\{Y = g(x_j)\right\} = p_{j_1} + p_{j_2} + \cdots$。

证明. $\mathsf{P}\{Y = g(x_j)\} = \mathsf{P}(X \in \{x_{j_1}, x_{j_2}, \cdots\}) = p_{j_1} + p_{j_2} + \cdots$。 □

例 2.10. 已知随机变量 X 的分布列如下：

X	-2	-1	0	1
概率	$1/4$	$3/16$	$1/2$	$1/16$

按照上述定理，随机变量 $Y = X^2$ 的分布列是：

Y	0	1	4
概率	$1/2$	$1/4$	$1/4$

求连续型随机变量 X 的函数 $g(X)$ 的分布函数可从分布函数的原始定义出发，也可以用下面的方法。

定理 2.10. 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 $f_Y(y) = |h'(y)|f_X[h(y)]$。

证明. 不妨设 g 单调不减，对任意的实数 y，则 Y 的分布函数为

$$\begin{aligned} F_Y(y) &= \mathsf{P}\{Y \le y\} = \mathsf{P}\{g(X) \le y\} = \mathsf{P}\{X \le h(y)\} \\ &= \int_{-\infty}^{h(y)} f_X(x)dx = \int_{-\infty}^{y} f_X[h(y)]h'(y)dy \end{aligned}$$

请读者仿此考虑 g 单调不增的情形，由密度函数的定义得证。 □

例 2.11. 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 $a \neq 0$，则 Y 的密度函数为

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

- 非线性映射 $Y = X^2$ 的逆映射为 $X = \pm\sqrt{Y}$，则 Y 的分布函数为

$$F_Y(y) = \begin{cases} 0 & \text{如果 } y \le 0 \\ \mathsf{P}(Y \le y) = \mathsf{P}(-\sqrt{y} \le X \le \sqrt{y}) & \text{如果 } y > 0 \\ = F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \end{cases}$$

于是，$f_Y(y) = F_Y'(y) = [f_X(\sqrt{y}) + f_X(-\sqrt{y})]/(2\sqrt{y})$。

例 2.12. 令 $X \sim \mathrm{U}(0,1)$，则 $Y = \exp(X)$ 的密度函数为

$$f_Y(y) = \begin{cases} 1/y & \text{当 } 1 < y < e \\ 0 & \text{其他} \end{cases}$$

2.2　随机向量及其基本性质

对随机现象的描述有时需要向量，譬如炮弹落点、身体状况（心率、血压、血糖等）、学习成绩、股票行情等。随机向量也是定义随机变量之间独立关系的工具。

定义 2.6 (随机向量)**.** 如果 $X_1, X_2, \cdots, X_n$ 是定义在同一概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 上的 n 个随机变量，则称 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathsf{T}}$ 是一个 n 维随机向量，其中符号 T 表示转置，称下面的 n 元函数为随机向量 $\boldsymbol{X}$ 的分布函数。

$$F_{\boldsymbol{X}}(x_1, \cdots, x_n) = \mathsf{P}\left(\bigcap_{j=1}^{n} \{X_j \le x_j\}\right) \tag{2.22}$$

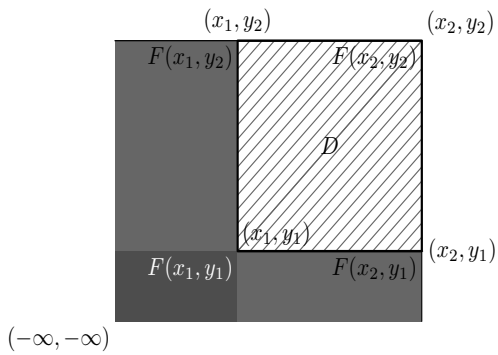
有时也称之为随机变量 $X_1, \cdots, X_n$ 的联合分布函数。

性质 2.4. 以二维随机向量 $(X, Y)^{\mathsf{T}}$ 为例，它的分布函数 $F(x, y) = \mathsf{P}(X \le x, Y \le y)$ 具有以下的性质*。

❶ $F(-\infty, y) = F(x, -\infty) = 0$ 并且 $F(+\infty, +\infty) = 1$。

❷ 对 x 来说，$F(x, y)$ 是非减、右连续的。对 y 来说，亦是如此。

❸ 二维随机向量 $(X, Y)^{\mathsf{T}}$ 落于区域 $D = (x_1, x_2] \times (y_1, y_2]$（见右图斜线阴影部分）里的概率是



$$\begin{aligned} &\mathsf{P}(x_1 < X \le x_2, y_1 < Y \le y_2) \\ &\quad = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \end{aligned} \tag{2.23}$$

定理 2.11. 若二元函数 $F(x, y)$ 满足性质 2.4 中的 ❶❷，则 $F(x, y)$ 是某个

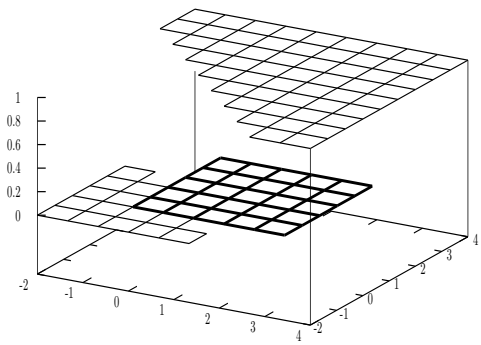
*对于 n 维随机向量，其分布函数的性质与二维的情形类似，可推广得到。

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

$$F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \geq 0 \tag{2.24}$$

条件式 (2.24) 是必要的，下面构造一个二元函数 $F(x, y)$ 满足性质 2.4 中的条件 ❶❷，但不满足条件式 (2.24)，$F(x, y)$ 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 (2.23) 则有 $\mathsf{P}(-1 < X \leq 3, -1 < Y \leq 3) = F(3, 3) - F(3, -1) - F(-1, 3) + F(-1, -1) = -1$，矛盾！

定义 2.7. 已知随机向量 $(X, Y)^{\mathsf{T}}$ 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 $(X, Y)^{\mathsf{T}}$ 所有可能的取值是可数的，概率函数为

$$\mathsf{P}(X = x_i, Y = y_j) = p_{ij} \tag{2.25}$$

$$\text{满足 } \sum_{i,j=1}^{\infty} p_{ij} = 1 \text{ 且 } F(x, y) = \sum_{x_i \leq x, y_j \leq y} p_{ij} \tag{2.26}$$

连续型： 如果存在非负函数 $f(x, y)$ 使得

$$\forall (x, y) \in D, F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t) dt ds \tag{2.27}$$

此处 $f(x, y)$ 称为 $(X, Y)^{\mathsf{T}}$ 的（联合）密度函数，满足如下性质。

$$\iint_{\mathbb{R}^2} f(x, y) dy dx = F(+\infty, +\infty) = 1 \text{ 且 } \frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y) \tag{2.28}$$

例 2.13 (二元正态分布)**.** 如果随机向量 $(X,Y)^\mathsf{T}$ 的密度函数是

$$\phi(x,y|\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{g(x,y)}{2(1-\rho^2)}\right\} \tag{2.29}$$

$$\text{其中，}\ g(x,y)=\frac{(x-\mu_X)^2}{\sigma_X^2}-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}$$

则称 $(X,Y)^\mathsf{T}$ 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 $(X,Y)^\mathsf{T}\sim \mathrm{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$，其中参数 $-1<\rho<1$ 是随机变量 X 与 Y 的相关系数，§2.3.3 的例 2.35 将给出详细论证。

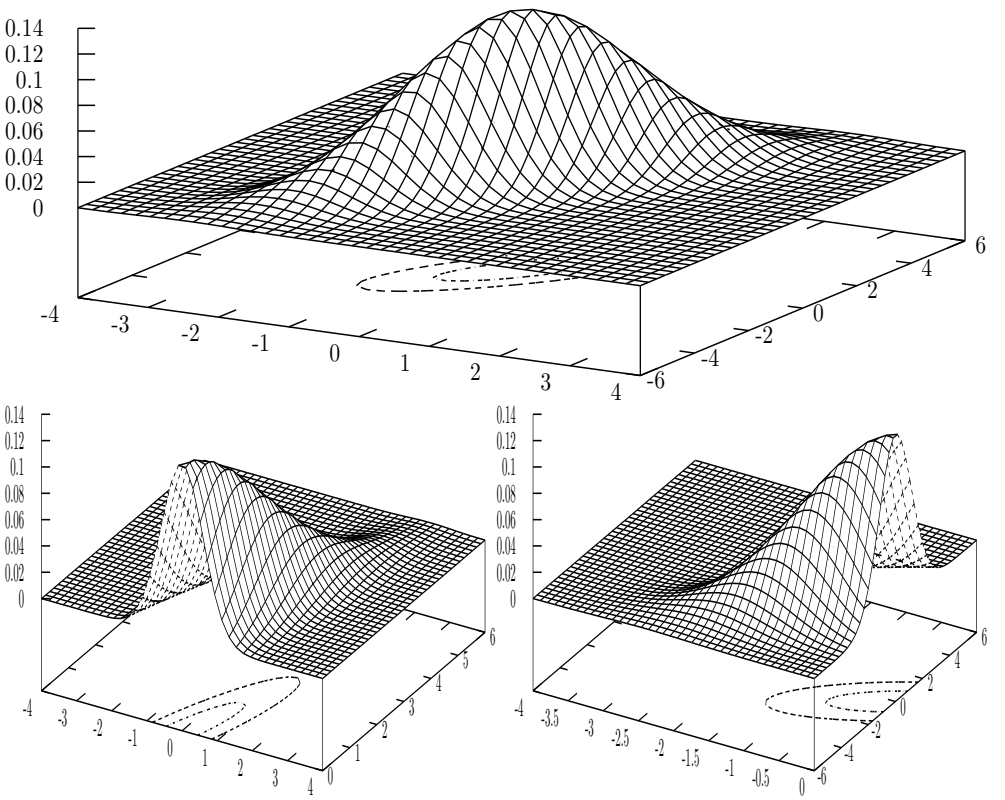


图 2.8: 二元正态分布的密度函数曲面呈钟形，例如 $z=\phi(x,y|0,0,1,4,0.8)$。不难发现，如果其他参数不变，$|\rho|$ 越接近 1，在 xoy 平面上曲面等高线就越“扁”。第二行左图：用 xoz 平面截取，剖面曲线与 $\phi(x)$ 相差一个常因子。第二行右图：用 yoz 平面截取，剖面曲线与 $\phi(x|0,4)$ 相差一个常因子。

n 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 可以理解为映射 $\boldsymbol{X}:\Omega\to\mathbb{R}^n$，即 $\omega\overset{X}{\mapsto}(X_1(\omega),X_2(\omega),\cdots,X_n(\omega))^{\mathsf{T}}$，使得 $\forall B\in\mathfrak{B}_n$ 皆有 $\boldsymbol{X}^{-1}(B)\in\mathscr{S}$，实质上就是样本空间 $(\Omega,\mathscr{S})$ 到 n 维 Borel 空间 $(\mathbb{R}^n,\mathfrak{B}_n)$ 的可测函数。所以 n 维随机向量可看作取值为 n 维向量的随机变量，有时也称作 n 维随机变量。对于事件 $\boldsymbol{X}^{-1}(-\infty,\boldsymbol{x}]=\bigcap_{j=1}^n\{X_j\le x_j\}\in\mathscr{S}$，其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathsf{T}}$ 是任意 n 维实向量 且 $(-\infty,\boldsymbol{x}]$ 表示笛卡尔积 (Cartesian product) $(-\infty,x_1]\times(-\infty,x_2]\times\cdots\times(-\infty,x_n]\subseteq\mathbb{R}^n$，我们约定用 $\{\boldsymbol{X}\in(-\infty,\boldsymbol{x}]\}$ 或 $\boldsymbol{X}\in(-\infty,\boldsymbol{x}]$ 来表示。于是 n 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 的分布函数 (2.22) 有时也记作 $F_{\boldsymbol{X}}(\boldsymbol{x})=\mathsf{P}\{\boldsymbol{X}\in(-\infty,\boldsymbol{x}]\}$。

定理 2.12. 类似定理 2.8，已知 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 是一个 n 维随机向量，若 $g:\mathbb{R}^n\to\mathbb{R}^m$ 是 Borel 可测函数（即 $\forall B\in\mathfrak{B}_m$，有 $g^{-1}(B)\in\mathfrak{B}_n$），则 $g(\boldsymbol{X})$ 是一个 m 维的随机向量。

证明. 对于 Borel 集 $B\in\mathfrak{B}_m$，$\{g(\boldsymbol{X})\in B\}=\{\boldsymbol{X}\in g^{-1}(B)\}$。因为 $g^{-1}(B)\in\mathfrak{B}_n$，于是 $\{\boldsymbol{X}\in g^{-1}(B)\}\in\mathscr{S}$。 □

本节内容 由随机向量的联合分布可以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。更重要的是，边缘分布可用来刻画随机变量之间的独立性（第二小节），条件分布是贝叶斯分析必不可少的工具（第十一章）。由已知的随机向量通过变换可以构造出新的随机向量或随机变量，第三小节重点讨论如何计算它们的分布。另外，重温了两个函数的卷积，它将用于第三章说明为何要定义特征函数。

学习目标 (1) 理解随机向量的联合分布、边缘分布和条件分布；(2) 掌握随机变量之间的独立性及其性质；(3) 计算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　边缘分布与条件分布

定义 2.8 (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，称之为**边缘分布** (marginal distribution)*。

- 如果 $(X,Y)^{\mathsf{T}}$ 是离散型的随机向量，满足 $\mathsf{P}(X=x_i,Y=y_j)=p_{ij}$，定义 X 的边缘分布的概率函数为 $\mathsf{P}(X=x_i)=\sum_{j=1}^{\infty}p_{ij}=p_{i\cdot}$，即“矩阵”$(p_{ij})$ 逐行求和“抹掉”了 Y 的信息而得到的分布列。

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 $(X,Y)^{\mathsf{T}}\sim p_{11}\langle x_1,y_1\rangle+p_{12}\langle x_1,y_2\rangle+\cdots+p_{ij}\langle x_i,y_j\rangle+\cdots$ 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

类似地，Y 的边缘分布定义为 $\mathsf{P}(Y=y_j)=\sum_{i=1}^{\infty}p_{ij}=p_{\cdot j}$，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，

$$\sum_{i=1}^{\infty}p_{i\cdot}=\sum_{j=1}^{\infty}p_{\cdot j}=\sum_{i,j=1}^{\infty}p_{ij}=1 \tag{2.30}$$

- 如果 $(X,Y)^{\mathsf{T}}$ 是连续型的随机向量，密度函数为 $f(x,y)$，定义 X 的边缘分布的密度函数（简称边缘密度）$f_X(x)$ 如下：

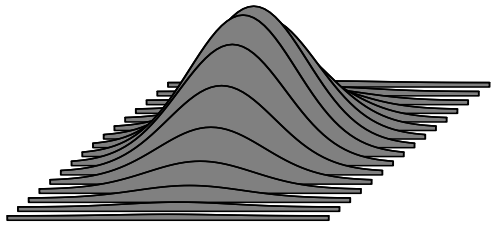
$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy \tag{2.31}$$

类似地，Y 的边缘分布的密度函数定义为 $f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx$。

---

* “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。

X 的边缘密度 $f_X(x)$ 的几何意义：暂时固定 x，曲线 $z = f(x,y)$ 与 $z = 0$ 所围成切片之面积（见右图），随机变量 Y 的信息被积分“抹掉”了。



例 2.14. 假设 $\{1,2,\cdots,21\}$ 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: 边缘分布的信息全部来自联合分布，但从边缘分布不能重构联合分布。请读者构造一个不同的联合分布，但边缘分布与此例相同。

	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

例 2.15. 已知 $(X,Y)^{\mathsf{T}} \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$，利用配方法（类似例 1.31），可得 X 的边缘分布 $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ 和 Y 的边缘分布 $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$。

定义 2.9 (条件分布)**.** 已知随机向量 $(X,Y)^{\mathsf{T}}$ 的概率函数 $\mathsf{P}(X = x_j, Y = y_j) = p_{ij}$，在给定 $Y = y_j$ 的条件下 $X = x_i$ 的概率为

$$\mathsf{P}(X = x_i | Y = y_j) = \frac{p_{ij}}{p_{\cdot j}} \tag{2.32}$$

称之为“在给定 Y 的条件下 X 的分布”，或简称 $(X|Y)$ 的条件分布。类似地，可定义 $(Y|X)$ 的条件分布

$$\mathsf{P}(Y = y_j | X = x_i) = \frac{p_{ij}}{p_{i\cdot}} \tag{2.33}$$

由条件分布的定义，显然有

$$\sum_{i=1}^{\infty} \mathsf{P}(X = x_i|Y = y_j) = 1 \text{ 且 } \sum_{j=1}^{\infty} \mathsf{P}(Y = y_j|X = x_i) = 1 \tag{2.34}$$

$$\sum_{j=1}^{\infty} \mathsf{P}(X = x_i|Y = y_j)p_{\cdot j} = p_{i\cdot} \text{ 且 } \sum_{i=1}^{\infty} \mathsf{P}(Y = y_j|X = x_i)p_{i\cdot} = p_{\cdot j} \tag{2.35}$$

上式即是古典概率的全概率公式 $\mathsf{P}(B) = \sum_{j=1}^{\infty} \mathsf{P}(A_j)\mathsf{P}(B|A_j)$ 的“离散型随机变量版”：$\sum_{j=1}^{\infty} \mathsf{P}(X = x_i|Y = y_j)\mathsf{P}(Y = y_j) = \mathsf{P}(X = x_i)$ 和 $\sum_{i=1}^{\infty} \mathsf{P}(Y = y_j|X = x_i)\mathsf{P}(X = x_i) = \mathsf{P}(Y = y_j)$。

例 2.16. 接着例 2.14，$(X|Y)$ 的条件分布为

$$\mathsf{P}(X = 1|Y = 1) = \frac{p_{11}}{p_{\cdot 1}} = \frac{3}{10}, \quad \mathsf{P}(X = 0|Y = 1) = \frac{p_{21}}{p_{\cdot 1}} = \frac{7}{10}$$
$$\mathsf{P}(X = 1|Y = 0) = \frac{p_{12}}{p_{\cdot 2}} = \frac{4}{11}, \quad \mathsf{P}(X = 0|Y = 0) = \frac{p_{22}}{p_{\cdot 2}} = \frac{7}{11}$$

$\mathsf{P}(X = 1|Y = 1) = 3/10$ 的含义是，从 $\{1, 2, \cdots, 21\}$ 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 2.10. 已知连续型的随机向量 $(X, Y)^{\mathsf{T}}$ 的密度函数为 $f(x, y)$，在给定 X 的条件下 Y 的分布函数，或 $(Y|X)$ 的条件分布函数 $F_{Y|X}(y|x)$ 定义为

$$\begin{aligned} F_{Y|X}(y|x) &= \lim_{\Delta x\to 0} \mathsf{P}(Y \le y|x < X \le x + \Delta x) = \lim_{\Delta x\to 0} \frac{\mathsf{P}(Y \le y, x < X \le x + \Delta x)}{\mathsf{P}(x < X \le x + \Delta x)} \\ &= \lim_{\Delta x\to 0} \frac{\frac{1}{\Delta x}\int_x^{x+\Delta x}\int_{-\infty}^{y} f(s,t)dtds}{\frac{1}{\Delta x}\int_x^{x+\Delta x}\int_{-\infty}^{+\infty} f(s,y)dyds} = \frac{\int_{-\infty}^{y} f(x,t)dt}{\int_{-\infty}^{+\infty} f(x,y)dy} = \frac{\int_{-\infty}^{y} f(x,t)dt}{f_X(x)} \end{aligned} \tag{2.36}$$

式 (2.36) 两边对 y 求导进而得到 $(Y|X)$ 的条件密度函数

$$f_{Y|X}(y|x) = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy} = \frac{f(x,y)}{f_X(x)} \tag{2.37}$$

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 $(X|Y)$ 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 $f_{X|Y}(x|y) = f(x,y)/f_Y(y)$。

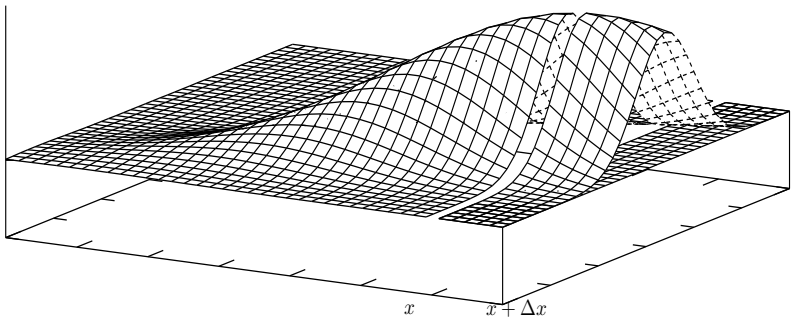


图 2.9: 条件分布函数 $F_{Y|X}(y|x)$ 的含义是随机向量 $(X,Y)^\mathsf{T}$ 落于 xoy 平面的子区域 $(x, x+\Delta x] \times (-\infty, y]$ 的概率和落于 $(x, x+\Delta x] \times \mathbb{R}$ 的概率之比，即曲面 $z = f(x,y)$ 在这两个区域上所围成的体积之比，在 $\Delta x \to 0$ 时的极限。

性质 2.5. 从式 (2.36) 和式 (2.37)，可直接得到全概率公式 $\mathsf{P}(B) = \sum_{j=1}^{\infty} \mathsf{P}(A_j)\mathsf{P}(B|A_j)$ 的“连续型随机变量版”。

$$F_Y(y) = \int_{-\infty}^{+\infty} f_X(x)F(y|x)dx \text{ 且 } f_Y(y) = \int_{-\infty}^{+\infty} f_X(x)f(y|x)dx \qquad (2.38)$$

例 2.17. 接着例 2.13，从 $(X,Y)^\mathsf{T} \sim \mathrm{N}(0,0,1,4,0.8)$ 得到条件密度函数 $f(x|y) = \phi(x,y|0,0,1,4,0.8)/\phi(y|0,4)$ 和 $f(y|x) = \phi(x,y|0,0,1,4,0.8)/\phi(x)$。

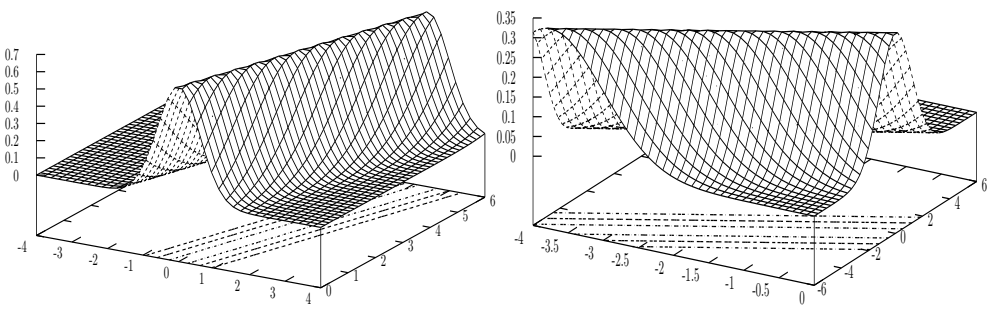


图 2.10: 左图：用 xoz 平面截曲面 $f(x|y) = \phi(x,y|0,0,1,4,0.8)/\phi(y|0,4)$，剖面曲线为 $\phi(x|0,9/16)$。右图：用 yoz 平面截曲面 $f(y|x) = \phi(x,y|0,0,1,4,0.8)/\phi(x)$，剖面曲线为 $\phi(x|0,9/4)$。

2.2.2　随机变量间的独立性

定义 2.11 (独立性)**.** 已知 $F(x,y)$ 是随机向量 $(X,Y)^{\mathsf{T}}$ 的分布函数，$F_X(x)$ 和 $F_Y(y)$ 分别是 X 和 Y 的边缘分布函数，随机变量 X,Y 独立当且仅当

$$\forall (x,y)\in\mathbb{R}^2 \text{ 皆有 } F(x,y)=F_X(x)F_Y(y) \tag{2.39}$$

一般地，随机变量 $X_1,X_2,\cdots,X_n$ 相互独立当且仅当 $F(x_1,x_2\cdots,x_n)=F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n)$，即联合分布函数能分解为边缘分布函数之积。特别地，当独立的随机变量 $X_1,X_2,\cdots,X_n$ 服从相同的分布，譬如正态分布 $\mathrm{N}(\mu,\sigma^2)$，则称之为**独立同分布的** (independent and identically distributed, *iid*)，记作 $X_1,X_2,\cdots,X_n \overset{iid}{\sim} \mathrm{N}(\mu,\sigma^2)$。

定义 2.12. 随机变量间独立性的定义可自然推广到随机向量上：$\boldsymbol{X}=(X_1,\cdots,X_n)^{\mathsf{T}}$ 与 $\boldsymbol{Y}=(Y_1,\cdots,Y_m)^{\mathsf{T}}$ 相互独立当且仅当它们的联合分布函数，即随机向量 $\boldsymbol{Z}=(X_1,\cdots,X_n,Y_1,\cdots,Y_m)^{\mathsf{T}}$ 的分布函数满足 $F_{\boldsymbol{Z}}(x_1,\cdots,x_n,y_1,\cdots,y_m)=F_{\boldsymbol{X}}(x_1,\cdots,x_n)F_{\boldsymbol{Y}}(y_1,\cdots,y_m)$。注意：$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立并不能推出 $X_1,\cdots,X_n$ 相互独立。

如果把随机变量看成随机向量的特殊情况，下面的两个有关随机向量间独立性的定理虽然简单，但很实用也很重要。

定理 2.13. 定义 2.12 中的随机向量 $\boldsymbol{X},\boldsymbol{Y}$ 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 $\mathsf{P}(\boldsymbol{X}\in B_1,\boldsymbol{Y}\in B_2)=\mathsf{P}(\boldsymbol{X}\in B_1)\mathsf{P}(\boldsymbol{Y}\in B_2)$。例如，$\mathsf{P}(x_1<X\le x_2,y_1<Y\le y_2)=\mathsf{P}(x_1<X\le x_2)\mathsf{P}(y_1<Y\le y_2)$。

证明. “$\Leftarrow$”是显然的。往证“$\Rightarrow$”：根据性质 2.1，事件 $\boldsymbol{X}^{-1}(B_1)$ 和事件 $\boldsymbol{Y}^{-1}(B_2)$ 分别是某些形如 $\boldsymbol{X}^{-1}(-\infty,\boldsymbol{x}]$ 和形如 $\boldsymbol{Y}^{-1}(-\infty,\boldsymbol{y}]$ 的事件经过可数次交、并、差运算得到。由已知条件，事件 $\boldsymbol{X}^{-1}(-\infty,\boldsymbol{x}]$ 与 $\boldsymbol{Y}^{-1}(-\infty,\boldsymbol{y}]$ 独立，于是事件 $\boldsymbol{X}^{-1}(B_1)$ 和 $\boldsymbol{Y}^{-1}(B_2)$ 也是独立的。 □

定理 2.14. 若随机向量 $\boldsymbol{X},\boldsymbol{Y}$ 相互独立，令 g_1,g_2 是 Borel 可测函数，则随机向量 $g_1(\boldsymbol{X})$ 与 $g_2(\boldsymbol{Y})$ 也相互独立。

证明. 由定理 2.12 知 $g_1(\boldsymbol{X}), g_2(\boldsymbol{Y})$ 是随机向量，利用定理 2.13 有

$$\begin{aligned}\mathsf{P}\{g_1(\boldsymbol{X}) \le \boldsymbol{x}, g_2(\boldsymbol{Y}) \le \boldsymbol{y}\} &= \mathsf{P}\left\{\boldsymbol{X} \in g_1^{-1}(-\infty, \boldsymbol{x}], \boldsymbol{Y} \in g_2^{-1}(-\infty, \boldsymbol{y}]\right\} \\ &= \mathsf{P}\left\{\boldsymbol{X} \in g_1^{-1}(-\infty, \boldsymbol{x}]\right\} \mathsf{P}\left\{\boldsymbol{Y} \in g_2^{-1}(-\infty, \boldsymbol{y}]\right\} \\ &= \mathsf{P}\{g_1(\boldsymbol{X}) \le \boldsymbol{x}\} \mathsf{P}\{g_2(\boldsymbol{Y}) \le \boldsymbol{y}\}\end{aligned}$$ □

性质 2.6. 若 X, Y 是相互独立的离散型随机变量，则

$$p_{ij} = \mathsf{P}(X = x_i, Y = y_j) = \mathsf{P}(X = x_i)\mathsf{P}(Y = y_j) = p_{i\cdot} p_{\cdot j} \tag{2.40}$$

$$\mathsf{P}(X = x_i | Y = y_j) = p_{i\cdot} \text{ 且 } \mathsf{P}(Y = y_j | X = x_i) = p_{\cdot j} \tag{2.41}$$

若 X, Y 是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 $F(x|y) = F_X(x)$。

表 2.4: 随机变量 X, Y 有相同的分布，它们有可能不相互独立。

	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

例 2.18. 已知 $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ 和 $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ 相互独立，则它们的联合密度函数为 $\phi(x, y | \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, 0)$，其中 X, Y 的相关系数为零。

1925 年，现代统计学奠基人之一、英国著名数学家 Ronald Aylmer Fisher (1890-1962) 证得下面重要的结果。

定理 2.15 (Fisher, 1925)**.** 已知 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，则随机变量 $\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$ 与 $S^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2$ 相互独立。

证明. 本书 §3.1.1 的定理 3.13 断言 $\overline{X}$ 与 $(X_1 - \overline{X}, \cdots, X_n - \overline{X})^\intercal$ 相互独立（证明过程要用到特征函数这一工具），利用定理 2.14 即证得。 □

2.2.3　随机向量的函数

已知随机向量 $(X,Y)^{\mathsf{T}}$ 的密度函数为 $f(x,y)$。假设变换 g_1,g_2 都是连续的一一映射，具有连续的偏导数，把随机向量 $(X,Y)^{\mathsf{T}}$ 变为随机向量 $(U,V)^{\mathsf{T}}$，其中 $U=g_1(X,Y),V=g_2(X,Y)$，问 $(U,V)^{\mathsf{T}}$ 的密度函数？

定理 2.16. 若映射 $\begin{cases} u=g_1(x,y) \\ v=g_2(x,y) \end{cases}$ 在 $(a,b]\times(c,d]$ 上有 $\begin{vmatrix} \partial g_1/\partial x & \partial g_1/\partial y \\ \partial g_2/\partial x & \partial g_2/\partial y \end{vmatrix} \neq 0$，则存在逆映射 $x=h_1(u,v),y=h_2(u,v)$ 使得

$$\begin{aligned} \mathsf{P}(a<X\le b,c<Y\le d) &= \int_a^b\int_c^d f(x,y)dydx \\ &= \iint_S f[h_1(u,v),h_2(u,v)]|J|dudv \end{aligned} \tag{2.42}$$

其中，J 为雅可比行列式 $J=\begin{vmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{vmatrix}$，$S$ 是矩形 $(a,b]\times(c,d]$ 在 g_1,g_2 下的像。$(U,V)^{\mathsf{T}}$ 的密度函数为 $f[h_1(u,v),h_2(u,v)]|J|$。

例 2.19. 已知 $f(x,y)$ 是随机向量 $(X,Y)^{\mathsf{T}}$ 的密度函数，分别求随机变量 $Z=X+Y,XY,X-Y,X/Y$ 的密度函数。

解. 变换 $\begin{cases} X=X \\ Z=X+Y \end{cases}$ 的逆变换为 $\begin{cases} X=X \\ Y=Z-X \end{cases}$，由定理 2.16，$(X,Z)^{\mathsf{T}}$ 的密度函数为 $f(x,z-x)$，于是

$$f_Z(z)=\int_{-\infty}^{+\infty} f(x,z-x)dx \tag{2.43}$$

变换 $\begin{cases} X=X \\ Z=XY \end{cases}$ 的逆变换为 $\begin{cases} X=X \\ Y=Z/X \end{cases}$ 且 $J=\begin{vmatrix} 1 & 0 \\ -z/x^2 & 1/x \end{vmatrix}=1/x$，由定理 2.16，$(X,Z)^{\mathsf{T}}$ 的密度函数为$f(x,z/x)/|x|$，于是

$$f_Z(z)=\int_{-\infty}^{+\infty} f(x,z/x)/|x|dx$$

作为练习，请读者验证随机变量 $Z = X - Y$ 和 $Z = X/Y$ 的密度函数分别为 $f_Z(z) = \int_{-\infty}^{+\infty} f(x, x-z)dx$ 和 $f_Z(z) = \int_{-\infty}^{+\infty} f(yz, y)|y|dy$。

例 2.20 (卷积)**.** 在例 2.19 中，如果加上 X, Y 相互独立这一额外条件，则 $Z = X + Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x)dx \tag{2.44}$$

上式右边恰是 $\mathbb{R}$ 上可积函数 $f_X(x)$ 和 $f_Y(x)$ 的卷积 (convolution)*，记作 $f_X * f_Y$。请读者验证卷积满足交换律、结合律、对加法的分配律。

例 2.21. 已知随机变量 $X \sim \mathrm{U}[0,1]$ 与 $Y \sim \mathrm{N}(0,1)$ 相互独立，则 $Z = X+Y$ 的密度函数为 $f_Z(z) = \int_0^1 \phi(z-x)dx = \Phi(z) - \Phi(z-1)$，该曲线关于 $z = 1/2$ 对称，也呈现钟形，但不是正态分布（见下图）。

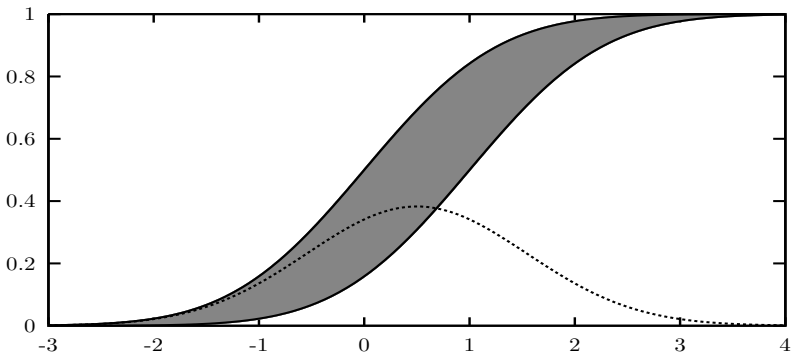


图 2.11: 阴影部分由 $\Phi(z)$ 和 $\Phi(z-1)$ 围成，虚线是 $\Phi(z) - \Phi(z-1)$。

练习 2.2. 如果 $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ 与 $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ 相互独立，试求 $X + Y$ 的分布？答案：$X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$。类似例 1.31 的解法，验证

$$\phi(x|\mu_X, \sigma_X^2)\phi(z-x|\mu_Y, \sigma_Y^2) \propto \phi\left(x \,\middle|\, \frac{\mu_X}{\sigma_X^2} + \frac{z-\mu_Y}{\sigma_Y^2}, \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)\phi(z|\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$$

更简单的证法见定理 4.6 的证明，用到了第三章即将介绍的特征函数。

*在物理学中，任何一个符合叠加原理的线性系统都存在卷积。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。

2.3　随机变量的数字特征

期望 (expectation)、方差 (variance)、矩 (moment)、分位数等是随机变量/向量概率分布的重要数字特征，通过它们可以勾勒出随机变量/向量的“大致轮廓”。例如，

定义 2.13. 已知随机变量 X 的分布函数为 $F(x)$，对于 $0 \le \alpha \le 1$，若实数 q_α 满足 $F(q_\alpha) \le \alpha$ 且 $F(q_\alpha + 0) \ge \alpha$，则称 q_α 为 α-分位数 (α-th quantile)*。特别地，分位数 $q_{1/2}$ 称为中位数 (median)，记作 $\mathsf{M}(X)$ 或 $\mathsf{M}X$ 或 m_X。只要选择合适的 p，通过一系列的 α-分位数就可以了解分布函数的大致情况。R 语言提供了常见分布的分位数函数，如正态分布的 `qnorm`。

不计较具体的分布，数字特征及其之间的基本关系揭示了随机变量的内在规律，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。

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 $\mathsf{E}(X)$ 或 $\mathsf{E}X$，有时也简记为 μ_X 或 μ。通俗地讲，随机变量 X 的期望就是 X 的平均取值。拿离散型的随机变量为例，期望值就是随机变量取值的加权平均，其中权重是每个取值对应的概率值——期望可看作是算术平均的推广。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。

问题 2.1 (赌资分配)**.** 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？

*按照定义，α-分位数总是存在的，但不一定唯一。当 $F(x)$ 严格单调增时，则 α-分位数 q_α 唯一，就是方程 $F(x) = \alpha$ 的解 $x = q_\alpha$。当 $F(x)$ 连续时，对于 $0 \le \alpha_1 < \alpha_2 \le 1$ 有 $\mathsf{P}(q_{\alpha_1} < X \le q_{\alpha_2}) = \alpha_2 - \alpha_1$。

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，而期望的概念则是这两位天才讨论所得的副产品，它在随机变量明确定义之前就出现了。

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 $X \sim \frac{15}{16}\langle 64\rangle + \frac{1}{16}\langle 0\rangle$，在现实中分配给甲的赌资恰是 X 的均值 $64 \times \frac{15}{16} + 0 \times \frac{1}{16} = 60$ 个金币。

受随机因素或缺失信息的影响，我们经常遇到这样的决策问题：有若干可选的行为 $a_1, \cdots, a_n$，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 $p_i (i = 1, 2, \cdots)$，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》(Pensées, 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 2.22. 现有私人财产 100 元人民币，若存银行可稳赚利息 20 元；若用于投资，有 10% 的可能血本无归，也有 90% 的可能赢得 50 元。试

*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。

问这笔钱该用于投资还是存银行？

表 2.5: “投资”有两个不确定的结果，它的期望损失 $= -50 \times 0.9 + 100 \times 0.1 = -35 <$ “存银行”的期望损失 $= -30$，所以选“投资”行为。

行为	投资成功 (90%)	投资失败 (10%)	期望损失
投资	−50	100	−35
存银行	−30	−30	−30

把该例中的货币单位“元”改为“亿元”后情况会怎样？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》[16] 和 M. H. DeGroot 的《最优统计决策》[31]。

本节内容　随机变量 X 的期望 $\mathsf{E}(X)$ 是首先考虑的数字特征，它刻画了随机变量的平均取值。第一小节还定义了条件期望，并得到颇有意思的“双期望定理”。方差是 $[X - \mathsf{E}(X)]^2$ 的期望，它描述的是随机变量 X 的取值在 $\mathsf{E}(X)$ 的分散程度。期望和方差引出了优美且应用广泛的 Markov 不等式、Chebyshev 不等式、Kolmogorov 不等式等结果，第二小节主要介绍它们的由来。还有一些数字特征，如原点矩、中心矩、绝对矩、偏度系数、峰度系数、变异系数、协方差等，在第三小节给出了定义和讨论。两个随机变量之间的线性相关程度可以由相关系数来衡量，这是由方差和协方差定义的一个数字特征，在统计学的回归分析中有重要的应用。第四小节重点讨论最小二乘法和回归 (regression)，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。

学习目标　(1) 理解期望、方差、协方差、相关系数的概率含义以及有关它们的诸多性质；(2) 熟练掌握 Hölder 不等式、Markov 不等式、Chebyshev 不等式、Kolmogorov 不等式、Lévy 不等式等结果。(3) 了解最小二乘法和回归直线。

2.3.1 期望与方差的定义与基本性质

定义 2.14 (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 $\mathsf{P}(X = x_j) = p_j, j = 1, 2, \cdots$ 满足下面的条件

$$\sum_{j=1}^{\infty} |x_j| p_j < \infty \tag{2.45}$$

则称以下级数为离散型随机变量 X 的期望，记作 $\mathsf{E}(X)$ 或 $\mathsf{E}X$。

$$\mathsf{E}(X) = \sum_{j=1}^{\infty} x_j p_j \tag{2.46}$$

条件式 (2.45) 说明了级数 (2.46) 是绝对收敛的，于是它的值与求和次序无关，这样式 (2.46) 才是有意义的。另外，式 (2.46) 亦可理解为 $x_1, x_2, \cdots$ 的加权平均，权重分别为 $p_1, p_2, \cdots$。算术平均值 $\sum_{j=1}^{n} y_j/n$（假设 $y_1, y_2, \cdots, y_n$ 两两不等）就是 $Y \sim \frac{1}{n}\langle y_1\rangle + \cdots + \frac{1}{n}\langle y_n\rangle$ 的期望。

例 2.23. 单点分布 $X \sim \langle c\rangle$ 的期望 $\mathsf{E}(X) = c$。两点分布 $X \sim p\langle a\rangle + (1-p)\langle b\rangle$ 的期望 $\mathsf{E}(X) = ap + b(1-p)$。

例 2.24. 二项分布 $X \sim \mathrm{B}(n, p)$ 的期望是

$$\begin{aligned}\mathsf{E}(X) &= \sum_{k=0}^{n} k \cdot \mathsf{P}(X = k) = \sum_{k=0}^{n} k C_n^k p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} C_{n-1}^{k-1} p^{k-1} (1-p)^{n-k} = np[p + (1-p)]^{n-1} = np\end{aligned}$$

定义 2.15 (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 $f(x)$ 满足下面的条件

$$\int_{-\infty}^{+\infty} |x| f(x) dx < \infty \tag{2.47}$$

则称如下积分为连续型随机变量 X 的期望，记作 $\mathsf{E}(X)$ 或 $\mathsf{E}X$。

$$\mathsf{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx \tag{2.48}$$

绝对可积条件 (2.47) 保证了式 (2.48) 的存在性。

例 2.25. 均匀分布 $X \sim \mathrm{U}[a,b]$（其中 $-\infty < a < b < +\infty$）的期望是

$$\mathsf{E}(X) = \int_a^b x \cdot \frac{1}{b-a}dx = \frac{b+a}{2}$$

例 2.26. 正态分布 $X \sim \mathrm{N}(\mu,\sigma^2)$ 的期望就是密度函数的对称位置，即

$$\mathsf{E}(X) = \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = \mu$$

若条件式 (2.45) 或式 (2.47) 不成立，即便式 (2.46) 或式 (2.48) 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 2.27. 考察离散型随机变量 X，设它的概率质量函数为

$$\mathsf{P}\left\{X = (-1)^k 2^k/k\right\} = 1/2^k，\text{其中 } k = 1,2,\cdots$$

显然 $\sum_{k=1}^{\infty} |x_k| p_k = \sum_{k=1}^{\infty} 1/k = \infty$ 使得 X 的期望不存在，此时算得 $\sum_{k=1}^{\infty} x_k p_k = \sum_{k=1}^{\infty} (-1)^k/k = -\ln 2$ 是没有任何意义的。

例 2.28 (Cauchy 分布)**.** 如果一个连续型随机变量 X 的密度函数为

$$f(x) = \frac{\lambda}{\pi[(x-\mu)^2 + \lambda^2]}，\text{其中 } \lambda > 0, -\infty < x, \mu < \infty \tag{2.49}$$

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 称为位置参数，λ 称为尺度参数。Cauchy 分布的分布函数为

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{x-\mu}{\lambda} \tag{2.50}$$

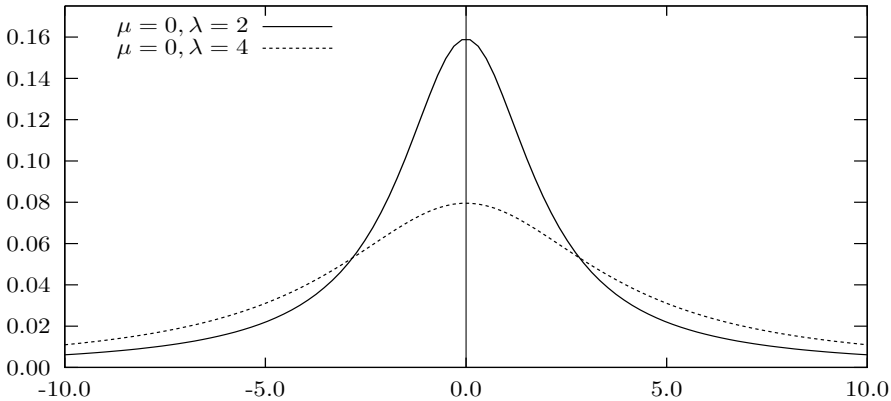


图 2.12: Cauchy 分布的密度函数关于 $x = \mu$ 对称。尺度参数 λ 越小，密度函数越“高瘦”。对于 Cauchy 分布，请读者验证条件式 (2.47) 不成立，这说明 Cauchy 分布的期望不存在。

已知随机变量 X 的分布函数为 $F(x)$，则不论 X 是离散型的还是连续型的，它的期望可统一表示成 Riemann-Stieltjes 积分或 Lebesgue-Stieltjes 积分*

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x dF(x) \tag{2.51}$$

这能带来书写和讨论的便捷，譬如

$$\mathsf{P}(X \in A) = \int_A dF(x)，\text{其中 } A \subseteq \mathbb{R} \tag{2.52}$$

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，指代式式 (2.46) 和式式 (2.48)。

定理 2.17. 设 h 是一个 Borel 可测函数，若 $Y = h(X)$ 的期望存在，则

$$\mathsf{E}(Y) = \int_{-\infty}^{\infty} h(x) dF(x) = \begin{cases} \sum_{j=1}^{\infty} h(x_j) p_j & \text{离散的情形} \\ \int_{-\infty}^{\infty} h(x) f(x) dx & \text{连续的情形} \end{cases} \tag{2.53}$$

*见附录 E 和附录 F 的简介或 W. Rudin 的名著《数学分析原理》 [79] 第六章。

证明. 见 Michel Loève (1907-1979) 的《概率论》第三章第十节。 □

例 2.29. 设信号为一个离散型随机变量 $X \sim p_1\langle x_1\rangle + p_2\langle x_2\rangle + \cdots + p_n\langle x_n\rangle$，为了提高通信效率，$x_j$ 出现的概率 p_j 越大，在通信中对 x_j 的编码长度就越短*。1948 年，信息论之父、美国电子工程师 Claude Elwood Shannon (1916-2001) 在《通信的数学理论》 [82] 一文中定义了离散型随机变量 $X \sim p_1\langle x_1\rangle + p_2\langle x_2\rangle + \cdots + p_n\langle x_n\rangle$ 的熵 (entropy) 如下：

$$\mathsf{H}(X) = -\sum_{j=1}^{n} p_j \ln p_j \tag{2.54}$$

其中 $-\ln p_j$ 是 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 $\mathsf{H}(X)$ 也可看作是平均编码长度（为方便讨论，我们约定 $0 \ln 0 = 0$）。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，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》 [26] 中 Huffman 编码的贪心算法）。

定义 2.16. 已知二维随机向量 $(X, Y)^{\mathsf{T}}$ 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

$$\mathsf{E}[g(X, Y)] = \int_{\mathbb{R}^2} g(x, y) dF(x, y) \tag{2.55}$$

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

$$\mathsf{E}[g(X, Y)] = \sum_{i,j=1}^{\infty} g(x_i, y_j) p_{ij} \tag{2.56}$$

连续型： 若 $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g(x, y)| f(x, y) dxdy < \infty$，则定义 $g(X, Y)$ 的期望为

$$\mathsf{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dxdy \tag{2.57}$$

*类似地，在自然语言中一般来说越常用的单词其音节也越短。

例 2.30. 设离散型随机向量 $(X,Y)^{\mathsf{T}}$ 的分布列为 $\mathsf{P}(X=x_i,Y=y_j)=p_{ij}$，其中 $i=1,2,\cdots,m$, $j=1,2,\cdots,n$。仿照式 (2.54)，随机向量 $(X,Y)^{\mathsf{T}}$ 的熵（也称之为 X 与 Y 的联合熵）定义如下：

$$\mathsf{H}(X,Y)=-\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}\ln p_{ij} \tag{2.58}$$

为了刻画 X 与 Y 相互依赖的程度，Shannon 还定义了 X 与 Y 的互信息 (mutual information)：

$$\mathsf{MI}(X,Y)=\mathsf{H}(X)+\mathsf{H}(Y)-\mathsf{H}(X,Y) \tag{2.59}$$

显然，如果 X 与 Y 独立，则互信息为 0（请读者验证）。

性质 2.7. 如果随机变量 X,Y 的期望都存在，则 $\forall a,b\in\mathbb{R}$ 有

$$\mathsf{E}(b)=b \tag{2.60}$$

$$\mathsf{E}(aX+b)=a\mathsf{E}(X)+b，\text{特别地 } \mathsf{E}[X-\mathsf{E}(X)]=0 \tag{2.61}$$

$$\mathsf{E}(X+Y)=\mathsf{E}(X)+\mathsf{E}(Y) \tag{2.62}$$

证明. 由单点分布 $\mathsf{P}(X=b)=1$ 易证式 (2.60)。式 (2.61) 由式 (2.51) 和 Riemann-Stieltjes 积分的性质（见附录 E）导出。作为练习，读者也可以从定义 2.14 和定义 2.15 出发给出式 (2.61) 和式 (2.62) 的证明，就像下面证明式 (2.62) 的离散情形。

$$\begin{aligned}\mathsf{E}(X+Y)&=\sum_{i,j}p_{ij}(x_i+y_j)=\sum_i x_i\sum_j p_{ij}+\sum_j y_j\sum_i p_{ij}\\&=\sum_i x_i p_{i\cdot}+\sum_j y_j p_{\cdot j}=\mathsf{E}(X)+\mathsf{E}(Y)\end{aligned}$$ □

✎公式 $\mathsf{E}[X-\mathsf{E}(X)]=0$ 是直观的：把 $X-\mathsf{E}(X)$ 解释为随机变量 X 偏离均值 $\mathsf{E}(X)$ 的误差，这误差可正可负，但其均值为零。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（留作课后习题）要用到即将介绍的 Chebyshev 不等式。

性质 2.8. 随机变量 X 满足 $\mathsf{E}(X^2)=0 \Leftrightarrow X$ 服从单点分布 $\mathsf{P}(X=0)=1$。

性质 2.9. 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 $\mathsf{E}(XY)=\mathsf{E}(X)\mathsf{E}(Y)$。

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。

$$\mathsf{E}(XY)=\sum_{i,j}p_{ij}x_iy_j=\sum_{i,j}p_{i\cdot}p_{\cdot j}x_jy_j=\sum_i x_ip_{i\cdot}\sum_j y_jp_{\cdot j}=\mathsf{E}(X)\mathsf{E}(Y) \qquad \square$$

定理 2.18 (Hölder 不等式)**.** 若 $r>1$ 且 $1/r+1/s=1$，则

$$\mathsf{E}|XY|\le\{\mathsf{E}|X|^r\}^{1/r}\{\mathsf{E}|Y|^s\}^{1/s} \tag{2.63}$$

证明. 利用数学分析中的 Hölder 不等式：对任意实数 x,y 皆有

$$|xy|\le\frac{|x|^r}{r}+\frac{|y|^s}{s}，\text{其中 } r>1 \text{ 且 } \frac{1}{r}+\frac{1}{s}=1 \tag{2.64}$$

令 $x=X\{\mathsf{E}|X|^r\}^{-1/r}, y=Y\{\mathsf{E}|Y|^s\}^{-1/s}$，有 $|XY|\le r^{-1}|X|^r\{\mathsf{E}|X|^r\}^{1/r-1}\{\mathsf{E}|Y|^s\}^{1/s}+s^{-1}|Y|^s\{\mathsf{E}|Y|^s\}^{1/s-1}\{\mathsf{E}|X|^s\}^{1/s}$，两边求期望便得证。 $\square$

推论 2.1. 下面两个结果是 Hölder 不等式的推论，留作习题。

❶ Minkowski 不等式：若 $r\ge1$，则 $\{\mathsf{E}|X+Y|^r\}^{1/r}\le\{\mathsf{E}|X|^r\}^{1/r}+\{\mathsf{E}|Y|^r\}^{1/r}$。

❷ Cauchy-Schwarz 不等式：$\{\mathsf{E}|XY|\}^2\le\mathsf{E}(X^2)\mathsf{E}(Y^2)$。

定义 2.17. 分别定义 Y 的条件期望和 X 的条件期望如下：

$$\mathsf{E}(Y|X=x)=\int_{-\infty}^{\infty}ydF(y|x),\qquad \mathsf{E}(X|Y=y)=\int_{-\infty}^{\infty}xdF(x|y)$$

$$\mathsf{E}(Y|X=x_i)=\sum_j y_j\frac{p_{ij}}{p_{i\cdot}},\qquad \mathsf{E}(X|Y=y_j)=\sum_i x_i\frac{p_{ij}}{p_{\cdot j}} \tag{2.65}$$

$$\mathsf{E}(Y|X=x)=\int_{-\infty}^{+\infty}y\frac{f(x,y)}{f_X(x)}dy,\qquad \mathsf{E}(X|Y=y)=\int_{-\infty}^{+\infty}x\frac{f(x,y)}{f_Y(y)}dx \tag{2.66}$$

定理 2.19 (双期望)**.** 按照条件期望的定义，$\mathsf{E}(Y|X)$ 是关于 X 的函数，也是一个随机变量，它的期望是 $\mathsf{E}(Y)$。即

$$\mathsf{E}[\mathsf{E}(Y|X)] = \mathsf{E}(Y) \tag{2.67}$$

证明. 我们证离散的情形，连续的情形类似，留作练习。

$$\begin{aligned}\mathsf{E}[\mathsf{E}(Y|X)] &= \sum_i \mathsf{P}(X = x_i)\mathsf{E}(Y|X = x_i) \\ &= \sum_i p_{i\cdot} \sum_j y_j \frac{p_{ij}}{p_{i\cdot}} = \sum_i \sum_j y_j p_{ij} = \sum_j y_j p_{\cdot j} = \mathsf{E}(Y)\end{aligned}$$ □

双期望定理的含义是：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。

对于一个随机变量 X，我们可以用 $[X - \mathsf{E}(X)]^2$ 来刻画随机变量 X 离开均值 $\mathsf{E}(X)$ 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 $\mathsf{E}(X)$ 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 2.18 (方差)**.** 若 $\mathsf{E}(X)$ 和 $\mathsf{E}(X^2)$ 皆存在，随机变量 X 的*方差*†定义为

$$\begin{aligned}\mathsf{V}(X) &= \mathsf{E}[X - \mathsf{E}(X)]^2 &(2.68)\\ &= \mathsf{E}\left\{X^2 - 2X \cdot \mathsf{E}(X) + [\mathsf{E}(X)]^2\right\} = \mathsf{E}(X^2) - 2[\mathsf{E}(X)]^2 + [\mathsf{E}(X)]^2 \\ &= \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 &(2.69)\end{aligned}$$

有时 X 的方差 $\mathsf{V}(X)$ 又记为 σ_X^2，或者 σ^2。方差的算术平方根，$\sigma_X = \sqrt{\mathsf{V}(X)}$ 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。式 (2.69) 常用作方差定义的另一个等价形式，由它可直接推得下面的式 (2.71)。

*当然，我们也可以用 $|X - \mathsf{E}(X)|$ 来衡量随机变量 X 的取值在 $\mathsf{E}(X)$ 周围的分散程度，并称 $\mathsf{E}|X - \mathsf{E}(X)|$ 为平均差异。

†有些文献中把 X 的方差记作 $\mathrm{var}(X)$、$D(X)$ 等，我们采用最流行的记号 $\mathsf{V}(X)$。

性质 2.10. 若随机变量 X 的方差 $\mathsf{V}(X)$ 存在，则有

$$\mathsf{V}(aX+b)=a^2\mathsf{V}(X)，\text{其中 } \forall a,b\in\mathbb{R} \tag{2.70}$$

$$\mathsf{V}(X)\le\mathsf{E}(X^2) \tag{2.71}$$

$$\mathsf{V}(X)\le\mathsf{E}(X-c)^2，\text{其中 } \forall c\in\mathbb{R} \tag{2.72}$$

证明. $\mathsf{V}(aX+b)=\mathsf{E}[aX+b-\mathsf{E}(aX+b)]^2=a^2\mathsf{E}[X-\mathsf{E}(X)]^2=a^2\mathsf{V}(X)$。

$$\begin{aligned}\mathsf{E}(X-c)^2&=\mathsf{E}[X-\mathsf{E}(X)+\mathsf{E}(X)-c]^2\\&=\mathsf{E}[X-\mathsf{E}(X)]^2+2\mathsf{E}[X-\mathsf{E}(X)]\cdot[\mathsf{E}(X)-c]+[\mathsf{E}(X)-c]^2\\&=\mathsf{V}(X)+[\mathsf{E}(X)-c]^2\ge\mathsf{V}(X)\end{aligned}$$ □

从 $\mathsf{V}(X+b)=\mathsf{V}(X)$ 看出，方差是一个平移不变量：平移一个随机变量，丝毫不能改变其取值的分散程度。不等式 (2.72) 意味着函数 $f(c)=\mathsf{E}(X-c)^2$ 在 $c=\mathsf{E}(X)$ 处取最小值 $\mathsf{V}(X)$。

性质 2.11. 若随机变量 $X_1,X_2,\cdots,X_n$ 相互独立且方差都存在，则

$$\mathsf{V}\left(\sum_{j=1}^{n}X_j\right)=\sum_{j=1}^{n}\mathsf{V}(X_j) \tag{2.73}$$

证明. 只证 $n=2$ 的情形：由式 (2.69) 和性质 2.9，

$$\begin{aligned}&\mathsf{V}(X_1+X_2)=\mathsf{E}[(X_1+X_2)^2]-[\mathsf{E}(X_1+X_2)]^2\\&=\mathsf{E}(X_1^2)+\mathsf{E}(X_2^2)+2\mathsf{E}(X_1)\mathsf{E}(X_2)-[\mathsf{E}(X_1)]^2-[\mathsf{E}(X_2)]^2-2\mathsf{E}(X_1)\mathsf{E}(X_2)\\&=\mathsf{V}(X_1)+\mathsf{V}(X_2)\end{aligned}$$ □

性质 2.12. $\mathsf{V}(X)=0$ 当且仅当 X 服从单点分布 $\mathsf{P}\{X=\mathsf{E}(X)\}=1$。

证明. 令 $Y=X-\mathsf{E}(X)$，利用性质 2.8 可证。 □

例 2.31. 考察 $X \sim \mathrm{B}(n,p)$ 的方差。利用式 (2.69)，首先计算

$$\mathsf{E}(X^2) = \sum_{k=0}^{n} k^2 C_n^k p^k (1-p)^{n-k} = np(1-p+pn)$$

于是，$\mathsf{V}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 = np(1-p+pn) - n^2p^2 = np(1-p)$。

练习 2.3. 请读者验证：正态分布 $\mathrm{N}(\mu,\sigma^2)$ 的方差是 σ^2，并且 σ^2 越小，$\mathsf{P}\{X \in (\mu-\epsilon, \mu+\epsilon)\}$ 越大，其中 ϵ 是一个给定的正数。

定义 2.19. 如果随机变量 Y 满足 $\mathsf{E}(Y) = 0$ 且 $\mathsf{V}(Y) = 1$，我们称 Y 为标准化的随机变量。已知随机变量 X 的期望和标准差分别为 $\mu = \mathsf{E}(X)$ 和 $\sigma = \sqrt{\mathsf{V}(X)}$，则 $Y = (X-\mu)/\sigma$ 是 X 经过标准化得到的随机变量。

2.3.2　Chebyshev 不等式和 Kolmogorov 不等式

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、函数逼进理论等方面颇多建树。他对概率论的贡献包括：(1) 1866 年利用 Chebyshev 不等式证明了 Chebyshev 弱大数律。(2) 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。

引理 2.1 (Markov 不等式)**.** 令非负随机变量 Y 有期望 $\mathsf{E}(Y)$，则 $\forall k > 0$，下面的不等式成立。

$$\mathsf{P}(Y \geq k) \leq \frac{\mathsf{E}(Y)}{k} \text{ 或 } \mathsf{P}(Y \leq k) \geq 1 - \frac{\mathsf{E}(Y)}{k} \tag{2.74}$$

证明. 设 Y 的分布函数为 $F(y)$，则

$$\mathsf{E}(Y) = \int_0^{+\infty} y dF(y) \geq \int_k^{+\infty} y dF(y) \geq k \int_k^{+\infty} dF(y) \geq k\mathsf{P}(Y \geq k) \quad \square$$

推论 2.2 (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 $\mathsf{E}h(X)$ 存在，则 $\forall k > 0$ 有

$$\mathsf{P}\{h(X) \geq k\} \leq \frac{\mathsf{E}h(X)}{k} \tag{2.75}$$

例 2.32. 假设随机变量 X 的期望 $\mathsf{E}(X)=0$，方差 $\mathsf{V}(X)=\sigma^2$，试证明：

$$\mathsf{P}(X\geq x)\leq\frac{\sigma^2}{x^2+\sigma^2}\qquad\text{如果 } x>0$$
$$\mathsf{P}(X\geq x)\geq\frac{x^2}{x^2+\sigma^2}\qquad\text{如果 } x<0$$

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 $h(X)\geq(x+c)^2$。

$$\mathsf{P}\{X\geq x\}\leq\mathsf{P}\left\{h(X)\geq(x+c)^2\right\}\leq\frac{\mathsf{E}(X+c)^2}{(x+c)^2}=\frac{\sigma^2+c^2}{(x+c)^2}$$

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 2.20 (Chebyshev 不等式, 1866)**.** 若随机变量 X 的期望 $\mathsf{E}(X)$ 和方差 $\mathsf{V}(X)$ 都存在，则 $\forall\epsilon>0$，下面的不等式成立。

$$\mathsf{P}\left\{|X-\mathsf{E}(X)|\geq\epsilon\sqrt{\mathsf{V}(X)}\right\}\leq\frac{1}{\epsilon^2},\text{ 或等价于}\tag{2.76}$$
$$\mathsf{P}\{|X-\mathsf{E}(X)|\geq\epsilon\}\leq\frac{\mathsf{V}(X)}{\epsilon^2}\text{ 或 }\mathsf{P}\{|X-\mathsf{E}(X)|<\epsilon\}\geq1-\frac{\mathsf{V}(X)}{\epsilon^2}\tag{2.77}$$

证明. 令 $Y=[X-\mathsf{E}(X)]^2$ 且 $k=\epsilon^2\mathsf{V}(X)$，将之代入 Markov 不等式 (2.74) 便可证得 Chebyshev 不等式 (2.76)。它的等价形式 (2.77) 是显然的。 □

例 2.33. 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. 假设在 n 重 Bernoulli 试验中，事件 A 出现了 m 次。由 Chebyshev 不等式，$\mathsf{P}\{|m/n-p|<\epsilon\}\geq1-p(1-p)/(n\epsilon^2)\geq0.95$，其中 $\epsilon=0.01, p=1/2$，解此不等式得到 $n\geq5\times10^4$。

类似于正态分布的 3σ 原则，Chebyshev 不等式 (2.76) 揭示了 X 距离 $\mathsf{E}(X)$ 不超过 3 个标准差的概率大于 8/9。不等式 (2.77) 揭示了方差越小，X 落于 $(\mathsf{E}X-\epsilon, \mathsf{E}X+\epsilon)$ 的概率就越大，至少为 $1-\mathsf{V}(X)/\epsilon^2$。简洁的 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 5.3），而不是概率的精确估计。Chebyshev 不等式的一个非平凡推广是 Kolmogorov 不等式（在第五章将被用于证明 Kolmogorov 强大数律，见定理 5.8 的证明），而 Lévy 不等式又是对 Kolmogorov 不等式的推广。

定理 2.21 (Kolmogorov 不等式, 1928-1929)**.** 随机变量 $X_1, X_2, \cdots, X_n$ 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。

$$\mathsf{P}\left(\bigcup_{i=1}^{n}\left\{\left|\sum_{j=1}^{i} X_j - \mathsf{E}X_j\right| \geq \epsilon\right\}\right) \leq \frac{\sum_{i=1}^{n}\mathsf{V}(X_i)}{\epsilon^2} \tag{2.78}$$

如果 X_j 有界*，不妨设 $|X_j| \leq c$，其中 $j = 1, 2, \ldots, n$，则还有

$$1 - \frac{(\epsilon+2c)^2}{\sum_{i=1}^{n}\mathsf{V}(X_i)} \leq \mathsf{P}\left(\bigcup_{i=1}^{n}\left\{\left|\sum_{j=1}^{i} X_j - \mathsf{E}X_j\right| \geq \epsilon\right\}\right) \tag{2.79}$$

证明. 令 $Y_j = X_j - \mathsf{E}X_j$ 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 $i = 1, 2, \cdots, n$。显然，事件 $A_1, A_2, \cdots, A_n$ 是两两互斥的。

下面的事件是等价的：“至少有一个 i $(1 \leq i \leq n)$ 使得 $|Z_i| \geq \epsilon$” $\Leftrightarrow$ “$\max\limits_{1\leq i\leq n}|Z_i| \geq \epsilon$” $\Leftrightarrow$ “$\sum_{i=1}^{n} A_i$”。因为 $\mathsf{V}(Z_n) = \sum_{i=1}^{n}\mathsf{V}(X_i)$，所以往证 (2.78) 即往证 $\sum_{i=1}^{n}\mathsf{P}(A_i) \leq \mathsf{V}(Z_n)/\epsilon^2$。

首先，$\mathsf{V}(Z_n) = \sum_{i=0}^{n}\mathsf{P}(A_i)\mathsf{E}(Z_n^2|A_i) \geq \sum_{i=1}^{n}\mathsf{P}(A_i)\mathsf{E}(Z_n^2|A_i)$，若能证得 $\mathsf{E}(Z_n^2|A_i) \geq \epsilon^2$ 便万事大吉。将 Z_n 分解为 $Z_n = Z_i + (Y_{i+1} + \cdots + Y_n)$，进而 $Z_n^2 = Z_i^2 + (Y_{i+1} + \cdots + Y_n)^2 + 2Z_i(Y_{i+1} + \cdots + Y_n)$，下面验证 ϵ^2 的确是

*对于随机变量 X，如果存在常数 $c > 0$ 使得 $\mathsf{P}(|X| \leq c) = 1$，则称 X 为有界的，在不引起歧义的情况下也常记作 $|X| < c$。

$\mathsf{E}(Z_n^2|A_i)$ 的下界。

$$\begin{aligned}\mathsf{E}(Z_n^2|A_i) &= \mathsf{E}\left(Z_i^2 + \sum_{j>i} Y_j^2 + 2\sum_{j>i} Z_i Y_j + 2\sum_{k>j>i} Y_k Y_j \middle| A_i\right) \\ &\geq \mathsf{E}\left(Z_i^2 + 2\sum_{j>i} Z_i Y_j + 2\sum_{k>j>i} Y_k Y_j \middle| A_i\right) \geq \epsilon^2\end{aligned}$$

最后一步因为 $\mathsf{E}(Z_i Y_j|A_i) = \mathsf{E}(Z_i|A_i)\mathsf{E}(Y_j|A_i) = 0$ 且 $\mathsf{E}(Y_k Y_j|A_i) = 0$。不等式 (2.79) 的证明见 M. Loève 的《概率论》 [61] 第 248 页。 □

练习 2.4. 验证 Kolmogorov 不等式 (2.78) 还等价于

$$\mathsf{P}\left(\max_{1\leq i\leq n}\left|\sum_{j=1}^{i} X_j - \mathsf{E}X_j\right| \geq \epsilon\right) \leq \frac{\sum_{i=1}^n \mathsf{V}(X_i)}{\epsilon^2} \tag{2.80}$$

$$\mathsf{P}\left(\bigcap_{i=1}^{n}\left\{\left|\sum_{j=1}^{i} X_j - \mathsf{E}X_j\right| < \epsilon\right\}\right) \geq 1 - \frac{\sum_{i=1}^n \mathsf{V}(X_i)}{\epsilon^2} \tag{2.81}$$

并说明 Kolmogorov 不等式是 Chebyshev 不等式的推广。

推论 2.3. 设 $X_1, X_2, \cdots, X_n$ 是期望为 0 且具有有限方差的独立随机变量，令 $S_k = \sum_{j=1}^k X_j, k = 1, 2, \cdots, n$，则 $\forall \epsilon > 0$ 有以下不等式成立。

$$\mathsf{P}\left\{\max_{1\leq k\leq n}|S_k| \geq \epsilon\right\} \leq \frac{\mathsf{E}(S_n^2)}{\epsilon^2} \tag{2.82}$$

下面不加证明地给出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》 [61] 或 W. Feller 的《概率论及其应用》下卷 [36]。

定理 2.22 (Lévy 不等式, 1937)**.** 随机变量 $X_1, X_2, \cdots, X_n$ 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。

$$\mathsf{P}\left\{\max_{1\le k\le n}[S_k - \mathsf{M}(S_k - S_n)] \ge x\right\} \le 2\mathsf{P}\{S_n \ge x\} \tag{2.83}$$

$$\mathsf{P}\left\{\max_{1\le k\le n}|S_k - \mathsf{M}(S_k - S_n)| \ge x\right\} \le 2\mathsf{P}\{|S_n| \ge x\} \tag{2.84}$$

定理 2.23. 接着定理 2.22 的条件，如果 $g(x) \ge 0$ 是凸的单调函数且 $\mathsf{E}g(|S_n|) < \infty$，则

$$\mathsf{P}\left\{\max_{1\le k\le n}|S_k| \ge x\right\} \le \frac{\mathsf{E}g(|S_n|)}{g(x)} \tag{2.85}$$

定理 2.24 (Bernstein-Kolmogorov 不等式, 1911, 1929)**.** 已知随机变量 $X_1, X_2, \cdots, X_n$ 的期望都为 0 且都有界（不妨设 $|X_j| \le c$，其中 $c > 0$ 为常数）。令 $\sigma^2 = \mathsf{V}(X_1 + X_2 + \cdots + X_n)$，则 $\forall \epsilon > 0$ 下面的不等式成立。

$$\mathsf{P}\{|X_1 + X_2 + \cdots + X_n| \ge \epsilon\} \le 2\exp\left\{-\frac{\epsilon^2}{2(\sigma^2 + c\epsilon/3)}\right\} \tag{2.86}$$

练习 2.5. 利用 Bernstein-Kolmogorov 不等式 (2.86) 试给出定理 1.1 的另一个证明。

定理 2.25 (Hoeffding 不等式 [50], 1963)**.** 已知随机变量 $X_1, X_2 \cdots, X_n$ 独立且 $\mathsf{P}\{X_j - \mathsf{E}(X_j) \in [a_j, b_j]\} = 1$，其中 $j = 1, 2, \cdots, n$。令 $S_n = X_1 + X_2 + \cdots + X_n$，则对任意 $t > 0$ 有下面的不等式成立。

$$\mathsf{P}\{S_n - \mathsf{E}(S_n) \ge nt\} \le \exp\left\{-\frac{2n^2t^2}{\sum_{j=1}^{n}(b_j - a_i)^2}\right\} \tag{2.87}$$

$$\mathsf{P}\{|S_n - \mathsf{E}(S_n)| \ge nt\} \le 2\exp\left\{-\frac{2n^2t^2}{\sum_{j=1}^{n}(b_j - a_i)^2}\right\} \tag{2.88}$$

具体应用见定理 1.1 的证明。

2.3.3 矩、协方差与相关系数

矩是概率分布的数字特征，它是期望和方差的自然推广。为了表述的方便，假设此小节中的定义所涉及的期望都存在。

定义 2.20 (矩)**.** 随机变量 X 的 k 阶矩（k 为自然数）定义为

$$m_k = \mathsf{E}(X^k) = \int_{-\infty}^{\infty} x^k dF_X(x) = \begin{cases} \sum_j x_j^k p_j & \text{离散型} \\ \int_{-\infty}^{+\infty} x^k f_X(x)dx & \text{连续型} \end{cases} \tag{2.89}$$

也称为 X 的 k 阶原点矩。下面的数字特征被称为 X 的 k 阶中心矩，

$$\mu_k = \mathsf{E}[X - \mathsf{E}(X)]^k \tag{2.90}$$

显然 2 阶中心矩就是方差 $\mathsf{V}(X)$。由定义不难得出以下关系式：

$$\mu_1 = 0, \quad \mu_2 = m_2 - m_1^2, \quad \mu_3 = m_3 - 3m_1m_2 + 2m_1^3, \cdots \tag{2.91}$$

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 [36] 中指出对数正态分布（见本书第四章）不被它的各阶矩唯一决定。

定义 2.21. 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：

- 变异系数 (coefficient of variation) $c_v = \sigma/\mu$ 是随机变量取值分散程度的（无量纲的）相对度量。
- 偏度系数 (coefficient of skewness) $\gamma_1 = \mu_3/\sigma^3$ 刻画了随机变量关于期望的对称程度。显然，若 X 的密度函数或概率函数关于 $\mathsf{E}(X)$ 对称，则 $\gamma_1 = 0$。

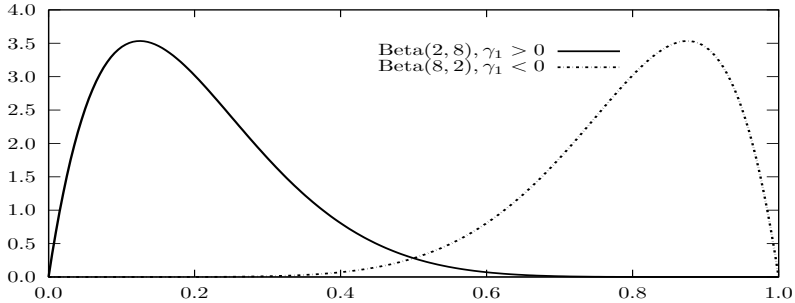


图 2.13: 单峰的密度函数：偏度系数大于 0，“高密度区”（即随机变量最有可能的取值）偏左且右尾长；偏度系数小于 0，“高密度区”偏右且左尾长。K. Pearson 建议这样定义偏度系数：$3[\mathsf{E}(X) - \mathsf{M}(X)]/\sqrt{\mathsf{V}(X)}$，显得更直观些。

❑ 峰度系数 (coefficient of kurtosis) $\gamma_2 = \mu_4/\sigma^4 - 3$ 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。

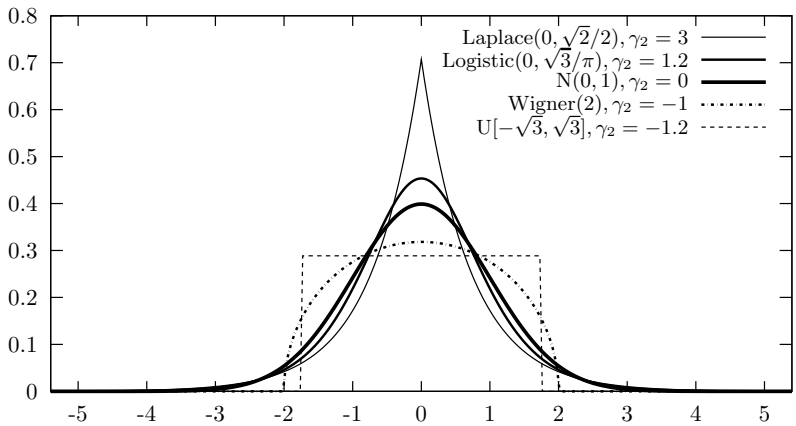


图 2.14: 图中显示了一些常见分布（详见第四章）的密度函数及其峰度系数。正态分布的峰度系数为 0（图中最粗的实线），它是约定的参考标准，其他密度函数曲线顶部的陡峭程度都是相对它而言的（虚线：峰度系数为负）。

练习 2.6. 验证正态分布 $\mathrm{N}(\mu, \sigma^2)$ 的偏度系数和峰度系数分别为 $\gamma_1 = 0$ 和 $\gamma_2 = 0$。更多分布的偏度系数和峰度系数见第四章。

定义 2.22 (绝对矩)**.** 我们称 $\beta_k = \mathsf{E}(|X|^k)$ 为随机变量 X 的 k 阶绝对矩。

性质 2.13. 如果 X 的 k 阶绝对矩 β_k 存在，则 $\beta_1, \cdots, \beta_{k-1}$ 也存在。

证明. 由 $|x|^{k-1} < |x|^k + 1$ 可证。　□

定理 2.26 (Lyapunov 不等式)**.** 假设随机变量 X 的 n 阶绝对矩 $\mathsf{E}(|X|^n)$ 存在，则对 $k = 1, 2, \cdots, n-1$ 有以下不等式成立，

$$\sqrt[k]{\beta_k} \leq \sqrt[k+1]{\beta_{k+1}} \quad \text{或者} \quad \beta_1 \leq \sqrt{\beta_2} \leq \sqrt[3]{\beta_3} \leq \cdots \leq \sqrt[n]{\beta_n} \tag{2.92}$$

证明. 令 r 是任意实数，则有

$$\int_{-\infty}^{+\infty} \left[r|x|^{\frac{k-1}{2}} + |x|^{\frac{k+1}{2}} \right]^2 dF_X(x) = \beta_{k-1} r^2 + 2r\beta_k + \beta_{k+1} \geq 0$$

于是根的判别式非正，得到 $\beta_k^2 \leq \beta_{k-1}\beta_{k+1}$，进而 $\beta_k^{2k} \leq \beta_{k-1}^k \beta_{k+1}^k, k = 1, 2, \cdots, n-1$。即 $\beta_1^2 \leq \beta_0^1\beta_2^1, \beta_2^4 \leq \beta_1^2\beta_3^2, \cdots, \beta_{n-1}^{2(n-1)} \leq \beta_{n-2}^{n-1}\beta_n^{n-1}$，其中 $\beta_0 = 1$。将前 k 个不等式相乘便得到 $\beta_{k-1}^k \leq \beta_k^{k-1}$。 □

定义 2.23 (协方差)**.** 已知随机向量 $(X, Y)^\mathsf{T}$，我们称 $m_{s,t} = \mathsf{E}(X^s Y^t)$ 为 $(X, Y)^\mathsf{T}$ 的 (s, t) 阶矩，其中 s, t 为自然数。类似地，(s, t) 阶中心矩定义为 $\mu_{s,t} = \mathsf{E}[(X - \mathsf{E}(X))^s (Y - \mathsf{E}(Y))^t]$。例如，$m_{1,0} = \mathsf{E}(X), m_{2,0} = \mathsf{E}(X^2), \mu_{1,0} = 0, \mu_{2,0} = \mathsf{V}(X)$。特别地，$\mu_{1,1} = \mathsf{E}[(X - \mathsf{E}X)(Y - \mathsf{E}Y)] = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$ 被称为 X, Y 的 *协方差* (covariance)，记作 $\mathsf{Cov}(X, Y)$。根据式 (2.69) 很容易证得下面的结果，

$$\mathsf{V}(X + Y) = \mathsf{V}(X) + \mathsf{V}(Y) + 2\mathsf{Cov}(X, Y) \tag{2.93}$$

于是性质 2.11 可以换成下面的方式来叙述。

性质 2.14. 如果随机变量 X, Y 独立，则 $\mathsf{Cov}(X, Y) = 0$。

定义 2.24 (相关系数)**.** 已知随机向量 $(X, Y)^\mathsf{T}$，随机变量 X 与 Y 之间的*相关系数* (coefficient of correlation) 定义为

$$\rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{V}(X)\mathsf{V}(Y)}} = \frac{\mu_{1,1}}{\sigma_X \sigma_Y} \tag{2.94}$$

其中 $\sigma_X > 0, \sigma_Y > 0$ 分别为随机变量 X 和 Y 的标准差。有时 $\rho(X, Y)$ 也

记作 $\rho_{X,Y}$ 或 ρ。

性质 2.15. 随机变量 X, Y 之间的相关系数 $\rho(X, Y)$ 满足

$$-1 \leq \rho(X, Y) \leq 1 \tag{2.95}$$

证明. 由 $\mathsf{E}[r(X - \mathsf{E}X) + (Y - \mathsf{E}Y)]^2 \geq 0$，于是对于任意实数 r 总有

$$r^2\sigma_X^2 + 2r\mu_{1,1} + \sigma_Y^2 \geq 0 \tag{2.96}$$

因此根的判别式非正，得到 $\mu_{1,1}^2 \leq \sigma_X^2\sigma_Y^2$。 □

定义 2.25. 如果随机变量 X, Y 满足 $\rho(X, Y) = 0$，则称它们是不相关的。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 2.34. 已知 $\theta \sim \mathrm{U}[-\pi, \pi]$，往证 $X = \sin\theta, Y = \cos\theta$ 是不相关的。

$$\left.\begin{array}{r}\mathsf{E}(X) = 0 \\ \mathsf{E}(XY) = 0\end{array}\right\} \Rightarrow \mathsf{Cov}(X, Y) = 0 \Rightarrow \rho(X, Y) = 0$$

然而，由 $X^2 + Y^2 = 1$ 可推出 X, Y 不独立，这说明“不相关”比“独立”要弱些。

练习 2.7. 令随机变量 U, V 具有相同的期望和方差，请验证：$X = U + V$ 和 $Y = U - V$ 不相关。

定理 2.27 (线性相关的判定)**.** $\rho^2(X, Y) = 1$ 当且仅当 $\exists (a, b) \in \mathbb{R}^2$ 使得 $\mathsf{P}(Y = aX + b) = 1$。

证明. 往证“⇐”：由式 (2.67)，$\mathsf{E}(Y) = \mathsf{P}(Y = aX + b)\mathsf{E}(Y|Y = aX + b) + \mathsf{P}(Y \neq aX + b)\mathsf{E}(Y|Y \neq aX + b) = a\mathsf{E}(X) + b$，于是

$$\sigma_Y^2 = \mathsf{E}(Y - \mathsf{E}Y)^2 = \mathsf{E}[aX - a\mathsf{E}(X)]^2 = a^2\sigma_X^2$$
$$\mu_{1,1} = \mathsf{E}[(X - \mathsf{E}X)(aX - a\mathsf{E}X)] = a\sigma_X^2$$

由定义 2.24 得出结论 $\rho^2(X,Y)=1$。

往证“$\Rightarrow$”：由 $\mu_{1,1}^2-\sigma_X^2\sigma_Y^2=0$ 和式 (2.96) 以及事实 $\sigma_X\sigma_Y>0$ 得知，$\exists!r_0\neq 0$ 使得 $\mathsf{E}[r_0(X-\mathsf{E}X)+(Y-\mathsf{E}Y)]^2=0$。由性质 2.8 可得 $\mathsf{P}\{r_0(X-\mathsf{E}X)+(Y-\mathsf{E}Y)=0\}=1$，稍加整理便得证。 □

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 $\rho=\pm1$ 时，X 与 Y 之间存在线性关系的概率是 100%。

例 2.35. 已知 $(X,Y)^{\mathsf{T}}\sim \mathrm{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$，计算相关系数 $\rho(X,Y)$。

解. 二元正态分布 $(X,Y)^{\mathsf{T}}\sim \mathrm{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ 的详情见例 2.13，其密度函数 $\phi(x,y|\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ 见式 (2.29)。按照定义 2.24，有 $\rho(X,Y)=[\mathsf{E}(XY)-\mu_X\mu_Y]/(\sigma_X\sigma_Y)=\rho$，这是因为有下面的积分（请读者验证）。

$$\mathsf{E}(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xy\phi(x,y|\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)dxdy=\rho\sigma_X\sigma_Y+\mu_X\mu_Y$$

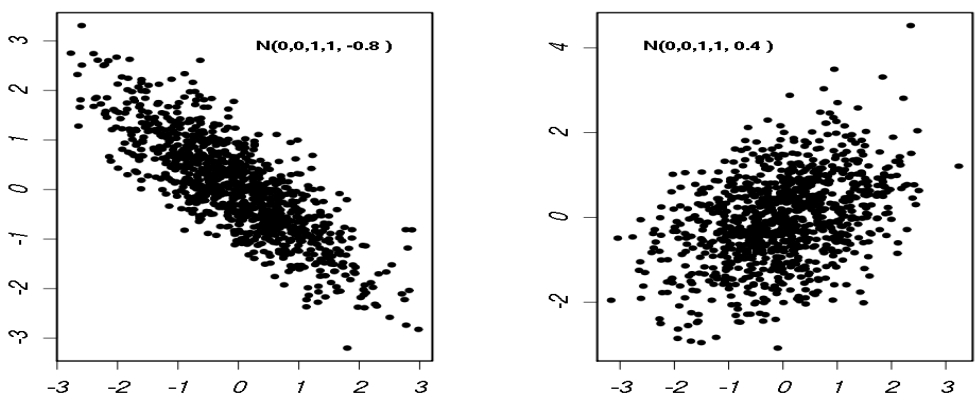


图 2.15: 二元正态分布 $\mathrm{N}(0,0,1,1,-0.8)$（左图）和 $\mathrm{N}(0,0,1,1,0.4)$（右图）分别产生 $n=1000$ 个随机数对。通过对比发现 $|\rho|$ 越接近 1，散点越集中在某直线周围，换句话说，线性关系越明显。

练习 2.8. 由 $(X,Y)^{\mathsf{T}}\sim \mathrm{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,0)$ 可推得 X,Y 相互独立。

定义 2.26 (随机向量的期望与协方差矩阵)**.** n 维随机向量 $\boldsymbol{X}=(X_1,\cdots,X_n)^{\mathsf{T}}$ 的期望定义为 $\mathsf{E}\boldsymbol{X}=(\mathsf{E}X_1,\cdots,\mathsf{E}X_n)^{\mathsf{T}}$，它的*协方差矩阵*定义为

$$\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})=\mathsf{E}[(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})^{\mathsf{T}}]=\Big(\mathsf{Cov}(X_i,X_j)\Big)_{n\times n} \tag{2.97}$$

性质 2.16. 已知 $\boldsymbol{X}$ 是一个 n 维随机向量，则对于任意列向量 $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^n$，下述结果成立。

$$\mathsf{E}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X})=\boldsymbol{a}^{\mathsf{T}}\mathsf{E}(\boldsymbol{X})=\mathsf{E}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{a})=\mathsf{E}(\boldsymbol{X}^{\mathsf{T}})\boldsymbol{a} \tag{2.98}$$

$$\mathsf{Cov}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{b}^{\mathsf{T}}\boldsymbol{X})=\boldsymbol{a}^{\mathsf{T}}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{b} \tag{2.99}$$

$$\mathsf{V}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X})=\mathsf{V}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{a})=\boldsymbol{a}^{\mathsf{T}}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{a} \tag{2.100}$$

证明. 请读者验证式 (2.98)。往证式 (2.99)：$\mathsf{Cov}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{b}^{\mathsf{T}}\boldsymbol{X})=\mathsf{E}\{\boldsymbol{a}^{\mathsf{T}}[\boldsymbol{X}-\mathsf{E}(\boldsymbol{X})]\cdot[\boldsymbol{X}^{\mathsf{T}}-\mathsf{E}(\boldsymbol{X}^{\mathsf{T}})]\boldsymbol{b}\}=\boldsymbol{a}^{\mathsf{T}}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{b}$。式 (2.100) 是式 (2.99) 的特例：$\mathsf{V}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X})=\mathsf{Cov}(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X})=\boldsymbol{a}^{\mathsf{T}}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{a}$。 □

练习 2.9. 利用式 (2.100) 将结果 (2.93) 推广为：

$$\mathsf{V}(X_1+X_2+\cdots+X_n)=\sum_{i=1}^{n}\mathsf{V}(X_i)+2\sum_{1\le j<k\le n}\mathsf{Cov}(X_j,X_k) \tag{2.101}$$

定理 2.28. 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（半正定矩阵的性质见附录 G）。

证明. 往证“$\Rightarrow$”：若方阵 $\Sigma=(\sigma_{ij})_{n\times n}$ 是随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 的协方差矩阵，则 $\sigma_{ij}=\sigma_{ji}=\mathsf{Cov}(X_i,X_j)$，并且对任意的非零向量 $\boldsymbol{x}\in\mathbb{R}^n$ 皆有 $\boldsymbol{x}^{\mathsf{T}}\Sigma\boldsymbol{x}=\boldsymbol{x}^{\mathsf{T}}\mathsf{E}[(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})^{\mathsf{T}}]\boldsymbol{x}=\mathsf{E}[\boldsymbol{x}^{\mathsf{T}}(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})]^2\ge 0$。
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 $\Sigma=AA^{\mathsf{T}}$，其中 $1\le k\le n$。令 $X_1,X_2,\cdots,X_k\overset{iid}{\sim}\mathrm{N}(0,1)$ 且 $\boldsymbol{X}=(X_1,X_2,\cdots,X_k)^{\mathsf{T}}$。定义 $\boldsymbol{Y}=A\boldsymbol{X}$，显然 $\mathsf{E}(\boldsymbol{Y})=\boldsymbol{0}$ 并且 $\boldsymbol{Y}$ 的协方差矩阵为 $\mathsf{E}(\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}})=\mathsf{E}[(A\boldsymbol{X})(A\boldsymbol{X})^{\mathsf{T}}]=A\mathsf{E}(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})A^{\mathsf{T}}=AIA^{\mathsf{T}}=AA^{\mathsf{T}}=\Sigma$。 □

2.3.4 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 $y = h(x)$，如何找出这个函数呢？我们必须依靠“最小二乘法”这件工具。1794-1795 年，伟大的天才数学家 C. F. Gauss 首次提出了*最小二乘法* (method of least squares)*，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。

本着*最小二乘原则*，即 h 的选取要使得 $\mathsf{E}[Y - h(X)]^2$ 达到最小，这里假设 $\mathsf{E}Y^2$ 和 $\mathsf{E}[h(X)]^2$ 都存在。以连续型的随机向量 $(X, Y)^\mathsf{T}$ 为例，

$$\begin{aligned}\mathsf{E}[Y - h(X)]^2 &= \iint_{\mathbb{R}^2} [y - h(x)]^2 f(x, y)dxdy \\ &= \int_{-\infty}^{\infty} f_X(x) \left\{ \int_{-\infty}^{\infty} [y - h(x)]^2 f_{Y|X}(y|x)dy \right\} dx\end{aligned}$$

根据式 (2.72)，当 $h(x) = \mathsf{E}(Y|X = x)$ 时上式花括号内的积分达到最小，进而 $\mathsf{E}[Y - h(X)]^2$ 达到最小。离散型的情形也是类似的。

定义 2.27 (回归)**.** 函数 $y = \mathsf{E}(Y|X = x)$ 称为 Y 关于 X 的*回归* (regression)，曲线 $l_{Y|X} = \{(x, \mathsf{E}(Y|X = x))\}$ 称为 Y 关于 X 的*回归曲线*。类似地，函数 $x = \mathsf{E}(X|Y = y)$ 称为 X 关于 Y 的回归，曲线 $l_{X|Y} = \{(\mathsf{E}(X|Y = y), y)\}$ 称为 X 关于 Y 的回归曲线。

两个随机变量之间不存在严格的函数关系，搞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。

*1805-1806 年，法国数学家 Adrien-Marie Legendre (1752-1833) 独立发现了最小二乘法，数学史把 Legendre 也列作最小二乘法的创立者之一。

例 2.36. 接着例 2.17 考虑二元正态分布 $(X,Y)^{\mathsf{T}} \sim \mathrm{N}(0,0,1,4,0.8)$。(1) Y 关于 X 的回归曲线是 $y = \int_{-\infty}^{\infty} yf(y|x)dy = 1.6x$；(2) X 关于 Y 的回归曲线是 $x = \int_{-\infty}^{\infty} xf(x|y)dx = 0.4y$。见图 2.36 中的左图部分。

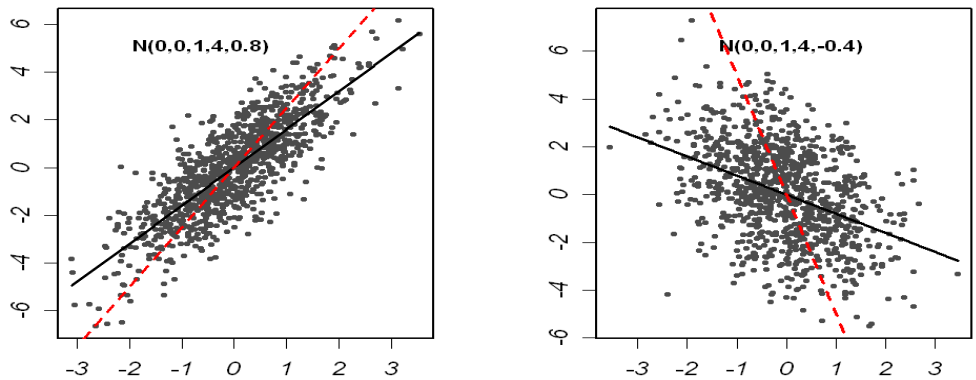


图 2.16:（左图）二元正态分布 $\mathrm{N}(0,0,1,4,0.8)$ 产生的 1000 个随机数对和两条回归曲线：实线是 $y = 1.6x$（给定 $X = x$，随机变量 Y 的平均取值是 $1.6x$），虚线是 $x = 0.4y$（给定 $Y = y$，随机变量 X 的平均取值是 $0.4y$）。（右图）二元正态分布 $\mathrm{N}(0,0,1,4,-0.4)$ 的两条回归曲线，请读者写出它们的方程。

在实践中，线性关系 $y = \alpha x + \beta$ 常被首先考虑到。使得 $L(\alpha,\beta) = \mathsf{E}(Y - \alpha X - \beta)^2 = \mathsf{E}Y^2 + \alpha^2\mathsf{E}X^2 + \beta^2 - 2\alpha\mathsf{E}(XY) + 2\alpha\beta\mathsf{E}X - 2\beta\mathsf{E}Y$ 达到最小的 α,β 即是下面线性方程的解。

$$\begin{cases} \partial L/\partial\alpha = 2\alpha\mathsf{E}X^2 - 2\mathsf{E}(XY) + 2\beta\mathsf{E}X = 0 \\ \partial L/\partial\beta = 2\beta + 2\alpha\mathsf{E}X - 2\mathsf{E}Y = 0 \end{cases}$$
$$\Rightarrow \quad \begin{cases} \alpha = \mathsf{Cov}(X,Y)/\mathsf{V}(X) \\ \beta = \mathsf{E}Y - \mathsf{E}X \cdot \mathsf{Cov}(X,Y)/\mathsf{V}(X) \end{cases} \tag{2.102}$$

定理 2.29 (回归直线方程)**.** 根据 (2.102) 所示的推导，Y 关于 X 的回归可用如下的回归直线 $l_{Y|X}$ 来刻画：

$$y - \mathsf{E}Y = \frac{\mathsf{Cov}(X,Y)}{\mathsf{V}(X)}(x - \mathsf{E}X) \tag{2.103}$$

练习 2.10. 请读者验证，X 关于 Y 的回归直线 $l_{X|Y}$ 如下：

$$x - \mathsf{E}X = \frac{\mathsf{Cov}(X,Y)}{\mathsf{V}(Y)}(y - \mathsf{E}Y) \tag{2.104}$$

约定记号：$\gamma_{Y|X} = \mathsf{Cov}(X,Y)/\mathsf{V}(X)$ 和 $\gamma_{X|Y} = \mathsf{Cov}(X,Y)/\mathsf{V}(Y)$，分别称之为 Y 在 X 上和 X 关于 Y 的回归系数。显然，

$$\rho^2(X,Y) = \gamma_{Y|X}\gamma_{X|Y} \tag{2.105}$$

练习 2.11. 回归直线 $l_{X|Y}$ 和 $l_{Y|X}$ 在什么条件下重合？答案：参考式 (2.105)，当 $\rho^2 = 1$ 时两条回归直线重合。

性质 2.17. 已知 $y = \alpha x + \beta$ 是通过 (2.102) 得到的 Y 关于 X 的回归直线，请读者验证下面的性质。

$$\mathsf{E}(Y - \alpha X - \beta) = 0 \tag{2.106}$$

$$\mathsf{E}[Y - (\alpha X + \beta)]^2 = (1 - \rho^2)\mathsf{V}(Y) \tag{2.107}$$

$$\mathsf{E}[(X - \mathsf{E}X)(Y - \alpha X - \beta)] = 0 \tag{2.108}$$

式 (2.107) 说明 ρ^2 越接近 1，随机变量 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 2.27）。而式 (2.108) 意味着 $X - \mathsf{E}X, Y - \alpha X - \beta$ 是不相关的，进而误差 $X - \mathsf{E}X$ 与误差 $Y - \alpha X - \beta$ 之间不存在线性关系。

注记 2.1. 1865 年，英国人类学家 Francis Galton (1822-1911) 受其表兄、《物种起源》的作者 Charles Robert Darwin (1809-1882) 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 X 和子代身高 Y 呈现出一定的规律性：高的父代产生的子代平均也高，但有向父代均值退化的趋势（设父代身高均值为 μ_X，某一父代的身高 $x > \mu_X$，其子代身高的均值 y ）。1886 年，Galton 发表论文《遗传结构中向中心的回归》。

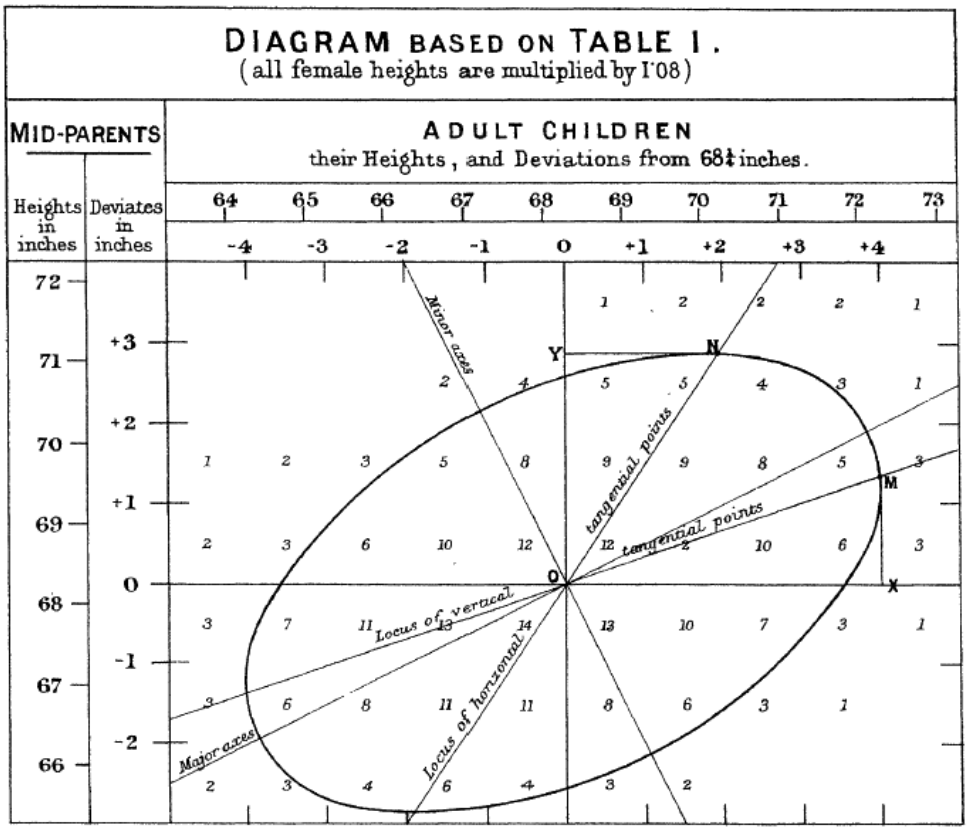


图 2.17: 因为女性的身高一般低于男性，Galton 利用男女平均身高之比把女性的身高乘以 1.08 换算成男性身高。为刻画父代的身高，Galton 定义了中亲 (min-parents) 身高 = $\frac{1}{2}$(父亲的身高 + 1.08 × 母亲的身高) 来。此图来自 K. Pearson 的文章《对相关性的历史注记》 [68]。

2.4 习题

2.1. 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。

☆ 2.2. 将 3 个球随机地放入编号为 $1,2,3,4$ 的 4 个盒子中，请给出有球的盒子的最大编号 X 的分布列。

2.3. 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，直到取得 1 个白球为止。请写出所抽取的球的个数 X 的分布列。

2.4. 对于二项分布 $X \sim \mathrm{B}(n,p)$，问 k 取何值时 $\mathsf{P}\{X=k\}$ 最大？

2.5. 设随机变量 $X \sim \mathrm{B}(2,p)$，随机变量 $Y \sim \mathrm{B}(3,p)$。若 $\mathsf{P}\{X \geq 1\} = 5/9$，求 $\mathsf{P}\{Y \geq 1\}$。

2.6. 设随机变量 X 的分布列为 $\mathsf{P}(X=k) = \lambda^k e^{-\lambda}/k!$，其中 $k = 0,1,2,\cdots$。(1) 求 X 取偶数的概率；(2) 若 $\mathsf{P}(X=2) = \mathsf{P}(X=3)$，求 X 取偶数的概率。

2.7. 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。

2.8. 设 $F_1(x)$ 与 $F_2(x)$ 都是分布函数，a 和 b 是非负常数且 $a+b=1$。试证明：函数 $F(x) = aF_1(x) + bF_2(x)$ 也是分布函数。

2.9. 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。

2.10. 向区间 $[0,1]$ 上任意投掷一点，以 X 表示该点的坐标，并设该点落在 $[0,1]$ 中任意小区间内的概率与这个小区间的长度成正比。试求 X 的分布。

2.11. 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) $\mathsf{P}\{-1 < X < 2\}$。

2.12. 设离散型随机变量的概率分布为 $\mathsf{P}\{X = k\} = 1/(ck!)$，其中 $k = 1, 2, \cdots$，求常数 c。

2.13. 设连续型随机变量 X 的分布函数为 $F(x) = \begin{cases} a + be^{-x^2/2} & \text{当 } x > 0 \\ 0 & \text{当 } x \le 0 \end{cases}$
试求：(1) 常数 a, b；(2) X 的概率密度 $f(x)$；(3) $\mathsf{P}(1 < X < 2)$。

2.14. 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 $Y = 2X + 1$ 的概率密度函数 $f_Y(y)$。

2.15. 设 $X \sim \mathrm{N}(0, 1)$，求下面随机变量函数的概率密度函数：(1) $Y = e^X$；(2) $Z = \sqrt{|X|}$。

2.16. 若随机变量 X 服从正态分布，其概率密度为 $f(x) = k\exp\{-x^2/2 - x\}$，问 k 为多少？

2.17. 若 $c > 0$，试证明 $f(x) = \begin{cases} c^{-1}xe^{-x^2/(2c)} & \text{当 } x > 0 \\ 0 & \text{当 } x \le 0 \end{cases}$ 是密度函数。

2.18. 已知随机变量 X 的分布函数为 $F_X(x) = \begin{cases} 1 - e^{-2x} & \text{当 } x > 0 \\ 0 & \text{当 } x \le 0 \end{cases}$
试证明：$Y = 1 - e^{-2X} \sim \mathrm{U}(0, 1)$。

☆ 2.19. 假设随机变量 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{U}[0, a]$，其中 $a > 0$，试求 $Y = \max(X_1, X_2, \cdots, X_n)$ 的概率密度函数。

2.20. 如果随机变量 $X \sim \mathrm{U}(-r, r)$，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. 设二维随机向量 $(X, Y)^\mathsf{T}$ 在矩形区域 $D = \{(x, y) : 1 \le x \le 3, 1 \le y \le 3\}$ 上服从均匀分布，求 $Z = |X - Y|$ 的概率密度。

2.22. 已知 $(X,Y)^{\mathsf{T}}$ 的分布列为

X \ Y	-1	1	2
-1	$1/10$	$2/10$	$3/10$
2	$-2/10$	$1/10$	$1/10$

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) $Z=\max(X,Y)$。

2.23. 设随机变量 X 与 Y 都服从正态分布 $\mathrm{N}(0,\sigma^2)$，且 $\mathsf{P}\{X\le 0,Y\ge 0\}=1/3$，求 $\mathsf{P}\{X>0,Y<0\}$。

2.24. 设二维随机变量 $(X,Y)^{\mathsf{T}}$ 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) $\mathsf{P}\{X\ge 0,Y\ge 0\}$。

2.25. 设随机向量 $(X,Y)^{\mathsf{T}}$ 的概率密度为 $f(x,y)=\begin{cases} e^{-x-y}/k & \text{当 } x>0,y>0 \\ 0 & \text{其他} \end{cases}$

试求：(1) 常数 k；(2) $\mathsf{P}\{X<2,Y<2\}$。

2.26. 设随机向量 $(X,Y)^{\mathsf{T}}$ 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 是否相互独立；(2) 求 $Z=X+Y$ 的分布函数 $F_Z(z)$；(3) 求 $\mathsf{P}\{Z>3\}$。

2.27. 设随机变量 $X_1,X_2,\cdots,X_n$ 独立同分布且取值总为正数，试证明：$\mathsf{E}(\sum_{j=1}^{k}X_j/\sum_{j=1}^{n}X_j)=k/n$，其中 $k=1,2,\cdots,n$。

☆ 2.28. 设随机变量 X 只取 [0,1] 上的值，试证明不等式 $\mathsf{V}(X)\le 1/4$ 并指出何时取等号。

2.29. 设连续型的随机变量 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：$\mathsf{P}\{X\ge a\}\le e^{-\lambda a}\mathsf{E}(Y)$，其中 a 为任意实数。

2.30. 已知随机变量 $X \sim \mathrm{N}(\mu, \sigma^2)$，试证明 $\mathsf{E}(|X-\mu|) = \sigma\sqrt{2/\pi}$。

☆ 2.31. 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 $p_1 = p_2 = \cdots = p_n = 1/n$ 时，熵 $\mathsf{H}(X)$ 最大。

2.32. 试证明书中的性质 2.8。

✯ 2.33. 试用其他方法证明例 2.32 中的第二个不等式。

2.34. 若 $X \sim \mathrm{N}(1,2)$ 和 $Y \sim \mathrm{N}(0,1)$ 独立，求 $Z = 2X - Y + 3$ 的分布。

2.35. 已知 $X, Y \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$，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？

☆ 2.36. 设随机变量 X 与 Y 相互独立，密度函数分别为 $f_X(x) = f(x)$ 和 $f_Y(y) = f(y)$，其中 $f(\cdot)$ 如下定义

$$f(t) = \begin{cases} e^{-t} & \text{当 } t > 0 \\ 0 & \text{当 } t \le 0 \end{cases}$$

试证明：随机变量 $U = X + Y$ 与 $V = X/Y$ 也是相互独立的。

2.37. 设随机变量 X 的密度函数为 $f(x) = x^m e^{-x}/m!$，其中 $x \ge 0$。试证明：$\mathsf{P}\{0 < X < 2(m+1)\} \ge m/(m+1)$。

☆ 2.38. 已知随机变量 $\{X_n : n = 1, 2, \cdots\}$ 相互独立，且 X_n 的概率函数为 $\mathsf{P}(X_n = \pm\sqrt{n+1}) = 1/(n+1), \mathsf{P}(X_n = 0) = 1 - 2/(n+1)$。试用 Chebyshev 不等式证明：$\forall \epsilon > 0$ 有 $\lim\limits_{n\to\infty} \mathsf{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n} X_j\right| < \epsilon\right) = 1$。

☆ 2.39. 求证：随机变量 X 的中位数 $\mathsf{M}(X)$ 满足 $|\mathsf{M}(X) - \mathsf{E}(X)| \le \sqrt{2\mathsf{V}(X)}$。

✯ 2.40. 利用 Hölder 不等式证明推论 2.1 中的两个不等式。

2.41. 已知随机变量 X 的 k 阶绝对矩存在，试证明：$\forall \epsilon > 0$，不等式 $\mathsf{P}\{|X| \ge \epsilon\} \le \mathsf{E}|X|^k/\epsilon^k$ 成立。

2.42. 设 $(X,Y)^{\mathsf{T}}$ 在圆盘 $x^2+y^2\le r^2$ 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？

2.43. 设随机向量 $(X,Y)^{\mathsf{T}}$ 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 $\gamma_{Y|X},\gamma_{X|Y}$。

2.44. 设随机变量 X 和 Y 的相关系数为 0.9，若 $Z=X-0.4$，问 Y 与 Z 的相关系数为多少？

2.45. 设随机事件 A 和 B 发生的概率分别为 $\mathsf{P}(A)>0,\mathsf{P}(B)>0$，定义随机变量 X 和 Y 如下：

$$X=\begin{cases} 1 & \text{若事件 } A \text{ 发生} \\ 0 & \text{若事件 } A \text{ 不发生} \end{cases} \text{ 和 } Y=\begin{cases} 1 & \text{若事件 } B \text{ 发生} \\ 0 & \text{若事件 } B \text{ 不发生} \end{cases}$$

试证明：$\rho(X,Y)=0$ 当且仅当 A 与 B 相互独立。

✯ 2.46. 已知随机变量 X,Y 的期望和方差分别为 $\mathsf{E}(X)=\mathsf{E}(Y)=0,\mathsf{V}(X)=\mathsf{V}(Y)=1$ 且 $\mathsf{Cov}(X,Y)=\rho$。试证明：$\mathsf{E}[\max(X^2,Y^2)]\le 1+\sqrt{1-\rho^2}$。

2.47. 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。

★ 2.48. 设连续型随机变量 X 的概率密度函数为 $f_\theta(x)$，其中 $\theta\in\Theta$ 为未知参数，则

$$\mathsf{E}_\theta\left[\frac{\partial \ln f_\theta(X)}{\partial\theta}\right]=0 \text{ 且 } \mathsf{E}_\theta\left[\frac{\partial \ln f_\theta(X)}{\partial\theta}\right]^2=-\mathsf{E}_\theta\left[\frac{\partial^2 \ln f_\theta(X)}{\partial\theta^2}\right] \quad (2.109)$$

第三章

特征函数

为了方便推理或运算，我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。在一定条件下，Fourier 变换*把函数 $g(x)$ 变为函数 $\varphi_{a,b}(t)$，具体定义为

$$\mathcal{F}(g) = a\int_{-\infty}^{+\infty} e^{bitx} g(x)dx = \varphi_{a,b}(t) \tag{3.1}$$

其中，a, b 是两个非零常数，$i = \sqrt{-1}$ 是虚数单位。若 $\varphi_{a,b}(t)$ 在 $\mathbb{R}$ 上可积，则几乎处处成立下面的关系，称为 $\varphi_{a,b}(t)$ 的 Fourier 逆变换：

$$g(x) = \frac{b}{2a\pi}\int_{-\infty}^{\infty} e^{-bitx}\varphi_{a,b}(t)dt \tag{3.2}$$

为了研究的方便，我们约定以后所考虑的 Fourier 变换 (3.1) 都满足 $a = b = 1$，即 $\mathcal{F}(g) = \int_{-\infty}^{+\infty} e^{itx} g(x)dx$，这样的 Fourier 变换有一个神奇之处就是下面的卷积定理。

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

*Fourier 变换是复 Fourier 级数的一般化，有关这种积分变换的性质详见华罗庚的《高等数学引论》 [1] 第十九章，或 W. Rudin 的《泛函分析》 [81] 第七章。

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 $f_i(x), i = 1, 2, \cdots, n$，则随机变量 $X_1 + X_2 + \cdots + X_n$ 的密度函数为 $(f_1 * f_2 * \cdots * f_n)(x)$。但是，做多次卷积是相当麻烦的。正是出于简化计算的目的，我们需要利用卷积定理并据此提出特征函数的概念。

定义 3.1 (随机变量的特征函数)**.** 已知随机变量 X 的分布函数为 $F(x)$，具有分布列 $X \sim p_1\langle x_1\rangle + p_2\langle x_2\rangle + \cdots$，或者密度函数 $f(x)$。定义随机变量 X 的特征函数 (characteristic function) 如下，

$$\varphi(t) = \mathsf{E}\left(e^{itX}\right) = \left\{ \begin{array}{ll} \displaystyle\sum_{n=1}^{\infty} e^{itx_n} p_n & \text{离散型} \\ \displaystyle\int_{-\infty}^{+\infty} e^{itx} f(x)dx & \text{连续型} \end{array} \right\} = \int_{-\infty}^{+\infty} e^{itx} dF(x) \qquad (3.3)$$

由于 $e^{itx} = \cos(tx) + i\sin(tx)$ 的模长等于 1，所以 $\sum_{n=1}^{\infty} e^{itx_n} p_n$ 绝对收敛或 $e^{itx} f(x)$ 在 $\mathbb{R}$ 上绝对可积，因此特征函数 $\varphi(t)$ 总是存在的*。

例 3.1. 两点分布 $X \sim p\langle a\rangle + (1-p)\langle b\rangle$ 的特征函数为 $pe^{ita} + (1-p)e^{itb}$。

例 3.2. 均匀分布 $X \sim \mathrm{U}[a, b]$ 的特征函数为 $(e^{itb} - e^{ita})/[it(b-a)]$。

例 3.3. 标准正态分布 $X \sim \mathrm{N}(\mu, \sigma^2)$ 的特征函数为

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left\{ -\frac{(x - \mu - it\sigma^2)^2}{2\sigma^2} + it\mu - \frac{\sigma^2 t^2}{2} \right\} dx = \exp\left\{ it\mu - \frac{\sigma^2 t^2}{2} \right\}$$

练习 3.1. 分布 $X \sim \mathrm{Cauchy}(\mu, \lambda)$ 的特征函数为 $\exp\{it\mu - \lambda|t|\}$。

定义 3.2 (随机向量的特征函数)**.** 已知 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathsf{T}}$ 为一个 n 维随机向量，我们称下面的函数 $\varphi : \mathbb{R}^n \to \mathbb{C}$ 为随机向量 $\boldsymbol{X}$ 的特征函数，

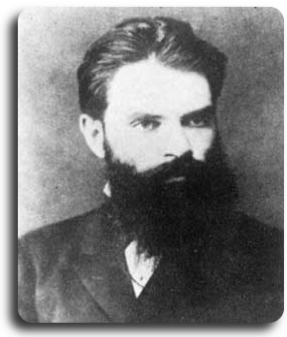
$$\varphi(\boldsymbol{t}) = \mathsf{E}\left\{ e^{i\boldsymbol{t}^{\mathsf{T}}\boldsymbol{X}} \right\} = \mathsf{E}\left\{ e^{i(t_1X_1 + t_2X_2 + \cdots + t_nX_n)} \right\} \qquad (3.4)$$

*与特征函数类似的工具还有矩母函数 (moment-generating function) $\mathsf{E}(e^{tX})$，但矩母函数并不像特征函数那样总是存在的 [78]，本书不作深入介绍。

其中，系数 $\boldsymbol{t} = (t_1, t_2, \cdots, t_n)^{\mathsf{T}} \in \mathbb{R}^n$。

例 3.4. 二元正态分布 $(X, Y)^{\mathsf{T}} \sim \mathrm{N}(\mu_1, \mu_2, \sigma^2, \sigma^2, \rho)$ 的特征函数为

$$\begin{aligned}\varphi(s,t) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(isx + ity)\phi(x, y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)dxdy \\ &= \exp\left\{i\mu_1 s + i\mu_2 t - \frac{1}{2}\left(\sigma_1^2 s^2 + 2\sigma_1\sigma_2\rho st + \sigma_2^2 t^2\right)\right\}\end{aligned}$$



1900-1901 年杰出的俄国数学家 A. M. Lyapunov (1857-1918，右图) 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) 计算各阶原点矩；(2) 揭示分布函数与特征函数之间关系的反演 (inversion) 公式*和 Lévy 连续性定理等。

*国内的文献也有将之译为“逆转公式”、“反转公式”等。

3.1 特征函数的基本性质

对任意随机变量 X 而言，其特征函数 $\varphi_X(t)$ 总是存在的。特征函数是概率函数的离散 Fourier 变换或密度函数的 Fourier 变换之结果。积分变换 (3.3) 也称为分布函数 $F(x)$ 的 Fourier-Stieltjes 变换。

定理 3.2. 特征函数 $\varphi(t)$ 在 $\mathbb{R}$ 上是一致连续的*且 $\varphi(0) = 1, |\varphi(t)| \leq 1$。

证明. 由特征函数的定义，$\varphi(0) = \int_{-\infty}^{\infty} 1 \cdot dF(x) = 1$ 且

$$|\varphi(t)| = \left| \int_{-\infty}^{\infty} e^{itx} \cdot dF(x) \right| \leq \int_{-\infty}^{\infty} \left| e^{itx} \right| dF(x) = 1$$

下面往证一致连续性：

$$|\varphi(t+h) - \varphi(t)| = \left| \int_{-\infty}^{\infty} e^{itx}(e^{ihx} - 1) dF(x) \right| \leq \int_{-\infty}^{\infty} \left| e^{ihx} - 1 \right| dF(x)$$

$\forall \epsilon > 0$，选足够大的 $a > 0$ 使得 $\int_{|x| \geq a} dF(x) < \epsilon/4$；同时选择足够小的 h 使得对所有 $x \in [-a, a]$ 皆有 $|e^{ihx} - 1| = 2|\sin \frac{hx}{2}| < \epsilon/2$。于是，

$$|\varphi(t+h) - \varphi(t)| \leq \int_{-a}^{a} \left| e^{ihx} - 1 \right| dF(x) + 2 \int_{|x| \geq a} dF(x) < \epsilon \qquad \square$$

练习 3.2. 若 $\varphi_X(t)$ 是随机变量 X 的特征函数，试证明：

- ❑ $\varphi_X(-t) = \overline{\varphi_X(t)}$。
- ❑ 随机变量 $Y = aX + b$ 的特征函数为 $\varphi_Y(t) = e^{itb}\varphi_X(at)$。

练习 3.3. 二维随机向量 $(X, Y)^\mathsf{T}$ 的特征函数 $\varphi(s, t)$ 具有性质：$\varphi(0, 0) = 1, |\varphi(s, t)| \leq 1$，且 $\varphi(-s, -t) = \overline{\varphi(s, t)}$。还有 $\varphi(s, 0) = \varphi_X(s), \varphi(0, t) = \varphi_Y(t)$。

*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 [27]。

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。

- 由随机向量 $(X,Y)^{\mathsf{T}}$ 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 $\varphi_Y(t)=\varphi(0,t)$。
- 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 $\sum_{n=1}^{\infty}\alpha_n=1$。留作练习。
- 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。
- 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。
- Bochner 准则：函数 $\varphi:\mathbb{R}^n\to\mathbb{C}$ 是某随机向量的特征函数当且仅当 φ 半正定（见附录 G），且在原点连续并取值为 1。
- Khinchin 准则：绝对连续的复值函数 φ 满足 $\varphi(0)=1$，它是特征函数当且仅当它能够表示为

$$\varphi(t)=\int_{-\infty}^{+\infty}g(t+\theta)\overline{g(\theta)}d\theta \tag{3.5}$$

- 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。

  ❶ $\varphi(0)=1$ 且 $\varphi(\infty)=0$。

  ❷ $\varphi(t)$ 是偶函数，且当 $t>0$ 时为凸函数。

本节内容　“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”，但反之不成立，第一小节给出了证明和反例。第二小节展示了如何用特征函数计算各阶原点矩。

学习目标　(1) 掌握特征函数的定义和性质，特别是求解独立随机变量之和的特征函数；(2) 会利用特征函数计算各阶原点矩。

3.1.1 特征函数与独立性

定理 3.4. 已知随机变量 $X_1, X_2, \cdots, X_n$ 是独立的，且特征函数分别为 $\varphi_1(t), \varphi_2(t), \cdots, \varphi_n(t)$，则 $Y = X_1 + X_2 + \cdots + X_n$ 的特征函数为

$$\varphi_Y(t) = \prod_{i=1}^{n} \varphi_i(t) \tag{3.6}$$

证明. 由 $\varphi_Y(t) = \mathsf{E}\left[e^{it(X_1+X_2+\cdots+X_n)}\right] = \prod_{i=1}^{n} \mathsf{E}\left[e^{itX_i}\right] = \prod_{i=1}^{n} \varphi_i(t)$，得证。 □

例 3.5. 已知随机变量 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} p\langle 1\rangle + (1-p)\langle 0\rangle$，由定理 3.4 和例 3.1 知，随机变量 $Y = X_1 + X_2 + \cdots + X_n$ 的特征函数为 $\varphi_Y(t) = [1 + p(e^{it} - 1)]^n$。我们称 Y 服从二项分布 (binomial distribution)，记作 $Y \sim \mathrm{B}(n, p)$。

例 3.6. 已知两个独立的随机变量 X_1, X_2 的概率函数分别为 $\mathsf{P}(X_1 = k) = \lambda_1^k e^{-\lambda_1}/k!$ 和 $\mathsf{P}(X_2 = k) = \lambda_2^k e^{-\lambda_2}/k!$，计算 $Y = X_1 - X_2$ 的特征函数。

解. X_1 的特征函数为 $\exp\{\lambda_1(e^{it} - 1)\}$，$-X_2$ 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 $\exp\{\lambda_1 e^{it} + \lambda_2 e^{-it} - \lambda_1 - \lambda_2\}$。

人们很自然地要问定理 3.4 的逆命题是否成立，即如果 $Y = X_1 + X_2 + \cdots + X_n$ 的特征函数为 $\varphi(t) = \prod_{i=1}^{n} \varphi_i(t)$，能否判定随机变量 $X_1, X_2, \cdots, X_n$ 独立？答案是“不能”，构造两个反例如下。

例 3.7. 已知随机变量 X 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 $\varphi_Y(t) = \exp(-c|t|)$ 并且 $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$。

例 3.8. 随机向量 $(X, Y)^{\mathsf{T}}$ 的密度函数为

$$f(x, y) = \begin{cases} \frac{1}{4}\left[1 + xy(x^2 - y^2)\right] & \text{如果 } |x| \le 1 \text{ 且 } |y| \le 1 \\ 0 & \text{否则} \end{cases} \tag{3.7}$$

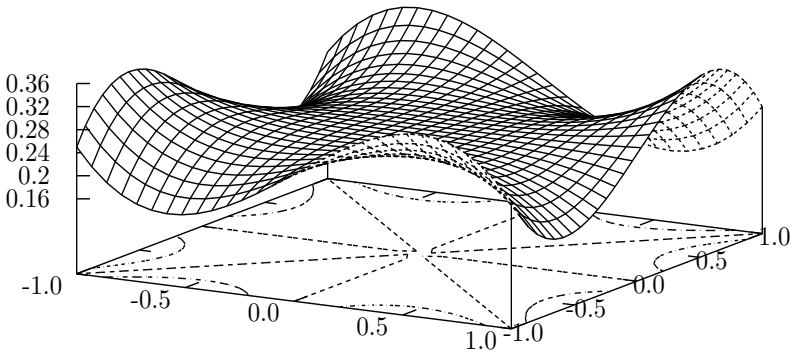


图 3.1: 由式 (3.7) 定义的密度函数曲面。求得 X 和 Y 的边缘分布为 $f_X(x) = f_Y(y) = 1/2$，显然 $f(x,y) \neq f_X(x)f_Y(y)$，这意味着随机变量 X, Y 不独立。

随机变量 $Z = X + Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z-x)dx = \begin{cases} \int_{-1}^{z+1} f(x,z-x)dx = \frac{1}{4}(2+z) & \text{当 } -2 \le z \le 0 \\ \int_{z-1}^{1} f(x,z-x)dx = \frac{1}{4}(2-z) & \text{当 } 0 < z \le 2 \\ 0 & \text{当 } |z| > 2 \end{cases}$$

随机变量 X, Y, Z 的特征函数为

$$\varphi_X(t) = \frac{1}{2}\int_{-1}^{1} e^{itx}dx = \frac{\sin t}{t}，\text{同理，} \varphi_Y(t) = \frac{\sin t}{t}$$

$$\varphi_Z(t) = \frac{1}{4}\int_{-2}^{0}(2+z)e^{itz}dz + \frac{1}{4}\int_{0}^{2}(2-z)e^{itz}dz = \frac{\sin^2 t}{t^2}$$

显然，$\varphi_Z(t) = \varphi_X(t)\varphi_Y(t)$，但 X, Y 不独立。

性质 3.1. 若随机变量 X, Y 相互独立，则随机向量 $(X,Y)^\mathsf{T}$ 的特征函数 $\varphi(s,t) = \varphi_X(s)\varphi_Y(t)$。

证明. $\varphi(s,t) = \mathsf{E}(e^{isX+itY}) = \mathsf{E}(e^{isX})\mathsf{E}(e^{itY}) = \varphi_X(s)\varphi_Y(t)$，得证。 □

问题 3.1. 在例 3.4 中，若 $\varphi(s,t) = \varphi_X(s)\varphi_Y(t)$，则 X, Y 相互独立。问一般情况下，性质 3.1 的逆命题成立吗？

答案：成立，见定理 3.12。证明要用到 §3.2.2 将介绍的唯一性定理。

3.1.2 利用特征函数计算原点矩

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 $k \le n$ 时

$$\frac{d^k}{dt^k}\varphi(t) = \left\{ \begin{array}{ll} \displaystyle\sum_n i^k x_n^k e^{itx_n} p_n & \text{离散情形} \\ \\ \displaystyle\int_{-\infty}^{+\infty} i^k x^k e^{itx} f(x)dx & \text{连续情形} \end{array} \right\} = \mathsf{E}\left(i^k X^k e^{itX}\right) \tag{3.8}$$

特别地，X 的 k 阶矩 $m_k = \dfrac{\varphi^{(k)}(0)}{i^k}$ (3.9)

证明. 因为 $\left|\int_{-\infty}^{\infty} x^k e^{itx} dF(x)\right| \le \int_{-\infty}^{\infty} |x|^k dF(x) < \infty$，所以 $\varphi^{(k)}(t)$ 存在。由 $\varphi^{(k)}(0) = i^k \mathsf{E}(X^k) = i^k m_k$ 可证得式 (3.9)。 □

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，式 (3.9) 必须在保证 k 阶矩 m_k 存在的前提下才可以使用。

例 3.9. 设随机变量 X 的密度函数为 $f(x) = \left\{ \begin{array}{ll} 0 & \text{当 } |x| \le 2 \\ \dfrac{c}{x^2 \ln|x|} & \text{当 } |x| > 2 \end{array} \right.$ 其中常数 c 是使得 $\int_{-\infty}^{\infty} f(x)dx = 1$ 的归一因子。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 $s \to \infty$ 而趋向无穷，即 X 的期望不存在。

例 3.10. 对于标准正态分布 $X \sim \mathrm{N}(0,1)$，其特征函数为 $\varphi(t) = \exp\{-t^2/2\}$。由式 (3.9) 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 的各阶矩都存在，则特征函数 $\varphi_X(t)$ 可表示为

$$\varphi_X(t) = \varphi_X(0) + \sum_{k=1}^{\infty} \frac{\varphi_X^{(k)}(0)}{k!} t^k = 1 + \sum_{k=1}^{\infty} \frac{m_k}{k!} (it)^k \tag{3.10}$$

定义 3.3. 若随机变量 X 的各阶矩都存在，令 $z=\sum_{k=1}^{\infty} m_k(it)^k/k!$，定义

$$\psi_X(t)=\ln\varphi_X(t)=\ln(1+z)=\frac{z}{1}-\frac{z^2}{2}+\frac{z^3}{3}-\cdots=\sum_{k=1}^{\infty}\frac{\varkappa_k}{k!}(it)^k \tag{3.11}$$

其中，$\varkappa_k=i^{-k}\psi_X^{(k)}(0)$ 被称作 k 阶半不变量。请读者验证 $\mathsf{E}(X)=-i\psi_X'(0)=\varkappa_1, V(X)=-\psi_X''(0)=\varkappa_2$。

定理 3.6. 接着定理 3.4，随机变量 $Y=X_1+X_2+\cdots+X_n$ 的 k 阶半不变量等于 $X_1,X_2,\cdots,X_n$ 各自的 k 阶半不变量之和。

例 3.11. 如果随机变量 X,Y 具有关系 $Y=X+b(b\neq 0)$，一般地 X 的 k 阶矩不等于 Y 的 k 阶矩。由 $\ln\varphi_Y(t)=bit+\ln\varphi_X(t)$ 易见，$\varkappa_2,\varkappa_3,\cdots$ 却是不变量。

✂ **例 3.12.** 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 $e^z=1+z+z^2/2!+\cdots+z^n/n!+\cdots$ 寻找矩与半不变量之间的关系。

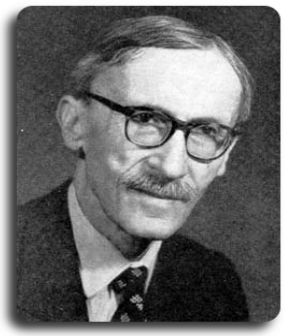
$$\begin{aligned}\varphi(t)&=1+\sum_{k=1}^{\infty}\frac{m_k}{k!}(it)^k=\exp\left\{\sum_{k=1}^{\infty}\frac{\varkappa_k}{k!}(it)^k\right\}\\&=1+\sum_{k=1}^{\infty}\frac{\varkappa_k}{k!}(it)^k+\frac{1}{2!}\left[\sum_{k=1}^{\infty}\frac{\varkappa_k}{k!}(it)^k\right]^2+\frac{1}{3!}\left[\sum_{k=1}^{\infty}\frac{\varkappa_k}{k!}(it)^k\right]^3+\cdots\end{aligned}$$

通过对比系数，我们得到

$$\begin{cases}m_1=\varkappa_1\\m_2=\varkappa_2+\varkappa_1^2\\m_3=\varkappa_3+3\varkappa_1\varkappa_2+\varkappa_1^3\\\cdots\end{cases}\quad\text{或者}\quad\begin{cases}\varkappa_1=m_1\\\varkappa_2=m_2-m_1^2\\\varkappa_3=m_3-3m_1m_2+2m_1^3\\\cdots\end{cases}$$

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 P. Lévy (1886-1971) 继 A. M. Lyapunov 之后于 1919-1925 年系统地建立了特征函数理论，特别是反演公式和连续性定理揭示了分布函数与特征函数之间内在联系。此后，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（详见第五章）和独立增量过程的研究。在介绍 Lévy 反演公式和连续性定理这两个结果之前，我们先定义随机变量序列及其一种收敛方式——依分布收敛*。



定义 3.4 (随机变量序列)**.** 若 $X_1, X_2, \cdots, X_n, \cdots$ 是定义在同一概率空间 $(\Omega, \mathscr{S}, \mathsf{P})$ 上的随机变量，则称 $X_1, X_2, \cdots, X_n, \cdots$ 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 近似地为怎样的分布。

定义 3.5. 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 $n = 2, 3, \cdots$ 皆有 $X_1, X_2, \cdots, X_n$ 相互独立。

例 3.13. 随机试验中事件 A 发生的概率 $\mathsf{P}(A)$ 的直观含义是：在相同条件下的多次重复试验中 A 发生的频率之稳定值（见例 1.34）。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，

$$X_j = \begin{cases} 1 & \text{若第 } j \text{ 次试验 } A \text{ 发生} \\ 0 & \text{若第 } j \text{ 次试验 } A \text{ 不发生} \end{cases}$$

则随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，事件 A 在 n 次重复试验中发生的

*第五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。

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 $\langle \mathsf{P}(A)\rangle$。

定义 3.6 (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 $\lim\limits_{n\to\infty} F_n(x) = F_X(x)$，则称分布函数序列 $\{F_n(x)\}$ *弱收敛*于分布函数 $F_X(x)$，或称 $\{X_n\}$ *依分布收敛* (converge in law/distribution) 于随机变量 X，记作 $X_n \xrightarrow{L} X$。

定理 3.7 (Helly, 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 $f(x)$ 皆有

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f(x)dF_n(x) = \int_{-\infty}^{\infty} f(x)dF(x) \tag{3.12}$$

证明. 见 Gnedenko 的《概率论教程》 [40] 第七章。此定理也称为第二 Helly 定理。 □

性质 3.2. 依分布收敛还有一个等价的定义：$X_n \xrightarrow{L} X$ 当且仅当对于任意有界连续函数 g 皆有 $\mathsf{E}g(X_n) \to \mathsf{E}g(X)$。

证明. 往证“$\Rightarrow$”：利用定理 3.7。“$\Leftarrow$”的证明见 W. Feller 的《概率论及其应用》下卷第八章第一节。 □

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。

学习目标　掌握 Lévy 反演公式和 Lévy 连续性定理。

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 $\mathsf{P}(x_1 < X \le x_2)$ 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。

定理 3.8 (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

$$F(x_0+h) - F(x_0-h) = \lim_{s\to\infty} \frac{1}{\pi} \int_{-s}^{s} \frac{\sin(ht)}{t} e^{-itx_0} \varphi(t) dt \tag{3.13}$$

证明. 在求解式 (3.13) 右边的极限式之前，先对它做一些整理。

$$\begin{aligned} J_s &= \frac{1}{\pi} \int_{-s}^{s} \frac{\sin(ht)}{t} e^{-itx_0} \varphi(t) dt = \frac{1}{\pi} \int_{-s}^{s} \left\{ \int_{-\infty}^{+\infty} \frac{\sin(ht)}{t} e^{-itx_0} e^{itx} dF(x) \right\} dt \\ &= \frac{1}{\pi} \int_{-\infty}^{+\infty} \left\{ \int_{-s}^{s} \frac{\sin(ht)}{t} e^{itx-itx_0} dF(x) \right\} dt \\ &= \int_{-\infty}^{+\infty} \left\{ \frac{2}{\pi} \int_{0}^{s} \frac{\sin(ht)}{t} \cos[(x-x_0)t] dt \right\} dF(x) = \int_{-\infty}^{+\infty} g_s(x) dF(x) \end{aligned}$$

其中，$g_s(x)$ 代表花括号内的算式。上面的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 $\int_{-\infty}^{\infty} \left| e^{itx-itx_0} \sin(ht)/t \right| dF(x) \le h$。对 $g_s(x)$ 进一步做整理得

$$g_s(x) = \frac{1}{\pi} \int_{0}^{s} \left\{ \frac{\sin[(x-x_0+h)t]}{t} - \frac{\sin[(x-x_0-h)t]}{t} \right\} dt$$

由数学分析知，$\int_0^s \frac{\sin x}{x} dx$ 对所有 $s > 0$ 是有界的，且 $\lim_{s\to\infty} \int_0^s \frac{\sin x}{x} dx = \pi/2$（读者也可以用 Maxima 验证之），于是 $|g_s(x)|$ 有界并且

$$\lim_{s\to\infty} \frac{1}{\pi} \int_{0}^{s} \frac{\sin(\beta t)}{t} dt = \begin{cases} \frac{1}{2} & \text{当 } \beta > 0 \\ 0 & \text{当 } \beta = 0 \\ -\frac{1}{2} & \text{当 } \beta < 0 \end{cases} \Rightarrow \lim_{s\to\infty} g_s(x) = \begin{cases} 0 & \text{当 } |x-x_0| > h \\ \frac{1}{2} & \text{当 } |x-x_0| = h \\ 1 & \text{当 } |x-x_0| < h \end{cases}$$

积分 $\int_{-\infty}^{+\infty} g_s(x)dF(x)$ 对于参数 s 一致收敛，因此式 (3.13) 右边的求极限可与求积分可交换次序，于是式 (3.13) 右边为

$$\lim_{s\to\infty} J_s = \int_{-\infty}^{+\infty} \lim_{s\to\infty} g_s(x)dF(x) = \int_{x_0-h}^{x_0+h} dF(x) = F(x_0+h) - F(x_0-h) \quad \square$$

推论 3.2. 如果 x_1, x_2 是分布函数 $F(x)$ 的两个连续点，不妨设 $x_1 < x_2$，反演式 (3.13) 有下面的等价形式：

$$F(x_2) - F(x_1) = \lim_{s\to\infty} \frac{1}{2\pi} \int_{-s}^{s} \frac{e^{-itx_1} - e^{-itx_2}}{it} \varphi(t)dt \tag{3.14}$$

$$F(x) = \lim_{y\to-\infty} \lim_{s\to\infty} \frac{1}{2\pi} \int_{-s}^{s} \frac{e^{-ity} - e^{-itx}}{it} \varphi(t)dt \tag{3.15}$$

其中，x 是 $F(x)$ 的连续点，y 沿着 $F(x)$ 的连续点趋向 $-\infty$。

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 $f(x)$ 如下：

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi(t)dt \tag{3.16}$$

证明. 由于 $\varphi(t)$ 在 $\mathbb{R}$ 上绝对可积，Lévy 反演式 (3.13) 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 $\Delta x \to 0$，因为 $\varphi(t)$ 绝对可积，所以求极限可与求积分交换次序，

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \lim_{\Delta x\to 0} \frac{\sin(t\Delta x/2)}{t\Delta x/2} e^{-it(x+\Delta x/2)} \varphi(t)dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi(t)dt \quad \square$$

例 3.14. 利用式 (3.16) 算得特征函数 $\varphi(t) = \begin{cases} 1-|t| & \text{当 } |t| \le 1 \\ 0 & \text{当 } |t| > 1 \end{cases}$ 对应着密度函数 $f(x) = (1-\cos x)/(\pi x^2)$。特征函数 $\varphi(t) = \exp\{-|t|\}$ 对应着密度

函数 $f(x) = 1/[\pi(x^2+1)]$，即分布 Cauchy(0, 1) 的密度函数。

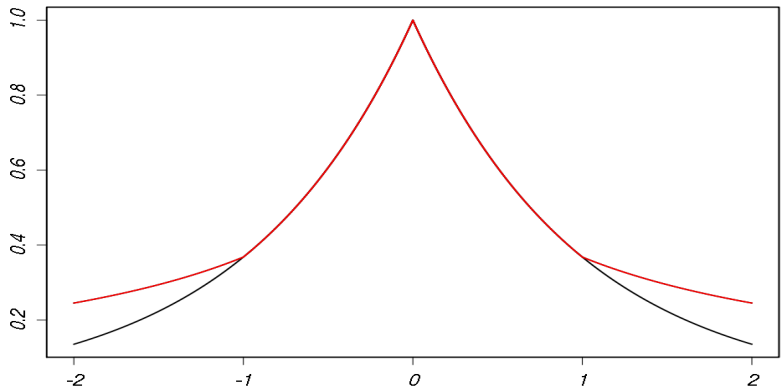


图 3.2: 按照 Pólya 准则，$\varphi_1(t) = \exp\{-|t|\}$ 和 $\varphi_2(t) = \max\{\varphi_1(t), 2/[e(1+|t|)]\}$ 都是特征函数。这两个特征函数仅在区间 $[-1, 1]$ 上重叠，它们对应的密度函数不同。此例表明特征函数的局部性质无法决定密度函数。

定理 3.9 (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 $p_k = \mathsf{P}(X = k)$，其中 $k \in \mathbb{Z}$，则 X 的特征函数是 $\varphi(t) = \sum_{k=-\infty}^{+\infty} p_k e^{itk}$，相应的反演公式为

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \varphi(t) dt \tag{3.17}$$

证明. 根据 $\int_{-\pi}^{\pi} e^{-it(k'-k)} dt = 0$，其中 $k \neq k'$，以及事实 $e^{-itk'}\varphi(t) = p_{k'} + \sum_{k \neq k'} p_k e^{-it(k'-k)}$ 可证。 □

定理 3.10 (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 $\boldsymbol{X}$ 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 (3.15) 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，$F_1(y) = \lim\limits_{n\to\infty} F_1(x_n) = \lim\limits_{n\to\infty} F_2(x_n) = F_2(y)$。 □

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

$$F(x_0) - F(x_0-) = \lim_{s\to\infty} \frac{1}{2s} \int_{-s}^{s} e^{-itx_0}\varphi(t)dt \tag{3.18}$$

❑ J. Gil-Pelaez 于 1951 年证得：若 $F(x)$ 在 $x = x_0$ 处连续，则

$$F(x_0) = \frac{1}{2} - \frac{1}{\pi} \int_0^{\infty} \frac{\Im[e^{-itx_0}\varphi(t)]}{t} dt \tag{3.19}$$

其中，$\Im(z)$ 表示复数 $z \in \mathbb{C}$ 的虚部。

定理 3.12 (用特征函数判定独立性)**.** 随机变量 X, Y 相互独立当且仅当 $(X, Y)^\mathsf{T}$ 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“⇒”：由性质 3.1 可证。往证“⇐”：只证连续的情形，离散的情形留作练习。

$$\begin{aligned}\iint_{\mathbb{R}^2} e^{isx+ity} f(x,y)dxdy &= \int_{-\infty}^{\infty} e^{isx} f_X(x)dx \int_{-\infty}^{\infty} e^{ity} f_Y(y)dy \\ &= \iint_{\mathbb{R}^2} e^{isx+ity} f_X(x)f_Y(y)dxdy\end{aligned}$$

由唯一性定理知，$f(x,y) = f_X(x)f_Y(y)$。 □

定理 3.13 (Fisher, 1925)**.** 已知 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，则随机变量 $\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$ 与随机向量 $(X_1 - \overline{X}, \cdots, X_n - \overline{X})^\mathsf{T}$ 相互独立。

证明. 我们考虑随机向量 $(\overline{X}, X_1 - \overline{X}, \cdots, X_n - \overline{X})^\mathsf{T}$ 的特征函数，

$$\begin{aligned}&\varphi(t, t_1, \cdots, t_n) = \mathsf{E}\exp\left\{it\overline{X} + it_1(X_1 - \overline{X}) + \cdots + it_n(X_n - \overline{X})\right\} \\ &= \mathsf{E}\exp\left\{i\sum_{k=1}^n X_k\left(t_k - \frac{t_1 + \cdots + t_n - t}{n}\right)\right\} = \prod_{k=1}^n \mathsf{E}\exp\left\{\frac{iX_k[t + n(t_k - \bar{t})]}{n}\right\}\end{aligned}$$

其中 $\bar{t} = (t_1 + t_2 + \cdots + t_n)/n$。由于 $X_k \sim \mathrm{N}(\mu, \sigma^2)$，其特征函数为 $\varphi(t) = \exp\left\{it\mu - \sigma^2 t^2/2\right\}$，所以

$$\begin{aligned}
\varphi(t, t_1, \cdots, t_n) &= \prod_{k=1}^{n} \exp\left\{\frac{i[t + n(t_k - \bar{t})]\mu}{n} - \frac{\sigma^2[t + n(t_k - \bar{t})]^2}{2n^2}\right\} \\
&= \exp\left\{it\mu - \frac{\sigma^2 t^2}{2n}\right\} \exp\left\{-\frac{\sigma^2 \sum_{k=1}^{n}(t_k - \bar{t})^2}{2}\right\} \\
&= \varphi(t, 0, \cdots, 0)\varphi(0, t_1, \cdots, t_n) \\
&= \varphi_{\overline{X}}(t)\varphi_{X_1 - \overline{X}, \cdots, X_n - \overline{X}}(t_1, \cdots, t_n)
\end{aligned}$$

由定理 3.12 知，$\overline{X}$ 与 $(X_1 - \overline{X}, \cdots, X_n - \overline{X})^{\mathsf{T}}$ 相互独立。 □

3.2.2* Lévy 连续性定理

定理 3.14. 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。

- 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。

- 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 $X_n \xrightarrow{L} X$。

证明. 往证正极限定理：令 $a<0, b>0$ 是 $F(x)$ 的两个连续点。

$$\varphi_n(t)=\int_{-\infty}^{a} e^{itx}dF_n(x)+\int_{a}^{b} e^{itx}dF_n(x)+\int_{b}^{+\infty} e^{itx}dF_n(x)=C_{n1}+C_{n2}+C_{n3}$$
$$\varphi(t)=\int_{-\infty}^{a} e^{itx}dF(x)+\int_{a}^{b} e^{itx}dF(x)+\int_{b}^{+\infty} e^{itx}dF(x)=C_1+C_2+C_3$$

对于任意的 $\epsilon>0$，总可以令 $|a|$ 足够地大，使得

$$|C_{n1}-C_1|\le\int_{-\infty}^{a} dF_n(x)+\int_{-\infty}^{a} dF(x)=F_n(a)+F(a)<\epsilon/6+\epsilon/6=\epsilon/3$$

类似地，令 $|b|$ 足够地大，使得 $|C_{n3}-C_3|\le\epsilon/3$。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，

$$\begin{aligned}|C_{n2}-C_2|&\le|F_n(b)-F(b)|+|F_n(a)-F(a)|+|t|\int_a^b|F_n(x)-F(x)|dx\\&<\epsilon/9+\epsilon/9+\epsilon/9=\epsilon/3\end{aligned}$$

于是，在 t 的一个有限区域上，$\varphi_n(t)$ 一致收敛到 $\varphi(t)$。

往证逆极限定理：显然 $0\le F(x)\le 1$。为了证明 $F(+\infty)-F(-\infty)=1$，下面用归谬法，令 $a=F(+\infty)-F(-\infty)<1$。因为 $\varphi(0)=\lim\limits_{n\to\infty}\varphi_n(0)=1$

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 $\lim\limits_{k\to\infty}\varphi_{n_k}(t) = \varphi(t)$，对足够大的 k 有

$$\frac{1}{2T}\left|\int_{-T}^{T}\varphi_{n_k}(t)dt\right| > a + \frac{\epsilon}{2} \tag{3.20}$$

令 $b > 4/(T\epsilon)$，对足够大的 k 有 $a_k = F_{n_k}(b) - F_{n_k}(-b) < a + \epsilon/4$，进而

$$\begin{aligned}
\frac{1}{2T}\left|\int_{-T}^{T}\varphi_{n_k}(t)dt\right| &= \frac{1}{2T}\left|\int_{-T}^{T}\left\{\int_{-\infty}^{+\infty}e^{itx}dF_{n_k}(x)\right\}dt\right| \\
&= \frac{1}{2T}\left|\int_{-\infty}^{+\infty}\left\{\int_{-T}^{T}e^{itx}dt\right\}dF_{n_k}(x)\right| \\
&\le \frac{1}{2T}\int_{|x|\le b}\left|\int_{-T}^{T}e^{itx}dt\right|dF_{n_k}(x) + \frac{1}{2T}\int_{|x|>b}\left|\int_{-T}^{T}e^{itx}dt\right|dF_{n_k}(x) \\
&\le \int_{|x|\le b}dF_{n_k}(x) + \frac{1}{2T}\int_{|x|>b}\frac{2}{|x|}dF_{n_k}(x) \le a_k + \frac{1}{bT} \le a + \frac{\epsilon}{2}
\end{aligned} \tag{3.21}$$

由以下事实 $\left|\int_{-T}^{T}e^{itx}dt\right| = 2|\sin(Tx)|/|x| \le 2/|x|$，首先，(3.20)+(3.21) $\Rightarrow$ 矛盾！于是，$F(+\infty) - F(-\infty) = 1$，即 $F(x)$ 是一个分布函数。其次，如果 $\lim\limits_{n\to\infty}F_n(x) = F(x)$ 不真，则存在子序列 $\{F_{n_j}(x)\}$ 使得 $\lim\limits_{j\to\infty}F_{n_j}(x) = \tilde{F}(x) \ne F(x)$，二者却有相同的特征函数，矛盾！ □

✎粗略地说，Lévy 连续性定理 (Lévy's continuity theorem) 保证了下列图表交换。

$$\begin{array}{ccc}
X_n & \xrightarrow{L} & X \\
\Updownarrow & & \Updownarrow \\
\varphi_n(t) & \longrightarrow & \varphi(t)
\end{array}$$

例 3.15. 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，特征函数

$\varphi_n(t)=\sin(nt)/(nt), n=1,2,\cdots$ 所对应的分布函数是

$$F_n(x)=\begin{cases}0 & \text{当 } x\le -n\\ \dfrac{x+n}{2n} & \text{当 } -n<x<n\\ 1 & \text{当 } x\ge n\end{cases}$$

显然 $\lim\limits_{n\to\infty}F_n(x)=1/2$ 不是分布函数，而且 $\lim\limits_{n\to\infty}\varphi_n(t)=\begin{cases}0 & \text{若 } t\neq 0\\ 1 & \text{若 } t=0\end{cases}$ 不是特征函数。

3.3 习题

3.1. 令二维随机向量 $(X,Y)^{\mathsf{T}}$ 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. 已知特征函数，求概率分布：(1) $\varphi(t)=\cos t$，(2) $\varphi(t)=\cos^2 t$，(3) $\varphi(t)=\sin t/t$。

3.4. 设随机变量 X 有密度函数为 $f(x)=c\exp\{-a|x|\}$，其中 $a>0,c$ 为合适的常数，求它的特征函数。

3.5. 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。

3.6. 设随机变量 $X\sim \mathrm{B}(m,p)$ 和 $Y\sim \mathrm{B}(n,p)$ 相互独立，试证明：$Z=X+Y\sim \mathrm{B}(m+n,p)$。

3.7. 设随机变量 $X_1,X_2,\cdots$ 独立同分布，满足 $\mathsf{P}(X_1=k)=q^k p$，$k=0,1,\cdots$，其中 $0<p<1$ 且 $q=1-p$。试求 $X=\sum_{j=1}^{n}X_j$ 的分布。

3.8. 设 $\varphi(t)$ 为一个实值的特征函数，试证明：$1-\varphi(2t)\le 4[1-\varphi(t)]$。

3.9. 若 $t\to 0$ 时，特征函数 $\varphi(t)=1+o(t^2)$，试证明 $\varphi(t)\equiv 1$。

3.10. 已知 $\varphi(t)$ 是某随机变量的特征函数，对于任意的正整数 n 及任意的实数 $t_1,t_2,\cdots,t_n\in\mathbb{R}$ 和任意的复数 $z_1,z_2,\cdots,z_n\in\mathbb{C}$，试证明：$\sum_{k=1}^{n}\sum_{j=1}^{n}\varphi(t_k-t_j)z_k\bar{z}_j\ge 0$。

第四章

一些常见的分布

有了特征函数和期望、方差等数字特征作为工具，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个小节中。熟练掌握这些分布或有助于概率建模和统计计算，或有助于对中心极限定理（第五章的主要内容）的理解，例如若 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{U}[0,1]$，只要 $n \geq 4$ 就近似地有 $X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n/2, n/12)$。本章内容是已知工具的试验场，几乎全可付诸计算机实践。

出于计算机模拟的目的，随机数模拟器至今仍是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 $[0,1]$ 上均匀分布得到的随机数，必须反映出 $[0,1]$ 中任何实数都有相同的机会被抽取到。然而，通过确定的算法不能产生真正意义上的随机数，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像随机数，通常把这样的算法称作随机数产生器。

目前多数随机数产生器的设计都源自 1949 年 D. H. Lehmer 定义的下述递归关系，

$$z_{n+1} = (az_n + c) \mod m \tag{4.1}$$

其中 $m, a, c > 0$ 都是确定的整数且 $a, c < m$，初始整数 $0 \le z_0 < m$ 是用户设定的，称为“种子”。该递归关系所产生的 $z_0, z_1, \cdots, z_n, \cdots$ 称为线性同余序列，它总是有有限周期的且 $0 \le z_n < m - 1$。例如，式 (4.1) 中 $z_0 = 2, a = 3, c = 1, m = 7$ 所定义的线性同余序列的周期是 6。

```
(%i1) z[0] : 2 $
(%i2) z[n] := mod(3*z[n-1]+1,7) $
(%i3) makelist(z[n],n,0,30);
(%o3) [2, 0, 1, 4, 6, 5, 2, 0, 1, 4, 6, 5, 2, 0, 1, 4, 6, 5, 2, 0, 1, 4, 6, 5,
       2, 0, 1, 4, 6, 5, 2]
```

通常 m 选得很大，如 $m = 2^{32}$，使得 z_n/m 看起来像是 U[0, 1) 的随机数。递归式 (4.1) 产生全周期 m 的线性同余序列关键在于选择 a, c，譬如 $a = c = 1$，但这样的序列因不是随机的而没有意义。设定合适的参数 m, a, c 需要古典数论的技巧，如常用的一个结果：若 $m = 2^s$，则当 c 为奇数且 $a \equiv 1 \mod 4$ 时线性同余序列为满周期。感兴趣的读者可以参阅 D. Knuth 的名著《计算机程序设计艺术》第二卷《半数值算法》的第 3.2 节《产生均匀的随机数》。

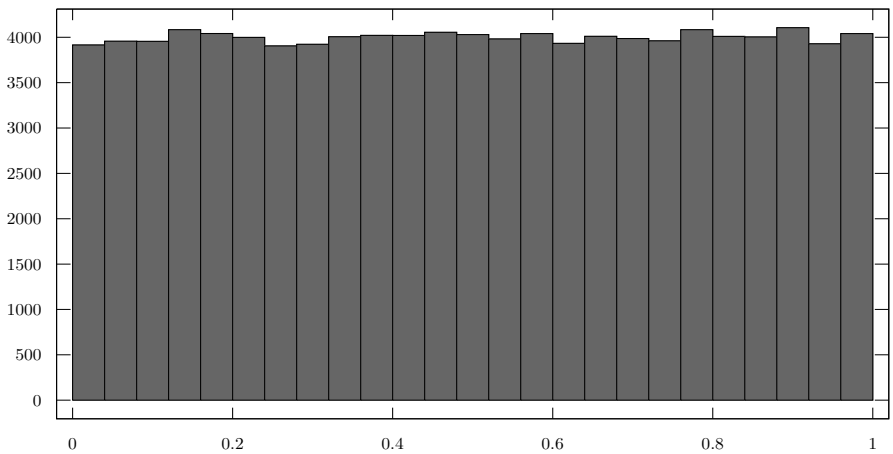
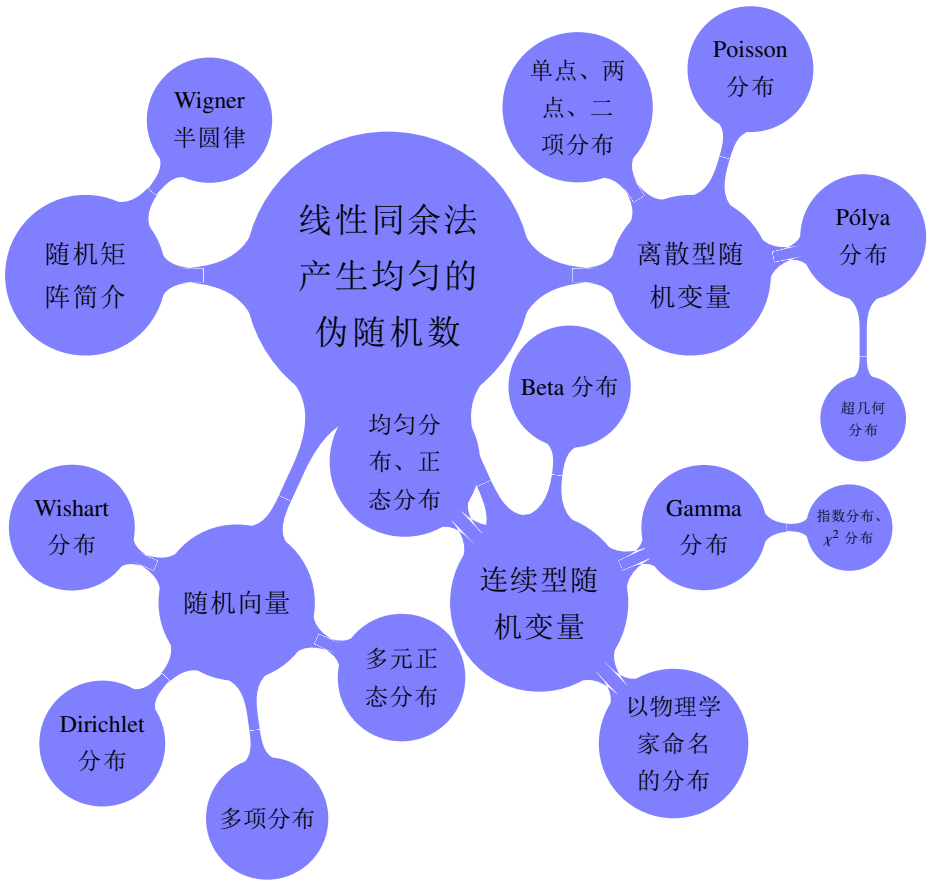


图 4.1: 以 $m = 2^{32}, a = 69069, c = 1$ 为例，递归式 (4.1) 产生的线性同余序列具有全周期。令 $z_0 = 2^{31}$，该图是前 10^5 个形如 z_n/m 的伪随机数的直方图，看起来像产自 U[0, 1)。当 $m = 2^{35}, a = 7, c = 1, z_0 = 1$ 时，效果也不错。伪随机数是否“合格”还要通过统计检验（详见第八章的拟合优度检验）。


线性同余法产生均匀的伪随机数
随机矩阵简介
Wigner 半圆律
离散型随机变量
单点、两点、二项分布
Poisson 分布
Pólya 分布
超几何分布
连续型随机变量
均匀分布、正态分布
Beta 分布
Gamma 分布
指数分布、χ^2 分布
以物理学家命名的分布
随机向量
Wishart 分布
多元正态分布
Dirichlet 分布
多项分布

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 $X \sim \frac{1}{m}\langle k_1\rangle + \frac{1}{m}\langle k_2\rangle + \cdots + \frac{1}{m}\langle k_m\rangle$，其中 $k_1, k_2, \cdots, k_m \in \mathbb{R}$ 两两不等，约定用符号 $X \sim \mathrm{U}\{k_1, k_2, \cdots, k_m\}$ 简记之。

本节内容 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 分布的特款，而 Pólya 分布在一定条件下以二项分布为“极限状态”。这些用处各异的离散型随机变量的分布之间存在着千丝万缕的联系，本节重点是第一小节的单点分布、两点分布和二项分布，以及第四小节的 Poisson 分布。

学习目标 (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，如二项分布与正态分布的关系。

4.1.1　单点分布、两点分布和二项分布

单点分布 $X \sim \langle c \rangle$ 的定义见例 2.5，其期望和方差分别为 $\mathsf{E}(X) = c, \mathsf{V}(X) = 0$，特征函数为 $\varphi(t) = e^{itc}$。经常利用性质 2.8 和性质 2.12 来判定随机变量服从单点分布。两点分布 $X \sim p\langle a \rangle + (1-p)\langle b \rangle$ 的定义见例 2.6，亦可看作是两个相异单点分布的凸组合。两点分布的特款 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 称作 0-1 分布。

性质 4.1. 0-1 分布 $X \sim p\langle 1 \rangle + (1-p)\langle 0 \rangle$ 的特征函数为 $\varphi(t) = 1 + p(e^{it} - 1)$，由式 (3.9) 求得各阶原点矩 $m_k = p$，其中 $k = 1, 2, \cdots$。进而，$\mathsf{V}(X) = m_2 - m_1^2 = p(1-p)$。偏度系数和峰度系数分别为 $\gamma_1 = (1-2p)/\sqrt{p(1-p)}, \gamma_2 = 1/[p(1-p)] - 6$。

```
p <- 0.8; num <- 100      # R 语言产生 num 个 0-1 分布 p<1>+(1-p)<0> 的随机数
x <- sample(c(1,0), num, replace=TRUE, prob=c(p,1-p))  # 产生 0-1 分布的随机数
```

性质 4.2. 如例 1.27 和例 3.5 所描述，二项分布 $X \sim \mathrm{B}(n, p)$ 的概率函数为 $\mathsf{P}(X = k) = C_n^k p^k (1-p)^{n-k}$，其中 $k = 0, 1, 2, \cdots, n$，特征函数为 $\varphi(t) = [1 + p(e^{it} - 1)]^n$。由式 (3.9) 求得 $\mathsf{E}(X) = np, \mathsf{V}(X) = np(1-p)$。$\mathsf{M}(X)$ 为 $\lfloor np \rfloor$ 或 $\lceil np \rceil$。偏度系数和峰度系数分别为 $\gamma_1 = (1-2p)/\sqrt{p(1-p)}, \gamma_2 = -6/n + 1/[np(1-p)]$。

```
n <- 5; p <- 0.6; num <- 70         # R 语言产生 num 个二项分布 B(n,p) 的随机数
x <- rbinom(num, size=n, prob=p)    # 产生二项分布的随机数
```

定义 4.1 (广义二项分布)**.** 已知随机变量 $X_k \sim p_k\langle 1 \rangle + (1-p_k)\langle 0 \rangle, k = 1, 2, \cdots, n$ 相互独立，则称 $X = X_1 + X_2 + \cdots + X_n$ 服从广义二项分布。

性质 4.3. 如上定义的广义分布的特征函数为 $\varphi(t) = \prod_{k=1}^n [1 + p_k(e^{it} - 1)]$，期望和方差分别为 $\mathsf{E}(X) = \sum_{k=1}^n p_k$ 和 $\mathsf{V}(X) = \sum_{k=1}^n p_k(1-p_k)$。

例 4.1 (用正态分布近似二项分布)**.** 对于 $X \sim \mathrm{B}(n, p)$，当 n 很大时，图 1.4 显示概率 $\mathsf{P}(X = k|n, p) = C_n^k p^k (1-p)^{n-k}, k = 0, 1, \cdots, n$ 呈现出对称

性，如果用一个连续函数来拟合，它会是一个怎样的函数呢？附录 C 用基于 Stirling 公式的初等方法推到出它是正态分布 $\mathrm{N}(np, npq)$ 的密度函数，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 $\mathrm{N}(n/2, n/4)$ 来近似，而且 n 越大近似程度越高。为了解 $\mathrm{B}(n, 1/2)$ 与 $\mathrm{N}(n/2, n/4)$ 的近似程度与 n 的关系，我们考察误差 $\varepsilon(n) = \sum_{k=0}^{n} |\mathrm{P}(X = k|n, 1/2) - \phi(k|n/2, n/4)|$。

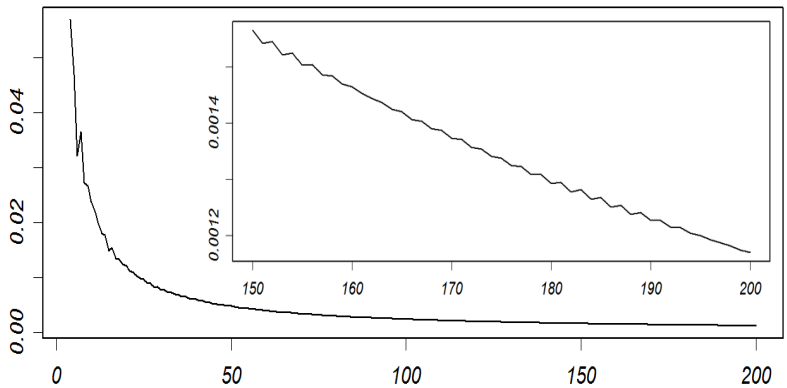


图 4.2: 不难发现 n 越大，误差 $\varepsilon(n)$ 反而越小，即在离散点 $k = 0, 1, \cdots, n$ 上 $\phi(k|n/2, n/4)$ 越全体地接近 $2^{-n}C_n^k$，有时也粗略地说 n 越大 $\mathrm{N}(n/2, n/4)$ 越逼进 $\mathrm{B}(n, 1/2)$。由 de Moivre-Laplace 中心极限定理可证得 $\lim_{n\to\infty} \varepsilon(n) = 0$。

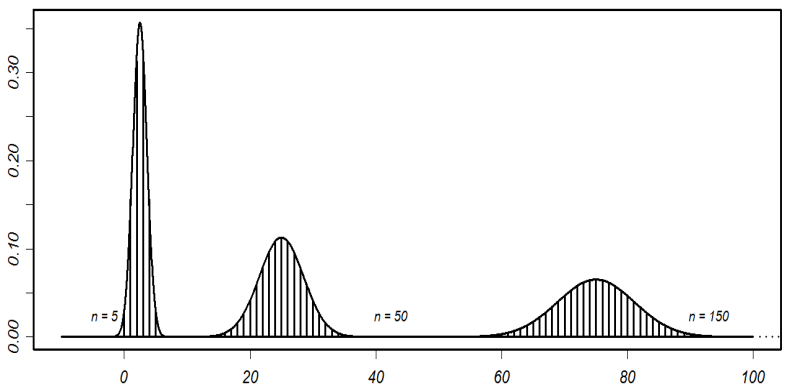


图 4.3: de Moivre 发现二项分布 $\mathrm{B}(n, 1/2)$ 可用正态分布 $\mathrm{N}(n/2, n/4)$ 来近似。该图同时显示了 $n = 5, 50, 150$ 时 $\mathrm{B}(n, 1/2)$ 的概率函数的竖线图和 $\mathrm{N}(n/2, n/4)$ 的概率密度曲线，二者的差距用 $\varepsilon(n)$ 来刻画，分别为 $\varepsilon(5) = 0.04700868, \varepsilon(50) = 0.004741236, \varepsilon(150) = 0.001565345$。

4.1.2　几何分布、负二项分布

定义 4.2 (几何分布)**.** 设 Bernoulli 试验中某事件 A 出现的概率为 $p \in (0,1)$，在 $n+1$ 重 Bernoulli 试验中 A 在第 $n+1$ 次试验中头次出现的概率为 pq^n，其中 $q = 1-p$。称随机变量 $X \sim p\langle 0\rangle + \cdots + pq^n\langle n\rangle + \cdots$ 服从几何分布 (geometric distribution)，记作 $X \sim \mathrm{Geom}(p)$。

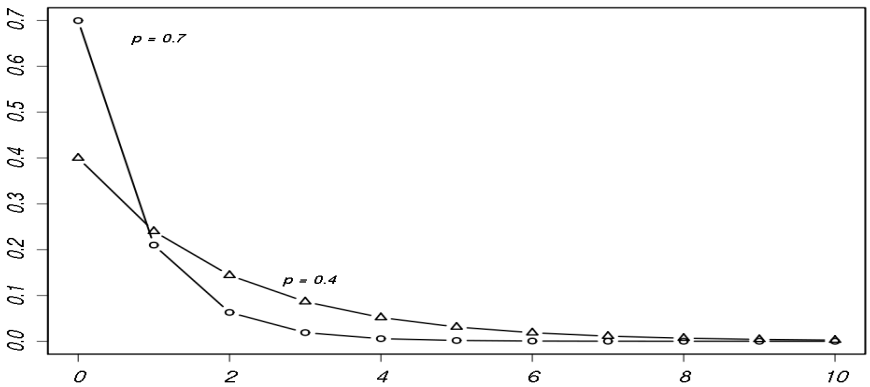


图 4.4: 几何分布 $X \sim \mathrm{Geom}(0.7)$ 与 $X \sim \mathrm{Geom}(0.4)$ 的比较。

练习 4.1. 几何分布 $X \sim \mathrm{Geom}(p)$ 的特征函数为 $p(1-qe^{it})^{-1}$，常见数字特征为 $\mathsf{E}(X) = q/p, \mathsf{V}(X) = q/p^2, \gamma_1 = (1+q)/\sqrt{q}, \gamma_2 = 6 + p^2/q$。

```
> num <- 30; p <- 0.4      # R 语言产生 num 个几何分布 Geom(p) 的随机数
> rgeom(num, prob=p)       # 产生几何分布 Geom(p) 的随机数
 [1] 0 0 0 0 0 7 4 2 0 0 6 0 0 1 0 2 2 1 2 1 3 0 0 0 5 1 0 0 0 6
```

定义 4.3 (负二项分布)**.** 如果随机变量 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{Geom}(p)$，则随机变量 $X = X_1 + X_2 + \cdots + X_n$ 服从负二项分布 (negative binomial distribution)*，记作 $X \sim \mathrm{NegB}(n,p)$。

练习 4.2. 负二项分布 $X \sim \mathrm{NegB}(n,p)$ 的概率函数为

$$\mathsf{P}(X=k) = C_{n+k-1}^{k} p^n q^k，\text{其中 } 0<p<1, q=1-p, k=0,1,2,\cdots \tag{4.2}$$

*负二项分布也称作 Pascal 分布。

它的特征函数为 $p^n(1-qe^{it})^{-n}$，常见数字特征为 $\mathsf{E}(X)=nq/p, \mathsf{V}(X)=nq/p^2, \gamma_1=(1+q)/\sqrt{nq}, \gamma_2=6/n+p^2/(nq)$。

✎按照定义 4.3，负二项分布 $X\sim \mathrm{NegB}(n,p)$ 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 $k=0,1,2,\cdots$。

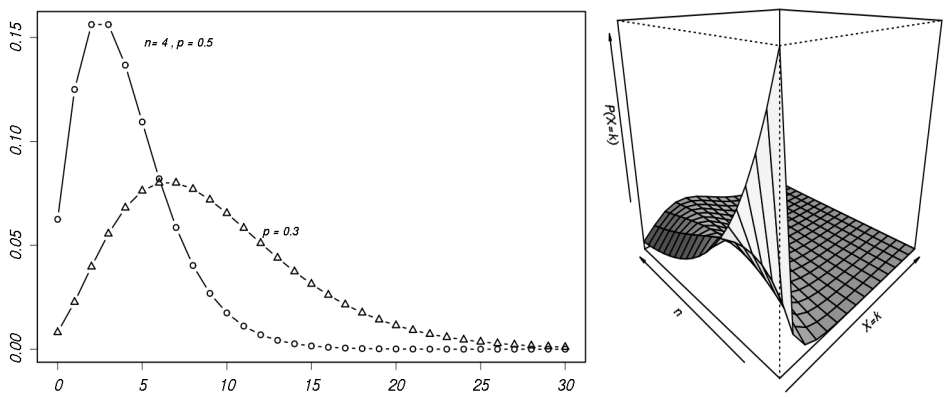


图 4.5: 左图是负二项分布 $X\sim \mathrm{NegB}(4,0.5)$ 与 $X\sim \mathrm{NegB}(4,0.3)$ 的比较。右图是负二项分布 $X\sim \mathrm{NegB}(n,p=0.5)$ 的概率函数 $\mathsf{P}(X=k)$ 在取不同的 k,n 时的变化情况。

4.1.3　Pólya 分布、超几何分布

问题 4.1. 1923 年，G. Pólya 考虑了下面的球-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 表示取出黑球的数量，试求 $\mathsf{P}(X=k)$。

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 $k=1,2,\cdots,n-1$ 有

$$\begin{aligned}\mathsf{P}(X=k) &= C_n^k \frac{b(b+s)\cdots[b+(k-1)s]w(w+s)\cdots[w+(n-k-1)s]}{N(N+s)\cdots[N+(n-1)s]} \\ &= C_n^k \frac{p(p+a)\cdots[p+(k-1)a]q(q+a)\cdots[q+(n-k-1)a]}{(1+a)\cdots[1+(n-1)a]} \\ \mathsf{P}(X=0) &= \frac{w(w+s)\cdots[w+(n-1)s]}{N(N+s)\cdots[N+(n-1)s]} = \frac{q(q+a)\cdots[q+(n-1)a]}{(1+a)\cdots[1+(n-1)a]} \\ \mathsf{P}(X=n) &= \frac{b(b+s)\cdots[b+(n-1)s]}{N(N+s)\cdots[N+(n-1)s]} = \frac{p(p+a)\cdots[p+(n-1)a]}{(1+a)\cdots[1+(n-1)a]}\end{aligned}$$

称 X 所服从的分布为 Pólya 分布，它常用作传染病模型。请读者验证 $\mathsf{E}(X)=np$ 且 $\mathsf{V}(X)=npq(1+na)/(1+a)$。

❑ 令 $s=-1$，Pólya 分布的特款称为超几何分布，概率函数为

$$\mathsf{P}(X=k)=\frac{C_b^k C_w^{n-k}}{C_N^n}=\frac{C_b^k C_{N-b}^{n-k}}{C_N^n} \tag{4.3}$$

期望和方差分别为 $\mathsf{E}(X)=nb/N, \mathsf{V}(X)=nb(1-b/N)(N-n)/[N(N-1)]$。偏度系数和峰度系数表达式较复杂，可通过 Maxima 查看。

❑ 令 $N\to\infty$，若 $p=b/N, q=1-p$ 为常数且 $\lim\limits_{N\to\infty} a=0$，则有

$$\lim_{N\to\infty}\mathsf{P}(X=k)=C_n^k p^k q^{n-k} \tag{4.4}$$

4.1.4 Poisson 分布

1837 年法国数学家 S. D. Poisson (1781-1840) 研究了随机变量序列 $X_n \sim \mathrm{B}(n, p_n), n = 1, 2, \cdots$ 在附加约束条件 $\lim\limits_{n\to\infty} np_n = \lambda > 0$（其中 λ 为常数）之下的“极限状态”：发现随着 $n \to \infty$，概率 $\mathsf{P}(X_n = k) = C_n^k p_n^k (1 - p_n)^{n-k}$ 具有如下变化趋势。

$$\frac{(np_n)^k}{k!}\left(1 - \frac{np_n}{n}\right)^n \frac{n(n-1)\cdots(n-k+1)}{(n-np_n)^k} \to \frac{\lambda^k}{k!}e^{-\lambda}$$

定义 4.4 (Poisson 分布)**.** 如果离散型随机变量 X 的概率函数为

$$\mathsf{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}，\text{其中 } k = 0, 1, 2, \cdots \tag{4.5}$$

则称 X 服从参数为 λ 的 Poisson 分布，记作 $X \sim \mathrm{Poisson}(\lambda)$。它是 n 很大，p 很小的二项分布的近似。Poisson 分布常用于描述一些物理现象*，如放射性分裂、定长时间段里电话被呼叫的次数（见例 1.46）等，Poisson 分布在排队论中也很有用。

练习 4.3. 验证 $X \sim \mathrm{Poisson}(\lambda)$ 满足递归关系 $\mathsf{P}(X = k+1) = \frac{\lambda}{k+1}\mathsf{P}(X = k)$。

性质 4.4. 分布 $X \sim \mathrm{Poisson}(\lambda)$ 的特征函数为 $\varphi(t) = \exp\{\lambda(e^{it} - 1)\}$，求得 $\mathsf{E}(X) = \lambda, \mathsf{V}(X) = \lambda, \gamma_1 = 1/\sqrt{\lambda}, \gamma_2 = 1/\lambda$。

练习 4.4. 若 $X \sim \mathrm{Poisson}(\lambda)$，证明：当 $\lambda \to \infty$ 时，随机变量 $Y = (X - \lambda)/\sqrt{\lambda}$ 的分布趋于 $\mathrm{N}(0, 1)$。提示：只要证明当 $\lambda \to \infty$ 时，Y 的特征函数 $\varphi_Y(t)$ 趋于 $e^{-t^2/2}$ 即可。

✂ **例 4.2.** 对于随机变量 $X \sim \mathrm{Poisson}(\lambda)$，有

$$\psi(t) = \ln \varphi(t) = \lambda \sum_{k=1}^{\infty} \frac{(it)^k}{k!}$$

*详见 W. Feller 的《概率论及其应用》上卷 [35] 第六章。

由式 (3.11) 可得 $\varkappa_k = \lambda$，其中 $k = 1, 2, 3, \cdots$。

性质 4.5. 如果随机变量 $X_j \sim \text{Poisson}(\lambda_j), j = 1, 2, \cdots, n$ 相互独立，则 $X_1 + X_2 + \cdots + X_n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$。

定理 4.1. 如果 $X_1 \sim \text{Poisson}(\lambda_1)$ 与 $X_2 \sim \text{Poisson}(\lambda_2)$ 相互独立，则

$$X_1 | (X_1 + X_2 = n) \sim \text{B}\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right) \tag{4.6}$$

定理 4.2 (D. A. 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 4.3 (复合分布)**.** 已知 $X \sim \text{B}(N, p)$ 且随机变量 $N \sim \text{Poisson}(\lambda)$，试证明：$X \sim \text{Poisson}(\lambda p)$。

证明. 对于任意的 $s \in \{0, 1, 2, \cdots\}$，皆有

$$\begin{aligned}\mathsf{P}(X = s) &= \sum_{n=0}^{\infty} \mathsf{P}(X = s | N = n)\mathsf{P}(N = n) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} C_n^s p^s (1-p)^{n-s} \\ &= \frac{e^{-\lambda} p^s \lambda^s}{s!} \sum_{n=s}^{\infty} \frac{(\lambda - \lambda p)^{n-s}}{(n-s)!} = \frac{e^{-\lambda} p^s \lambda^s}{s!} e^{\lambda(1-p)} = \frac{(\lambda p)^s}{s!} e^{-\lambda p}\end{aligned}$$ □

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述定理的特例。

定理 4.3. 已知 $N \sim \text{Poisson}(\lambda)$。假设随机变量 $X_1, \cdots, X_N$ 独立同分布，特征函数都是 $\varphi(t)$。若 $N, X_1, \cdots, X_N$ 相互独立，则随机变量 $Y = X_1 + \cdots + X_N$ 的特征函数为 $\varphi_Y(t) = \exp\{\lambda[\varphi(t) - 1]\}$。

证明. 由双期望定理得到 $\varphi_Y(t) = \mathsf{E}\{\mathsf{E}[e^{it(X_1+\cdots+X_N)}|N]\}$，并且 $\mathsf{E}[e^{it(X_1+\cdots+X_N)}|N = n] = [\varphi(t)]^n$，所以 $\varphi_Y(t) = \sum_{n=1}^{\infty} [\varphi(t)]^n \lambda^n e^{-\lambda}/n! = \exp\{\lambda[\varphi(t) - 1]\}$。 □

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中。标准正态分布的密度函数 $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ 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？原因是第五章即将介绍的中心极限定理，附录 C 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。

本节内容 均匀分布之所以重要，是因为其他分布的随机数都源于 U[0, 1] 的随机数，例如，若 $X \sim \mathrm{U}(0,1)$，则 $Y = -\ln X \sim \mathrm{Expon}(1)$；正态分布由于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 分布的特款；χ^2 分布、t 分布和 F 分布来自统计学；Weibull 分布、Rayleigh 分布、Maxwell 和 Wigner 半圆分布都来自物理学。

学习目标 (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和特征函数；(2) 掌握这些常见分布的性质。

4.2.1　均匀分布

均匀分布是“等概率”的连续情形。像例 2.7 描述的，服从区间 $[a,b]$ 上的均匀分布的随机变量 $X \sim \mathrm{U}[a,b]$ 的特征函数为 $\varphi(t) = [\exp(itb) - \exp(ita)]/[it(b-a)]$。请读者自己验证，$\mathsf{E}(X) = \mathsf{M}(X) = (a+b)/2, \mathsf{V}(X) = (b-a)^2/12, \gamma_1 = 0, \gamma_2 = -6/5$。

性质 4.6. 如果 $X \sim \mathrm{U}[0,1]$ 且 $b > a$，则 $a + (b-a)X \sim \mathrm{U}[a,b]$。

```
a <- 1; b <- 4; num <- 30        # R 语言产生 num 个均匀分布 U[a,b] 的随机数
x <- runif(num, min=a, max=b)    # 产生均匀分布 U[a,b] 的随机数
```

性质 4.7. 令随机变量 $X_1, X_2, \cdots \overset{iid}{\sim} \frac{1}{2}\langle 1\rangle + \frac{1}{2}\langle 0\rangle$，则 $X \sim \mathrm{U}[0,1]$ 的随机数可由按下述方式构造：$X = \sum_{k=1}^{\infty} 2^{-k} X_k$。

定理 4.4. 已知随机变量 X 的分布函数 $F_X(x)$ 连续，则新构造的随机变量 $Y = F_X(X) \sim \mathrm{U}[0,1]$。

证明. 对任意 $y \in [0,1]$，至少存在一个 x 使得 $y = F_X(x) = \mathsf{P}\{X \leq x\}$，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

$$F_Y(y) = \mathsf{P}\{F_X(X) \leq y\} = \begin{cases} 0 & \text{当 } y \leq 0 \\ \mathsf{P}\{X \leq F_X^{-1}(y)\} = y & \text{当 } 0 < y \leq 1 \\ 1 & \text{当 } y > 1 \end{cases}$$ □

注记 4.1. 定理 4.4 为随机数的产生提供了一个可行的方法：(1) 产生 $Y \sim [0,1]$ 的随机数 y，(2) 求得 $x = F_X^{-1}(y)$ 为 X 的随机数。步骤 (1) 是非常关键的，在用 Monte Carlo 方法产生随机数时也是如此，这部分内容将在第十四章“随机模拟技术”中再作介绍。多数编程语言都提供了 $\mathrm{U}[0,1]$ 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 `uniformdeviate` 函数来实现该功能（例如图 1.7 是利用 MetaPost 生成的）。

例 4.4. 随机变量 X 具有 Logistic 分布当且仅当其分布函数为

$$F_X(x) = \frac{1}{1 + \exp\{-(x-m)/s\}} \tag{4.7}$$

记作 $X \sim \text{Logistic}(m, s)$，其中 m 为位置参数，s 为尺度参数。所对应的密度函数为

$$f_X(x) = \frac{\exp\{-(x-m)/s\}}{s\left[1 + \exp\{-(x-m)/s\}\right]^2} \tag{4.8}$$

特别地，当 $m = 0, s = 1$ 时，分布函数 $F_X(x) = 1/[1 + \exp(-x)]$ 在人工神经网络、机器学习中俗称 sigmoid 函数。R 语言自带了 Logistic 分布的密度函数 `dlogis` 及其随机数产生函数 `rlogis`，下面用基于定理 4.4 的方法重新实现一个 $X \sim \text{Logistic}(0, 1)$ 的随机数产生器。

```
rsigm <- function(n) {     # rsigm(n) 产生 n 个 Logistic(0,1) 分布的随机数
y <- runif(n);             # 先产生 [0,1] 上均匀分布的 n 个随机数
x <- log(y/(1-y));         # 再利用分布函数的反函数求得随机数
}
n <- 100000; x <- rsigm(n); z <- seq(min(x),max(x),length.out=n);
hist(x, freq=FALSE, ylim=c(0,0.25)); points(z,dlogis(z), type="l");
```

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。下面介绍一个方法，从 U[0, 1] 的随机数很方便地“构造”出正态分布的随机数。

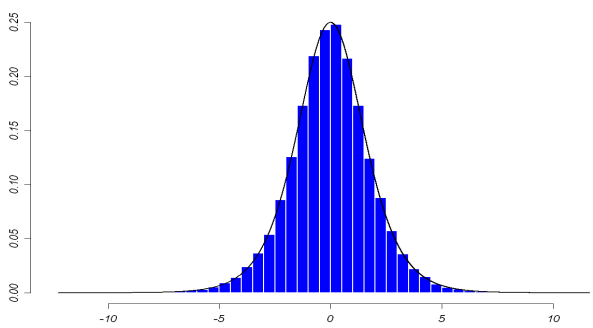


图 4.6: 用基于定理 4.4 的方法产生的 $n = 10^5$ 个随机数的直方图，看上去它与 Logistic(0, 1) 的密度函数曲线吻合得很好。

定理 4.5. 已知随机变量 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{U}[0,1]$，则随机变量 $X = X_1 + X_2 + \cdots + X_n$ 的特征函数为 $\varphi_X(t) = [i(1-e^{it})/t]^n$，期望和方差分别为 $\mathsf{E}(X) = n/2$ 和 $\mathsf{V}(X) = n/12$，且 X 具有密度函数

$$f(x,n) = \begin{cases} \sum_{k=0}^{n}(-1)^k C_n^k (x-k)_+^{n-1}/(n-1)! & \text{当 } x \in [0,n] \\ 0 & \text{当 } x \notin [0,n] \end{cases} \tag{4.9}$$

其中正截尾函数 x_+ 定义如下：

$$x_+ = \left\{ \begin{array}{ll} x, & \text{若 } x > 0 \\ 0, & \text{若 } x \le 0 \end{array} \right\} = xJ(x) = \frac{x + x \cdot \mathrm{sgn}(x)}{2} \tag{4.10}$$

其中符号函数 $\mathrm{sgn}(x)$ 定义如下：

$$\mathrm{sgn}(x) = \begin{cases} 1, & \text{若 } x > 0 \\ 0, & \text{若 } x = 0 \\ -1, & \text{若 } x < 0 \end{cases} \tag{4.11}$$

例 4.5. 定理 4.5 中，当 $n = 2$ 时 X 的分布称为三角分布，密度函数为

$$f(x,2) = \begin{cases} 1 - |x-1| & \text{当 } x \in [0,2] \\ 0 & \text{当 } x \notin [0,2] \end{cases} \tag{4.12}$$

$f(x,n)$ 关于 $x = n/2$ 对称，通过 Maxima 可以了解当 n 增大时，$f(x,n)$ 的变化情况：当 $n \ge 4$ 时，$f(x,n)$ 已经非常之接近正态分布 $\phi(x|n/2, n/12)$ 了。例如，把 U[0,1] 的随机数四个一组地相加，所构造出来的可视为 N(2, 1/3) 的随机数，再经过适当的变换就是 N(0, 1) 的随机数。

```
load(functs) $
g(x) := (x+x*signum(x))/2 $  /* 定义正截尾函数 */
f(x,n) := sum((-1)^k*combination(n,k)*(g(x-k))^(n-1),k,0,n)/(n-1)! $
plot2d([f(x,2),f(x,3),f(x,4),f(x,5)],[x,-1,6],
        [style, [lines,2]], [legend,"n=2","n=3","n=4","n=5"],[ylabel,"f(x,n)"]);
```

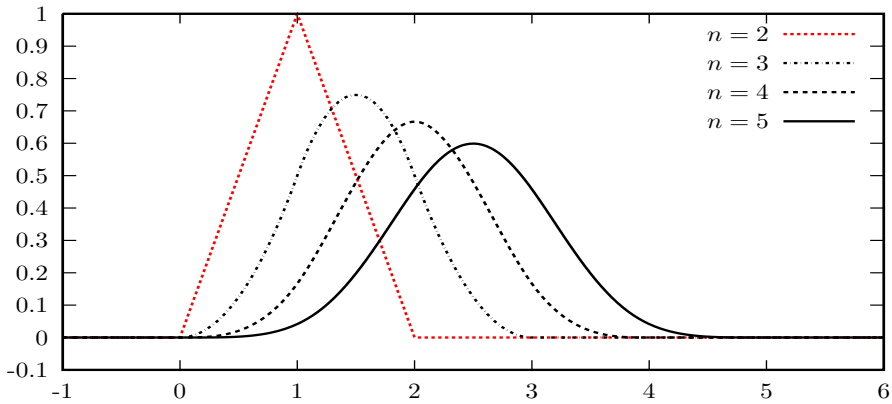


图 4.7: 随着 n 的增大，式 (4.9) 定义的密度函数 $f(x,n)$ 越来越逼近正态分布 $\phi(x|n/2,n/12)$，究其原因要用到 Lindeberg-Lévy 中心极限定理（见本书定理 5.14）。当 $n = 4$ 时，逼进的效果就非常好了，足以满足实际应用，譬如用此法产生标准正态分布 N(0, 1) 的随机数（R 代码如下）。

```
TotalNum <- 1000000;  RandNum <- runif(TotalNum)  # 产生 U[0,1] 随机数
n <- 4;   RandMat <- matrix(RandNum, nrow=n)       # n 个随机数一组排成一列
NewRandNum <- colSums(RandMat)          # 按列求和得到 TotalNum/n 个新的随机数
hist((NewRandNum-n/2)/sqrt(n/12), freq=FALSE, col="blue", border="white")
s <- seq(-3,3,by=0.01);  points(s,dnorm(s), type="l")
```

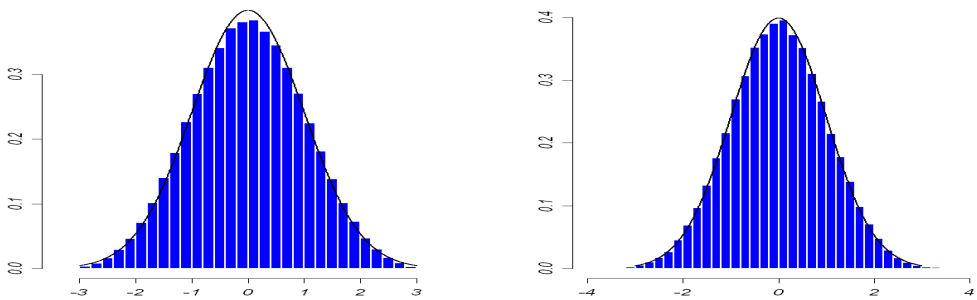


图 4.8: 产生 10^6 个 U[0, 1] 随机数，4 个一组相加得到 2.5×10^5 个新的随机数（左图），10 个一组相加得到 10^5 个新的随机数（右图），经过标准化后得到 N(0, 1) 的随机数的直方图。

4.2.2　正态分布、Laplace 分布、对数正态分布

正态分布的定义见例 2.8 和例 1.30，不再赘述。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，他于 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 $\phi(x|\mu,\sigma^2)$ 来逼进二项分布（见例 4.1 和附录 C）。

后来 Laplace 和 Gauss 把 $\phi(x|\mu,\sigma^2)$ 用于误差分析（误差被假定服从正态分布），因此以前的文献也把 $\phi(x|\mu,\sigma^2)$ 称为 Gauss-Laplace 分布。术语“正态分布”更晚些才出现，是统计学之父 K. Pearson 定名的。函数 $\phi(x)$ 在很多领域有着重要应用，譬如 $u(x,t)=\frac{1}{\sqrt{2\pi t}}\exp\{-\frac{x^2}{4t}\}$ 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。

正态分布 $X\sim \mathrm{N}(\mu,\sigma^2)$ 的期望和方差分别为 $\mathsf{E}(X)=\mu,\mathsf{V}(X)=\sigma^2$，偏度系数和峰度系数都是 0，特征函数为 $\varphi(t)=\exp(it\mu-\sigma^2t^2/2)$。除了性质 2.2，正态分布还有下面两个常见的漂亮结果。

定理 4.6. 如果随机变量 $X_j\sim \mathrm{N}(\mu_j,\sigma_j^2), j=1,2,\cdots,n$ 相互独立，则 $\sum_{j=1}^n X_j\sim \mathrm{N}(\sum_{j=1}^n\mu_j,\sum_{j=1}^n\sigma_j^2)$。

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 $\mathrm{N}(\sum_{j=1}^n\mu_j,\sum_{j=1}^n\sigma_j^2)$ 的特征函数。 □

下面的结果先被 P. Lévy 猜测到，后于 1936 年被 H. Cramér 证得。证明详见 W. Feller 的《概率论及其应用》下卷 [36] 第十五章第八节。

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。

Cramér-Lévy 定理中，X 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之结果。对正态分布更深刻的理解要等到学习中心极限定理（详见第五章）。

1812 年，P. S. Laplace 在《概率的分析理论》中提出了一个与正态分布类似的分布，它的密度函数是

$$f(x) = \frac{1}{2\sigma} \exp\left\{-\frac{|x-\mu|}{\sigma}\right\}, \sigma > 0 \qquad (4.13)$$

该分布记作 $X \sim \text{Laplace}(\mu, \sigma)$，称作 Laplace 分布。请读者验证它的特征函数、常见数字特征如下：

$$\varphi(t) = \frac{\exp(it\mu)}{1+\sigma^2 t^2} \text{ 并且 } \mathsf{E}(X) = \mathsf{M}(X) = \mu, \mathsf{V}(X) = 2\sigma^2, \gamma_1 = 0, \gamma_2 = 3 \quad (4.14)$$

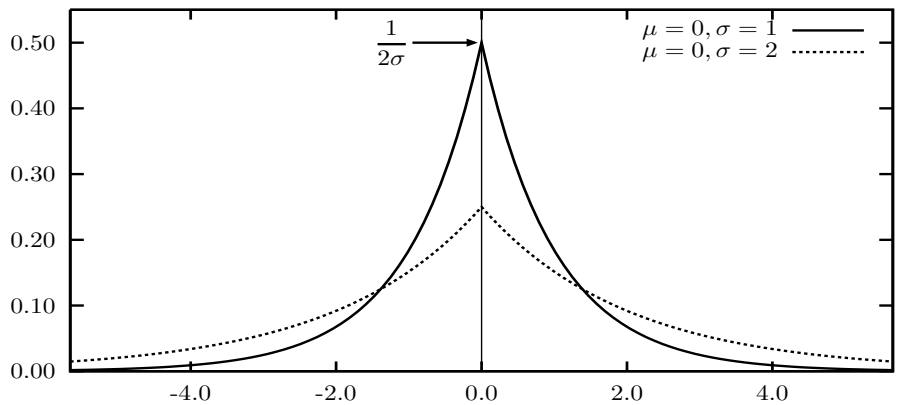


图 4.9: Laplace 分布的密度函数曲线，参数 σ 越小，曲线显得越“高瘦”。Laplace 分布又称为双侧指数分布，见下面的定理 4.8。

性质 4.8. 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。

$$\int_{-\infty}^{\infty} e^{itx} \frac{\exp(-|x|)}{2} dx = \frac{1}{1+t^2} \text{ 且 } \int_{-\infty}^{\infty} \frac{e^{-itx}}{\pi(1+t^2)} dt = \exp(-|x|)$$

定理 4.8. 如果 $X, Y \overset{iid}{\sim} \text{Expon}(\beta)$，则 $\mu + X - Y \sim \text{Laplace}(\mu, 1/\beta)$，其中指数分布 Expon($\beta$) 见定义 4.7。

证明. 先往证 $Z = X - Y \sim \text{Laplace}(0, 1/\beta)$：由例 2.19 知 Z 的密度函数为

$$f_Z(z) = \left\{ \begin{array}{ll} \int_0^\infty \beta e^{-\beta x} \beta e^{-\beta(x-z)} dx = \frac{\beta}{2} \exp(\beta z) & \text{当 } z < 0 \\ \int_z^\infty \beta e^{-\beta x} \beta e^{-\beta(x-z)} dx = \frac{\beta}{2} \exp(-\beta z) & \text{当 } z \geq 0 \end{array} \right\} = \frac{\beta}{2} \exp(-\beta |z|)$$

再由例 2.11 知 $\mu + Z$ 的密度函数为 $f_Z(z - \mu)$，得证。 □

定义 4.5 (对数正态分布)**.** 如果随机变量 $Y \sim \text{N}(\mu, \sigma^2)$，则称随机变量 $X = e^Y$ 服从对数正态分布，记作 $X \sim \text{logN}(\mu, \sigma^2)$，其中参数 $\sigma > 0$。该对数正态分布的密度函数为

$$f(x) = \left\{ \begin{array}{ll} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right\} & \text{当 } x > 0 \\ 0 & \text{当 } x \leq 0 \end{array} \right. \tag{4.15}$$

请读者验证 $\mathsf{E}(X) = \exp(\mu + \sigma^2/2), \mathsf{V}(X) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1], \gamma_1 = [\exp(\sigma^2)+2] \sqrt{\exp(\sigma^2) - 1}, \gamma_2 = \exp(4\sigma^2)+2\exp(3\sigma^2)+3\exp(2\sigma^2)-6, \mathsf{M}(X) = \exp(\mu)$，并证明以下性质。

性质 4.9. 对数正态分布的随机变量之积仍然是对数正态分布。

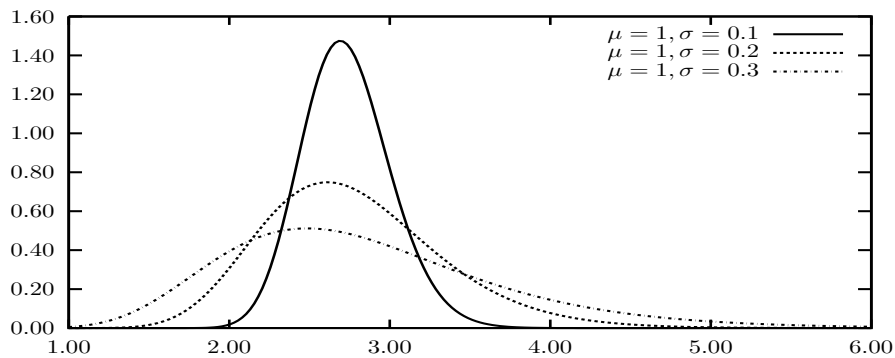


图 4.10: 对数正态分布的密度函数曲线，参数 σ 越小，曲线显得越“高瘦”。对数正态分布不能由各阶矩唯一确定。

4.2.3 Gamma 分布、χ^2 分布和指数分布

1729-1731 年，数学分析大师 L. Euler 研究了下面两个由积分定义的超越函数：

第一类 Euler 积分，$\forall \alpha > 0, \beta > 0$

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \qquad (4.16)$$

第二类 Euler 积分，$\forall s \in \mathbb{R}$ 且 $s \neq 0, -1, \cdots$

$$\Gamma(s) = \int_0^{+\infty} x^{s-1}e^{-x}dx \qquad (4.17)$$

为何要考虑 (4.17) 呢？这是因为 Euler 发现

$$n! = \int_0^1 (-\ln t)^n dt = \int_0^{\infty} x^n e^{-x}dx \text{（令 } t = -\ln x\text{）}$$

Euler 发现了性质 $\Gamma(s+1) = s\Gamma(s)$，特别地，当 $n \in \mathbb{N}$ 时 $\Gamma(n+1) = n!$，即超越函数 $\Gamma(s)$ 是对阶乘的推广。另外，$\Gamma(1/2) = \sqrt{\pi}$。

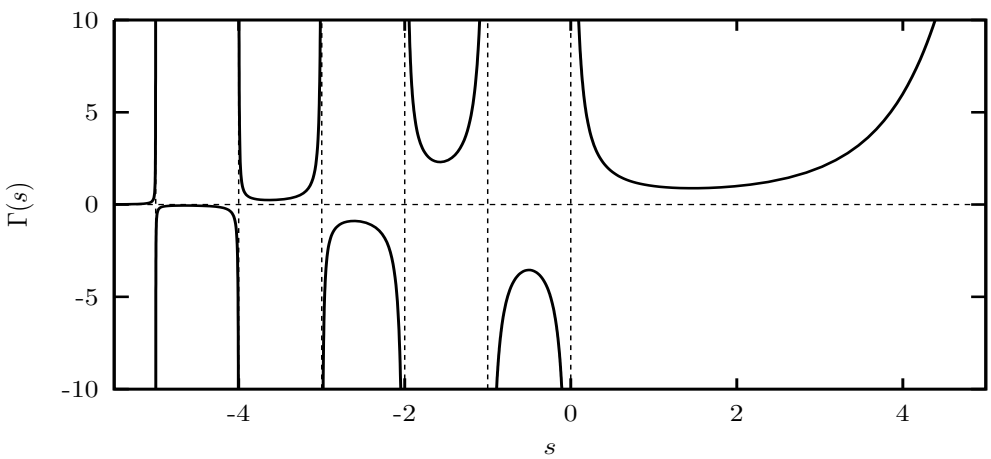


图 4.11: 1814 年，A. M. Legendre 将第二类 Euler 积分定名为 Gamma 函数并采用记号 $\Gamma(s)$。显然，$\Gamma(\alpha)/\beta^{\alpha} = \int_0^{+\infty} x^{\alpha-1}e^{-\beta x}dx$，其中 $\alpha > 0, \beta > 0$。

定义 4.6 (Gamma 分布)**.** 如果连续型随机变量 X 的密度函数为

$$g_{\alpha,\beta}(x)=\begin{cases}0 & \text{当 } x\le 0\\ \dfrac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & \text{当 } x>0,\alpha>0,\beta>0\end{cases}\tag{4.18}$$

则称 X 服从参数为 (α,β) 的 Gamma 分布，记作 $X\sim\text{Gamma}(\alpha,\beta)$。请验证 $\varphi(t)=(1-it/\beta)^{-\alpha},\mathsf{E}(X)=\alpha\beta^{-1},\mathsf{V}(X)=\alpha\beta^{-2},\gamma_1=2/\sqrt{\alpha},\gamma_2=6/\alpha$。

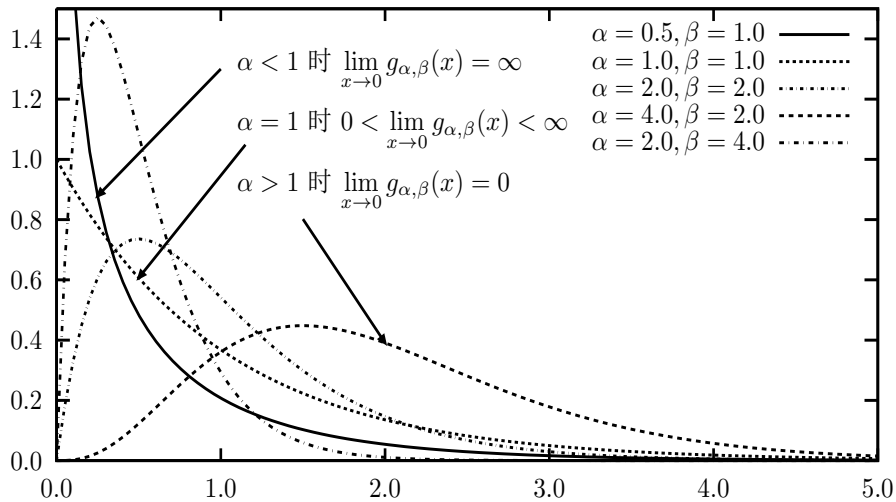


图 4.12: Gamma 分布的密度函数 $g_{\alpha,\beta}(x)$ 中，α 称为形状参数，它决定了 Gamma 分布的密度函数曲线的形状；β 称为尺度参数，当 $\alpha>0$ 固定时，β 越大曲线在 0 附近越“高瘦”。

练习 4.5. 试证明：$X\sim\text{Gamma}(\alpha,\beta)$ 的 k 阶矩为 $m_k=\beta^{-k}\alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k-1)$。

性质 4.10. 已知 $X_j\sim\text{Gamma}(\alpha_j,\beta),j=1,2,\cdots,n$ 相互独立，则 $\sum_{j=1}^{n}X_j\sim\text{Gamma}(\sum_{j=1}^{n}\alpha_j,\beta)$。

定义 4.7 (χ^2 分布与指数分布)**.** 把 Gamma 分布稍作限制，便可得到下面两个常见的分布，请读者写出它们的特征函数以及期望、方差等。

❑ 分布 $X\sim\text{Gamma}(\eta/2,1/2)$ 特称为自由度为 η 的 χ^2 分布*，记作

*国内的有些文献里把它译为“卡方”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。

$X \sim \chi^2_\eta$，其密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } x \le 0 \\ \dfrac{x^{\eta/2-1}e^{-x/2}}{2^{\eta/2}\Gamma(\eta/2)} & \text{当 } x > 0, \eta > 0 \end{cases} \tag{4.19}$$

❑ 分布 $X \sim \text{Gamma}(1,\beta)$ 特称为参数为 β 的指数分布，记作 $X \sim \text{Expon}(\beta)$，其分布函数为 $1 - \exp\{-\beta x\}$，密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } x \le 0 \\ \beta e^{-\beta x} & \text{当 } x > 0, \beta > 0 \end{cases} \tag{4.20}$$

性质 4.11. 若 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \text{N}(0,1)$，则 $X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2_n$。

证明. 往证 $Y = X_1^2$ 的密度函数为 $f(y) = \begin{cases} 0 & \text{若 } y \le 0 \\ e^{-x/2}/\sqrt{2\pi x} & \text{若 } y > 0 \end{cases}$

即 $Y \sim \chi^2_1$。再利用性质 4.10 可证得。 □

✎当 $n \to \infty$ 时，随机变量 χ^2_n, $\sqrt{2\chi^2_n}$ 和 $\sqrt[3]{\chi^2_n/n}$ 都趋向正态分布。χ^2_n 分布因 χ^2 统计量（见 §6.2.1）和拟合优度的 Pearson χ^2 检验（详见第 301 页的引理 8.2）而成为常见分布。有关 χ^2 分布的近似计算见定理 5.21。

性质 4.12. 对于指数分布 $X \sim \text{Expon}(\beta)$，有 $\mathsf{P}(X > t) = \exp\{-\beta t\}$。

指数分布常用于描述点子元器件的使用寿命，它具有“无记忆性”，即

↜ **定理 4.9** (指数分布无记忆)**.** 只取正值的连续型随机变量 X 服从指数分布当且仅当 $\mathsf{P}(X > s + t | X > s) = \mathsf{P}(X > t)$，其中 $s \ge 0, t \ge 0$。

证明. “⇐”的证明留作习题。下面往证“⇒”：

$$\mathsf{P}(X > s + t | X > s) = \frac{\mathsf{P}(X > s + t)}{\mathsf{P}(X > s)} = \frac{\exp\{-\beta(s+t)\}}{\exp\{-\beta s\}} = \mathsf{P}(X > t) \qquad \square$$

4.2.4　Beta 分布

第一类 Euler 积分 (4.16) 可由 Gamma 函数表示，即

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{4.21}$$

定义 4.8 (Beta 分布)**.** 如果连续型随机变量 X 的密度函数为

$$b_{\alpha,\beta}(x)=\begin{cases}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & \text{当 } 0<x<1,\alpha>0,\beta>0\\ 0 & \text{当 } x\le 0 \text{ 或 } x\ge 1\end{cases} \tag{4.22}$$

则称 X 服从参数为 (α,β) 的 Beta 分布，记作 $X\sim\text{Beta}(\alpha,\beta)$。

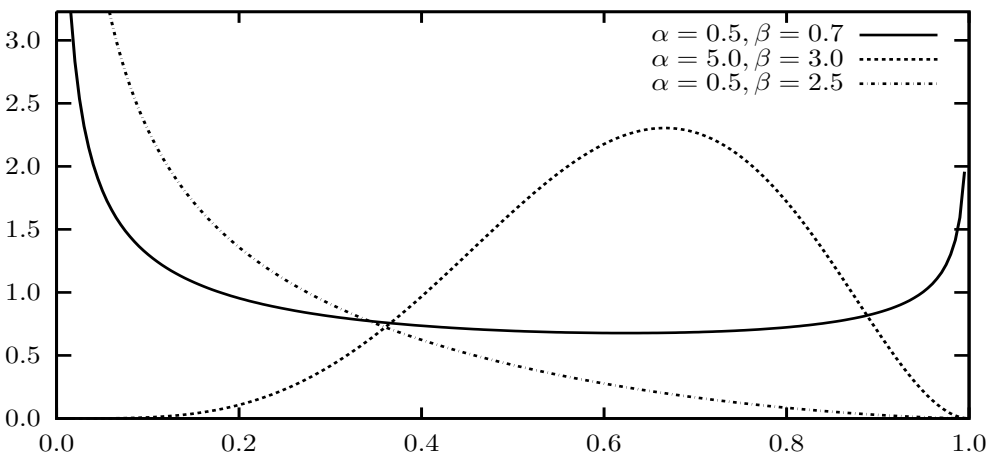


图 4.13: 当 $\alpha<1,\beta<1$ 时，密度曲线呈现 U 型。当 α,β 中只有一个小于 1 时，则密度曲线只有一头翘向无穷。极端的情形是：Beta(1, 1) = U(0, 1)。

按照定义，随机变量 $X\sim\text{Beta}(\alpha,\beta)$ 的 k 阶矩 $m_k=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}B(\alpha+k,\beta)=\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}$，于是期望 $\mathsf{E}(X)=\alpha/(\alpha+\beta)$，方差 $\mathsf{V}(X)=\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$。请读者利用 Maxima 验证 Beta$(\alpha,\beta)$ 分布的偏度系数和峰度系数分别为 $\gamma_1=2(\beta-\alpha)\sqrt{\alpha+\beta+1}/[(\alpha+\beta+2)\sqrt{\alpha\beta}]$ 和 $\gamma_2=3(\alpha+\beta+1)[2(\alpha+\beta)^2+\alpha\beta(\alpha+\beta-6)]/[\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)]$。

4.2.5 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》[42]，提出了 t 分布（亦称学生 t 分布），开小样本分析之先河（见第六章）。t 分布是统计学中最常用的分布之一。

定义 4.9 (t 分布)**.** 如果随机变量 $X \sim \mathrm{N}(0,1)$ 与 $Y \sim \chi^2_n$ 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 $t(n) = \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_n/n}}$。显然，$t(1)$ 就是 Cauchy$(0,1)$，见例 2.28。随机变量 $X \sim t(n)$ 的密度函数为

$$f_n(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \tag{4.23}$$

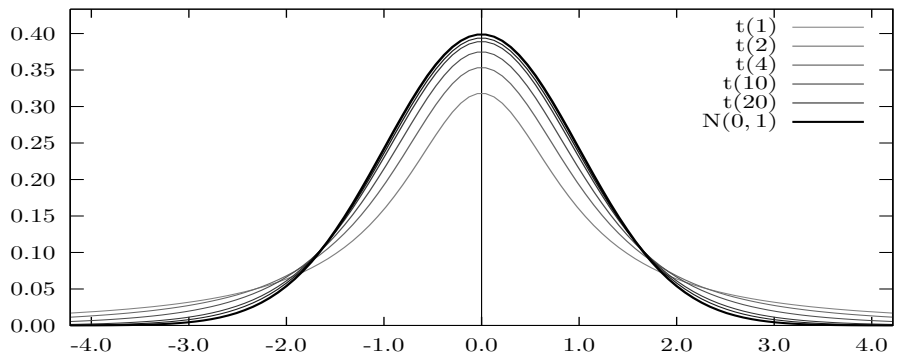


图 4.14: 随着 n 的增大，$t(n)$ 分布的密度函数越来越显得“高瘦”。当 n 趋向无穷时，$t(n)$ 分布的极限就是标准正态分布。

随机变量 $X \sim t(n)$ 的数字特征为 $\mathsf{E}(X) = \mathsf{M}(X) = 0, \mathsf{V}(X) = n/(n-2)$（其中 $n > 2$），$\gamma_1 = 0, \gamma_2 = 6/(n-4)$（其中 $n > 4$）。$t(n)$ 分布的 k 阶矩仅当 $k < n$ 时存在，且奇数阶矩都为 0。

定义 4.10 (F 分布)**.** 如果随机变量 $X \sim \chi^2_m$ 与 $Y \sim \chi^2_n$ 相互独立，则随机变量 $F = \frac{X/m}{Y/n}$ 的分布称为自由度为 (m,n) 的 F 分布，亦称 Fisher-

Snedecor 分布*，记作 $F \sim F_{m,n}$。为方便记忆，$F_{m,n}$ 分布的定义也简记作 $F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n}$。随机变量 $X \sim F_{m,n}$ 的密度函数为

$$f(x) = \frac{1}{xB(m/2,n/2)}\sqrt{\frac{(mx)^m n^n}{(mx+n)^{m+n}}}，\text{其中 } x>0，\text{参数 } m,n>0 \quad (4.24)$$

随机变量 $X \sim F_{m,n}$ 的期望、方差、偏度系数、峰度系数在一定条件下存在，用如下 Maxima 命令显示它们：$\mathsf{E}(X) = n/(n-2), \mathsf{V}(X) = 2n^2(n+m-2)/[m(n-4)(n-2)^2], \gamma_1 = (2m+n-2)\sqrt{8(n-4)}/[(n-6)\sqrt{m(m+n-2)}], \gamma_2 = 12[(n-2)^2(n-4)+m(m+n-2)(5n-22)]/[m(n-6)(n-8)(m+n-2)]$。

```
load(distrib) $
assume ( m > 0 and n > 2) $
mean_f(m,n) ;
assume ( m > 0 and n > 4) $
var_f(m,n) ;
assume ( m> 0 and n > 6 ) $
skewness_f(m,n) ;
assume ( m> 0 and n > 8 ) $
kurtosis_f (m,n) ;
```

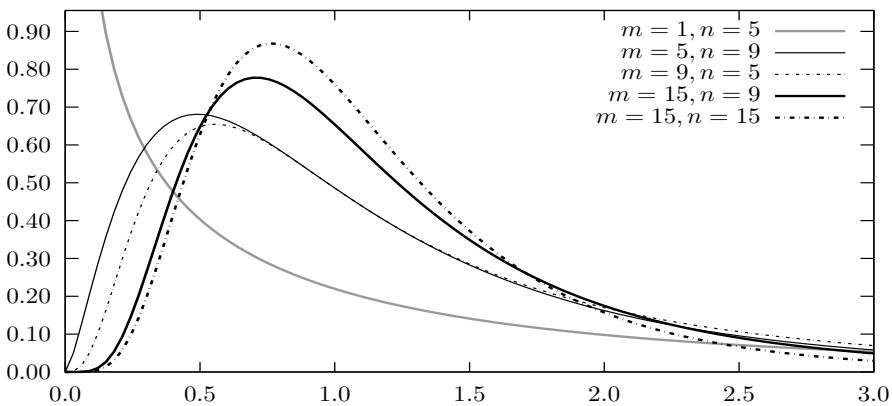


图 4.15: 不同自由度的 $F_{m,n}$ 分布之比较。从 F 分布的定义可知：若 $X \sim F_{m,n}$，则 $\frac{1}{X} \sim F_{n,m}$ 且 $\lim\limits_{n\to\infty} mX \sim \chi_m^2$。

*Fisher 在方差分析方面的研究工作与 F 分布有关，在此基础上美国生物统计学家 G. W. Snedecor 于 1934 年定义了 F 分布，并以 Fisher 的首字母命名了该分布。

4.2.6* 以物理学家命名的分布

本节所介绍的 Weibull 分布、Rayleigh 分布、Maxwell 分布和 Wigner 半圆分布都与物理学有关系。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》中用到了法国数学家 Maurice Fréchet (1878-1973) 于 1927 年提出的一个分布，后人将此分布称作 Weibull 分布，常用于可靠性分析和寿命数据分析 (life data analysis)。

定义 4.11 (Weibull 分布)**.** 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。它的密度函数是

$$f(x) = \begin{cases} 0 & \text{当 } x \le 0 \\ \alpha\lambda^{\alpha}x^{\alpha-1}\exp\{-(\lambda x)^{\alpha}\} & \text{当 } x > 0, \alpha > 0, \lambda > 0 \end{cases} \tag{4.25}$$

分布函数为 $F(x) = 1 - \exp\{-(\lambda x)^{\alpha}\}$，期望和方差分别为 $\mathsf{E}(X) = \lambda^{-1}\Gamma(1 + \alpha^{-1}), \mathsf{V}(X) = \lambda^{-2}[\Gamma(1 + \alpha^{-2}) - \Gamma^2(1 + \alpha^{-1})]$。

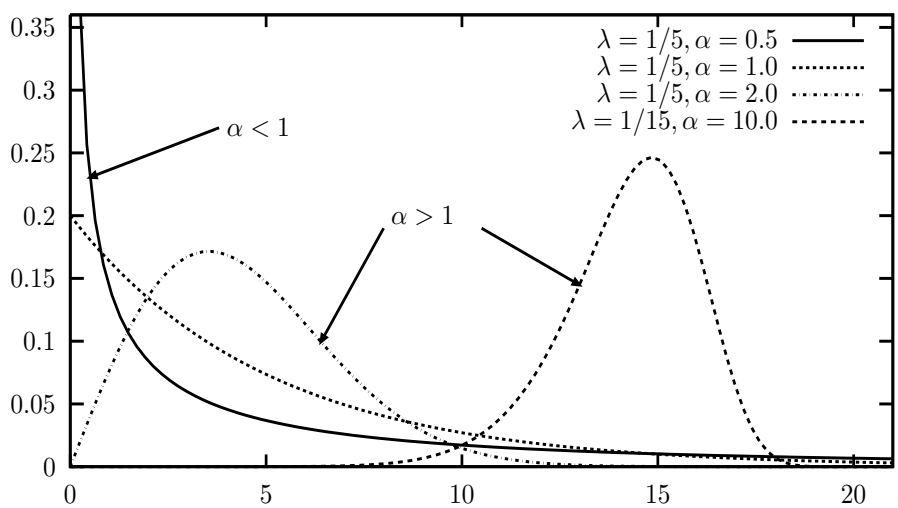


图 4.16: 取不同参数的 Weibull 分布的密度函数曲线之比较。显然，指数分布是 Weibull 分布的特例，因为 Expon(β) 就是 Weibull(β, 1)。

定义 4.12 (Rayleigh 分布)**.** 如果随机变量 $X_1, X_2 \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$，则随机变量 $X = \sqrt{X_1^2 + X_2^2}$ 服从的分布称为 Rayleigh 分布，记作 $X \sim \mathrm{Rayleigh}(\sigma)$，其中 $\sigma > 0$ 称为尺度参数。它的密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } x \le 0 \\ \dfrac{x}{\sigma^2} \exp\left(-\dfrac{x^2}{2\sigma^2}\right) & \text{当 } x > 0, \sigma > 0 \end{cases} \tag{4.26}$$

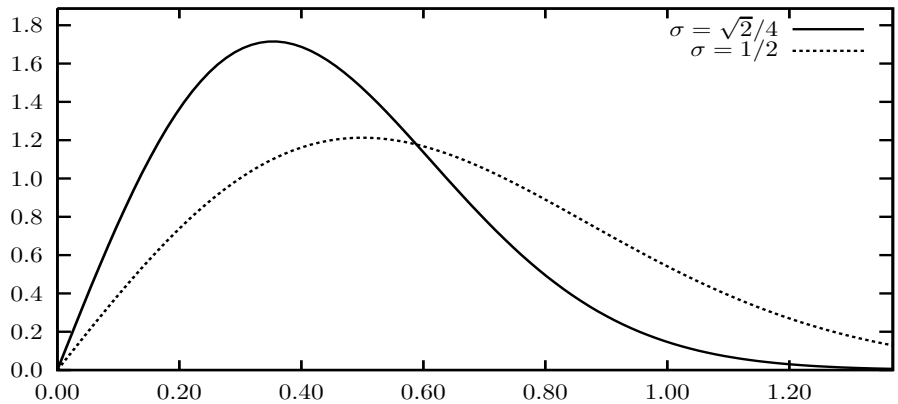


图 4.17: 取不同参数的 Rayleigh 分布的密度函数曲线之比较。显然，Rayleigh 分布也是 Weibull 分布的特例，因为 Rayleigh(σ) 就是 Weibull($\frac{1}{\sqrt{2}\sigma}, 2$)。

Rayleigh 分布是英国物理学家 Lord Rayleigh (1842-1919) 于 1880 年提出的。从它的定义可得到 $\mathsf{E}(X) = \sigma\sqrt{\pi/2}$ 和 $\mathsf{V}(X) = (4-\pi)\sigma^2/2$。请读者验证 Rayleigh 分布的下述性质。

性质 4.13. (1) 如果 $X \sim \mathrm{Rayleigh}(1)$，则 $X^2 \sim \chi_2^2$。(2) 若复数的实部和虚部独立同分布于 $\mathrm{N}(0, \sigma^2)$，则复数的模长服从 $\mathrm{Rayleigh}(\sigma)$。(3) 如果 $X \sim \mathrm{Expon}(\lambda)$，则 $Y = \sqrt{2\lambda\sigma^2 X} \sim \mathrm{Rayleigh}(\sigma)$。

定义 4.13 (Maxwell 分布)**.** 如果随机变量 $X_1, X_2, X_3 \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$，则随机变量 $X = \sqrt{X_1^2 + X_2^2 + X_3^2}$ 服从的分布称为 Maxwell 分布，记作 $X \sim$

Maxwell(σ)，其中 $\sigma > 0$ 称为尺度参数。它的密度函数为

$$f(x) = \begin{cases} 0 & \text{当 } x \le 0 \\ \sqrt{\dfrac{2}{\pi}} \dfrac{x^2}{\sigma^3} \exp\left(-\dfrac{x^2}{2\sigma^2}\right) & \text{当 } x > 0, \sigma > 0 \end{cases} \tag{4.27}$$

期望与方差分别为 $\mathsf{E}(X) = 2\sqrt{2}\sigma/\sqrt{\pi}, \mathsf{V}(X) = 3\sigma^2 - 8\sigma^2/\pi$。

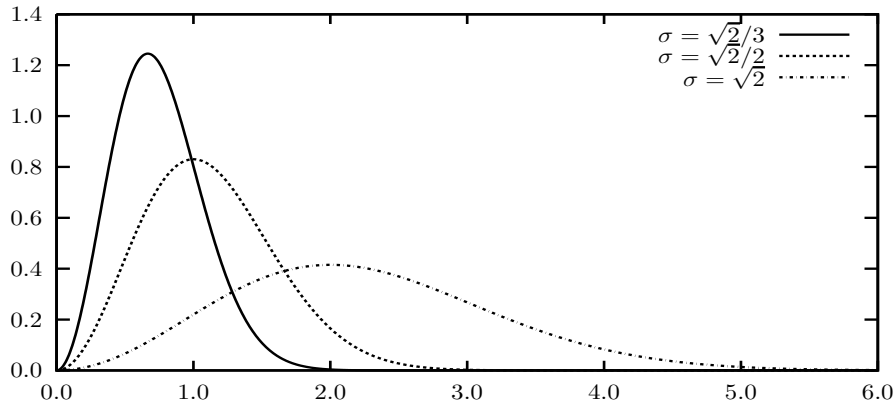


图 4.18: 取不同参数的 Maxwell 分布的密度函数曲线之比较。1859 年，英国著名的物理学家 J. C. Maxwell 证明了平衡态下理想气体分子的速率服从 Maxwell 分布，因此得名。

定义 4.14 (Wigner 半圆分布)**.** 当阶数趋近无穷时，许多随机对称阵的特征值 X 的极限分布的密度函数为

$$f(x) = \begin{cases} \dfrac{2}{\pi r^2} \sqrt{r^2 - x^2} & \text{若 } |x| < r \\ 0 & \text{若 } |x| \ge r \end{cases} \tag{4.28}$$

其中 $r > 0$。该分布以美籍匈牙利物理学家、数学家 Eugene Paul Wigner (1902-1995) 命名，称为 Wigner 半圆分布。X 服从 Wigner 半圆分布记作 $X \sim \text{Wigner}(r)$，它的期望和方差分别为 $\mathsf{E}(X) = 0, \mathsf{V}(X) = r^2/4$。

4.3　随机向量的分布

随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 实质上就是样本空间 $(\Omega,\mathscr{S})$ 到 n 维 Borel 空间 $(\mathbb{R}^n,\mathfrak{B}_n)$ 的可测函数，使得 $\boldsymbol{X}^{-1}(-\infty,\boldsymbol{x}]\in\mathscr{S}$，其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathsf{T}}$ 是任意 n 维向量，$(-\infty,\boldsymbol{x}]$ 表示区域 $(-\infty,x_1]\times(-\infty,x_2]\times\cdots\times(-\infty,x_n]\subseteq\mathbb{R}^n$。所以，随机向量也可看作取值为向量的随机变量。随机向量 $\boldsymbol{X}$ 的分布函数定义为 $F(\boldsymbol{x})=\mathsf{P}\{\boldsymbol{X}^{-1}(-\infty,\boldsymbol{x}]\}$。

矩阵是研究随机向量的常用工具。为了行文的方便，约定以分号表示换行，矩阵 $A=\left(\begin{smallmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\end{smallmatrix}\right)$ 和分块矩阵 $B=\left(\begin{smallmatrix}B_{11}&B_{12}\\B_{21}&B_{22}\end{smallmatrix}\right)$ 有时也记作 $A=(a_{11},a_{12},a_{13};a_{21},a_{22},a_{23})$ 和 $B=(B_{11},B_{12};B_{21},B_{22})$。下面由分块矩阵求逆的方法非常实用。

本节内容　第一、二小节介绍了多项分布和 Dirichlet 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。

学习目标　(1) 了解多项分布和 Dirichlet 分布的关系。(2) 掌握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。有关矩阵论的知识可参考 [41,51]。

4.3.1 多元正态分布

定义 4.15 (多元正态分布)**.** 如果连续型随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 使得内积 $\langle \boldsymbol{t},\boldsymbol{X}\rangle=\boldsymbol{t}^{\mathsf{T}}\boldsymbol{X}=t_1X_1+t_2X_2+\cdots+t_nX_n$ 是正态分布或常数，其中 $\boldsymbol{t}=(t_1,t_2,\cdots,t_n)^{\mathsf{T}}\in\mathbb{R}^n$ 是任意 n 维列向量，则称 $\boldsymbol{X}$ 服从 n 维正态分布。

性质 4.14. 正态分布的随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 的所有分量 $X_j,\ j=1,2,\cdots,n$ 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。由定义 3.2 和例 3.3，以及式 (2.100)，得到 $\boldsymbol{X}$ 的特征函数为

$$\begin{aligned}\varphi(\boldsymbol{t})&=\mathsf{E}\exp\{i\boldsymbol{t}^{\mathsf{T}}\boldsymbol{X}\}=\exp\left\{i\mathsf{E}(\boldsymbol{t}^{\mathsf{T}}\boldsymbol{X})-\frac{1}{2}\mathsf{V}(\boldsymbol{t}^{\mathsf{T}}\boldsymbol{X})\right\}\\&=\exp\left\{i\mathsf{E}(\boldsymbol{t}^{\mathsf{T}}\boldsymbol{X})-\frac{1}{2}\boldsymbol{t}^{\mathsf{T}}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{t}\right\}\end{aligned}$$

定理 4.10. 假设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 服从 n 维正态分布。约定：记 $\boldsymbol{X}$ 的均值为 $\mathsf{E}\boldsymbol{X}=\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_n)^{\mathsf{T}}=(\mathsf{E}X_1,\mathsf{E}X_2,\cdots,\mathsf{E}X_n)^{\mathsf{T}}$，记协方差矩阵为 $\Sigma=(\sigma_{ij})_{n\times n}=\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})$。则 $\boldsymbol{X}$ 的特征函数为

$$\varphi(\boldsymbol{t})=\exp\left\{i\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{t}^{\mathsf{T}}\Sigma\boldsymbol{t}\right\}\tag{4.29}$$

当 Σ 为正定矩阵时，行列式 $|\Sigma|>0$。利用 Lévy 的特征函数反演公式，得到 n 维正态分布随机向量 $\boldsymbol{X}$ 的密度函数如下：

$$\phi(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)=\frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}}{\sqrt{(2\pi)^n|\Sigma|}}\tag{4.30}$$

其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathsf{T}}$。该多元正态分布记作 $\boldsymbol{X}\sim\mathrm{N}(\boldsymbol{\mu},\Sigma)$，有时为强调维数也记作 $\boldsymbol{X}\sim\mathrm{N}_n(\boldsymbol{\mu},\Sigma)$。

练习 4.6. 例 2.13 所描述的随机向量 $(X,Y)^{\mathsf{T}}\sim\mathrm{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ 的密度函数 (2.29) 可表示为 (4.30) 的形式，其中 $\boldsymbol{\mu}=(\mu_X,\mu_Y)^{\mathsf{T}}$ 且 $(X,Y)^{\mathsf{T}}$ 的协方差矩阵 Σ 及其逆矩阵分别为

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \text{ 和 } \Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \sigma_X^{-2} & -\rho\sigma_X^{-1}\sigma_Y^{-1} \\ -\rho\sigma_X^{-1}\sigma_Y^{-1} & \sigma_Y^{-2} \end{pmatrix}$$

定理 4.11. 正态分布的随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T} \sim \mathrm{N}_n(\boldsymbol{\mu}, \Sigma)$ 经过线性变换所得的新的随机向量 $A_{m\times n}\boldsymbol{X}$ 仍为正态分布，其中 $m \leq n$。

$$\boldsymbol{Y} = A_{m\times n}\boldsymbol{X} \sim \mathrm{N}_m(A\boldsymbol{\mu}, A\Sigma A^\mathsf{T}) \tag{4.31}$$

证明. 在式 (4.29) 中令 $\boldsymbol{t} = A^\mathsf{T}\boldsymbol{s}$ 便得到随机向量 $\boldsymbol{Y}$ 的特征函数为

$$\begin{aligned}\varphi(s_1, s_2, \cdots, s_m) &= \mathsf{E}\left[\exp\left(i\boldsymbol{s}^\mathsf{T}\boldsymbol{Y}\right)\right] = \mathsf{E}\left[\exp\left(i\boldsymbol{s}^\mathsf{T}A\boldsymbol{X}\right)\right] = \mathsf{E}\left[\exp\left(i\boldsymbol{t}^\mathsf{T}\boldsymbol{X}\right)\right] \\ &= \exp\left\{i\boldsymbol{t}^\mathsf{T}\boldsymbol{X} - \frac{1}{2}\boldsymbol{t}^\mathsf{T}\Sigma\boldsymbol{t}\right\} = \exp\left\{i\boldsymbol{s}^\mathsf{T}(A\boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{s}^\mathsf{T}(A\Sigma A^\mathsf{T})\boldsymbol{s}\right\}\end{aligned}$$

具有如此特征函数的分布只有多元正态分布 $\mathrm{N}_m(A\boldsymbol{\mu}, A\Sigma A^\mathsf{T})$，得证。 □

性质 4.14 的逆命题不成立，即随机向量 $\boldsymbol{X}$ 的边缘分布都是正态分布并不能推导出 $\boldsymbol{X}$ 也服从正态分布。譬如，

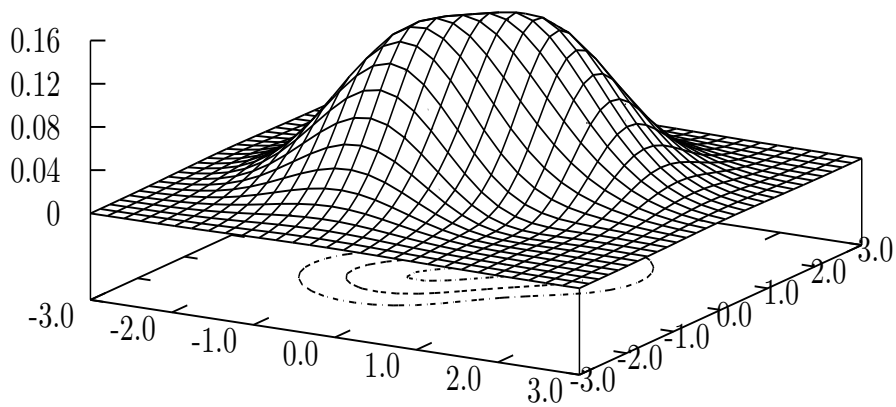


图 4.19: 密度函数为 $f(x,y) = \frac{1}{2\pi}\exp\left\{-\frac{x^2+y^2}{2}\right\}(1+\sin x \sin y)$ 的随机向量 $(X,Y)^\mathsf{T}$ 非正态分布，但它的任一边缘分布都是标准正态分布。

例 4.6. 设 n 维随机向量 $\boldsymbol{X} = (X_1, \cdots, X_m, X_{m+1}, \cdots, X_n)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ 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：$\boldsymbol{X}_{(1)} = (X_1, \cdots, X_m)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}_{(1)}, \Sigma_1)$ 且 $\boldsymbol{X}_{(2)} = (X_{m+1}, \cdots, X_n)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}_{(2)}, \Sigma_2)$。

证明. $\boldsymbol{X} = (X_1, \cdots, X_m, X_{m+1}, \cdots, X_n)^{\mathsf{T}}$ 的概率密度函数为

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma_1|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_{(1)} - \boldsymbol{\mu}_{(1)})^{\mathsf{T}} \Sigma_1^{-1} (\boldsymbol{x}_{(1)} - \boldsymbol{\mu}_{(1)})\right\} \times \frac{1}{\sqrt{(2\pi)^{n-m} |\Sigma_2|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)})^{\mathsf{T}} \Sigma_2^{-1} (\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)})\right\}$$

其中，$\boldsymbol{x} = (x_1, \cdots, x_n)^{\mathsf{T}}, \boldsymbol{x}_{(1)} = (x_1, \cdots, x_m)^{\mathsf{T}}, \boldsymbol{x}_{(2)} = (x_{m+1}, \cdots, x_n)^{\mathsf{T}}$。于是，$\boldsymbol{X}_{(1)} = (X_1, \cdots, X_m)^{\mathsf{T}}$ 的密度函数为

$$f(\boldsymbol{x}_{(1)}) = \int_{\mathbb{R}^{n-m}} f(\boldsymbol{x}) d\boldsymbol{x}_{(2)} = \frac{1}{\sqrt{(2\pi)^m |\Sigma_1|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_{(1)} - \boldsymbol{\mu}_{(1)})^{\mathsf{T}} \Sigma_1^{-1} (\boldsymbol{x}_{(1)} - \boldsymbol{\mu}_{(1)})\right\}$$

即 $(X_1, \cdots, X_m)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}_{(1)}, \Sigma_1)$，同理 $(X_{m+1}, \cdots, X_n)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}_{(2)}, \Sigma_2)$。 □

性质 4.15. 设 $(X_1, \cdots, X_m, \cdots, X_n)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$，试证明：对于 $1 < m < n$，皆有 $(X_1, \cdots, X_m)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}_{(1)}, \Sigma_{11})$，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 $\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} I_m & O \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$，其中 I_m 是 m 阶单位矩阵，则 $(Y_1, \cdots, Y_m) = (X_1, \cdots, X_m)$。由定理 4.11 和例 4.6 可证得，这是因为

$$\begin{pmatrix} I_m & O \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_m & O \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix}^{\mathsf{T}} = \begin{pmatrix} \Sigma_{11} & O \\ O & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix} \square$$

4.3.2　多项分布

多项分布是二项分布向高维的推广，它来自这样的概率模型：在 n 次重复的独立试验中，令 X_j 表示事件 A_j 发生的次数，其中事件 $A_j, j = 1, 2, \cdots, k$ 两两互斥且每次试验中 $\mathsf{P}(A_j) = p_j$，多项分布刻画了事件 A_j 发生了 n_j 次的概率。

定义 4.16 (多项分布)**.** 如果离散型随机变量 $X_1, X_2, \cdots, X_k$ 的联合分布的概率函数为

$$\mathsf{P}(X_1 = n_1, X_2 = n_2, \cdots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \tag{4.32}$$

即 $(p_1 + p_2 + \cdots + p_k)^n$ 多项式展开中的一般项，其中 $n_1 \geq 0, n_2 \geq 0, \cdots, n_k \geq 0$ 且 $n_1 + n_2 + \cdots + n_k = n, p_1 + p_2 + \cdots + p_k = 1$，则称随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_k)^{\mathsf{T}}$ 服从多项分布 (multinomial distribution)，记作 $\boldsymbol{X} \sim \text{Multin}(n; p_1, p_2, \cdots, p_k)$。

性质 4.16. 多项分布 $\boldsymbol{X} \sim \text{Multin}(n; p_1, p_2, \cdots, p_k)$ 的特征函数为

$$\varphi(t_1, t_2, \cdots, t_k) = (p_1 e^{it_1} + p_2 e^{it_2} + \cdots + p_k e^{it_k})^n \tag{4.33}$$

期望和协方差阵分别为 $\mathsf{E}(\boldsymbol{X}) = (np_1, np_2, \cdots, np_k)^{\mathsf{T}}$ 和 $\Sigma = (\sigma_{ij})$，其中

$$\sigma_{ij} = \begin{cases} np_i(1 - p_i) & \text{若 } i = j \\ -np_i p_j & \text{若 } i \neq j \end{cases}$$

例 4.7. 设随机向量 $(X_1, X_2, \cdots, X_k)^{\mathsf{T}} \sim \text{Multin}(n; p_1, p_2, \cdots, p_k)$，对 $1 \leq r \leq k$，求 $Y = X_1 + X_2 + \cdots + X_r$ 的概率分布。

解. 离散型随机变量 Y 的所有可能的取值是 $0, 1, \cdots, n$，其对应的概率就是事件 $A = \sum_{j=1}^{r} A_j$ 分别发生 $0, 1, \cdots, n$ 次的概率，所以 $Y \sim \mathrm{B}(n, p_1 + p_2 + \cdots + p_r)$。

4.3.3* Dirichlet 分布

Dirichlet*分布是 Beta 分布向高维的推广，它对非参数统计学中次序统计量理论非常重要（见第十章），也常作为多项分布的共轭先验分布出现在贝叶斯数据分析中（见第十一章）。

定义 4.17 (Dirichlet 分布)**.** 如果连续型随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 的密度函数为

$$f(x_1, x_2, \cdots, x_n) = \begin{cases} \dfrac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)} \prod_{j=1}^{n} x_j^{\alpha_j - 1} & \text{其中 } x_j \geq 0 \text{ 满足 } \sum_{j=1}^{n} x_j = 1, \alpha_j > 0 \\ 0 & \text{其他} \end{cases}$$

则称 $\boldsymbol{X}$ 服从 Dirichlet 分布，记作 $\boldsymbol{X} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \cdots, \alpha_n)$。显然，$n = 2$ 时就是 $\text{Beta}(\alpha_1, \alpha_2)$ 分布。

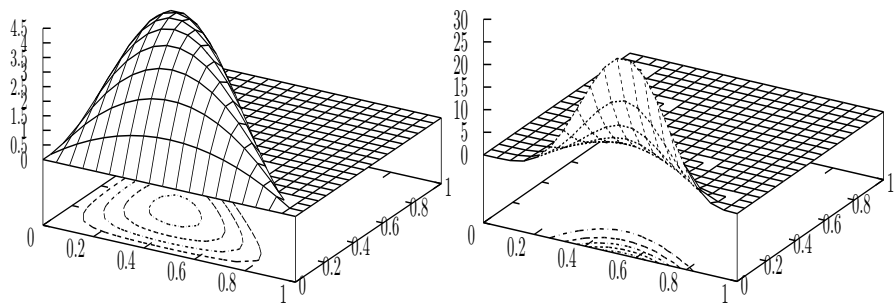


图 4.20: Dirichlet 分布 $(X, Y, 1 - X - Y)^\mathsf{T} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ 的密度函数曲面 $z = f(x, y, 1 - x - y)$。左图参数 $(\alpha_1, \alpha_2, \alpha_3)$ 为 $(1, 1, 1)$，右图参数为 $(5, 1/3, 4)$。

性质 4.17. 已知 $(X_1, X_2, \cdots, X_n)^\mathsf{T} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \cdots, \alpha_n)$，则边缘分布 $X_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j), j = 1, 2, \cdots, n$，其中 $\alpha_0 = \sum_{j=1}^{n} \alpha_j$，并且 $\text{Cov}(X_j, X_k) = -\alpha_j \alpha_k / [\alpha_0^2(\alpha_0 + 1)]$。

定理 4.12. 已知 $Y_j \sim \text{Gamma}(\alpha_j, \beta), j = 1, 2, \cdots, n$ 相互独立，定义 $Y = \sum_{j=1}^{n} Y_j, X_j = Y_j / Y$，则 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \cdots, \alpha_n)$。

*Gustav Lejeune Dirichlet (1805-1859) 是德国数学家，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。

证明. 变换 $X_j = Y_j/Y, j = 1,2,\cdots,n$ 的逆变换是 $Y_1 = YX_1,\cdots,Y_{n-1} = YX_{n-1}, Y_n = Y(1 - X_1 - X_2 - \cdots - X_{n-1})$，其雅可比行列式为

$$J\left(\frac{y_1,y_2,\cdots,y_n}{y,x_1,\cdots,x_{n-1}}\right) = \begin{vmatrix} x_1 & x_2 & \cdots & x_{n-1} & 1-\sum_{j=1}^{n-1} x_j \\ y & 0 & \cdots & 0 & -y \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & y & -y \end{vmatrix} = y^{n-1}$$

从 $\boldsymbol{Y} = (Y_1,Y_2,\cdots,Y_n)^\mathsf{T}$ 的联合密度函数 $f(y_1,y_2,\cdots,y_n) = \prod_{j=1}^n \frac{\beta^{\alpha_j}}{\Gamma(\alpha_j)} y_j^{\alpha_j-1} e^{-\beta y_j}$ 得到 $(Y,X_1,X_2,\cdots,X_{n-1})^\mathsf{T}$ 的密度函数为

$$\prod_{j=1}^{n-1} \frac{1}{\Gamma(\alpha_j)} x_j^{\alpha_j-1} \left(1-\sum_{j=1}^{n-1} x_j\right)^{\alpha_n-1} \beta^{\sum_{j=1}^n \alpha_j} y^{\sum_{j=1}^n \alpha_j - 1} e^{-\beta y}$$

由性质 4.10 知 $Y = \sum_{j=1}^n Y_j \sim \text{Gamma}(\sum_{j=1}^n \alpha_j, \beta)$，所以 $(X_1,X_2,\cdots,X_{n-1})^\mathsf{T}$ 的密度函数为

$$\begin{aligned} &\prod_{j=1}^{n-1} \frac{1}{\Gamma(\alpha_j)} x_j^{\alpha_j-1} \left(1-\sum_{j=1}^{n-1} x_j\right)^{\alpha_n-1} \int_0^\infty \beta^{\sum_{j=1}^n \alpha_j} y^{\sum_{j=1}^n \alpha_j - 1} e^{-\beta y} dy \\ &= \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \prod_{j=1}^n x_j^{\alpha_j-1}，\text{其中 } x_j \geq 0 \text{ 满足 } \sum_{j=1}^n x_j = 1 \end{aligned}$$ □

4.3.4* Wishart 分布

1928 年，英国统计学家 John Wishart (1898-1956) 首次提出了 Wishart 分布，可把它比喻成 χ_n^2 分布的矩阵推广。Wishart 分布是由多元正态分布导出的，在多元统计学中非常重要，因为它恰是多元正态分布协方差阵最大似然估计的概率分布 [53,74]。该分布的发现揭开了多元统计学的篇章。

定义 4.18. 已知 d 维随机向量 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n \overset{iid}{\sim} \mathrm{N}_d(\boldsymbol{0}, \Sigma)$，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 $\boldsymbol{W} = \sum_{j=1}^n \boldsymbol{X}_j \boldsymbol{X}_j^{\mathsf{T}}$ 服从自由度为 n 的 Wishart 分布*，记作 $\boldsymbol{W} \sim \mathrm{Wishart}_n(\Sigma, d)$，当维数 d 众所周知时也简记作 $\boldsymbol{W} \sim \mathrm{Wishart}_n(\Sigma)$。显然，当 Σ 退化为 1 时，$\boldsymbol{W} \sim \chi_n^2$。当 $n \geq d$ 且 Σ 正定时，称分布 $\mathrm{Wishart}_n(\Sigma, d)$ 是非退化的，此时它具有密度函数

$$f_W(W) = \frac{|W|^{(n-d-1)/2} \exp[-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}W)]}{2^{nd/2}\pi^{d(d-1)/4}|\Sigma|^{n/2}\prod_{j=1}^d \Gamma\left(\frac{n+1-j}{2}\right)}, \text{ 其中 } W \text{ 正定} \tag{4.34}$$

性质 4.18. 如果随机矩阵 $\boldsymbol{W}_1 \sim \mathrm{Wishart}_{n_1}(\Sigma)$ 与 $\boldsymbol{W}_2 \sim \mathrm{Wishart}_{n_2}(\Sigma)$ 相互独立，则 $\boldsymbol{W}_1 + \boldsymbol{W}_2 \sim \mathrm{Wishart}_{n_1+n_2}(\Sigma)$。

定理 4.13. 已知 $\boldsymbol{W} \sim \mathrm{Wishart}_n(\Sigma, d)$ 非退化且 C 是一个 $p \times d$ 矩阵，则

$$C\boldsymbol{W}C^{\mathsf{T}} \sim \mathrm{Wishart}_n(C\Sigma C^{\mathsf{T}}, p) \tag{4.35}$$

特别地，令 $\boldsymbol{c} \in \mathbb{R}^d$ 使得 $\sigma^2 = \boldsymbol{c}^{\mathsf{T}}\Sigma\boldsymbol{c} \neq 0$，则 $\boldsymbol{c}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{c}/\sigma^2 \sim \chi_n^2$。

证明. $C\boldsymbol{W}C^{\mathsf{T}} = \sum_{j=1}^n C\boldsymbol{X}_j(C\boldsymbol{X}_j)^{\mathsf{T}}$，因为 $C\boldsymbol{X}_j \sim \mathrm{N}_p(\boldsymbol{0}, C\Sigma C^{\mathsf{T}})$ 得证。 □

性质 4.19. 分布 $\boldsymbol{W} \sim \mathrm{Wishart}_n(\Sigma, d)$ 的特征函数为 $\varphi(T_{d\times d}) = \mathsf{E}\{i\langle T, \boldsymbol{W}\rangle\} = \mathsf{E}\{i \cdot \mathrm{tr}(T\boldsymbol{W})\} = |I_d - 2iT\Sigma|^{-n/2}$，其中 I_d 是 $d \times d$ 单位阵。

*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。

4.4　习题

4.1. 设一个人在一年中患感冒的次数 $X \sim \text{Poisson}(5)$，某药品对 75% 的人来说可将上述参数减小为 3，而对另外 25% 的人则是无效的。若某人试用此药一年，在试用期间患了两次感冒，问此药对他有效的概率是多少？

4.2. 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。

4.3. 设随机变量 $X \sim \text{U}(0,1)$。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？

4.4. 设随机变量 $X \sim \text{U}[-1/2, 1/2]$，令 $g(x) = \begin{cases} \ln x & \text{当 } x > 0 \\ 0 & \text{当 } x \le 0 \end{cases}$，求 $Y = g(X)$ 的期望与方差。

4.5. 设 $X \sim \text{U}[0,1]$，求单调增函数 $h(x)$ 使得 $Y = h(X) \sim \text{Expon}(\beta)$。

☆ 4.6. 已知 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \text{Expon}(1)$，试证明 $2(X_1 + X_2 + \cdots + X_n) \sim \chi^2_{2n}$。

☆ 4.7. 已知 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \text{U}(0,1)$，试证明 $-2\ln(X_1 X_2 \cdots X_n) \sim \chi^2_{2n}$。

4.8. 设随机变量 $X \sim \text{N}(0,1)$，求 X^n 的数学期望和方差，其中 $n \in \mathbb{N}$。

★ 4.9. 设随机变量 $X, Y \overset{iid}{\sim} \text{N}(\mu, \sigma^2)$，求 $\mathsf{E}[\max(X,Y)]$ 和 $\mathsf{E}[\min(X,Y)]$。

4.10. 验证图 4.3.1 中给出的联合密度 $f(x,y)$ 使得边缘分布为正态分布。

4.11. 设随机变量 X 的密度函数为 $f(x) = e^{-|x|}/2$，其中 $-\infty < x < +\infty$。求 $\mathsf{E}(|X|)\mathsf{V}(|X|)$ 和 $\mathsf{Cov}(X, |X|)$，并判定 X 与 $|X|$ 是否独立。

4.12. 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。

☆ 4.13. 设随机变量 X 与 Y 独立， $X \sim \text{Gamma}(p,\beta), Y \sim \text{Gamma}(q,\beta), \beta > 0, p > 0, q > 0$, 求 X/Y 与 $X/(X+Y)$ 的密度函数。

★ 4.14. 证明定理 4.9。

☆ 4.15. 已知随机变量 $X_1, \cdots, X_n \overset{iid}{\sim} \text{Expon}(\beta)$，令 $U = \max(X_1, \cdots, X_n)$ 和 $V = \min(X_1, \cdots, X_n)$，(1) 试求 U 的密度函数；(2) 试证 $V \sim \text{Expon}(\beta n)$。

4.16. 已知 $X \sim \text{Expon}(\beta)$，如果 $\mathsf{P}\{X \geq 1\} = \mathsf{P}\{X \leq 1\}$，求 $\sum_{k=1}^{\infty} \mathsf{P}\{X \geq k\}$。

4.17. 设 $X \sim \chi_m^2$ 与 $Y \sim \chi_n^2$ 相互独立，求 $Z = X/Y$ 的密度函数。

4.18. 求证：若 $T \sim t(n)$，则 $T^2 \sim F(1,n)$。

4.19. 设随机变量 $X \sim \text{Rayleigh}(\sigma)$，求 $\mathsf{E}(1/X)$。

☆ 4.20. 设 $(X,Y)^{\mathsf{T}}$ 服从二维正态分布，且有 $\mathsf{V}(X) = \sigma_X^2$，$\mathsf{V}(Y) = \sigma_Y^2$。证明：当 $a^2 = \sigma_X^2/\sigma_Y^2$ 时，$W = X - aY$ 与 $V = X + aY$ 相互独立。

4.21. 设 $(X,Y,Z)^{\mathsf{T}} \sim \text{N}(\boldsymbol{\mu}, \Sigma)$,其中 $\boldsymbol{\mu} = (3,5,7)^{\mathsf{T}}, \Sigma = (8,3,2;3,4,1;2,1,2)$，求 $X+Y$ 的密度函数。

4.22. 设 $\boldsymbol{X} \sim \text{N}(\boldsymbol{\mu}, \Sigma)$ ，给出 $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{\alpha}$ 服从的分布。

4.23. 如果 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathsf{T}} \sim \text{N}(\boldsymbol{\mu}, \Sigma)$ ，则 $X_1, X_2, \cdots, X_n$ 相互独立的充分必要条件是 $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$，其中 σ_j^2 是 X_j 的方差，$j = 1, 2, \cdots, n$。

4.24. 已知 $(X_1, X_2, \cdots, X_n)^{\mathsf{T}} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \cdots, \alpha_n)$，试证明：$\mathsf{E}(X_j) = \alpha_j/\alpha_0, \mathsf{V}(X_j) = \alpha_j(\alpha_0 - \alpha_j)/[\alpha_0^2(\alpha_0+1)]$，其中 $\alpha_0 = \sum_{j=1}^n \alpha_j$。

4.25. 求分布 $\boldsymbol{W} \sim \text{Wishart}_n(\Sigma, d)$ 的期望。

第五章

大数律与中心极限定理

十七世纪末，瑞士数学家 J. Bernoulli 发现了 Bernoulli 弱大数律*，第一次对随机事件 A 的概率 $\mathsf{P}(A)$ 给出了频率解释，即在 n 次独立的重复试验中 A 出现的频率 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

定理 5.1 (Poisson, 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

$$\lim_{n\to\infty} \mathsf{P}\left\{\left|\frac{m}{n} - \frac{p_1 + \cdots + p_n}{n}\right| \le \epsilon\right\} = 1 \tag{5.1}$$

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。如果定义随机变量 X_j 如下：

$$X_j = \begin{cases} 1 & \text{若事件 } A \text{ 在第 } j \text{ 次试验中出现} \\ 0 & \text{若事件 } A \text{ 在第 } j \text{ 次试验中不出现} \end{cases} \tag{5.2}$$

则在 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：$\lim_{n\to\infty} \mathsf{P}\{|\frac{1}{n}\sum_{j=1}^{n} X_j - \frac{1}{n}\sum_{j=1}^{n} \mathsf{E}X_j| \le \epsilon\} = 1$。1867

*见第一章的定理 1.1 和例 1.34 对该定理给出的模拟试验解释。

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)**.** 已知 $\{X_n\}_{n=1}^{\infty}$ 是一个随机变量序列，如果 $\forall \epsilon > 0$，

$$\lim_{n\to\infty} \mathsf{P}\left\{\left|\frac{1}{n}\sum_{j=1}^{n} X_j - \frac{1}{n}\sum_{j=1}^{n} \mathsf{E}X_j\right| \le \epsilon\right\} = 1 \tag{5.3}$$

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，当且仅当 $\forall \epsilon > 0$，$\lim\limits_{n\to\infty} \mathsf{P}\{|X_n - X| \le \epsilon\} = 1$，记为 $X_n \xrightarrow{P} X$。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 $Y_n = \frac{1}{n}\sum_{j=1}^{n} X_j - \frac{1}{n}\sum_{j=1}^{n} \mathsf{E}X_j \xrightarrow{P} 0$。

练习 5.1. 离散型随机变量 X_n 的概率函数为 $\mathsf{P}(X_n = k/n - 1/2) = C_n^k/2^n$，其中 $k = 0, 1, \cdots, n-1, n$。试证明：$X_n \xrightarrow{P} 0$（参考定理 1.1 的证明）。

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，气体压力来自单位时间内撞击单位面积上的分子的总体效果，分子的撞击次数和速度都是随机的，但由于大数律压力表现得几乎为常数。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 $\frac{1}{n}\sum_{j=1}^{n} X_j - \frac{1}{n}\sum_{j=1}^{n} \mathsf{E}X_j \xrightarrow{P} 0$。在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。

定理 5.2. 如果 $Y_n \xrightarrow{P} 0$，则 $Y_n \xrightarrow{L} 0$，即 $\forall y \in \mathbb{R}\backslash\{0\}$ 有

$$\lim_{n\to\infty} F_n(y) = \begin{cases} 0 & \text{当 } y < 0 \\ 1 & \text{当 } y \geq 0 \end{cases}$$

证明. 由 $Y_n \xrightarrow{P} 0$ 有 $\forall \epsilon > 0$，当 $n \to \infty$时，$\mathsf{P}(Y_n \leq -\epsilon) = F_n(-\epsilon) \to 0$ 且 $\mathsf{P}(Y_n \geq \epsilon) = 1 - F_n(\epsilon) + \mathsf{P}(Y_n = \epsilon) \to 0$，于是 $F_n(\epsilon) \to 1$。 □

性质 5.1. 依概率收敛、依分布收敛具有以下性质：

❶ 若 $X_n \xrightarrow{P} X$，则 $X_n \xrightarrow{L} X$。由定理 5.2 可证，留作练习。

❷ 若 $X_n \xrightarrow{L} x_0$（其中 x_0 是常数），且函数 $g(x)$ 在 $x = x_0$ 处连续，则 $g(X_n) \xrightarrow{P} g(x_0)$。特别地，$X_n \xrightarrow{L} x_0 \Rightarrow X_n \xrightarrow{P} x_0$。

❸ Slutsky 定理：若 $X_n \xrightarrow{L} X$ 且 $Y_n \xrightarrow{P} y_0$（常数），则有 $X_n + Y_n \xrightarrow{L} X + y_0$ 且 $Y_n X_n \xrightarrow{L} y_0 X$。

*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。

证明. 性质 ❷、❸ 的证明见俄国数学家 A. N. Shiryayev 的《概率论》[84] 第二章第十节。 □

由性质 5.1 中的 ❷，要证明 $\frac{1}{n}\sum_{j=1}^{n}X_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}X_j\xrightarrow{P}0$，只需证明 $\frac{1}{n}\sum_{j=1}^{n}X_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}X_j\xrightarrow{L}0$ 即可。很自然地，利用 Lévy 连续性定理，把证明转移到特征函数的收敛性上。

例 5.1. 设 $\{X_j\}_{j=1}^{\infty},\{Y_j\}_{j=1}^{\infty}$ 为两个满足弱大数律的随机变量序列，令 $Z_j=X_j+Y_j, j=1,2,\cdots$。试证明：$\{Z_j\}_{j=1}^{\infty}$ 也满足弱大数律。

证明. 根据 Slutsky 定理，从 $\frac{1}{n}\sum_{j=1}^{n}X_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}X_j\xrightarrow{P}0$ 和 $\frac{1}{n}\sum_{j=1}^{n}Y_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}Y_j\xrightarrow{P}0$ 可得 $\frac{1}{n}\sum_{j=1}^{n}Z_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}Z_j\xrightarrow{L}0$。 □

例 5.2. 设随机变量 $X,X_1,X_2,\cdots\overset{iid}{\sim}\frac{1}{2}\langle 1\rangle+\frac{1}{2}\langle 0\rangle$，因此 $X_n\xrightarrow{L}X$。下面说明 $\{X_n\}$ 不依概率收敛于 X。$\forall\epsilon\in(0,1)$ 皆有

$$\begin{aligned}\mathsf{P}(|X_n-X|\geq\epsilon)&=\mathsf{P}(X_n=1,X=0)+\mathsf{P}(X_n=0,X=1)\\&=\mathsf{P}(X_n=1)\mathsf{P}(X=0)+\mathsf{P}(X_n=0)\mathsf{P}(X=1)=\frac{1}{2}\end{aligned}$$

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 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。最后介绍了比强大数律更精细的重对数律。

学习目标 (1) 掌握依分布收敛、依概率收敛、几乎必然收敛的性质；(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法。

5.1.1　弱大数律

定理 5.3 (Chebyshev)**.** 对任意 $n \in \mathbb{N}$，如果随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立且 $\mathsf{V}(X_j) \leq c, j = 1, 2, \cdots, n, \cdots$，其中 c 为一个有限的常数（即 X_j 的方差一致有界），则随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 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

$$\mathsf{V}(Y_n) = \mathsf{V}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{j=1}^{n}\mathsf{V}(X_j) \leq \frac{c}{n}$$

由 Chebyshev 不等式得到 $\mathsf{P}\{|Y_n - \mathsf{E}(Y_n)| \leq \epsilon\} \geq 1 - \mathsf{V}(Y_n)/\epsilon^2$，令 $n \to \infty$ 即得证。 □

定理 5.4 (Markov)**.** 如果随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足条件 $\lim\limits_{n\to\infty} \frac{1}{n^2}\mathsf{V}\left(\sum_{j=1}^{n} X_j\right) = 0$，则 $\{X_j\}_{j=1}^{\infty}$ 满足弱大数律。

练习 5.2. 仿照 Chebyshev 弱大数律的证明给出 Markov 弱大数律的证明，并说明 Markov 弱大数律 $\Rightarrow$ Chebyshev 弱大数律 $\Rightarrow$ Poisson 弱大数律 $\Rightarrow$ Bernoulli 弱大数律。

例 5.3. 设随机变量序列 $\{X_k\}_{k=1}^{\infty}$ 的方差一致有界，即存在常数 $c > 0$ 使得 $\mathsf{V}(X_k) \leq c, k = 1, 2, \cdots$，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。

解. 由于当 $|k-j| \geq 2$ 时，X_k 和 X_j 不相关，故当 $|k-j| \geq 2$ 时 $\mathsf{Cov}(X_k, X_j) = 0$。所以，$|\mathsf{Cov}(X_k, X_{k+1})| = |\rho(X_k, X_{k+1})|\sqrt{\mathsf{V}(X_k)\mathsf{V}(X_{k+1})} \leq c$。从而有

$$\begin{aligned}\frac{1}{n^2}\mathsf{V}\left(\sum_{k=1}^{n} X_k\right) &= \frac{1}{n^2}\left[\sum_{k=1}^{n}\mathsf{V}(X_k) + 2\sum_{k=1}^{n-1}\mathsf{Cov}(X_k, X_{k+1})\right] \\ &\leq \frac{1}{n^2}[nc + 2(n-1)c] \leq 3c/n\end{aligned}$$

利用 Markov 弱大数律可证 $\{X_k\}_{k=1}^{\infty}$ 满足弱大数律。

由 Lévy 连续性定理，我们可以轻易证得 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。

定理 5.5 (Khinchin, 1928)**.** 已知 $\{X_j\}_{j=1}^{\infty}$ 是独立同分布的随机变量序列，满足 $\mathsf{E}X_j = \mu < \infty$，则 $\{X_j\}_{j=1}^{\infty}$ 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 $Y_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ 的特征函数为 $[\varphi(t/n)]^n = [1 + \mu it/n + o(t/n)]^n$。利用开圆盘 $|z| < 1$ 上的 $\ln(1+z) = z - z^2/2 + z^3/3 - z^4/4 + \cdots$，其中 $z \in \mathbb{C}$，对每个暂时固定的 t，总存在足够大的 n 使得 $|\mu it/n + o(t/n)| < 1$，进而得到

$$\ln[\varphi(t/n)]^n = n \ln\left[1 + \frac{\mu it}{n} + o(t/n)\right] = \mu it + no(t/n)$$

于是，$\lim\limits_{n\to\infty}[\varphi(t/n)]^n = e^{\mu it}$，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。 □

例 5.4. 设 $X_1, X_2, \cdots, X_n, \cdots$ 独立同分布，且 $\mathsf{E}(X_j) = 0, \mathsf{V}(X_j) = \sigma^2 < \infty, j = 1, 2, \cdots, n, \cdots$。试证明：$\forall \epsilon > 0$ 有 $\lim\limits_{n\to\infty} \mathsf{P}\{|\frac{1}{n}\sum_{j=1}^{n} X_j^2 - \sigma^2| \le \epsilon\} = 1$。

证明. 因为 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，所以 $X_1^2, X_2^2, \cdots, X_n^2, \cdots$ 也独立同分布。由于 $\mathsf{E}(X_j) = 0$，因此 $\mathsf{E}(X_j^2) = \sigma^2, j = 1, 2, \cdots$。由 Khinchin 弱大数律可证得随机变量序列 $\{X_j^2\}_{j=1}^{\infty}$ 满足弱大数律。 □

定理 5.6 (Kolmogorov, 1926)**.** 随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足弱大数律当且仅当

$$\lim_{n\to\infty} \mathsf{E}\left\{\frac{\left[\sum_{j=1}^{n}(X_j - \mathsf{E}X_j)\right]^2}{n^2 + \left[\sum_{j=1}^{n}(X_j - \mathsf{E}X_j)\right]^2}\right\} = 0 \tag{5.4}$$

✂ **证明**. 往证“⇐”：令 $G_n(x)$ 为 $Y_n = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mathsf{E}X_j)$ 的分布函数。

$$\begin{aligned}\mathsf{P}(|Y_n| \geq \epsilon) &= \int_{|x|\geq\epsilon} dG_n(x) \leq \frac{1+\epsilon^2}{\epsilon^2}\int_{|x|\geq\epsilon}\frac{x^2}{1+x^2}dG_n(x)\\ &\leq \frac{1+\epsilon^2}{\epsilon^2}\int_{\mathbb{R}}\frac{x^2}{1+x^2}dG_n(x) = \frac{1+\epsilon^2}{\epsilon^2}\mathsf{E}\left\{\frac{Y_n^2}{1+Y_n^2}\right\} \to 0\end{aligned}$$

往证“⇒”：当 $n \to \infty$ 时，

$$\begin{aligned}\mathsf{P}(|Y_n| \geq \epsilon) &= \int_{|x|\geq\epsilon} dG_n(x) \geq \int_{|x|\geq\epsilon}\frac{x^2}{1+x^2}dG_n(x)\\ &= \int_{\mathbb{R}}\frac{x^2}{1+x^2}dG_n(x) - \int_{|x|<\epsilon}\frac{x^2}{1+x^2}dG_n(x)\\ &\geq \mathsf{E}\left\{\frac{Y_n^2}{1+Y_n^2}\right\} - \frac{\epsilon^2}{1+\epsilon^2}\int_{\mathbb{R}}dG_n(x) \geq \mathsf{E}\left\{\frac{Y_n^2}{1+Y_n^2}\right\} - \epsilon^2 \to 0\end{aligned}$$

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 $\mathsf{P}(|Y_n|\geq\epsilon)<\eta$，但不保证 $\mathsf{P}\{(|Y_{N+1}|\geq\epsilon)\cup(|Y_{N+2}|\geq\epsilon)\cup\cdots\}<\eta$。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

$$\frac{Y_n^2}{1+Y_n^2} \leq Y_n^2 = \left[\frac{1}{n}\sum_{j=1}^{n}(X_j - \mathsf{E}X_j)\right]^2，\text{ 于是 } \mathsf{E}\left\{\frac{Y_n^2}{1+Y_n^2}\right\} \leq \frac{1}{n^2}\mathsf{V}\left(\sum_{j=1}^{n}X_j\right)$$

注记 5.1. 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。

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 5.7 (Borel, 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 $\mathsf{P}(A)$，即

$$\mathsf{P}\left\{\lim_{n\to\infty}\left[\frac{m}{n}-\mathsf{P}(A)\right]=0\right\}=1 \tag{5.5}$$

按照式 (5.2) 的定义可给出 Borel 强大数律的“随机变量版”的表达形式：$\mathsf{P}\{\lim_{n\to\infty}(\frac{1}{n}\sum_{j=1}^{n}X_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}X_j)=0\}=1$，它比弱大数律的条件 (5.3) 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

$$\mathsf{P}\left\{\lim_{n\to\infty}\left(\frac{1}{n}\sum_{j=1}^{n}X_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}X_j\right)=0\right\}=1 \tag{5.6}$$

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 A. Ya. Khinchin（见右图）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，当且仅当 $\mathsf{P}\{\lim_{n\to\infty}X_n-X=0\}=1$，记为 $X_n\overset{a.s.}{\to}X$。譬如，$\{X_j\}_{j=1}^{\infty}$ 满足强大数律意味着 $\frac{1}{n}\sum_{j=1}^{n}X_j-\frac{1}{n}\sum_{j=1}^{n}\mathsf{E}X_j\overset{a.s.}{\to}0$。

性质 5.2. 若 $X_n\overset{a.s.}{\to}X$，则 $X_n\overset{P}{\to}X$。

注记 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，它们为“多次重复观测的算术平均为 $\mathsf{E}(X)$ 的点估计”提供了理论依据。

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 $N \in \mathbb{N}$ 使得事件 “$|\frac{1}{n}\sum_{j=1}^{n} X_j - \frac{1}{n}\sum_{j=1}^{n} \mathsf{E}X_j| \le \epsilon, n = N, N+1, \cdots$” 以概率 1 发生。利用 Borel-Cantelli 引理（见第一章引理 1.1 或 Borel 0-1 律），若能证得 $\sum_{n=1}^{\infty} \mathsf{P}\{|\frac{1}{n}\sum_{j=1}^{n} X_j - \frac{1}{n}\sum_{j=1}^{n} \mathsf{E}X_j| \ge \epsilon\} < \infty$，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 (2.78) 可以证明下面的 Kolmogorov 强大数律。

定理 5.8 (Kolmogorov, 1930)**.** 独立的随机变量 $X_j, j = 1, 2, \cdots$ 满足 $\sum_{j=1}^{\infty} \mathsf{V}(X_j)/j^2 < \infty$，则随机变量序列 $\{X_j\}_{j=1}^{\infty}$ 满足强大数律。

证明. 令 $Z_n = \sum_{j=1}^{n} X_j - \mathsf{E}X_j$ 且 $Y_n = \frac{1}{n} Z_n$，由 Kolmogorov 不等式可得

$$\begin{aligned} p_m &= \mathsf{P}(\max |Y_n| \ge \epsilon, 2^m \le n < 2^{m+1}) \\ &\le \mathsf{P}(\max |Z_n| \ge 2^m \epsilon, 2^m \le n < 2^{m+1}) \le \frac{1}{(2^m\epsilon)^2} \sum_{j<2^{m+1}} \mathsf{V}(X_j) \end{aligned}$$

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 $A_m =$ “$\max |Y_n| \ge \epsilon, 2^m \le n < 2^{m+1}$”，$m = 1, 2, \cdots$ 中有无穷多个发生的概率为 0。事实上，

$$\begin{aligned} \sum_{m=1}^{\infty} p_m &\le \sum_{m=1}^{\infty} \frac{1}{(2^m\epsilon)^2} \sum_{j<2^{m+1}} \mathsf{V}(X_j) = \frac{1}{\epsilon^2} \sum_{j=1}^{\infty} \mathsf{V}(X_j) \sum_{m:\, j<2^{m+1}} 2^{-2m} \\ &= \frac{1}{\epsilon^2} \sum_{j=1}^{\infty} \mathsf{V}(X_j) \sum_{m=\rho}^{\infty} 2^{-2m} \le \frac{16}{3\epsilon^2} \sum_{j=1}^{\infty} \frac{\mathsf{V}(X_j)}{j^2} < \infty \end{aligned}$$

最后一步是因为 $2^\rho \le j < 2^{\rho+1}$ 才有 $\sum_{m=\rho}^{\infty} 2^{-2m} = \frac{4}{3} \cdot 2^{-2\rho} \le \frac{16}{3j^2}$。 □

练习 5.3. 独立的随机变量 $X_j, j = 1, 2, \cdots$ 满足 $\mathsf{V}(X_j) \le c$，则 $\{X_j\}_{j=1}^{\infty}$ 满足强大数律。

定理 5.9 (Kolmogorov, 1933)**.** 已知随机变量 $X_j, j = 1, 2, \cdots$ 独立同分布且 $\mathsf{E}(X_j) = \mu < \infty$（与 Khinchin 弱大数律条件相同，见定理 5.5），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 5.4. 试说明 Borel 强大数律是 Kolmogorov 强大数律的特例。

设 $X_1, X_2, \cdots, X_n, \cdots \overset{iid}{\sim} p\langle 1\rangle + (1-p)\langle 0\rangle$，令 $S_n = \sum_{j=1}^{n} X_j$，强大数律说 $\mathsf{P}\{\lim_{n\to\infty} \frac{1}{n} S_n = p\} = 1$，即 $\forall \epsilon > 0$，不等式 $|\frac{1}{n} S_n - p| \le \epsilon$ 除了有限个 n 外以概率 1 成立。Khinchin 于 1924 年证得一个更强的结果，

定理 5.10 (Khinchin, 1924)**.** 已知随机变量 $X_1, X_2, \cdots, X_n, \cdots \overset{iid}{\sim} p\langle 1\rangle + (1-p)\langle 0\rangle$，则有

$$\mathsf{P}\left\{\overline{\lim_{n\to\infty}} \frac{|S_n - np|}{\sqrt{2p(1-p)n \ln\ln n}} = 1\right\} = 1 \tag{5.7}$$

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。

$$\left|\frac{1}{n} S_n - p\right| \le (1+\epsilon)\sqrt{\frac{2p(1-p)}{n} \ln\ln n} \tag{5.8}$$

这个结果被称为 Khinchin 重对数律 (law of the iterated logarithm)，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（实例参见问题 1.6）。

Kolmogorov (1929)、P. Hartman 和 A. Wintner (1941) 在更一般的条件下发现了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》 [70] 第七章或《独立随机变量之和》 [69] 第十章。

定理 5.11 (Kolmogorov, 1929)**.** 若 $\{X_n\}_{n=1}^{\infty}$ 为独立随机变量序列，且 $\mathsf{E}(X_n) = 0, \mathsf{V}(X_n) = \sigma_n^2$。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 $|X_n| \le c_n\sqrt{\tau_n^2(\ln\ln\tau_n^2)^{-1}}$ 成立，则有

$$\mathsf{P}\left\{\overline{\lim_{n\to\infty}}\frac{|S_n|}{\sqrt{2\tau_n^2\ln\ln\tau_n^2}}=1\right\}=1 \tag{5.9}$$

定理 5.12 (Hartman-Wintner, 1941)**.** 若随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 独立同分布，且 $\mathsf{E}(X_n)=0, \mathsf{V}(X_n)=\sigma^2$，则有

$$\mathsf{P}\left\{\overline{\lim_{n\to\infty}}\frac{|S_n|}{\sqrt{2\sigma^2 n\ln\ln n}}=1\right\}=1 \text{ 当且仅当 } \sigma^2<\infty \tag{5.10}$$

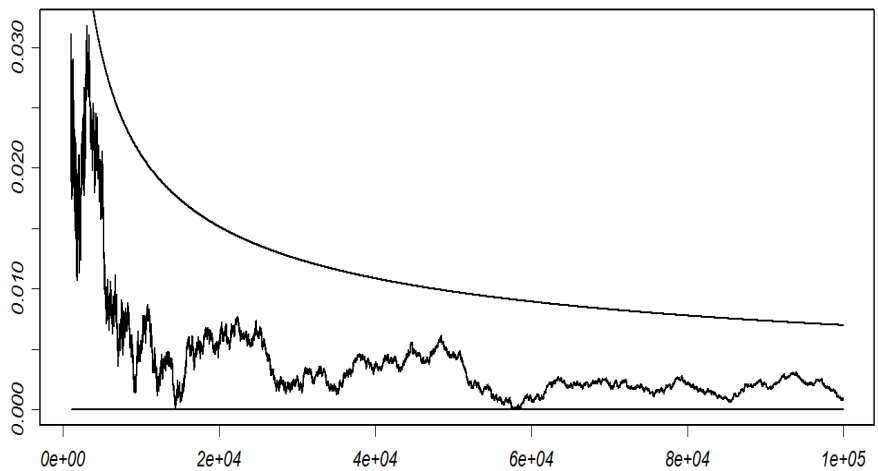


图 5.1: 直观了解 Hartman-Wintner 重对数律：从 N(0, 1) 产生 $N = 10^5$ 个随机数，观察 $\frac{1}{n}|\sum_{j=1}^{n}X_j|, n = 1, 2, \cdots, N$，发现当 n 增大时，它们被 $\sqrt{(2/n)\ln\ln n}$ “控制”着趋向于 0。

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 C）。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理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*。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。

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。de Moivre-Laplace 中心极限定理是古典概率论的巅峰，附录 C 曾基于 Stirling 公式给出过证明，下面的证法基于特征函数。

定理 5.13 (de Moivre-Laplace, 1733, 1801)**.** 已知随机变量 $X_n \sim \mathrm{B}(n,p)$，其中 $n = 1, 2, \cdots$，于是

$$Y_n = \frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{L} \mathrm{N}(0,1) \tag{5.11}$$

*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。

证明. X_n 和 Y_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 $e^z = \sum_{j=0}^{\infty} z^j/j!$，可得到

$$q\exp\left\{-\frac{pit}{\sqrt{npq}}\right\} = q - it\sqrt{\frac{pq}{n}} - \frac{pt^2}{2n} + o\left(\frac{t^2}{n}\right)，\text{并且}$$
$$p\exp\left\{\frac{qit}{\sqrt{npq}}\right\} = p + it\sqrt{\frac{pq}{n}} - \frac{qt^2}{2n} + o\left(\frac{t^2}{n}\right)$$

仿照仿照 Khinchin 弱大数律（定理 5.5）的证明，我们有

$$\ln\tilde{\varphi}_n(t) = n\ln\left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right] = -\frac{t^2}{2} + no\left(\frac{t^2}{n}\right)$$

因此，$\lim\limits_{n\to\infty}\tilde{\varphi}_n(t) = \exp(-t^2/2)$，它是标准正态分布的特征函数。根据 Lévy 连续性定理，$Y_n \xrightarrow{L} \mathrm{N}(0,1)$ 得证。 □

对于上述证明，读者也可用 Maxima 进行一下简单的验证，不难发现结果是正确的。回顾例 4.1 和图 1.4 所示 n 很大时的二项分布 $\mathrm{B}(n,p)$，二项分布与正态分布的关系就更明晰了。

```
(%i1) q : 1-p $
(%i2) taylor(log(exp(-n*p*%i*t/sqrt(n*p*q)) * (q + p*exp(%i*t/sqrt(n*p*q)))^n),
       [t,n], 0, 3) $  limit(%, n, inf);
                                    2
                                   t
(%o3)                            - --
                                   2
```

de Moivre-Laplace 中心极限定理也可以表述为：已知随机变量序列 $Z_1, Z_2, \cdots, Z_n, \cdots \overset{iid}{\sim} p\langle 1\rangle + (1-p)\langle 0\rangle$，令 $X_n = \sum_{j=1}^{n} Z_j$，则式 (5.11)

成立。当 n 很大时，每个 $Z_j, j = 1, 2, \cdots, n$ 对 X_n 的影响都不是至关重要的。从 de Moivre-Laplace 中心极限定理可提炼出以下概念：

定义 5.5. 随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立且每个 $X_j, j = 1, 2, \cdots$ 有有限的期望 $\mu_j = \mathsf{E}X_j$ 和方差 $\sigma_j^2 = \mathsf{V}(X_j)$，令 $\tau_n = \sqrt{\sum_{j=1}^n \sigma_j^2}$，如果

$$Y_n = \frac{1}{\tau_n} \sum_{j=1}^{n} (X_j - \mu_j) \xrightarrow{L} \mathrm{N}(0, 1) \tag{5.12}$$

则称随机变量序列 $\{X_n\}$ 满足**中心极限定理**。显然，Y_n 就是随机变量 $\sum_{j=1}^n X_j$ 的标准化。

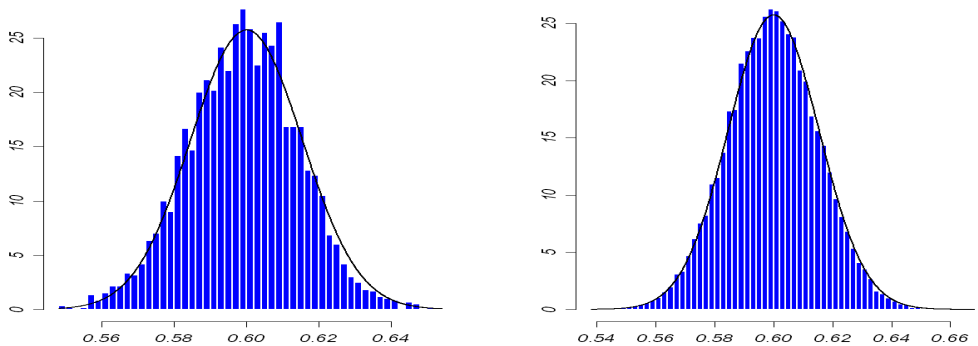
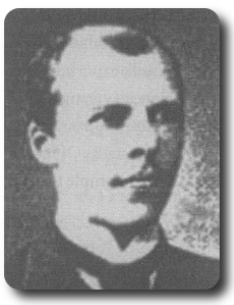


图 5.2: 通过模拟试验了解 de Moivre-Laplace 中心极限定理：再次重复例 1.34 中的随机试验，抛那枚不均匀的硬币（正面出现的概率为 $p = 0.6$）$n = 1000$ 次，记录下频率并重复此过程 3000 遍（左图）和 30000 遍（右图），得到正面频率的直方图，将之与密度函数 $\phi(x|p, p(1-p)/n)$ 进行比较，显然重复遍数越多拟合效果越好。

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。第二小节是中心极限定理在近似计算方面的应用。

学习目标 (1) 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) 会利用中心极限定理做近似计算。

5.2.1　Lindeberg-Feller 中心极限定理



1920 年，芬兰数学家 Jarl Waldemar Lindeberg (1876-1932) 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。二人的工作合称为 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 5.14 (Lindeberg-Lévy)**.** 如果随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 独立同分布，具有有限期望和方差 $\mathsf{E}(X_n) = \mu, \mathsf{V}(X_n) = \sigma^2 > 0$，则随机变量序列 $\{X_n\}$ 满足中心极限定理，即

$$Y_n = \frac{\sum_{j=1}^{n}(X_j - \mu)}{\sigma\sqrt{n}} \xrightarrow{L} \mathrm{N}(0,1) \tag{5.13}$$

证明. 随机变量 $X_n - \mu$ 和 Y_n 的特征函数分别是

$$\varphi(t) = \varphi(0) + \frac{\varphi'(0)}{1!}t + \frac{\varphi''(0)}{2!}t^2 + o(t^2) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)$$
$$\tilde{\varphi}_n(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n$$

仿照仿照定理 5.5 的证明，于是 $\ln\tilde{\varphi}_n(t) = -t^2/2 + no(t^2/n)$，显然 $\lim_{n\to\infty}\tilde{\varphi}_n(t) = \exp(-t^2/2)$。根据 Lévy 连续性定理，$Y_n \xrightarrow{L} \mathrm{N}(0,1)$ 得证。 □

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：

$$\frac{1}{n}\sum_{j=1}^{n} X_j \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{5.14}$$

Lindeberg-Lévy 中心极限定理也可轻易地推广到高维的情形。

定理 5.15 (高维中心极限定理)**.** 已知 d 维随机向量 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n, \cdots$ 独立同分布，期望 $\boldsymbol{\mu}$ 的每个分量都有限，协方差阵 Σ 正定且 $|\Sigma| < \infty$，则

$$\frac{\sum_{j=1}^{n} \boldsymbol{X}_j - \boldsymbol{\mu}}{\sqrt{n}} \xrightarrow{L} \mathrm{N}_d(\mathbf{0}, \Sigma) \tag{5.15}$$

根据 Lindeberg-Lévy 中心极限定理，例 4.5 所示的由 U[0, 1] 的随机数构造 N(0, 1) 的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，X_n 的分布不是常见的。

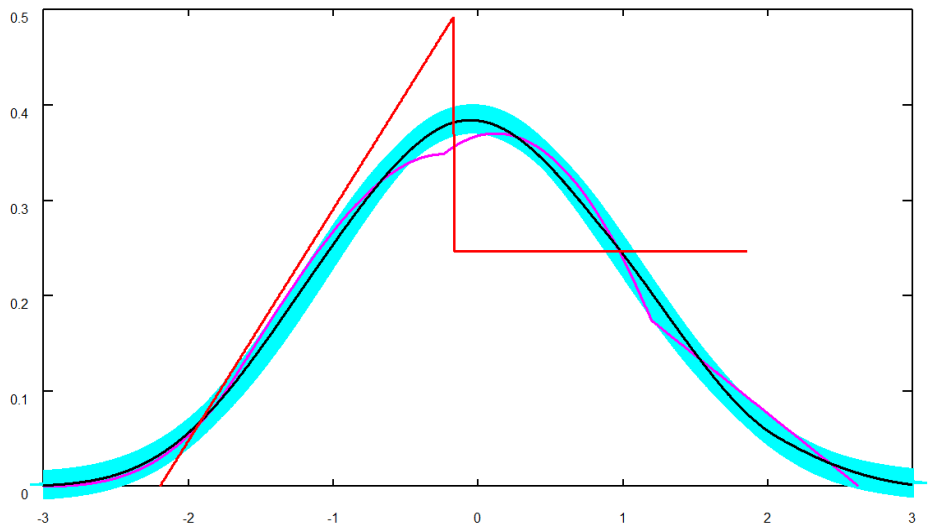


图 5.3: 独立同分布的随机变量 X_1, X_2, X_3 的密度函数是图中的折线，期望为 0，方差为 1。随机变量 $\frac{1}{\sqrt{2}}(X_1 + X_2)$ 的密度函数已基本进入 $\phi(x)$ 的一条窄带之中，更不用说 $\frac{1}{\sqrt{3}}(X_1 + X_2 + X_3)$ 可视为服从标准正态分布。

Lindeberg-Lévy 中心极限定理中方差 $\mathrm{V}(X_n) < \infty$ 的条件必不可少，譬如 $X_1, X_2, \cdots, X_n, \cdots \overset{iid}{\sim} \mathrm{Cauchy}(0, 1)$，每个随机变量的特征函数都是 $\varphi(t) = e^{-|t|}$，故 $\frac{1}{n} \sum_{j=1}^{n} X_j \sim \mathrm{Cauchy}(0, 1)$。

注记 5.3. 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 $n \to \infty$ 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 $X_1, X_2, \cdots$ 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 $\mathsf{P}\{|\frac{1}{n}S_n - \mu| \le \epsilon\}$ 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 (5.14) 的方式进行的，并且指出 n 足够大时，

$$\mathsf{P}\left\{\left|\frac{1}{n}S_n - \mu\right| \le \epsilon\right\} \approx 2\Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right) - 1$$

定理 5.16 (Lindeberg-Feller, 1922, 1935)**.** 已知随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，且满足 $\mathsf{E}(X_n) = \mu_n, \mathsf{V}(X_n) = \sigma_n^2 > 0$。令 $Y_n = \sum_{j=1}^{n}(X_j - \mu_j)/\tau_n$ 和 $\tau_n = \sqrt{\sum_{j=1}^{n}\sigma_j^2}$。$\lim\limits_{n\to\infty}\max\limits_{1\le j\le n}\sigma_j/\tau_n = 0$ 且 $Y_n \xrightarrow{L} \mathrm{N}(0,1)$ 当且仅当 $\forall\epsilon > 0$

$$\lim_{n\to\infty}\frac{1}{\tau_n^2}\sum_{j=1}^{n}\int_{|x-\mu_j|>\epsilon\tau_n}(x-\mu_j)^2 dF_j(x) = 0 \tag{5.16}$$

条件式 (5.16) 通常称为 Lindeberg 条件。

证明. 详见 A. N. Shiryayev 的《概率论》[84] 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 $F_j(x), j = 1, 2, \cdots$ 往往是未知的，即便 $F_j(x)$ 已知，Lindeberg 条件式 (5.16) 中求极限的过程也是非常复杂的。下面给出的练习 5.5 以及 Lyapunov 定理 5.17 和例 5.5 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 $A_j = \{|X_j - \mu_j| > \epsilon\tau_n\}, j = 1, 2, \cdots, n$，则

$$\begin{aligned}\mathsf{P}\left\{\max_{1\le j\le n}|X_j-\mu_j|>\epsilon\tau_n\right\} &= \mathsf{P}\left\{\bigcup_{j=1}^{n}A_j\right\} \le \sum_{j=1}^{n}\mathsf{P}(A_j)\\ &= \sum_{j=1}^{n}\int_{|x-\mu_j|>\epsilon\tau_n} dF_j(x) \le \frac{1}{(\epsilon\tau_n)^2}\sum_{j=1}^{n}\int_{|x-\mu_j|>\epsilon\tau_n}(x-\mu_j)^2dF_j(x)\end{aligned}$$

由 Lindeberg 条件式 (5.16)，于是

$$\lim_{n\to\infty}\mathsf{P}\left\{\max_{1\le j\le n}\left|\frac{X_j-\mu_j}{\tau_n}\right|>\epsilon\right\}=0 \tag{5.17}$$

这说明每个随机变量 $X_j, j = 1, 2, \cdots, n$ 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 $\sigma_j^2 > 0, j = 1, 2, \cdots$ 减弱为 $\sigma_1^2, \sigma_2^2, \cdots$ 不全为零。

练习 5.5. 试说明：Lindeberg-Feller 中心极限定理 ⇒ Lindeberg-Lévy 中心极限定理 ⇒ de Moivre-Laplace 中心极限定理。提示：Lindeberg-Lévy 中心极限定理的条件使得 Lindeberg 条件式 (5.16) 成立。

$$\lim_{n\to\infty}\int_{|\frac{x-\mu}{\sigma}|>\epsilon\sqrt{n}}\left(\frac{x-\mu}{\sigma}\right)^2dF(x)=0$$

定理 5.17 (Lyapunov, 1901)**.** 对定义 5.5 描述的独立随机变量序列 $\{X_n\}$，如果能找到正数 $\delta > 0$ 使得下述所谓的“Lyapunov 条件”成立，

$$\lim_{n\to\infty}\frac{1}{\tau_n^{2+\delta}}\sum_{j=1}^{n}\mathsf{E}|X_j-\mu_j|^{2+\delta}=0 \tag{5.18}$$

则随机变量序列 $\{X_n\}$ 满足中心极限定理。

证明. 只需验证 Lyapunov 条件式 (5.18) 能推出 Lindeberg 条件式 (5.16)。

$$\frac{1}{\tau_n^2}\sum_{j=1}^{n}\int_{|x-\mu_j|>\epsilon\tau_n}(x-\mu_j)^2dF_j(x) \le \frac{1}{\tau_n^2(\epsilon\tau_n)^\delta}\sum_{j=1}^{n}\int_{|x-\mu_j|>\epsilon\tau_n}|x-\mu_j|^{2+\delta}dF_j(x)$$
$$= \frac{1}{\epsilon^\delta}\left[\frac{1}{\tau_n^{2+\delta}}\sum_{j=1}^{n}\mathsf{E}|X_j-\mu_j|^{2+\delta}\right] \qquad \square$$

例 5.5. 设 $\{X_n\}_{n=1}^{\infty}$ 是独立随机变量序列，并且 $X_n \sim \frac{1}{2}\langle -n^\alpha\rangle + \frac{1}{2}\langle n^\alpha\rangle, n = 1,2,\cdots$，其中 $\alpha > -1/2$，试证明：$\{X_n\}$ 满足中心极限定理。

解. 算得 $\mathsf{E}(X_j) = 0, \mathsf{V}(X_j) = j^{2\alpha}$，于是 $\tau_n^2 = \sum_{j=1}^n j^{2\alpha} > \sum_{j=1}^n \int_{j-1}^{j} z^{2\alpha}dz = n^{2\alpha+1}/(2\alpha+1)$。$\forall\epsilon > 0$，当 $\epsilon\tau_n > n^\alpha$ 时 Lindeberg 条件式 (5.16) 成立，因此只需 $\epsilon^2 n^{2\alpha+1}/(2\alpha+1) > n^{2\alpha}$，即 $n > (2\alpha+1)/\epsilon^2$。


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。

定理 5.18 (Bárány-Vu, 2007)**.** 已知 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n \overset{iid}{\sim} \mathrm{N}_d(\mathbf{0}, I)$，其中 I 是 $d\times d$ 的单位阵。$\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ 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

$$\frac{X_n - \mathsf{E}(X_n)}{\sqrt{\mathsf{V}(X_n)}} \xrightarrow{L} \mathrm{N}(0,1)$$

定理 5.19. 对于多项分布 $\boldsymbol{X} = (X_1, X_2, \cdots, X_k)^\mathsf{T} \sim \mathrm{Multin}(n; p_1, p_2, \cdots, p_k)$，当 $n\to\infty$ 时，$\boldsymbol{X}$ 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 $\boldsymbol{Y}$ 依分布收敛于一个多元正态分布。同时也有

$$\sum_{j=1}^{k}(1-p_j)Y_j^2 \xrightarrow{L} \chi^2_{k-1} \tag{5.19}$$

5.2.2 中心极限定理的应用

定理 5.20. 对于二项分布 $X \sim \mathrm{B}(n,p)$，当 n 很大时，有近似计算公式

$$\mathsf{P}(a < X \le b) \approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right) \tag{5.20}$$

例 5.6. 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：$\mathsf{P}(50 < X_{100} \le 60)$。

解. $\mathsf{P}(50 < X_{100} \le 60) \approx \Phi(2) - \Phi(0)$，利用 R 计算具体结果。

```
> pnorm(2, mean=0, sd=1) - pnorm(0, mean=0, sd=1)
[1] 0.4772499
```

练习 5.6. 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 5.7. 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 $X \sim \mathrm{B}(n, 0.1)$。这批产品被认为不合格的概率为 $\mathsf{P}\{10 < X \le n\} \approx \Phi(3\sqrt{n}) - \Phi[(10 - 0.1n)/(0.3\sqrt{n})]$，依题意知 $n > 10$，故 $\Phi(3\sqrt{n}) \approx 1$ 且 $10 - 0.1n < 0$，故 $\mathsf{P}\{10 < X \le n\} \approx \Phi[(0.1n - 10)/(0.3\sqrt{n})] \ge 0.9$，用 Maxima 求得 $n \ge 147$。

```
(%i1) load (distrib) $
(%i2) find_root((0.1*n-10)-(0.3*sqrt(n))*quantile_normal(0.9,0,1), n, 0, 10^10);
(%o2)                          146.5411535301901
```

定理 5.21 (Fisher, 1925)**.** 如果 $X \sim \chi^2_n$，则

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{2X} - \sqrt{2n-1} \le z\right\} = \Phi(z) \tag{5.21}$$

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，当 $n \to \infty$ 时，渐近地有

$$Z_n = \frac{X-n}{\sqrt{2n}} \sim \mathrm{N}(0,1)，\text{即 } \lim_{n\to\infty} \mathsf{P}\left\{\frac{X-n}{\sqrt{2n}} \le z\right\} = \Phi(z)$$

同时，我们有

$$\begin{aligned}
\lim_{n\to\infty} \mathsf{P}\left\{\frac{X-n}{\sqrt{2n}} \le z\right\} &= \lim_{n\to\infty} \mathsf{P}\left\{\frac{X-n}{\sqrt{2n}} \le z + \frac{z^2-1}{2\sqrt{2n-1}}\right\} \\
&= \lim_{n\to\infty} \mathsf{P}\left\{\frac{X-n}{\sqrt{2n-1}} \le z + \frac{z^2-1}{2\sqrt{2n-1}}\right\} \\
&= \lim_{n\to\infty} \mathsf{P}\left\{X \le n + z\sqrt{2n-1} + \frac{z^2-1}{2}\right\} \\
&= \lim_{n\to\infty} \mathsf{P}\left\{2X \le 2n - 1 + 2z\sqrt{2n-1} + z^2\right\} \\
&= \lim_{n\to\infty} \mathsf{P}\left\{\sqrt{2X} \le z + \sqrt{2n-1}\right\}
\end{aligned}$$ □

当 $n \ge 30$ 时，Fisher 定理有如下近似计算的应用：

$$\begin{aligned}
\mathsf{P}(\chi_n^2 \le z) &= \mathsf{P}\left(\sqrt{2\chi_n^2} - \sqrt{2n-1} \le \sqrt{2z} - \sqrt{2n-1}\right) \\
&\approx \Phi(\sqrt{2z} - \sqrt{2n-1})
\end{aligned} \tag{5.22}$$

例 5.8. 某测量值 $X_j \sim \mathrm{N}(\mu, \sigma^2), j = 1, 2, \cdots$，其中 $\sigma^2 = 10^{-2}$。需要测量多少次才能使得 $\mathsf{P}\{|\frac{1}{n}\sum_{j=1}^n X_j - \mu| < 10^{-4}\} = 0.95$？

解. 由 Lindeberg-Lévy 中心极限定理知，当 n 足够地大时，近似地 $\sum_{j=1}^n X_j/n - \mu \sim \mathrm{N}(0, \sigma^2/n)$，所以 $\Phi(10^{-4}\sqrt{n}/\sigma) - \Phi(-10^{-4}\sqrt{n}/\sigma) = 0.95$ 或 $\Phi(\sqrt{n}/100) = 0.975$。仿照上例用 Maxima 求解，或者用 R 来算。

```
> ceiling((qnorm(0.975,mean=0,sd=1)*100)^2)
[1] 38415
```

故需要 $n \ge 38415$ 次测量。

练习 5.7. 在 n 次 Bernoulli 试验中，每次试验事件 A 发生的概率均为 0.7，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 p 是多少？请分别用 (i) Chebyshev 不等式；(ii) 中心极限定理来估计。答案：(1) $n \geq 5250, p = 0.916$；(2) $n \geq 1429, p = 0.99954$。

5.3　习题

5.1. 已知 $\{X_k\}_{k=1}^{\infty}$ 为独立随机变量序列，并且 $X_k \sim \frac{1}{k+1}\langle -\sqrt{k+1}\rangle + (1-\frac{2}{k+1})\langle 0\rangle + \frac{1}{k+1}\langle \sqrt{k+1}\rangle, k=1,2,\cdots$。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 $X_k \sim \frac{1}{2}\langle k^s\rangle + \frac{1}{2}\langle -k^s\rangle, k=1,2,\cdots,n,\cdots$。试证明：当 $s<1/2$ 时，$\{X_k\}$ 满足弱大数律。

☆ 5.3. 将编号为 $1,2,\cdots,n$ 的 n 个球随机放入编号为 $1,2,\cdots,n$ 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 $\lim\limits_{n\to\infty}\mathsf{P}\{|\frac{1}{n}S_n-\frac{1}{n}|\geq\epsilon\}=0$。

★ 5.4. 设随机变量 $X_1,X_2,\cdots$ 的方差都不超过 $c>0$，当 $|k-j|\to\infty$ 时 X_k 和 X_j 的相关系数 $\rho_{kj}\to 0$。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 $\mathsf{V}(X_k), k=1,2,\cdots$ 且 $\sum_{k=1}^{\infty}\mathsf{V}(X_k)/k^2<\infty$，试证明：$\{X_k\}$ 满足弱大数律。

5.6. 设独立随机变量序列 $\{X_n\}$ 满足：(1) 存在常数 $c>0$ 使得 $|X_n|<c, n=1,2,\cdots$；(2) $\mathsf{V}(X_n)$ 存在，但是 $\sum\limits_{n=1}^{\infty}\mathsf{V}(X_n)=\infty$。试问强大数律和中心极限定理是否成立？为什么？

5.7. 设随机变量序列 $X_1,X_2,\cdots\overset{iid}{\sim}\mathrm{U}[0,1]$，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。

✯ 5.8. 设 $h(x)$ 是在 $(0,\infty)$ 上的连续单调增函数，且 $\lim\limits_{x\to 0}h(x)=0, \sup h(x)<\infty$，求证：随机变量序列 $X_n\xrightarrow{P}0$ 的充要条件是 $\lim\limits_{n\to\infty}\mathsf{E}[h(|X_n|)]=0$。

5.9. 设 $X_1,X_2,\cdots,X_n,\cdots$ 为独立同分布的随机变量序列，$\mathsf{E}(X_n)=\mu, \mathsf{V}(X_n)=\sigma^2$，试证明：$\frac{2}{n(n+1)}\sum_{k=1}^{n}kX_k\xrightarrow{P}\mu$。

☆ 5.10. 已知独立随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 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？

5.11. 已知随机变量 $X_1, X_2, \cdots, X_{100}$ 独立同分布且 $\mathsf{E}(X_1) = 1, \mathsf{V}(X_1) = 2.4$，计算 $\mathsf{P}\{\sum_{i=1}^{100} X_i \geq 90\}$。

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，问优等品的个数在 18 个到 25 个之间的概率是多少？

☆ 5.14. 已知连续型随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 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 $\lim\limits_{n\to\infty} \mathsf{P}\{\frac{1}{n}\sum_{i=1}^{n} F(X_i) \leq \frac{1}{2}\}$。

5.15. 已知随机变量 $X_1, X_2, \cdots, X_n, \cdots \overset{iid}{\sim} \mathrm{Expon}(\lambda)$，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？

5.16. 设随机变量 $X_1, X_2, \cdots, X_n$ 独立同分布，$\mathsf{E}(X_1^k) = m_k, k = 1, 2, 3, 4$ 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？

5.17. 设 $\{X_n\}_{n=1}^{\infty}$ 是独立同分布的随机变量序列且 $\mathsf{E}(X_n) = \mathsf{V}(X_n) = 1$，求常数 c 使得 $\lim\limits_{n\to\infty} \mathsf{P}\{\frac{c}{\sqrt{n}}\sum_{j=1}^{n}(X_{2j} - X_{2j-1}) \leq x\} = \phi(x)$。

✯ 5.18. 利用中心极限定理证明：$\lim\limits_{n\to\infty} e^{-n}\sum_{k=0}^{n} n^k/k! = 1/2$。

第二部分

数理统计学初步

统计学（或数理统计学）是一个数学分支，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，都需要科学系统方法的指导。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，无需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，统计学在某种意义下可看作是概率论的一个应用领域：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，统计学伴随概率论的发展而变得更数学化，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 $(\Omega, \mathscr{S}, \mathsf{P})$ 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 $\mathrm{N}(\mu, \sigma^2)$，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确定了的，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。下面简要地介绍统计学的发展历史，它一般被划分为三个阶段 [4,5,10]：

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，以及上述统计学的基本认识的确立。(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 提出 t 分布和正态总体均值的 t 检验。(iii) R. A. Fisher 是一位在统计学发展史上举足轻重的天才统计学家，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 [37]；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。(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 [90]。(ix) 1946 年，瑞典统计学家 Harald Cramér (1893-1985) 发表著作《统计学数学方法》 [28] 总结了当时数理统计学的成果，标志着统计学走向成熟。

*详见 C. R. Rao 的纪念文章《R. A. Fisher：现代统计学的奠基人》 [75]。

二十世纪下半叶至今 计算机科学与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。(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（见第十一章）已成为经典统计学的强有力竞争者。(iv) 随机模拟技术（见第十四章）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大大削弱了理论模型的重要性，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、模式识别、机器学习、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。

本书第二部分仅是数理统计学的入门，为更好地了解统计思想和方法，由易及难，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》[38]（强调统计思想的入门书，像侦探小说一样引人入胜）。
- G. Casella 和 R. L. Berger 的《统计推断》[23]（例子丰富，预备知识仅需要数学分析和线性代数）。
- P. J. Bickel 与 K. A. Doksum 合著的《数理统计——基本概念及专题》[17]（2001 年的新版较 1977 年的旧版改动很大）。
- 陈希孺的《高等数理统计学》[9]（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，占了全书的近一半篇幅）。

第六章

数理统计学的一些基本概念

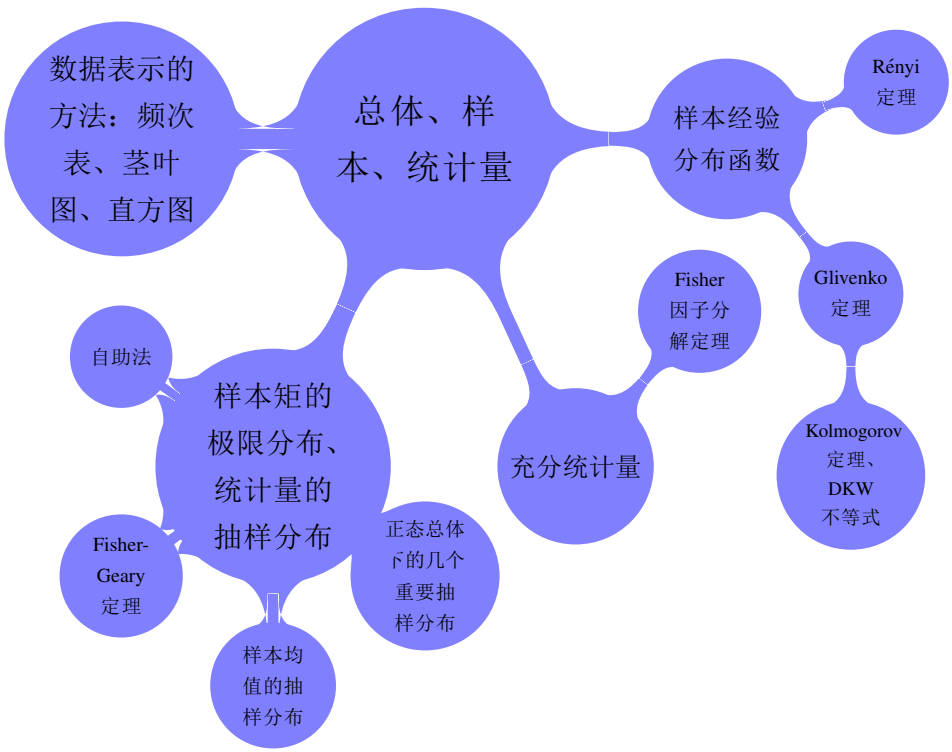
与所研究问题有关的全部个体的确定集合称作*总体* (population)，若其中个体数目有限，则称之为*有限总体*，例如，调查一个班级学生的身高状况，总体就是这个班的所有学生；调研国民的年收入情况，总体包括所有国民。有时总体只是一种想象中的存在，如测量某时刻某地点的温度，总体就是所有可能的观测值，即实数集 $\mathbb{R}$；再如，抛一枚硬币无穷多次所得结果的总体，像这种个体数目无穷多的总体称为*无限总体*。如果有限总体中个体数目足够大，也可近似地当作无限总体来处理，如某时间段内全球上网者的总体。

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经济新闻，对每个上网者可联系一个 0-1 分布的随机变量 X，若上网则取 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。

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*，如某时刻某地点温度测量值的总体分布就是以该时刻该地点的真实温度为均值的正态分布 $\mathrm{N}(\mu,\sigma^2)$，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 $\theta_1, \cdots, \theta_k$ 未知，这样的总体被称为参数总体。未知参数 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^{\mathsf{T}}$ 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 $\mathrm{N}(\mu, \sigma^2)$，但是参数 μ, σ^2 未知，参数空间是上半平面。显然，参数总体的所有可能分布的集合 $\mathscr{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ 是一个分布族*。例如，k-参数指数族（见定义 6.8）。



*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 $\mathscr{F} = \{F_\theta : \theta \in \Theta\}$。

6.1　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能的或不经济的，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取出若干个体（这些被抽取的个体称为样本点*，它们的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（由于总体已经数量化了，所以样本的观察结果也是数值，称为样本值），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。

由于抽样的随机性，每个样本点都是一个随机变量。样本量为 n 的样本记作 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$，其分布称为样本分布，取决于总体分布、抽样方式†和样本量。

定义 6.1 (简单随机样本)**.** 如果样本 $X_1, X_2, \cdots, X_n$ 独立同分布于总体分布 $F_\theta(x)$，则称样本 $X_1, X_2, \cdots, X_n$ 为独立同分布样本或简单随机样本‡，其分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{j=1}^{n} F_\theta(x_j) \tag{6.1}$$

如果没有特殊声明，后文中所提的样本都是指简单随机样本。

定义 6.2 (统计量)**.** 设 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 是从某总体抽得的样本，若 Borel 可测函数 $T = T(\boldsymbol{X})$ 不依赖于其他任何未知量，则称 T 为样本统

*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。

‡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。

计量或统计量*(statistic)，它就是由样本构造的随机变量。例如，

$$\text{样本均值} \qquad \overline{X} = \frac{\sum_{j=1}^{n} X_j}{n} \tag{6.2}$$

$$\text{样本方差} \qquad S^2 = \sum_{j=1}^{n} \frac{(X_j - \overline{X})^2}{n-1} = \frac{\sum_{j=1}^{n} X_j^2 - n\overline{X}^2}{n-1} \tag{6.3}$$

$$\text{样本 } k \text{ 阶矩、中心矩} \qquad A_k = \frac{\sum_{j=1}^{n} X_j^k}{n} \text{ 与 } B_k = \frac{\sum_{j=1}^{n} (X_j - \overline{X})^k}{n} \tag{6.4}$$

若总体 $X \sim \mathrm{N}(\mu, \sigma^2)$ 的期望 μ 已知，方差 σ^2 未知，则 $\overline{X} - \mu$ 是统计量，而 $\sum_{j=1}^{n} X_j/\sigma^2$ 不是统计量。

定义 6.3 (次序统计量)**.** 将样本 $X_1, X_2, \cdots, X_n$ 从小到大按升序排列 $X_{(1)} \le \cdots \le X_{(j)} \le \cdots \le X_{(n)}$，称 $X_{(j)}$ 为第 j 个次序统计量。其中，$X_{(1)}, X_{(n)}$ 称为极值，$X_{(n)} - X_{(1)}$ 称为极差。样本中位数 M 定义为

$$M = \begin{cases} X_{(\frac{n+1}{2})} & \text{若 } n \text{ 为奇数} \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}] & \text{若 } n \text{ 为偶数} \end{cases} \tag{6.5}$$

X_j 在 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ 中所在位置 R_j 称为 X_j 的秩 (rank)、显然，$R_1, R_2, \cdots, R_n$ 是 $1, 2, \cdots, n$ 的某一排列，称为秩统计量。这些统计量在非参数统计学中都是重要的工具（详见第十章）。

样本值 $x_1, \cdots, x_n$ 是样本 $X_1, X_2, \cdots, X_n$ 的具体观察结果，而这些样本值的均值 $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ 和方差 $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$ 则分别是样本均值 $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ 和样本方差 $S^2 = \sum_{j=1}^{n} \frac{(X_j - \overline{X})^2}{n-1} = \frac{\sum_{j=1}^{n} X_j^2 - n\overline{X}^2}{n-1}$ 的观察结果。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）。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 $x_1, \cdots, x_n$ 有一个直观的了解。

*搞清楚统计量的分布是统计学的一个基本问题：$T(X_1, \cdots, X_n)$ 的精确分布对小样本问题很重要，而 $n \to \infty$ 时 $T(X_1, \cdots, X_n)$ 的极限分布对大样本问题是至关重要的。

把样本值 $x_1, x_2, \cdots, x_n$ 中所有不同的值 $x_1^* < x_2^* < \cdots < x_k^*$ 都列出来并标出它们出现的频次 $n_1, n_2, \cdots, n_k$ 或频率 $f_1 = n_1/n, f_2 = n_2/n, \cdots, f_k = n_k/n$（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（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）。

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

若把 $x_1^*, x_2^*, \cdots, x_k^*$ 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 6.1. 考虑 Fisher 的 Iris 数据*中 setosa 类的花瓣长度的样本值，分别利用 R 语言的 summary、stem 函数给出它们的频次表和茎叶图。

```
> x <- subset(iris,Species=="setosa")$Petal.Length # 抽取 setosa 类的花瓣长度数据
> x                              # 显示 setosa 类的花瓣长度的 50 个样本值
 [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5 1.3 1.4
[19] 1.7 1.5 1.7 1.5 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4 1.5 1.2
[37] 1.3 1.4 1.3 1.5 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4
> summary(as.factor(sort(x)))   # 利用 summary 函数给出频次表
  1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.9
  1   1   2   7  13  13   7   4   2
> stem(x)                        # 利用 stem 函数绘制茎叶图
  The decimal point is 1 digit(s) to the left of the |

  10 | 0
  11 | 0
  12 | 00
  13 | 0000000
  14 | 00000000000000
  15 | 00000000000000
  16 | 0000000
  17 | 0000
  18 |
  19 | 00
```

*该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”。

直方图是直观显示数据聚散情况的最常见方法，第一章的例 1.34 中就有介绍。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。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，建议为 $k = \lceil \log_2 n + 1 \rceil$（样本量小于 30 时效果较差），它与区间宽度 h 得关系为 $\lceil \frac{1}{h}(\max x - \min x) \rceil$，其中 $\max x = \max(x_1, x_2, \cdots, x_n), \min x = \min(x_1, x_2, \cdots, x_n)$。也有人建议用 $h = 3.5n^{-1/3}s_n$，其中 $s_n = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2}$。

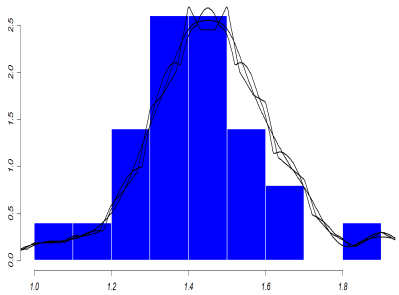


图 6.1: Iris 数据中 setosa 类的花瓣长度的直方图与核密度估计 (kernel density estimate) [85]。

将茎叶图逆时针旋转 90 度便是一个直方图，但它保留了（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察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（见图 6.1）。直方图的缺点是丢失了数据的很多原始信息，更新起来也不太容易。

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 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*。

学习目标 (1) 理解统计量等概念；(2) 数据表示的常见方法；(3) 熟悉经验分布函数的基本性质；(4) 了解 Glivenko 定理、Kolmogorov 定理和 DKW 不等式；(5) 掌握样本矩的基本性质。

*允许样本量趋向无穷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，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。

6.1.1　经验分布及其性质

通过样本 $X_1, X_2, \cdots, X_n$ 对总体 X 进行研究，无非是为了搞清楚 X 的分布 $F_X(x)$。在数学上 Glivenko 定理保证从样本到总体分布有一条“通途”*，它便是经验分布。

定义 6.4 (经验分布函数)**.** 样本 $X_1, X_2, \cdots, X_n$ 的经验分布函数定义为

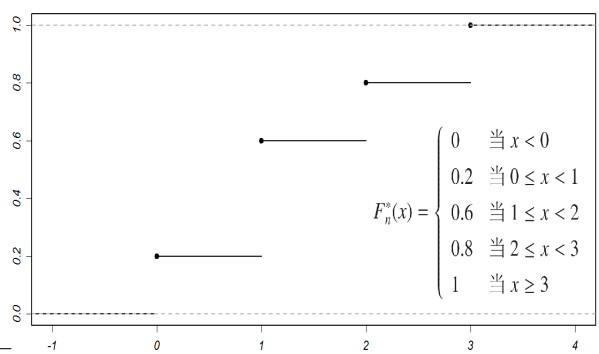
$$F_n^*(x) = \frac{1}{n}\sharp\{X_j \le x : j = 1, 2, \cdots, n\} = \frac{1}{n}\sum_{j=1}^{n} J(x - X_j) \tag{6.6}$$

其中，$\sharp\{X_j \le x : j = 1, 2, \cdots, n\}$ 表示 $X_1, X_2, \cdots, X_n$ 中不超过 x 的个数，$J(\cdot)$ 为式 (2.10) 定义的非负判定函数。

性质 6.1. 令 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为样本 $X_1, X_2, \cdots, X_n$ 的次序统计量 (order statistic)，则经验函数 $F_n^*(x)$ 可按如下方式构造：

$$F_n^*(x) = \begin{cases} 0 & \text{当 } x < X_{(1)} \\ \frac{k}{n} & \text{当 } X_{(k)} \le x < X_{(k+1)} \\ 1 & \text{当 } x \ge X_{(n)} \end{cases} \tag{6.7}$$

举个实例：从语料（即一些自然语言文本构成的集合）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 $0, 3, 2, 1, 1$，其经验分布函数如右图所示。



*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。Kant 的这一观点并非否认了认识过程的经验有助于确定认识能力，很多问题的确是求解屡遭挫折反而提醒人们关注解的存在性，如长期尝试从欧式几何的其他公设推导出第五公设都未成功，几何学家才会想到去论证第五公设的独立性，但经验不能替代对“认识能力”的理性认识。

性质 6.2. 总体分布函数 $F(x)$ 与经验分布函数 $F_n^*(x)$ 有如下关系：

$$\mathsf{P}\left\{F_n^*(x) = \frac{k}{n}\right\} = C_n^k[F(x)]^k[1-F(x)]^{n-k} \tag{6.8}$$

$$F_n^*(x) \xrightarrow{P} F(x) \tag{6.9}$$

$$\frac{\sqrt{n}[F_n^*(x) - F(x)]}{\sqrt{F(x)[1-F(x)]}} \xrightarrow{L} \mathrm{N}(0,1) \tag{6.10}$$

证明. $J(x-X_j), j = 1, 2, \cdots, n$ 独立同分布，并且 $\mathsf{P}\{J(x-X_j)=1\} = \mathsf{P}(x - X_j \geq 0) = F(x), \mathsf{P}\{J(x-X_j)=0\} = \mathsf{P}(x-X_j < 0) = 1 - F(x)$，由式 (6.6) 可看出 $nF_n^*(x) \sim \mathrm{B}(n, F(x))$，结果 (6.8) 得证。由弱大数律和中心极限定理，可证得结果 (6.9) 和结果 (6.10)。 □

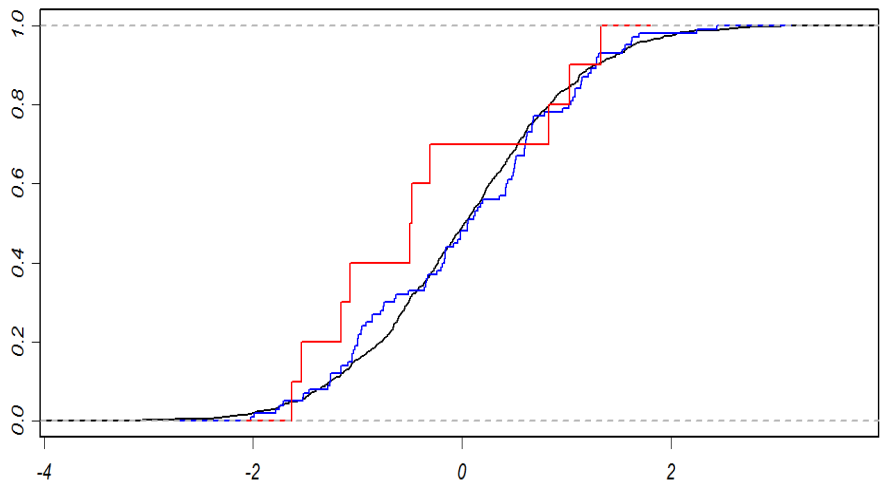


图 6.2: 样本来自正态总体 $\mathrm{N}(0,1)$，样本量分别为 $10, 10^2, 10^3$，相对应的经验分布函数 $F_n^*(x)$ 如图所示。经验分布函数是逐段线性函数，为了显示和识别起见，我们采用折线图来绘制它们。

例 6.2. 令 $F_n^*(x)$ 是从样本 $X_1, \cdots, X_n \overset{iid}{\sim} F(x)$ 得到的经验分布函数，要使得 $\forall x \in \mathbb{R}$ 皆有 $\mathsf{P}\{|F_n^*(x) - F(x)| \geq 0.1\} \leq 0.05$，样本量 n 至少该多大？

解. 由式 (6.10)，当 n 很大时有

$$\mathsf{P}\left\{\frac{\sqrt{n}|F_n^*(x) - F(x)|}{\sqrt{F(x)[1-F(x)]}} \leq z\right\} \approx 2\Phi(z) - 1$$

又因为 $F(x)[1-F(x)] \le 1/4$，所以 $\mathsf{P}\{|F_n^*(x)-F(x)| \le \frac{z}{2\sqrt{n}}\} \ge 2\Phi(z)-1$，根据条件解得 $n \ge 97$。

```
z <- qnorm((2-0.05)/2, mean = 0, sd = 1)  # qnorm: 正态分布分位数
n <- ceiling((z/0.2)^2)                   # ceiling: 上取整
```

1933 年，苏联数学家 Valery Ivanovich Glivenko (1896-1940) 得到一个比式 (6.9) 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 $F_n^*(x)$ 就能以任何要求的精度逼近总体分布 $F(x)$。

定理 6.1 (Glivenko, 1933)**.** 设样本 $X_1, X_2, \cdots, X_n$ 来自分布函数为 $F(x)$ 的总体，我们约定经验分布函数 $F_n^*(x)$ 与总体分布函数 $F(x)$ 的接近程度用离差 $D_n = \sup\limits_{x\in\mathbb{R}} |F_n^*(x)-F(x)|$ 来度量，则

$$\mathsf{P}\left\{\lim_{n\to\infty} D_n = 0\right\} = 1 \tag{6.11}$$

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，并未揭示离差 $D_n \le \epsilon$ 能以多大的概率发生。1933 年，Kolmogorov 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 §8.2.2）。

定理 6.2 (Kolmogorov, 1933)**.** 如果总体 X 的分布函数 $F(x)$ 是连续的，则有 $\lim\limits_{n\to\infty} \mathsf{P}\{\sqrt{n}D_n \le z\} = K(z)$，其中 Kolmogorov 分布函数 $K(z)$ 定义为

$$K(z) = \begin{cases} 0 & \text{当 } z \le 0 \\ \displaystyle\sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2z^2) & \\ \displaystyle = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2z^2) & \text{当 } z > 0 \end{cases} \tag{6.12}$$

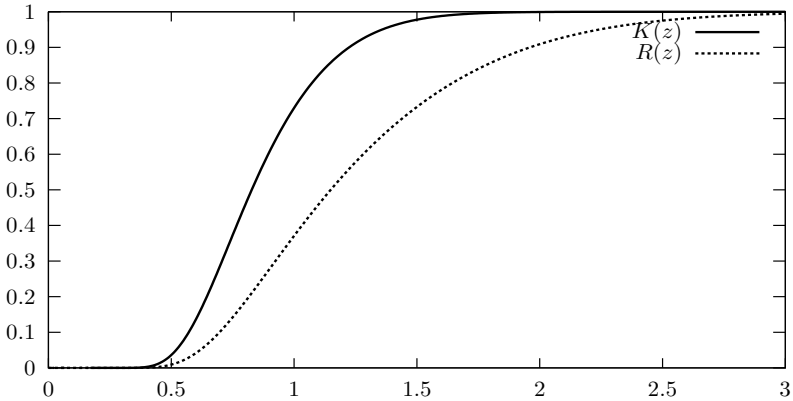


图 6.3: 式 (6.12) 定义的函数 $K(z)$ 被称为 Kolmogorov 分布函数（图中实线），用于 Kolmogorov 拟合优度检验。函数 $R(z)$ 是式 (6.16) 定义的 Rényi 分布函数（图中虚线）。

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依然能给出合理的估计方法和实验设计方法。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 $D_n = \sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)|$ 的收敛速度进行了描述。

定理 6.3 (DKW 不等式, 1956)**.** 对于任意 $\epsilon > 0$，D_n 满足

$$\mathsf{P}\{D_n > \epsilon\} \le 2\exp\{-2n\epsilon^2\} \tag{6.13}$$

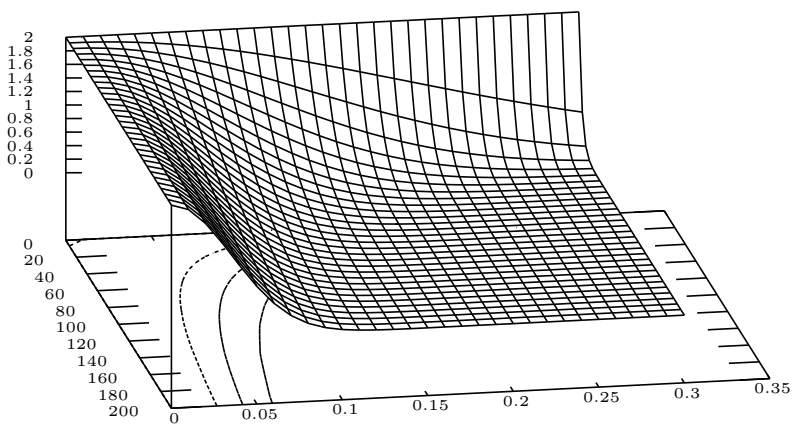


图 6.4: 曲面 $p = 2\exp\{-2n\epsilon^2\}$，其中 $\epsilon \in (0, 0.3], n = 1, 2, \cdots, 100$。

Kolmogorov 定理 6.2 和 DKW 不等式只考虑了绝对误差 $|F_n^*(x) - F(x)|$，1953 年匈牙利数学家 A. Rényi (1921-1970) 进一步研究了相对误差 $|F_n^*(x) - F(x)|/F(x)$ 的规律，得到下面的结果。

定理 6.4 (Rényi, 1953)**.** $\forall p \in (0,1)$，经验分布函数 F^* 与总体分布函数 $F(x)$ 具有如下关系：

$$\lim_{n\to\infty} \mathsf{P}\left\{ \sqrt{n} \sup_{\substack{F(x)\geq p \\ F(p)>0}} \frac{F_n^*(x) - F(x)}{F(x)} \leq z \right\} = \begin{cases} 0 & \text{当 } z < 0 \\ 2\Phi\left(z\sqrt{\frac{p}{1-p}}\right) - 1 & \text{当 } z \geq 0 \end{cases} \tag{6.14}$$

$$\lim_{n\to\infty} \mathsf{P}\left\{ \sqrt{n} \sup_{\substack{F(x)\geq p \\ F(p)>0}} \left|\frac{F_n^*(x) - F(x)}{F(x)}\right| \leq z \right\} = \begin{cases} 0 & \text{当 } z < 0 \\ R\left(z\sqrt{\frac{p}{1-p}}\right) & \text{当 } z \geq 0 \end{cases} \tag{6.15}$$

其中 $R(z)$ 被称为 Rényi 分布函数（见图 6.3），具体定义为

$$R(z) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{ -\frac{(2k+1)^2\pi^2}{8z^2} \right\} \tag{6.16}$$

当 $z \geq 2$ 时，$R(z) \approx 4\Phi(z) - 3$ 的精度在 4×10^{-9} 以内。

注记 6.1. 读者特别注意大样本和小样本不是指样本量的多少，而是区分样本量趋向无穷还是为有限固定值两个前提条件。譬如，在样本量 n 固定的前提下所讨论的统计量 $T_n = T(X_1, X_2, \cdots, X_n)$ 的性质都是小样本性质，如 DKW 不等式 (6.13)；而在 $n \to \infty$ 的时候 $T_n = T(X_1, X_2, \cdots, X_n)$ 的性质都是大样本性质或渐近性质（即随机变量序列 $\{T_n\}$ 的极限性质），如 Kolmogorov 定理 6.2 和 Rényi 定理 6.4。总体为一维随机变量时，经验告诉人们一般情况下样本量如果超过 45 就可以利用大样本性质做近似计算了。

6.1.2 样本矩及其极限分布

样本矩与总体矩（譬如，样本均值 $\overline{X}$ 与总体期望 μ）之间有什么样的关系？如果能从样本矩中挖掘出总体分布的数字特征*，将有助于了解总体分布（详见 §7.1.3 参数点估计的矩方法）。

性质 6.3. 令简单随机样本 $X_1, \cdots, X_n$ 来自的总体 X 具有期望 $\mathsf{E}(X) = \mu$，方差 $\mathsf{V}(X) = \sigma^2$，k 阶矩 $\mathsf{E}(X^k) = m_k$ 和 k 阶中心矩 $\mathsf{E}(X-\mu)^k = \mu_k$，则

$$\mathsf{E}(\overline{X}) = \mu \text{ 且 } \mathsf{V}(\overline{X}) = \sigma^2/n \tag{6.17}$$

$$\mathsf{E}(S^2) = \sigma^2 \text{ 且 } \mathsf{V}(S^2) = \frac{\mu_4}{n} + \frac{3-n}{n(n-1)}\mu_2^2 \tag{6.18}$$

$$A_k = \frac{1}{n}\sum_{j=1}^{n} X_j^k \xrightarrow{a.s.} m_k \text{ 且样本量足够大时，} A_k \sim \mathrm{N}\left(m_k, \frac{m_{2k} - m_k^2}{n}\right) \tag{6.19}$$

证明. 下面往证结果 (6.18)：利用下面的关系式（证明留给读者）

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n(n-1)}\sum_{i<j}(X_i-\mu)(X_j-\mu) \tag{6.20}$$

立即可得 $\mathsf{E}(S^2) = \mu_2 = \sigma^2$。下面求解 $\mathsf{V}(S^2)$：

$$\begin{aligned}
\mathsf{V}(S^2) &= \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n(n-1)}\sum_{i<j}(X_i-\mu)(X_j-\mu)\right]^2 - \mu_2^2 \\
&= \frac{1}{n^2}\mathsf{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]^2 + \frac{4}{n^2(n-1)^2}\mathsf{E}\left[\sum_{i<j}(X_i-\mu)(X_j-\mu)\right]^2 - \mu_2^2 \\
&= \frac{\mu_4}{n} + \frac{n-1}{n}\mu_2^2 + \frac{2}{n(n-1)}\mu_2^2 - \mu_2^2 = \frac{\mu_4}{n} + \frac{3-n}{n(n-1)}\mu_2^2
\end{aligned}$$

*利用样本均值 $\overline{X}$ 进行精确测量直到十七世纪才开始使用。根据性质 6.3，偏倚（或称系统误差）$\mathsf{E}(\overline{X}) - \mu$ 为零，并且 $\overline{X}$ 渐近服从于 $\mathrm{N}(\mu, \sigma^2/n)$。

由 Kolmogorov 强大数律（第 206 页的定理 5.9）证得结果 (6.19) 的前半部分。又因为 $X_1^k, \cdots, X_n^k$ 是独立同分布的且 $\mathsf{V}(X_1^k) = m_{2k} - m_k^2$，由 Lindeberg-Lévy 中心极限定理 5.14 证得结果 (6.19) 的后半部分。 □

例 6.3. 样本 $X_1, \cdots, X_{100}$ 来自总体 $0.3\langle 1\rangle + 0.7\langle 0\rangle$，求 $\mathsf{P}(|\overline{X} - 0.3| \leq 0.02)$。

解. $m_1 = m_2 = 0.3$，于是 $\sqrt{(m_2 - m_1^2)/100} \approx 0.0458$。利用结果 (6.19) 有

$$\mathsf{P}(|\overline{X} - 0.3| \leq 0.02) = \mathsf{P}\left(\frac{|\overline{X} - 0.3|}{0.0458} \leq 0.44\right) = 2\Phi(0.44) - 1 \approx 0.34$$

练习 6.1. 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 6.5. 仿照随机变量的变异系数、偏度系数和峰度系数（见定义 2.21），样本变异系数定义为 $C_v = S/\overline{X}$，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 $C_k = B_4/B_2^2 - 3$。

```
cv <- function (x){                          # 变异系数的 R 函数
    c <- sd(x)/mean(x)                       # 计算变异系数
    c                                        # 返回结果
}
skewness <- function (x){                    # 偏度系数的 R 函数
    n <- length(x)                           # 样本量
    x <- x - mean(x)                         #
    s <- sqrt(n) * sum(x^3)/(sum(x^2)^(3/2)) # 计算偏度系数
    s                                        # 返回结果
}
kurtosis <- function (x){                    # 峰度系数的 R 函数
    n <- length(x)                           # 样本量
    x <- x - mean(x)                         #
    k <- n * sum(x^4)/(sum(x^2)^2) - 3       # 计算峰度系数
    k                                        # 返回结果
}
```

利用上面定义的 R 的函数 `skewness` 和 `kurtosis` 算得 Iris 数据中 setosa 类的花瓣长度数据（见例 6.1）的变异系数 0.1187852，偏度系数 0.1031751 和峰度系数 0.8045921。

6.2 样本统计量及其性质

样本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。而有关未知参数的统计推断是通过统计量进行的（这个过程的图示如下），譬如，若总体期望 μ 未知，则可以通过样本均值 $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ 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。

$$\begin{array}{ccc} \text{总体 } X & \xrightarrow{\ ?\ } & \text{总体特征} \\ \text{随机抽样}\downarrow & & \uparrow \\ \text{简单随机样本 } X_1,\cdots,X_n & \xrightarrow{\text{统计量}} & \text{样本特征} \end{array}$$

定义 6.6 (抽样分布)**.** 统计量 $T = T(X_1, X_2, \cdots, X_n)$ 的分布称作 T 的抽样分布，它完全由样本 $X_1, X_2, \cdots, X_n$ 的分布唯一决定。

例 6.4. 下表给出了在若干不同总体分布之下，样本均值 $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ 或由它构造的新统计量的抽样分布。例如，倒数第二行是因为 Cauchy(μ, λ) 分布的特征函数为 $\exp\{i\mu t - \lambda|t|\}$；最后一行是因为 $2n\beta\overline{X}$ 与 χ^2_{2n} 的特征函数都为 $(1-2it)^{-n}$。

表 6.1: 在不同的总体之下，由样本均值 $\overline{X}$ 构造的统计量的抽样分布。

总体分布	抽样分布
$\mathrm{N}(\mu, \sigma^2)$	$\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n)$
$\mathrm{B}(m, p)$	$n\overline{X} \sim \mathrm{B}(mn, p)$
$\mathrm{Poisson}(\lambda)$	$n\overline{X} \sim \mathrm{Poisson}(n\lambda)$
$\mathrm{Cauchy}(\mu, \lambda)$	$\overline{X} \sim \mathrm{Cauchy}(\mu, \lambda)$
$\mathrm{Expon}(\beta)$	$2n\beta\overline{X} \sim \chi^2_{2n}$

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，无非是强调它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 $X_1^{(k)}, X_2^{(k)}, \cdots, X_n^{(k)}$，算出 $T^{(k)} = T(X_1^{(k)}, X_2^{(k)}, \cdots, X_n^{(k)})$，则 $T^{(k)}, k = 1, 2, \cdots, m$ 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 $T^{(1)}, T^{(2)}, \cdots, T^{(m)}$ 的经验分布近似得到。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 T 的经验分布（详见 §6.2.1）。

例 6.5. 样本均值 $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ 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：

$$\overline{X}_{\text{new}} = \overline{X} + \frac{1}{n+1}(X_{n+1} - \overline{X}) \tag{6.21}$$

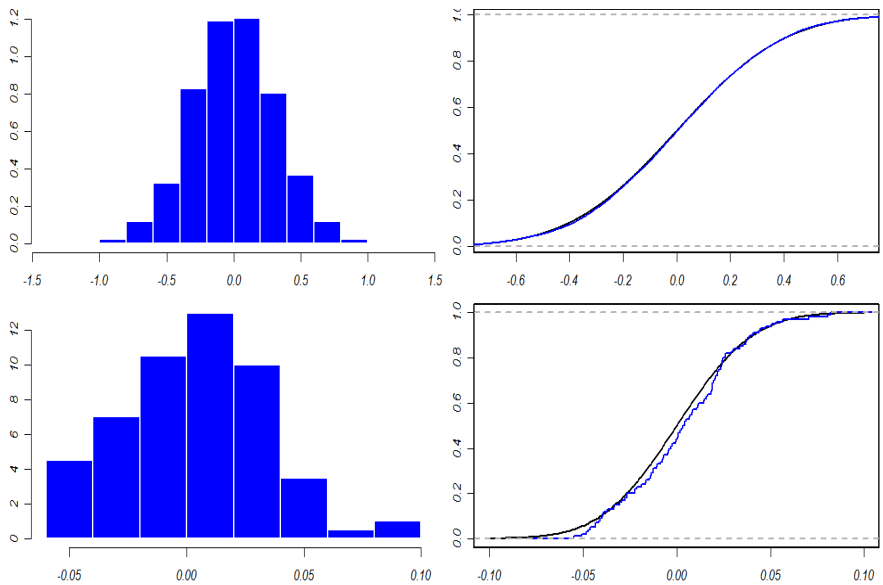


图 6.5: 容量为 n 的样本 $X_1, X_2, \cdots, X_n$ 来自从总体 $\mathrm{N}(0,1)$，样本均值 $\overline{X} \sim \mathrm{N}(0, 1/\sqrt{n})$。反复抽样 m 次，观察 $\overline{X}^{(1)}, \overline{X}^{(2)}, \cdots, \overline{X}^{(m)}$ 的直方图和经验分布函数在 n, m 的不同设置之下有何特点：第一行 $n = 10, m = 10^4$，第二行 $n = 10^3, m = 10^2$。显而易见，样本量越大，$\overline{X}$ 越紧密围绕在总体均值周围（第二行左图，样本方差很小）；反复抽样次数越多，$\overline{X}^{(1)}, \overline{X}^{(2)}, \cdots, \overline{X}^{(m)}$ 的经验分布越接近 $\overline{X}$ 的抽样分布（第一行右图，二者几乎重合）。

*自助法是美国统计学家 Bradley Efron (1938-) 于 1979 年提出的一种基于重抽样 (resampling) 的模拟方法 [29,32,33]，是很多统计推断问题的有效工具。

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 定理 2.15（或 Fisher-Geary 定理）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，通过自助法 (bootstrap method)*可获得统计量的经验分布。第二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。

学习目标 (1) 正态总体下几个常见统计量的抽样分布；(2) Fisher-Geary 定理；(3) 大致了解自助法；(4) 理解统计量的充分性并掌握 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。

6.2.1　统计量的抽样分布

已知总体分布，要搞清楚任一统计量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。但是对于正态总体来说情况要好些，有若干重要的统计量的抽样分布可以轻易求得，导致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*。当简单随机样本 $X_1, \cdots, X_n$ 来自正态总体 $\mathrm{N}(\mu, \sigma^2)$ 时，约定用 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$ 表示。

性质 6.4. 如果样本 $X_1, X_2, \cdots, X_n$ 来自正态总体 $\mathrm{N}(\mu, \sigma^2)$，则有

$$\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi_1^2 \tag{6.22}$$

↬ **定理 6.5** (Fisher-Geary, 1925, 1936)**.** 样本 $X_1, \cdots, X_n$ 来自一个正态总体当且仅当样本均值 $\overline{X}$ 与样本方差 S^2 独立。

证明. “⇒”即 Fisher 定理 2.15。“⇐”的证明由爱尔兰统计学家 Roy Charles Geary (1896-1983) 于 1936 年给出 [39]，已超出本书范围。 □

↬ **定理 6.6.** 已知样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，则有

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \tag{6.23}$$

证明. 我们知道

$$\sum_{j=1}^{n} \frac{(X_j - \mu)^2}{\sigma^2} \sim \chi_n^2 \text{ 并且 } \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi_1^2$$

*统计学在相当长的一段时间里把正态总体当作研究重点，为了能用上已有成果，统计推断的前提往往都是基于正态总体。

根据 $\sum_{j=1}^{n}(X_j-\overline{X})=0$，显然有

$$\sum_{j=1}^{n}\frac{(X_j-\mu)^2}{\sigma^2}=\sum_{j=1}^{n}\frac{(X_j-\overline{X}+\overline{X}-\mu)^2}{\sigma^2}=\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2+\frac{(n-1)S^2}{\sigma^2}$$

因为总体是正态分布，所以 $\overline{X}$ 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 $(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$。 □

定理 6.7. 已知样本 $X_1,\cdots,X_n\overset{iid}{\sim}\mathrm{N}(\mu,\sigma^2)$，样本方差为 S^2，则

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S}\sim t(n-1) \tag{6.24}$$

证明. 由 $\sqrt{n}(\overline{X}-\mu)/\sigma\sim\mathrm{N}(0,1)$ 和 $(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$，我们有

$$\frac{\sqrt{n}(\overline{X}-\mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}}=\frac{\sqrt{n}(\overline{X}-\mu)}{S}\sim t(n-1)$$ □

例 6.6. 设样本 $X_1,\cdots,X_n\overset{iid}{\sim}\mathrm{N}(\mu,\sigma^2)$ 的均值和方差分别为 $\overline{X}$ 和 S^2，若从总体再抽取一个样本点 X_{n+1}，问统计量 $Y=\sqrt{n/(n+1)}(X_{n+1}-\overline{X})/S$ 服从什么分布？

解. 由 $X_{n+1}-\overline{X}\sim\mathrm{N}(0,(n+1)\sigma^2/n)$ 和定理 6.6 得 $Y\sim t(n-1)$。

性质 6.5. 已知来自两个独立总体的样本 $X_1,\cdots,X_m\overset{iid}{\sim}\mathrm{N}(\mu_X,\sigma_X^2)$ 和 $Y_1,\cdots,Y_n\overset{iid}{\sim}\mathrm{N}(\mu_Y,\sigma_Y^2)$ 的样本均值和样本方差分别为 $\overline{X},S_X^2,\overline{Y},S_Y^2$，则

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}\sim F(m-1,n-1) \tag{6.25}$$

$$\frac{[\overline{X}-\overline{Y}-(\mu_X-\mu_Y)]\sqrt{\frac{m+n-2}{\sigma_X^2/m+\sigma_Y^2/n}}}{\sqrt{(m-1)S_X^2/\sigma_X^2+(n-1)S_Y^2/\sigma_Y^2}}\sim t(m+n-2) \tag{6.26}$$

证明. 因为这两个总体是独立的，于是

$$\overline{X}-\overline{Y}\sim \mathrm{N}\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}\right)\ \text{且}\ \frac{(m-1)S_X^2}{\sigma_X^2}+\frac{(n-1)S_Y^2}{\sigma_Y^2}\sim\chi^2_{m+n-2}$$

由定理 6.6 以及 F 分布、t 分布的定义，易证。 □

当统计量 $T=T(X_1,\cdots,X_n)$ 的抽样分布很难求得时，自助法是估计 T 的分布及方差的一个便捷的方法：(1) 有放回地从 $X_1,X_2,\cdots,X_n$ 中抽取 $X_*^{(1)},X_*^{(2)},\cdots,X_*^{(n)}$；(2) 计算 $T_*=T(X_*^{(1)},X_*^{(2)},\cdots,X_*^{(n)})$。重复步骤 (1) 和 (2) m 遍得到样本 $T_*^{(1)},\cdots,T_*^{(m)}$，估计 $\mathsf{V}(T)=\sigma^2$ 为

$$\hat{\sigma}^2=\frac{1}{m}\sum_{k=1}^{m}\left(T_*^{(k)}-\frac{1}{m}\sum_{j=1}^{m}T_*^{(j)}\right)^2 \tag{6.27}$$

例 6.7. 已知样本 $X_1,X_2,\cdots,X_n$ 来自正态总体 $\mathrm{N}(\mu,\sigma^2)$，其中参数 μ,σ^2 未知。利用自助法得到样本均值 $\overline{X}$ 的经验分布。

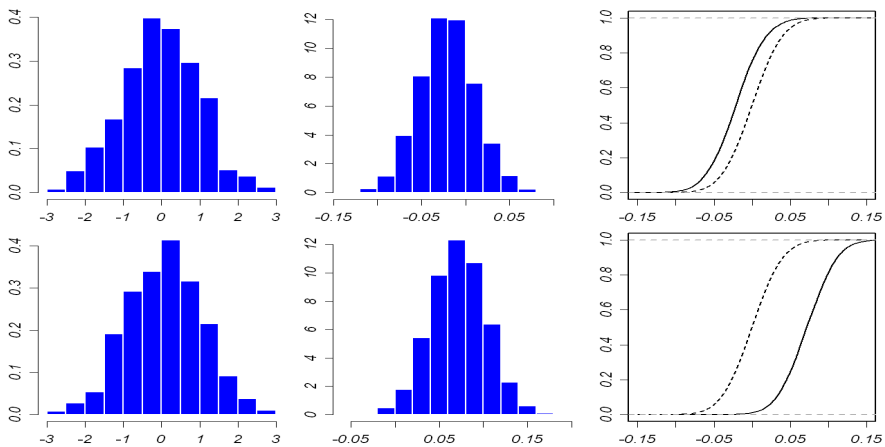
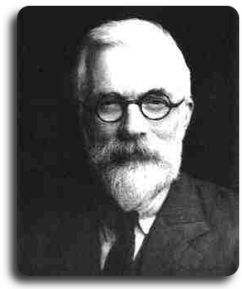


图 6.6: 自助法：从样本 $X_1,X_2,\cdots,X_n\overset{iid}{\sim}\mathrm{N}(0,1),n=10^3$（样本直方图见左列）中有放回地抽取得到新样本 $X_*^{(1)},X_*^{(2)},\cdots,X_*^{(n)}$ 算得均值 $\overline{X}_*$。重复此过程 $m=10^4$ 遍得到来自总体 $\overline{X}_*$ 的样本 $\overline{X}_*^{(1)},\overline{X}_*^{(2)},\cdots,\overline{X}_*^{(m)}$（直方图见中列），进而得到对应的经验分布函数（见右列图中的实线），其目标是逼近 $\overline{X}\sim\mathrm{N}(0,1/n)$（右列图中的虚线），最后的效果依赖于原始样本 $X_1,X_2,\cdots,X_n$。

6.2.2 统计量的充分性



由样本构造出的统计量与样本相比，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 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 6.7 (充分性)**.** 已知 $T = T(X_1, X_2, \cdots, X_n)$ 为一个统计量，如果 $\theta \in \Theta$ 样本的条件分布 $F_\theta(x_1, \cdots, x_n | T = t)$ 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 $\mathsf{P}_\theta\{X_1 = x_1, \cdots, X_n = x_n | T = t\}$ 或条件密度函数 $f_\theta(x_1, \cdots, x_n | T = t)$ 与 θ 无关。哪怕原始数据丢失了，凭借 $T = t$ 也能通过条件分布 $F(x_1, \cdots, x_n | T = t)$ 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 6.8. 设样本 $X_1, X_2 \overset{iid}{\sim} \text{Poisson}(\lambda)$，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。

$$\begin{aligned}&\mathsf{P}\{X_1 = x_1, X_2 = x_2 | X_1 + X_2 = t\}\\ &= \begin{cases} \dfrac{\mathsf{P}\{X_1 = x_1, X_2 = t - x_1\}}{\mathsf{P}\{X_1 + X_2 = t\}} = C_t^{x_1}/2^t & \text{如果 } x_1 + x_2 = t \\ 0 & \text{否则} \end{cases}\end{aligned}$$

*该结果于1935 年被 J. Neyman 重新发现，并于 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。Savage 是贝叶斯学派的代表人物之一。

上式的计算用到了性质 4.5。下面说明 $X_1 + 2X_2$ 不是充分统计量。

$$\begin{aligned}\mathsf{P}\{X_1 = 0, X_2 = 1 | X_1 + 2X_2 = 2\} &= \frac{\mathsf{P}\{X_1 = 0, X_2 = 1\}}{\mathsf{P}\{X_1 + 2X_2 = 2\}} \\ &= \frac{e^{-\lambda}(\lambda e^{-\lambda})}{\mathsf{P}\{X_1 = 0, X_2 = 1\} + \mathsf{P}\{X_1 = 2, X_2 = 0\}} \\ &= \frac{\lambda e^{-2\lambda}}{\lambda e^{-2\lambda} + (\lambda^2/2)e^{-2\lambda}} = \frac{2}{\lambda + 2}\end{aligned}$$

在例 6.8 中，假如原始数据 $\{x_1, x_2\}$ 丢失，利用 $Y_1 \sim \mathrm{B}(t, 1/2)$ 产生的随机数据 $\{y_1, y_2 = t - y_1\}$ 来“恢复”原始观察数据（假设数据都是从小到大排序），约定用欧氏距离 $\varepsilon = \sqrt{\sum_{j=1}^{2}(x_j - y_j)^2}$ 来看数据恢复的效果，显然 ε 越小效果越好。下面通过大量独立的重复试验来。

```
## 目的：利用充分统计量来“恢复”丢失数据
## 输出：恢复数据与真实数据欧氏距离的直方图
RepeatNum <- 1000          # 试验次数
dist <- rep(0,RepeatNum) # 距离的初始化
lambda <- 3                # 参数的设定

for (i in 1:RepeatNum){
x  <- sort(rpois(2,lambda))
t  <- sum(x)               # 充分统计量的值
y1 <- rbinom(1,t,1/2)
y <- sort(c(y1,t-y1))      # 产生“恢复”数据
dist[i] <- sum((x-y)^2)  # 欧氏距离
}
hist(dist,freq=FALSE)
```

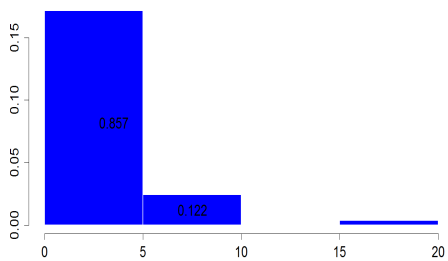


图 6.7: 通过 1000 次独立试验看例 6.8 中数据恢复的效果 ε。

例 6.9. 令样本 $X_1, X_2, \cdots X_n \overset{iid}{\sim} p\langle 1\rangle + (1-p)\langle 0\rangle$，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

$$\mathsf{P}\left\{X_1 = x_1, \cdots, X_n = x_n \middle| \sum_{j=1}^{n} X_j = t\right\}$$

$$= \begin{cases} \dfrac{\mathsf{P}\left\{X_1 = x_1, \cdots, X_n = x_n, \sum_{j=1}^{n} X_j = t\right\}}{\mathsf{P}\left\{\sum_{j=1}^{n} X_j = t\right\}} & \text{如果 } \sum_{j=1}^{n} x_j = t \\ = \dfrac{p^{\sum_{j=1}^{n} x_j}(1-p)^{n-\sum_{j=1}^{n} x_j}}{C_n^t p^t (1-p)^{n-t}} = \dfrac{1}{C_n^t} & \\ 0 & \text{否则} \end{cases}$$

通过充分性的定义 6.7 和上面两个离散型的例子可以看出：$\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}\}$ 可以分解为不含 θ 的有关 $\boldsymbol{x}$ 的某函数与 $\mathsf{P}_\theta\{T(\boldsymbol{X})=T(\boldsymbol{x})\}$ 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条件概率的繁琐计算，这就是著名的 Fisher 因子分解定理。

定理 6.8 (Fisher 因子分解定理*, 1925)**.** 设样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^\mathsf{T}$ 的概率密度函数为 $f_\theta(\boldsymbol{x})$ 或概率函数为 $f_\theta(\boldsymbol{x})=\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}\}$，其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^\mathsf{T}$，统计量 $T(\boldsymbol{X})$ 对未知参数 θ 而言是充分的当且仅当 $f_\theta(\boldsymbol{x})=h(\boldsymbol{x})g_\theta[T(\boldsymbol{x})]$，其中非负（可测）函数 $h(\boldsymbol{x})$ 不依赖于 θ，非负（可测）函数 $g_\theta[T(\boldsymbol{x})]$ 是关于 θ 和 $T(\boldsymbol{x})$ 的函数。

证明. 一般情况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》 [9] 第一章的附录。这里仅考虑总体是离散型的。往证“$\Rightarrow$”：令 $T(\boldsymbol{X})$ 是充分统计量，则 $\mathsf{P}\{\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=t\}$ 与参数 θ 无关。当 $T(\boldsymbol{x})=t$ 时，$\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x})=\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x},T(\boldsymbol{X})=t\}=\mathsf{P}\{\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=t\}\mathsf{P}_\theta\{T(\boldsymbol{X})=t\}$。对那些满足 $\forall\theta\in\Theta,\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x})=0$ 的 $\boldsymbol{x}$，定义 $h(\boldsymbol{x})=0$。对那些满足 $\exists\theta$ 使得 $\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x})>0$ 的 $\boldsymbol{x}$，定义 $h(\boldsymbol{x})=\mathsf{P}\{\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=t\}$，并定义 $g_\theta[T(\boldsymbol{x})]=\mathsf{P}_\theta\{T(\boldsymbol{X})=T(\boldsymbol{x})=t\}$。
下面往证“$\Leftarrow$”：对任意固定的 t_0 有

$$\mathsf{P}_\theta\{T(\boldsymbol{X})=t_0\}=\sum_{\{\boldsymbol{x}:T(\boldsymbol{x})=t_0\}}\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}\}=\sum_{\{\boldsymbol{x}:T(\boldsymbol{x})=t_0\}}h(\boldsymbol{x})g_\theta\left[T(\boldsymbol{x})\right]=g_\theta(t_0)\sum_{\{\boldsymbol{x}:T(\boldsymbol{x})=t_0\}}h(\boldsymbol{x})$$

若 $\mathsf{P}_\theta\{T(\boldsymbol{X})=t_0\}=0$，结果是平凡的。设 $\mathsf{P}_\theta\{T(\boldsymbol{X})=t_0\}>0$：若 $T(\boldsymbol{x})\neq t_0$，则 $\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=t_0\}=0$；若 $T(\boldsymbol{x})=t_0$，则

$$\begin{aligned}\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=t_0\}&=\frac{\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x},T(\boldsymbol{X})=t_0\}}{\mathsf{P}_\theta\{T(\boldsymbol{X})=t_0\}}=\frac{\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}\}}{\mathsf{P}_\theta\{T(\boldsymbol{X})=t_0\}}\\&=\frac{h(\boldsymbol{x})g_\theta(t_0)}{g_\theta(t_0)\sum\limits_{\{\boldsymbol{x}:T(\boldsymbol{x})=t_0\}}h(\boldsymbol{x})}=\frac{h(\boldsymbol{x})}{\sum\limits_{\{\boldsymbol{x}:T(\boldsymbol{x})=t_0\}}h(\boldsymbol{x})}\end{aligned}$$

不管怎样，$\mathsf{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=t_0\}$ 都不依赖于 θ，得证。 □

*有的文献中也称之为 Neyman 因子分解定理。

例 6.10. 已知样本 $X_1, X_2 \cdots, X_n \overset{iid}{\sim} \mathrm{U}[0,\theta]$，极值 $X_{(n)} = \max(X_1, X_2 \cdots, X_n)$ 对未知参数 θ 而言是充分的。这是因为样本的联合密度函数为

$$\begin{aligned} f_\theta(x_1, x_2, \cdots, x_n) &= \begin{cases} \theta^{-n} & \text{当 } 0 \le x_1, x_2, \cdots, x_n \le \theta \\ 0 & \text{其他} \end{cases} \\ &= J(x_{(1)})[\theta^{-n} J(\theta - x_{(n)})] \end{aligned}$$

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 6.8 可证得结论。

练习 6.2. 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 $\mathrm{U}\{1, 2, \cdots, m\}$，其中 m 未知，则 $X_{(n)}$ 对 m 而言是充分的。提示：$\mathsf{P}(X_1 = x_1, \cdots, X_n = x_n) = J(x_{(1)} - 1)[m^{-n} J(m - x_{(n)})]$。

例 6.11. 令简单随机样本 $X_1, X_2, \cdots, X_n$ 来自正态总体 $\mathrm{N}(\mu, \sigma^2)$，其中参数 μ, σ^2 未知，$\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 的概率密度函数为

$$\begin{aligned} f_{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{ -\frac{\sum_{j=1}^n (x_j - \mu)^2}{2\sigma^2} \right\} \\ &= \exp\left\{ \frac{\mu \sum_{j=1}^n x_j}{\sigma^2} - \frac{\sum_{j=1}^n x_j^2}{2\sigma^2} - \frac{n}{2}\left[\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2) \right] \right\} \end{aligned}$$

于是统计量 $(\overline{X}, A_2)^\mathsf{T}$ 与 $(\overline{X}, S^2)^\mathsf{T}$ 对 $\boldsymbol{\theta} = (\mu, \sigma^2)^\mathsf{T}$ 而言都是充分的。

定义 6.8. k-参数指数族 (exponential family) $\{f_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k, \boldsymbol{x} \in \mathbb{R}^d\}$ 中每个密度函数或概率函数 $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ 都具有如下形式：

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\left\{ \sum_{j=1}^k q_j(\boldsymbol{\theta}) T_j(\boldsymbol{x}) + g(\boldsymbol{\theta}) \right\} \tag{6.28}$$

其中 $g(\boldsymbol{\theta}), q_j(\boldsymbol{\theta}), j = 1, 2, \cdots, k$ 都是 Θ 上的实值函数，$h(\boldsymbol{x}), T_j(\boldsymbol{x}), j = 1, 2, \cdots, k$ 都是 $\mathbb{R}^d$ 上的实值函数。

例 6.12. 二项分布 B(m,p)、Poisson 分布 Poisson(λ)、正态分布 N(μ,σ^2)，还有例 6.11 中样本的概率密度函数 $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ 都属于指数族。

$$f(x)=C_m^x\exp\left\{x\ln\frac{p}{1-p}+m\ln(1-p)\right\}，\text{其中 } p\in(0,1),x\in\{0,\cdots,m\}$$

$$f(x)=\frac{1}{x!}\exp\{x\ln\lambda-\lambda\}，\text{其中 } \lambda>0 \text{ 且 } x\in\{0,1,2,\cdots\}$$

$$\phi(x|\mu,\sigma^2)=\exp\left\{\frac{\mu x}{\sigma^2}-\frac{x^2}{2\sigma^2}-\frac{1}{2}\left[\frac{\mu^2}{\sigma^2}+\ln(2\pi\sigma^2)\right]\right\}，\text{其中 } \mu\in\mathbb{R},\sigma^2>0$$

练习 6.3. 假设参数都是未知的，验证多项分布、Gamma 分布、Beta 分布都属于指数族。

定理 6.9. 如果总体分布属于指数族 (6.28)，设 $X_1,X_2,\cdots,X_n$ 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。

$$T(X_1,X_2,\cdots,X_n)=\left[\sum_{i=1}^n T_1(X_i),\cdots,\sum_{i=1}^n T_k(X_i)\right]^{\mathsf{T}} \tag{6.29}$$

证明. 样本 $X_1,X_2,\cdots,X_n$ 的密度函数为

$$f_{\boldsymbol{\theta}}(x_1,x_2,\cdots,x_n)=\prod_{i=1}^n h(x_i)\exp\left\{\sum_{j=1}^k q_j(\boldsymbol{\theta})\sum_{i=1}^n T_j(x_i)+ng(\boldsymbol{\theta})\right\}$$

由 Fisher 因子分解定理 6.8，得证。 □

练习 6.4. 样本 $X_1,X_2,\cdots,X_n$ 来自总体 N(μ,σ^2)，试证明：(1) 若 σ^2 已知，统计量 $\overline{X}$ 对未知参数 μ 而言是充分的。(2) 若 σ^2 未知，$\overline{X}$ 对未知参数 μ 而言不是充分的。(3) 若 μ 未知，S^2 对未知参数 σ^2 而言不是充分的。(4) 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（答案参见第 262 页的例 7.9）。

6.3　习题

6.1. 设 $\overline{X}$ 是样本 $X_1,\cdots,X_n$ 的均值，试证明：当 $c=\overline{X}$ 时，$\sum_{i=1}^n(X_i-c)^2$ 的值达到最小。

6.2. 设样本 $X_1,X_2,\cdots,X_n$ 与样本 $Y_1,Y_2,\cdots,Y_n$ 之间有关系 $Y_i=(X_i-a)/b$，其中 $b\neq 0,a$ 都是常数，求样本平均值 $\overline{Y}$ 与 $\overline{X}$，以及样本方差 S_Y^2 与 S_X^2 之间的关系。

6.3. 设简单随机样本 $X_1,X_2,\cdots,X_n$ 来自总体 $X\sim F(x)$。若 X 的二阶矩存在，$\overline{X}$ 为样本均值，试证明：$X_i-\overline{X}$ 与 $X_j-\overline{X}$ 的相关系数 $\rho=-(n-1)^{-1}$，其中 $i,j=1,2,\cdots,n$ 且 $i\neq j$。

6.4. 设样本 $X_1,X_2,\cdots,X_n\overset{iid}{\sim}p\langle 1\rangle+(1-p)\langle 0\rangle$。(1) 求样本均值 $\overline{X}$ 的分布列以及 $\mathsf{E}(\overline{X})$ 和 $\mathsf{V}(\overline{X})$；(2) 若 S^2 为样本方差，求 $\mathsf{E}(S^2)$；(3) 若样本值有 m 个 1，其余的为 0，求其经验分布函数。

6.5. 设样本 $X_1,X_2,\cdots,X_{10}\overset{iid}{\sim}\mathrm{N}(\mu,4^2)$，$S^2$ 为样本方差。若已知 $\mathsf{P}\{S^2>a\}=0.1,\chi_9^2(0.9)\approx 14.684$，求 a。

6.6. 已知样本 $X_1,\cdots,X_n\overset{iid}{\sim}\mathrm{Expon}(\lambda)$，试求极值 $X_{(1)}$ 的均值与方差。

6.7. 设 $\overline{X}_1$ 和 $\overline{X}_2$ 分别是取自正态总体 $\mathrm{N}(\mu,\sigma^2)$ 的容量为 n 的两个简单随机样本 $X_{11},X_{12},\cdots,X_{1n}$ 和 $X_{21},X_{22},\cdots,X_{2n}$ 的均值，试确定 n 使得 $\mathsf{P}(|\overline{X}_1-\overline{X}_2|>\sigma)=0.01$。

☆ 6.8. 设样本 $X_1,X_2\overset{iid}{\sim}\mathrm{N}(0,\sigma^2)$，求概率 $\mathsf{P}\{(X_1+X_2)^2/(X_1-X_2)^2<4\}$。

6.9. 设 $X_1,X_2,\cdots,X_9$ 是来自正态总体的简单随机样本，令 $Y_1=(X_1+X_2+\cdots+X_6)/6,Y_2=(X_7+X_8+X_9)/3$ 且 $S^2=\frac{1}{2}\sum_{k=7}^9(X_k-Y_2)^2$，试证明：$\sqrt{2}(Y_1-Y_2)/S\sim t(2)$。

6.10. 已知样本 $X_1,X_2,\cdots,X_n\overset{iid}{\sim}\mathrm{N}(\mu,2^2)$ 均值为 $\overline{X}$，要使 $\mathsf{E}(\overline{X}-\mu)^2\leq 0.1$ 成立，则样本量 n 不小于多少？

6.11. 已知两个总体 $X, Y \overset{iid}{\sim} \mathrm{N}(0, 4^2)$，简单随机样本 $X_1, X_2, \cdots, X_{16}$ 和 $Y_1, Y_2, \cdots, Y_{16}$ 分别来自总体 X 和 Y，问 $V = \sum_{i=1}^{16} X_i / \sqrt{\sum_{i=1}^{16} Y_i^2}$ 服从什么分布？

6.12. 设简单随机样本 $X_1, X_2, \cdots, X_n$ 和 $Y_1, Y_2, \cdots, Y_n$ 为分别来自总体 $X, Y \overset{iid}{\sim} \mathrm{N}(0, 1)$，问统计量 $W = (X_1+X_2+\cdots+X_n) / \sqrt{Y_1^2 + Y_2^2 + \cdots + Y_n^2}$ 服从什么分布？

6.13. 已知样本 $X_1, X_2, X_3, X_4 \overset{iid}{\sim} \mathrm{N}(0, 2^2)$ 且 $Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2 \sim \chi_2^2$，求 a, b 的值。

6.14. 已知样本 $X_1, X_2, \cdots, X_5 \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ 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？

6.15. 求总体 $\mathrm{N}(20, 3)$ 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。

6.16. 设样本 $X_1, X_2, \cdots, X_8 \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$，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 $Y_3 = \sqrt{2/3}(X_1 + X_2 + X_3)/|X_4 - X_5|$。

6.17. 设样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，样本均值 $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$，令 $Y_j = X_j + a\overline{X}$，其中 a 为常数。求 Y_j 的分布。

6.18. 不管简单随机样本 $X_1, \cdots, X_n$ 来自总体 $\mathrm{Expon}(\lambda)$ 还是总体 $\mathrm{Poisson}(\lambda)$，试证明：$T = \sum_{j=1}^{n} X_j$ 对未知参数 λ 而言是充分统计量。

☆ 6.19. 设简单随机样本 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ 来自二元正态总体 $\mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，其中参数 $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$ 都未知，试给出一个充分统计量。

第七章

参数估计理论

数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。譬如，已知总体 X 服从正态分布 $\mathrm{N}(\mu, \sigma^2)$，其中方差 σ^2 已知，而均值 μ 未知，人们可以从样本 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 出发“猜测” μ 的取值。频率派有两种方法来估计 μ：一种是直接给出 μ 的估计量 $\hat{\mu}(\boldsymbol{X})$，称为参数的点估计 (point estimation)；另一种是给出以某个概率覆盖住 μ 的区间表示 $[\underline{\mu}(\boldsymbol{X}), \overline{\mu}(\boldsymbol{X})]$，称为参数的区间估计 (interval estimation)，其中统计量 $\underline{\mu}(\boldsymbol{X}) < \overline{\mu}(\boldsymbol{X})$。

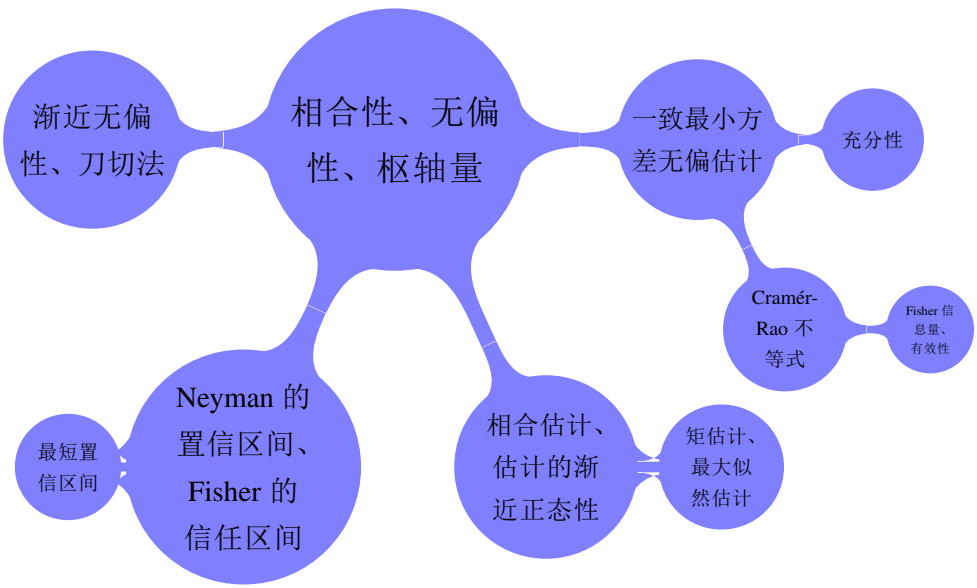
在频率派看来，参数都是固定值，不管它是已知的还是未知的。而贝叶斯学派则认为未知参数是随机变量（有先验分布和后验分布），根据这一观念上的差别可以区分经典统计方法和贝叶斯方法*。本章只关注频率派的参数估计方法，作为补充，§7.2.2 将介绍 Fisher 的信任区间估计，它有别于传统的置信区间方法和贝叶斯方法，一直备受争议。有趣的是 Fisher 信任推断得到的结果通过贝叶斯方法也多能得到，而 Fisher 本人自始至终是强烈反对贝叶斯学派的。

*第十一章将介绍贝叶斯参数估计，即可信区间 (credible interval) 估计。在信任区间 (fiducial interval) 估计中，Fisher 也把未知参数视作随机变量，具体内容见 §7.2.2。

点估计的第一步是由样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 构造具有一定“好品质”（如相合性、无偏性、有效性等）的统计量 $T(\boldsymbol{X})$ 来估算总体分布中的未知参数 θ，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 $\hat{\theta}(\boldsymbol{X})$ 或 $\hat{\theta}$，有时为了突出样本量 n，也记作 $\hat{\theta}_n(\boldsymbol{X})$ 或 $\hat{\theta}_n$。得到样本值 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathsf{T}}$ 后，经过计算所得的数值（或向量）$T(\boldsymbol{x})$ 称作 θ 的估计值，也常记作 $\hat{\theta}(\boldsymbol{x})$ 或 $\hat{\theta}_n(\boldsymbol{x})$，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。

有效估计是最好的无偏估计，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 7.4。针对有偏估计，刀切法有助于修正偏倚，§7.1.2 对它做了简介。

点估计的方法主要有 K. Pearson 提出的矩方法和 R. A. Fisher 提出的最大似然法，二者各有特点，在某些条件下最大似然法要略胜一筹（具体讨论见 §7.1.3）。



7.1　点估计及其优良性

点估计的目标就是构造统计量 $T = T(X_1, X_2, \cdots, X_n)$ 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) 和均方误差 (mean square error, MSE) 来评介统计量 T。

$$\begin{aligned} \text{BIAS}(\theta, T) &= \mathsf{E}_\theta(T) - \theta \qquad (7.1) \\ \text{MSE}(\theta, T) &= \mathsf{E}_\theta(T-\theta)^2 = \mathsf{E}_\theta\left[T - \mathsf{E}_\theta(T) + \mathsf{E}_\theta(T) - \theta\right]^2 \\ &= \mathsf{E}_\theta\left[T - \mathsf{E}_\theta(T)\right]^2 + \left[\mathsf{E}_\theta(T) - \theta\right]^2 \\ &= \mathsf{V}_\theta(T) + \left[\text{BIAS}(\theta, T)\right]^2 \qquad (7.2) \end{aligned}$$

显然，均方误差越小意味着估计的精度越高。为了刻画样本包含多少未知参数 θ 的信息，Fisher 提出了信息量*的概念 [37]。

定义 7.1 (Fisher 信息量)**.** 设连续型随机变量 X 的概率密度函数为 $f_\theta(x)$，其中 $\theta \in \Theta$ 为未知参数，Fisher 信息量 (Fisher information) $\mathcal{I}(\theta)$ 定义为

$$\mathcal{I}(\theta) = \mathsf{E}_\theta\left[\frac{\partial \ln f_\theta(X)}{\partial \theta}\right]^2 = \int_{-\infty}^{+\infty}\left[\frac{\partial \ln f_\theta(x)}{\partial \theta}\right]^2 f_\theta(x)dx \qquad (7.3)$$

根据式 (2.109)，Fisher 信息量具有另外一个等价的定义形式：

$$\mathcal{I}(\theta) = -\mathsf{E}_\theta\left[\frac{\partial^2 \ln f_\theta(X)}{\partial \theta^2}\right] = -\int_{-\infty}^{+\infty}\frac{\partial^2 \ln f_\theta(x)}{\partial \theta^2} f_\theta(x)dx \qquad (7.4)$$

对于离散型随机变量 X，Fisher 信息量定义为

$$\mathcal{I}(\theta) = \mathsf{E}_\theta\left[\frac{d \ln \mathsf{P}_\theta(X)}{d\theta}\right]^2 = \sum_j\left[\frac{d \ln \mathsf{P}_\theta(x_j)}{d\theta}\right]^2 \mathsf{P}_\theta(x_j) \qquad (7.5)$$

*Fisher 信息量在贝叶斯统计学中用于计算 Jeffreys 先验分布（见第十一章）。

一般地，对于向量参数 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^\mathsf{T} \in \Theta$，Fisher 信息阵 (Fisher information matrix) 定义为 $\mathcal{I}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}(\boldsymbol{Y}\boldsymbol{Y}^\mathsf{T})$，其中随机向量 $\boldsymbol{Y} = \nabla_{\boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(X)$（参见附录 G），即 $\mathcal{I}(\boldsymbol{\theta})$ 的第 (i, j) 元素定义为

$$\begin{aligned}\mathcal{I}_{ij}(\boldsymbol{\theta}) &= \mathsf{E}_{\boldsymbol{\theta}}\left\{\left[\frac{\partial \ln f_{\boldsymbol{\theta}}(X)}{\partial \theta_i}\right] \cdot \left[\frac{\partial \ln f_{\boldsymbol{\theta}}(X)}{\partial \theta_j}\right]\right\} \\ &= \int_{-\infty}^{\infty}\left[\frac{\partial \ln f_{\boldsymbol{\theta}}(x)}{\partial \theta_i}\right] \cdot \left[\frac{\partial \ln f_{\boldsymbol{\theta}}(x)}{\partial \theta_j}\right] f_{\boldsymbol{\theta}}(x)dx \end{aligned} \tag{7.6}$$

注记 7.1. Fisher 信息阵 $\mathcal{I}(\boldsymbol{\theta})$ 是一个 $k \times k$ 半正定对称阵，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) [13]，通过几何不变量来研究统计不变量。

例 7.1. 令总体为 $X \sim p\langle 1\rangle + (1-p)\langle 0\rangle$，其中参数 p 未知，则

$$\mathcal{I}(p) = -\mathsf{E}_p\left\{\frac{\partial^2 \ln[p^X(1-p)^{1-X}]}{\partial p^2}\right\} = \mathsf{E}_p\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p(1-p)}$$

当 $p = 1/2$ 时，Fisher 信息量达到最小，此时熵（见第 113 页的式 2.54）是最大的。

例 7.2. 令总体为 $X \sim \mathrm{N}(\mu, \sigma^2)$，其中参数 μ 未知，σ^2 已知，则

$$\mathcal{I}(\mu) = -\mathsf{E}_\mu\left\{\frac{\partial^2 \ln \phi(X|\mu, \sigma^2)}{\partial \mu^2}\right\} = \mathsf{E}_\mu\left(\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}$$

此例的直观含义是方差越小，有关参数 μ 的 Fisher 信息量越大。

例 7.3. 令简单随机样本 $X_1, \cdots, X_n$ 来自密度函数为 $f_\theta(x)$ 的总体，则样本的 Fisher 信息量为 $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$，这是因为

$$\mathcal{I}_n(\theta) = \mathsf{E}_\theta\left[\frac{\partial \sum_{j=1}^n \ln f_\theta(X_j)}{\partial \theta}\right]^2 = \mathsf{E}_\theta\left[\sum_{j=1}^n \frac{\partial \ln f_\theta(X_j)}{\partial \theta}\right]^2 = \sum_{j=1}^n \mathsf{E}_\theta\left[\frac{\partial \ln f_\theta(X_j)}{\partial \theta}\right]^2$$

最后一步是根据式 (2.109) 而得。该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。

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》 [59]。

学习目标　(1) 理解相合性、无偏性、有效性等基本概念的含义以及它们之间的关系；(2) 掌握 Fisher 信息量和 Cramér-Rao 不等式；(3) 大致了解刀切法；(4) 熟练掌握矩方法和最大似然法及其应用。

*在第十三章还将介绍另一常见的点估计方法——期望最大化算法。

7.1.1 相合性与渐近正态性

结果 (6.19) 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 7.2 (相合性)**.** 当 $n \to \infty$ 时如果 $T_n = T(X_1, X_2, \cdots, X_n) \xrightarrow{a.s.} \theta$，称 T_n 是参数 θ 的*强相合估计* (strong consistent estimator)。当 $n \to \infty$ 时如果 $T_n = T(X_1, X_2, \cdots, X_n) \xrightarrow{P} \theta$，称 T_n 是 θ 的*弱相合估计*或*相合估计*。

相合性并未描述收敛速度，但如果一个估计不具备相合性，样本量再大对改善估计的精度也无济于事。相合性是对点估计的最低要求，是大数律的一个应用。

例 7.4. 相合估计不一定是唯一的：令样本 $X_1, X_2, \cdots X_n \overset{iid}{\sim} p\langle 1\rangle + (1-p)\langle 0\rangle$，则 $T_n = \frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow{P} p$ 且 $T_n = \frac{1}{n+2}(\sum_{j=1}^{n} X_j + 1) \xrightarrow{P} p$。更一般地，$T_n' = T_n + c_n \xrightarrow{P} p$，其中 $c_n \to 0$。

例 7.5. 令样本 $X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，试证明：样本方差 S^2 是 σ^2 的相合估计。

证明. 由 $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ 可知，$\mathsf{V}(S^2) = \frac{2}{n-1}\sigma^4$ 且 $\mathsf{E}(S^2) = \sigma^2$。根据 Chebyshev 不等式有，$\forall \epsilon > 0$

$$\mathsf{P}\left\{\left|S^2 - \sigma^2\right| \geq \epsilon\right\} \leq \frac{V(S^2)}{\epsilon^2} = \frac{2\sigma^4}{(n-1)\epsilon^2} \qquad \square$$

定理 7.1. 令 $\{T_n = T(X_1, X_2, \cdots, X_n)\}_{n=1}^{\infty}$ 是一个统计量的序列，满足 $\lim\limits_{n\to\infty} \mathsf{E}(T_n) = \theta$ 且 $\lim\limits_{n\to\infty} \mathsf{V}(T_n) = 0$，则 T_n 是 θ 的相合估计。

证明. 由 Chebyshev 不等式，当 $n \to \infty$ 时有

$$\mathsf{P}\{|T_n - \theta| \geq \epsilon\} \leq \frac{\mathsf{E}(T_n - \mathsf{E}T_n + \mathsf{E}T_n - \theta)^2}{\epsilon^2} = \frac{\mathsf{V}(T_n) + (\mathsf{E}T_n - \theta)^2}{\epsilon^2} \to 0 \qquad \square$$

定理 7.2. 已知总体 X 的均值 μ 和方差 σ^2 都存在，设 $\overline{X}, S^2, B_2$ 分别是来自总体 X 的简单随机样本 $X_1, X_2, \cdots X_n$ 的样本均值、样本方差和样本二阶中心矩，则 $\overline{X}$ 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 $\overline{X} \xrightarrow{P} \mu$（即 $\overline{X}$ 是 μ 的相合估计）且 $\frac{1}{n}\sum_{j=1}^{n} X_j^2 \xrightarrow{P} \mathsf{E}(X^2)$。而 $B_2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - (\overline{X})^2$，由性质 5.1 可证 $B_2 \xrightarrow{P} \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 = \sigma^2$ 且 $S^2 = \frac{n}{n-1} B_2 \xrightarrow{P} \sigma^2$。 □

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，式 (6.19) 描述了 k 阶样本矩 A_k 以渐近正态的方式收敛于总体 k 阶矩 m_k，即

$$\frac{A_k - m_k}{\sqrt{(m_{2k} - m_k^2)/n}} \xrightarrow{L} \mathrm{N}(0,1)$$

定义 7.3 (渐近正态性)**.** 未知参数 θ 的估计量 $T_n = T(X_1, X_2, \cdots, X_n)$ 具有*渐近正态性* (asymptotic normality) 当且仅当存在一个与 θ 有关的量 $\sigma_n(\theta) > 0$ 使得 $(T_n - \theta)/\sigma_n(\theta) \xrightarrow{L} \mathrm{N}(0,1)$，即 n 很大时近似地有 $T_n \sim \mathrm{N}(\theta, \sigma_n^2(\theta))$。显然 $\mathsf{V}_\theta(T_n) \approx \sigma_n^2(\theta)$ 越小越好。

定义 7.4 (BAN 估计)**.** 未知参数 θ 的具有渐近正态性的估计量 T_n^* 称为*最优渐近正态估计* (best asymptotically normal estimator) 或简称为 BAN 估计，当且仅当对其他具有渐近正态性的估计量 T_n 而言，$\lim\limits_{n\to\infty} \sigma_n^*(\theta)/\sigma_n(\theta)$ 存在且小于 1。

在 §7.1.3，我们将以是否是 BAN 估计为标准来比较在一定条件下矩估计与最大似然估计孰优孰劣。

7.1.2 无偏性和有效性

定义 7.5 (无偏性)**.** 设 θ 是总体分布中的未知参数，若统计量 T 满足 $\mathsf{E}_\theta T = \theta$，则称 T 是参数 θ 的*无偏估计* (unbiased estimator)，否则称 T 是*有偏估计* (biased estimator)。

例 7.6. 如果总体 X 的期望和方差存在，由性质 6.3 知，样本均值 $\overline{X}$ 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 $\mathsf{E}(X^k) = m_k$ 存在，则 k 阶样本矩 A_k 是 m_k 的无偏估计。

参数 θ 的无偏估计 T 并不意味着精确估计，它只保证基于不同的样本用 T 在对 θ 进行多次重复的估计时 $T - \theta$ 或正或负相互抵消，平均意义上讲偏倚 $\textsc{bias}(\theta, T) = \mathsf{E}_\theta(T) - \theta = 0$。当 T 是 θ 的无偏估计时，均方误差 (7.2) 简化为 $\textsc{mse}(\theta, T) = \mathsf{V}_\theta(T)$。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。

定义 7.6 (UMVU 估计)**.** 如果对于任意的 $\theta \in \Theta$，θ 的无偏估计 T_* 满足 $\mathsf{V}_\theta(T_*) = \mathsf{E}_\theta(T_* - \theta)^2 \leq \mathsf{V}_\theta(T) = \mathsf{E}_\theta(T - \theta)^2$，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。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 7.6 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。

问题 7.1. 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 7.3 (Cramér-Rao 不等式, 1945, 1946)**.** 令简单随机样本 $X_1, \cdots, X_n$ 来自密度函数为 $f_\theta(x)$ 的总体 X，统计量 $T = T(X_1, \cdots, X_n)$ 满足

$$\mathsf{V}_\theta(T) < \infty \text{ 且 } \frac{d}{d\theta}\mathsf{E}_\theta(T) = \int_{\mathbb{R}^n}\cdots\int \frac{\partial}{\partial\theta}\left[T(x_1, \cdots, x_n)\prod_{j=1}^{n} f_\theta(x_j)\right]dx_1\cdots dx_n$$

记 $\psi(\theta) = \mathsf{E}_\theta(T)$，则 $\mathsf{V}_\theta(T)$ 满足下面的不等式：

$$\mathsf{V}_\theta(T) \geq \frac{[\psi'(\theta)]^2}{n\mathcal{I}(\theta)} \tag{7.7}$$

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 $\mathsf{V}_\theta(T)$ 满足下面的不等式：

$$\mathsf{V}_\theta(T) \geq \frac{1}{n\mathcal{I}(\theta)} \tag{7.8}$$

证明. 令 $Z = \sum_{j=1}^{n} \partial \ln f_\theta(X_j)/\partial\theta$，由式 (2.109) 的证明可知，

$$\begin{aligned}
&\mathsf{E}_\theta(Z) = 0\text{，并且 } \mathsf{V}_\theta(Z) = n\mathsf{E}\left[\frac{\partial \ln f_\theta(X)}{\partial\theta}\right]^2 = n\mathcal{I}(\theta)\\
&\psi'(\theta) = \int_{\mathbb{R}^n}\cdots\int T(x_1, \cdots, x_n)\left[\sum_{j=1}^{n}\frac{1}{f_\theta(x_j)}\frac{\partial f_\theta(x_j)}{\partial\theta}\right]\prod_{j=1}^{n} f_\theta(x_j)dx_1\cdots dx_n\\
&\quad = \int_{\mathbb{R}^n}\cdots\int T(x_1, \cdots, x_n)\left[\sum_{j=1}^{n}\frac{\partial \ln f_\theta(x_j)}{\partial\theta}\right]\prod_{j=1}^{n} f_\theta(x_j)dx_1\cdots dx_n\\
&\quad = \mathsf{E}_\theta(TZ)\\
&\rho^2(T, Z) = \left[\frac{\mathsf{E}_\theta(TZ) - \mathsf{E}_\theta(T)\mathsf{E}_\theta(Z)}{\sqrt{\mathsf{V}_\theta(T)\mathsf{V}_\theta(Z)}}\right]^2 \leq 1
\end{aligned} \tag{7.9}$$

由式 (7.9) 这一相关系数的事实即可得证。 □

定义 7.7 (有效性)**.** 如果 θ 的无偏估计 T_* 的方差 $\mathsf{V}_\theta(T_*)$ 达到了式 (7.8)

所示的这个下界，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 7.7. 已知简单随机样本 $X_1, \cdots, X_n$ 来自总体 $X \sim \mathrm{B}(m, p)$，其中 m 已知而 p 未知，则 $\overline{X}/m$ 是 p 的有效估计量。事实上，$\mathsf{V}_p(\overline{X}/m) = \frac{pq}{mn}$，其中 $q = 1 - p$。令 $p_k = \mathsf{P}(X = k)$，则参数 p 的 Fisher 信息量为

$$\begin{aligned}\mathcal{I}(p) &= \sum_{k=0}^{m} \left[\frac{d}{dp} \ln(C_m^k p^k q^{m-k})\right]^2 p_k = \sum_{k=0}^{m} \left(\frac{k}{p} - \frac{m-k}{1-p}\right)^2 p_k \\ &= \sum_{k=0}^{m} \left(\frac{k-mp}{pq}\right)^2 p_k = \frac{mpq}{p^2q^2} = \frac{m}{pq}，\text{经验证 } \mathsf{V}_p(\overline{X}/m) = \frac{1}{n\mathcal{I}(p)}\end{aligned}$$

例 7.8. 设样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{Poisson}(\lambda)$，其中参数 λ 未知，则 $\overline{X}$ 是 λ 的有效估计量。这是因为 $\mathsf{V}_\lambda(\overline{X}) = \lambda/n$，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

$$\mathcal{I}(\lambda) = \mathsf{E}_\lambda\left[\frac{\partial \ln f_\lambda(X)}{\partial\lambda}\right]^2 = \frac{\mathsf{E}_\lambda(X-\lambda)^2}{\lambda^2} = \frac{1}{\lambda}，\text{经验证 } \mathsf{V}_\lambda(\overline{X}) = \frac{1}{n\mathcal{I}(\lambda)}$$

练习 7.1. 已知样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中 μ 已知而 σ^2 未知，则 S^2 不是 σ^2 的有效估计量。

从 Cramér-Rao 不等式的证明还能得出什么结果？假设 T 是 θ 的有效估计量，从式 (7.9) 可知 $\mathsf{P}\{Z = cT + d\} = 1$，即几乎必然有 $Z = cT + d$，其中 c, d 为常数。因为 T 是无偏的且 $\mathsf{E}_\theta(Z) = 0$，所以 $\mathsf{E}_\theta(Z) = c\mathsf{E}_\theta(T) + d = 0 \Rightarrow d = -c\theta \Rightarrow \mathsf{P}\{Z = c(T - \theta)\} = 1$。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

$$Z=\frac{\partial \ln \prod_{j=1}^{n} f_\theta(X_j)}{\partial \theta}=c(T-\theta) \Rightarrow \prod_{j=1}^{n} f_\theta(X_j)=h(X_1,\cdots,X_n)g_\theta(T)$$
$$\Rightarrow \frac{\partial \ln g_\theta(T)}{\partial \theta}=c(T-\theta)，\text{其中 } g_\theta(T)>0$$

定理 7.4 (有效性的判定)**.** 条件与定理 7.3 相同，未知参数 θ 的无偏估计 $T=T(\boldsymbol{X})$ 是有效的当且仅当 (1) T 是充分的，即简单随机样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 的联合密度函数 $\prod_{j=1}^{n} f_\theta(x_j)=h(\boldsymbol{x})g_\theta[T(\boldsymbol{x})]$，其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathsf{T}}$；(2) 函数 $g_\theta(t)$ 对于 $g_\theta(t)>0$ 几乎必然满足方程 $\partial \ln g_\theta(t)/\partial\theta=c(t-\theta)$，其中 c 与 t 无关。

证明. “$\Rightarrow$”已证，现在往证“$\Leftarrow$”，即往证 $\mathsf{V}_\theta(T)=[n\mathcal{I}(\theta)]^{-1}$：

$$\begin{aligned}
T\text{ 是 }\theta\text{ 的无偏估计} &\Rightarrow \mathsf{E}_\theta T=\int_{\mathbb{R}^n} T(\boldsymbol{x})h(\boldsymbol{x})g_\theta[T(\boldsymbol{x})]d\boldsymbol{x}=\theta\\
&\Rightarrow \int_{\mathbb{R}^n} T(\boldsymbol{x})h(\boldsymbol{x})\frac{\partial g_\theta[T(\boldsymbol{x})]}{\partial\theta}d\boldsymbol{x}=1\\
h(\boldsymbol{x})g_\theta[T(\boldsymbol{x})]\text{ 是密度函数} &\Rightarrow \int_{\mathbb{R}^n} h(\boldsymbol{x})\frac{\partial g_\theta[T(\boldsymbol{x})]}{\partial\theta}d\boldsymbol{x}=0\\
\text{综合上述两个结果} &\Rightarrow \int_{\mathbb{R}^n} [T(\boldsymbol{x})-\theta]h(\boldsymbol{x})\frac{\partial g_\theta[T(\boldsymbol{x})]}{\partial\theta}d\boldsymbol{x}=1\\
&\Rightarrow \int_{\mathbb{R}^n} [T(\boldsymbol{x})-\theta]h(\boldsymbol{x})\frac{\partial \ln g_\theta[T(\boldsymbol{x})]}{\partial\theta}g_\theta[T(\boldsymbol{x})]d\boldsymbol{x}=1\\
&\Rightarrow c\int_{\mathbb{R}^n} [T(\boldsymbol{x})-\theta]^2h(\boldsymbol{x})g_\theta[T(\boldsymbol{x})]d\boldsymbol{x}=1\\
&\Rightarrow c\mathsf{V}_\theta(T)=1
\end{aligned}$$

由于 T 是充分统计量，于是样本的 Fisher 信息量为

$$\begin{aligned}
n\mathcal{I}(\theta)=\mathcal{I}_n(\theta)&=\mathsf{E}_\theta\left\{\frac{\partial \ln f_\theta(\boldsymbol{X})}{\partial\theta}\right\}^2=\mathsf{E}_\theta\left\{\frac{\partial \ln g_\theta[T(\boldsymbol{X})]}{\partial\theta}\right\}^2\\
&=c^2\mathsf{E}_\theta(T-\theta)^2=c^2\mathsf{V}_\theta(T)
\end{aligned}$$

与上式联立可得 $\mathsf{V}_\theta(T) = 1/[n\mathcal{I}(\theta)]$，达到了 Cramér-Rao 下界。 □

✂ **例 7.9.** 设样本 $X_1, X_2 \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中 μ 已知而 σ^2 未知。试证明 σ^2 的无偏估计 $V = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2$ 还是有效的，满足 $\mathsf{V}_{\sigma^2}(V) = 2\sigma^4/n$。

证明. 由 $nV/\sigma \sim \chi_n^2$ 及式 (4.19) 可得 V 的密度函数 $g_{\sigma^2}(v)$ 如下，

$$g_{\sigma^2}(v) = \frac{n^{n/2}v^{n/2-1}\exp\left\{-\frac{nv}{2\sigma^2}\right\}}{(2\sigma^2)^{n/2}\Gamma(n/2)}$$

然后验证定理 7.4 的必要条件成立：

$$\prod_{j=1}^{n}\phi(x_j|\mu,\sigma^2) = \frac{\Gamma(n/2)}{(n\pi)^{n/2}v^{n/2-1}}g_{\sigma^2}(v) \Rightarrow V \text{ 对 } \sigma^2 \text{ 而言是充分的}$$

$$\frac{\partial \ln g_{\sigma^2}(v)}{\partial\sigma^2} = \frac{n}{2\sigma^2}(v-\sigma^2)，\text{其中 } \tfrac{n}{2\sigma^2} \text{ 与 } v \text{ 无关}$$

为求得 $\mathsf{V}_{\sigma^2}(V) = [n\mathcal{I}(\sigma^2)]^{-1}$，只需求得未知参数 σ^2 的 Fisher 信息量

$$\begin{aligned}\mathcal{I}(\sigma^2) &= \int_{-\infty}^{\infty}\left[\frac{\partial \ln\phi(x|\mu,\sigma^2)}{\partial\sigma^2}\right]^2\phi(x|\mu,\sigma^2)dx \\ &= \int_{-\infty}^{\infty}\left[\frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right]^2\phi(x|\mu,\sigma^2)dx = \frac{1}{2\sigma^4}\end{aligned}$$ □

定义 7.8 (渐近无偏性)**.** 若统计量 $T_n = T(X_1, X_2, \cdots, X_n)$ 满足 $\lim\limits_{n\to\infty}\mathsf{E}(T_n) = \theta$，则称之为 θ 的*渐近无偏估计* (asymptotically unbiased estimator)。例如，样本二阶中心矩 $B_2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2 = \frac{n-1}{n}S^2$ 是总体方差的渐近无偏估计，而非无偏估计。把统计量 B_2 稍作改造变为 S^2 便能降低偏倚。

问题 7.2. 有没有一个普适的方法能由给定的统计量 $T_n = T(X_1, X_2, \cdots, X_n)$ 构造出具有更小偏倚的统计量？答案：刀切法 (jackknife method)*

*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 [72,73]，后由美国统计学家 John Wilder Tukey (1915-2000) 于 1958 年定名 [88]。刀切法是计算标准误差和置信区间 [91] 的非参数方法，还应用于方差估计 [93] 等。

[43,83]。

约定：将 n 维向量 $\boldsymbol{X}=(X_1,\cdots,X_j,\cdots,X_n)^{\mathsf{T}}$ 中第 j 个分量“切掉”后所得的 $(n-1)$ 维向量记作 $\boldsymbol{X}_{-j}=(X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_n)^{\mathsf{T}}$，并称之为 $\boldsymbol{X}$ 的弃一 (leave-one-out) 表示。

定义 7.9 (刀切估计量)**.** 如下构造的统计量称为 θ 的*刀切估计量*：

$$T_{\text{jack}}=nT(\boldsymbol{X})-\frac{n-1}{n}\sum_{j=1}^{n}T(\boldsymbol{X}_{-j}) \tag{7.10}$$

下面说明式 (7.10) 所定义的刀切估计量的偏倚比 $T(\boldsymbol{X})$ 的有所改善。不妨设 $\mathsf{E}_\theta[T(\boldsymbol{X})]-\theta$ 可用如下 n^{-1} 的幂级数展开：

$$\mathsf{E}_\theta[T(\boldsymbol{X})]-\theta=\frac{c_1}{n}+\frac{c_2}{n^2}+\frac{c_3}{n^3}+\cdots$$

$$\text{于是，}\ \mathsf{E}_\theta[T(\boldsymbol{X}_{-j})]-\theta=\frac{c_1}{n-1}+\frac{c_2}{(n-1)^2}+\frac{c_3}{(n-1)^3}+\cdots$$

$$\text{进而，}\ \mathsf{E}_\theta(T_{\text{jack}})-\theta\sim O(1/n^2)$$

比起偏倚 $\mathsf{E}_\theta[T(\boldsymbol{X})]-\theta\sim O(1/n)$ 收敛于 0 的速度更快些，当 $n\to\infty$ 时。

例 7.10. 假设总体方差 σ^2 存在且未知。样本二阶中心矩 $T(\boldsymbol{X})=\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X})^2$ 是 σ^2 的有偏估计，按照式 (7.10) 构造刀切估计量。

$$T(\boldsymbol{X}_{-j})=\frac{1}{n-1}\sum_{\substack{k=1\\k\neq j}}^{n}\left(X_k-\frac{n\overline{X}-X_j}{n-1}\right)^2\text{，代入到式 (7.10) 中}$$

$$\begin{aligned}\text{于是，}\ T_{\text{jack}}&=\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2+\frac{n}{n-1}S^2-\frac{1}{n}\sum_{k,j=1}^{n}\left(X_k-\frac{n\overline{X}-X_j}{n-1}\right)^2\\&=(n-1)S^2-\frac{1}{n}\sum_{k,j=1}^{n}\left(X_k-\overline{X}+\overline{X}-\frac{n\overline{X}-X_j}{n-1}\right)^2\\&=\frac{n}{n-1}S^2-\frac{1}{n-1}S^2=S^2\text{，为总体方差的无偏估计。}\end{aligned}$$

7.1.3 点估计的常用方法：矩方法、最大似然法

如果总体 X 的 k 阶矩 $m_k = \mathsf{E}(X^k)$ 存在，则样本 j 阶矩 $A_j = \frac{1}{n}\sum_{i=1}^n X_i^j, j = 1, 2, \cdots, k$ 是对 m_j 的相合的、无偏的估计。

定义 7.10 (矩估计)**.** 如果总体分布中的未知参数 θ 能整理成 $\theta = h(m_1, \cdots, m_k)$，其中 h 为 Borel 函数，这样就能保证

$$h(A_1, \cdots, A_k) = h\left(\frac{1}{n}\sum_{i=1}^n X_i, \cdots, \frac{1}{n}\sum_{i=1}^n X_i^k\right) \tag{7.11}$$

为一个统计量，称为 θ 的*矩估计*，记作 $\hat{\theta}$。

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，其优点是计算简单且矩估计在一般情况下是强相合的，缺点是当参数无法通过矩用 Borel 函数表示出来的时候矩方法无法使用。

例 7.11. 如果总体的方差 $\sigma^2 = m_2 - m_1^2$ 存在，从已知样本 $X_1, X_2, \cdots, X_n$ 可得到 σ^2 的矩估计 $\hat{\sigma}^2$，它是对 σ^2 的相合的、渐近无偏的估计。

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$$

例 7.12. 已知样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{B}(m, p)$，其中参数 m, p 都未知。由 $\mathsf{E}(X) = mp, \mathsf{E}(X^2) = \mathsf{V}(X) + [\mathsf{E}(X)]^2 = mp(1-p) + m^2p^2$，解如下方程组

$$\begin{cases} A_1 = mp \\ A_2 = mp(1-p) + m^2p^2 \end{cases}$$

得到 m 和 p 的矩估计 $\hat{m} = A_1^2/(A_1 + A_1^2 - A_2)$ 和 $\hat{p} = A_1/\hat{m}$，其中 A_1, A_2 分别是样本的一阶矩和二阶矩。请问：$\hat{m} \xrightarrow{P} m$ 是否成立？

定理 7.5. 如果总体分布中的未知参数 θ 能表示成有限个总体矩 $m_1, \cdots, m_k$ 的连续函数 $\theta = h(m_1, \cdots, m_k)$，则矩估计 $h(A_1, \cdots, A_k)$ 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。

练习 7.2. 已知样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{U}[\theta_1, \theta_2]$，其中参数 θ_1, θ_2 都未知，求它们的矩估计。答案：$\hat{\theta}_1 = A_1 - \sqrt{3B_2}, \hat{\theta}_1 = A_1 + \sqrt{3B_2}$。

练习 7.3. 设总体 X 的概率密度为 $f(x) = \begin{cases} \frac{1}{\theta_2}\exp\{-(x-\theta_1)/\theta_2\} & \text{当 } x \geq \theta_1 \\ 0 & \text{其他} \end{cases}$ 其中 $\theta_2 > 0$。已知 $X_1, X_2, \cdots, X_n$ 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。答案：由 $\mathsf{E}(X) = \theta_1 + \theta_2, \mathsf{E}(X^2) = 2\theta_2^2 + 2\theta_1\theta_2 + \theta_1^2$ 得到矩估计为 $\hat{\theta}_1 = \overline{X} - \sqrt{B_2}, \hat{\theta}_2 = \sqrt{B_2}$。

例 7.13. 已知样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{Poisson}(\lambda)$，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 $\overline{X}$ 和 $\sum_{i=1}^n (X_i - \overline{X})^2/n$ 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 $\hat{\lambda} = \overline{X}$。

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 年重新提及，接着 Fisher 于 1922 年在他的一篇重要论文《论理论统计学的数学基础》†中明确提出该方法（见 [37] 的第六节《估计问题的形式解》），所以人们常把最大似然法归功于 Fisher。

定义 7.11 (似然函数)**.** 设随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 的密度函数为 $f_{\boldsymbol{\theta}}(\boldsymbol{x})$，其中 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^\mathsf{T}$，称函数 $\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ 为*似然函数* (likelihood function)，突出它是关于参数 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^\mathsf{T} \in \Theta$ 的函数。并称 $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x})$ 为*对数似然函数* (log-likelihood function)。

*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 [58]。

例 7.14. 已知简单随机样本 $X_1, X_2, \cdots, X_n$ 来自于总体 $X \sim f_{\boldsymbol{\theta}}(x)$，则对数似然函数为 $\ell(\boldsymbol{\theta}; x_1, x_2, \cdots, x_n) = \sum_{j=1}^{n} \ln f_{\boldsymbol{\theta}}(x_j)$。

定义 7.12. 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：

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x}) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}) \tag{7.12}$$

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 (7.12) 中的 Θ。若在 Θ_1 的内点集 Θ_0 上，$\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x})$ 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* $\partial \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x})/\partial\theta_j = 0$ 或者（对数）*似然方程组* $\partial \ell(\boldsymbol{\theta}; \boldsymbol{x})/\partial\theta_j = 0, j = 1, 2, \cdots, k$ 得到（在解不唯一的时候，需要判定哪个是最大值点）。值得注意的是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 (7.12) 的最优化问题可能很复杂。

例 7.15. 若统计量 $T(\boldsymbol{X})$ 对 $\boldsymbol{\theta}$ 而言是充分的，并且 $\boldsymbol{\theta}$ 的最大似然估计通过对数似然方程组解得，那么它一定是 $T(\boldsymbol{X})$ 的函数。这是因为，由 $\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x}) = h(\boldsymbol{x}) g_{\boldsymbol{\theta}}[T(\boldsymbol{x})]$ 可得 $\partial \ln g_{\boldsymbol{\theta}}[T(\boldsymbol{x})]/\partial\theta_j = 0, j = 1, 2, \cdots, k$。

例 7.16. 已知样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中 $\sigma^2 > 0$ 且 $\boldsymbol{\theta} = (\mu, \sigma^2)^{\mathsf{T}}$ 未知。由对数似然函数 $\ell(\boldsymbol{\theta}; x_1, x_2, \cdots, x_n) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2 - \frac{n}{2}\ln(2\pi)$，求解下面的似然方程组：

$$\begin{cases} \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n}(x_j - \mu) = 0 \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{j=1}^{n}(x_j - \mu)^2 = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n}(X_j - \overline{X})^2 \end{cases}$$

$\forall \mu \neq \overline{X}$ 皆有 $\sum_{j=1}^{n}(X_j - \mu)^2 > \sum_{j=1}^{n}(X_j - \overline{X})^2$（见第六章课后习题），于是

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \overline{x})^2\right\} \geq \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2\right\}$$

而上式左端在 $\sigma^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2$ 取得最大值。经上述验证，μ 和 σ^2 的最大似然估计分别是 $\hat{\mu}$ 和 $\hat{\sigma}^2$。

定理 7.6. 如果参数空间 $\Theta_0 \subseteq \mathbb{R}^k$ 为开凸集，（对数）似然函数 $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ 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 $-\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}; \boldsymbol{x})$ 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 G 中的定理 G.5 及其证明。 □

定理 7.7. 设样本 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathsf{T}}$ 的密度函数为

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\left\{\sum_{j=1}^{k} \theta_j T_j(\boldsymbol{x}) + g(\boldsymbol{\theta})\right\} \tag{7.13}$$

其中参数空间 $\Theta \subseteq \mathbb{R}^k$ 为一个开凸集。若随机向量 $\boldsymbol{T} = (T_1(\boldsymbol{X}), \cdots, T_k(\boldsymbol{X}))^{\mathsf{T}}$ 的协方差阵正定，则似然方程组的解若存在即为 $\boldsymbol{\theta}$ 的最大似然估计。

证明. 陈希孺的《高等数理统计学》 [9] 的定理 1.1 保证了在本命题的条件之下有 $\mathsf{Cov}(\boldsymbol{T}, \boldsymbol{T}) = -\nabla^2_{\boldsymbol{\theta}} g(\boldsymbol{\theta})$，于是 $-\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}; \boldsymbol{x})$ 为正定矩阵，由定理 7.6 即可得证。 □

✂ **例 7.17.** 设简单随机样本 $(X_1, Y_1)^{\mathsf{T}}, (X_2, Y_2)^{\mathsf{T}}, \cdots, (X_n, Y_n)^{\mathsf{T}}$ 来自二元正态总体 $\mathrm{N}(0, 0, \sigma^2, \sigma^2, \rho)$，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

$$\left(2\pi\sigma^2\sqrt{1-\rho^2}\right)^{-n} \exp\left\{-\frac{\sum_{j=1}^{n}(x_j^2 + y_j^2 - 2\rho x_j y_j)}{2\sigma^2(1-\rho^2)}\right\}$$

引入新参数 $\theta_1 = -[2\sigma^2(1-\rho^2)]^{-1}, \theta_2 = \rho[\sigma^2(1-\rho^2)]^{-1}$，则似然函数简化为 $\exp\{\theta_1 T_1 + \theta_2 T_2 + g(\theta_1, \theta_2)\}$，其中 $g(\theta_1, \theta_2) = \ln[(2\pi)^{-n}(4\theta_1^2 - \theta_2^2)^{-n/2}]$, $T_1 = \sum_{j=1}^{n}(x_j^2 + y_j^2), T_2 = \sum_{j=1}^{n} x_j y_j$。经验证满足定理 7.7 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。

$$\begin{cases} -\dfrac{4n\theta_1}{4\theta_1^2-\theta_2^2} = T_1 \\ \dfrac{2n\theta_2}{4\theta_1^2-\theta_2^2} = T_2 \end{cases} + \begin{cases} \rho = -\dfrac{\theta_2}{\theta_1} \\ \sigma^2 = -\dfrac{2\theta_1}{4\theta_1^2-\theta_2^2} \end{cases} \Rightarrow \begin{cases} \hat{\sigma}^2 = \dfrac{1}{2n}\left(\sum\limits_{j=1}^{n} X_j^2 + \sum\limits_{j=1}^{n} Y_j^2\right) \\ \hat{\rho} = \dfrac{2\sum_{j=1}^{n} X_j Y_j}{\sum_{j=1}^{n} X_j^2 + \sum_{j=1}^{n} Y_j^2} \end{cases}$$

例 7.18. 设样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} f_\theta(x) = \begin{cases} 1/|\theta| & \text{当 } \theta \le x \le \theta + |\theta| \\ 0 & \text{其他} \end{cases}$ 其中参数 $\theta \in \Theta = (-\infty, 0)$ 未知，试求 θ 的最大似然估计。

解. 似然函数 $\mathscr{L}(\theta; x_1, \cdots, x_n) = \begin{cases} (-\theta)^{-n} & \text{当 } \theta \le x_1, \cdots, x_n \le 0 \\ 0 & \text{其他} \end{cases}$

当 θ 取 $\min\limits_{1\le j\le n} x_j$ 时，$\mathscr{L}$ 达到最大，于是 θ 的最大似然估计为 $\hat{\theta} = \min\limits_{1\le j\le n} X_j$。

例 7.19. 设样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{U}[\theta - 1/2, \theta + 1/2]$，其中 θ 是未知参数，试求 θ 的最大似然估计。

解. 似然函数 $\mathscr{L}(\theta; x_1, \cdots, x_n) = \begin{cases} 1 & \text{当 } \theta - 1/2 \le x_1, \cdots, x_n \le \theta + 1/2 \\ 0 & \text{其他} \end{cases}$

记 $x_{(1)} = \min(x_1, x_2, \cdots, x_n), x_{(n)} = \max(x_1, x_2, \cdots, x_n)$，则 $\theta - 1/2 \le x_{(1)} \le x_{(n)} \le \theta + 1/2$ 即是 $x_{(n)} - 1/2 \le \theta \le x_{(1)} + 1/2$，则满足 $X_{(n)} - 1/2 \le T(X_1, X_2, \cdots, X_n) \le X_{(1)} + 1/2$ 的每个统计量 $T(X_1, X_2, \cdots, X_n)$ 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 7.20. 求第 265 页的练习 7.3 中未知参数 θ_1, θ_2 的最大似然估计。

解. 似然函数为 $\mathscr{L}(\theta_1, \theta_2; x_1, \cdots, x_n) = \theta_2^{-n} \exp\{-\frac{1}{\theta_2}\sum_{k=1}^{n}(x_k - \theta_1)\}$，其中 $x_k \ge \theta_1, k = 1, 2, \cdots, n$。从而得到似然方程组

$$\begin{cases} \dfrac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_1} = \dfrac{n}{\theta_2} = 0 \\ \dfrac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_2} = -\dfrac{n}{\theta_2} + \dfrac{1}{\theta_2^2}\displaystyle\sum_{k=1}^{n}(x_k - \theta_1) = 0 \end{cases}$$

由第二式可得 $\hat{\theta}_2 = \overline{X} - \hat{\theta}_1$，但无论 θ_1 取何值都不能使第一式成立。为了使 $\mathscr{L}(\theta_1, \theta_2; x_1, \cdots, x_n)$ 达到最大就要选 $\hat{\theta}_1 = \min\limits_{1\le k\le n} X_k$。

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》 [28] 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。下面不加证明地介绍有关最大似然估计相合性和渐近正态性的 Cramér 定理 7.8。假设总体 X 的密度函数（或概率函数）为 $f_\theta(x)$，先给出一些条件：

❶ $\forall\theta\in\Theta$，偏导数 $\partial \ln f_\theta(x)/\partial\theta, \partial^2 \ln f_\theta(x)/\partial\theta^2, \partial^3 \ln f_\theta(x)/\partial\theta^3$ 皆存在，且

$$\int_{-\infty}^{\infty} \frac{\partial f_\theta(x)}{\partial\theta} dx = \mathsf{E}_\theta \frac{\partial \ln f_\theta(X)}{\partial\theta} = 0$$

❷ $\forall\theta\in\Theta$ 皆有 $\displaystyle\int_{-\infty}^{\infty} \frac{\partial^2 f_\theta(x)}{\partial\theta^2} dx = 0$。

❸ $\forall\theta\in\Theta$ 皆有 $\displaystyle -\infty < \int_{-\infty}^{\infty} \frac{\partial^2 f_\theta(x)}{\partial\theta^2} f_\theta(x) dx < 0$。

❹ 存在函数 $h(x)$ 使得 $\forall\theta\in\Theta$ 皆有

$$\left|\frac{\partial^3 f_\theta(x)}{\partial\theta^3}\right| < h(x) \text{ 并且 } \int_{-\infty}^{\infty} h(x) f_\theta(x) dx < \infty$$

④ 存在二阶可导函数 $g(\theta) > 0$ 使得 $\forall\theta\in\Theta$ 皆有

$$\left|\frac{\partial^2}{\partial\theta^2}\left[g(\theta)\frac{\partial f_\theta(x)}{\partial\theta}\right]\right| < h(x) \text{ 并且 } \int_{-\infty}^{\infty} h(x) f_\theta(x) dx < \infty$$

显然，条件 ❹ 就是条件 ④ 取 $g(\theta) = 1$ 的特殊情形。

定理 7.8 (Cramér, 1946)**.** 若条件 ❶、❸、❹ 成立，则最大似然估计是相合的。如果还满足条件 ❷，则最大似然估计 $\hat{\theta}_n$ 满足渐近正态性，即

$$\sqrt{n\mathcal{I}(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{L} \mathrm{N}(0,1)，\text{其中 } \mathcal{I}(\theta) \text{ 是 } \theta \text{ 的 Fisher 信息量} \qquad (7.14)$$

1957 年，G. Kulldorf 把 Cramér 定理 7.8 做了推广，把条件 ❹ 换成条件 ④，结论依然成立。具体证明见 P. J. Bickel 和 K. A. Doksum 的《数理统计：基本思想和选题》 [18] 第五章《渐近近似》或 Cramér 的《统计学数学方法》。有关最大似然估计强相合性的工作是由 Wald 于 1949 年做出的。

满足 Cramér 定理 7.8 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 $n \to \infty$ 时，$\mathsf{V}(\hat{\theta}_n)$ 趋近 Cramér-Rao 下界，于是 $\hat{\theta}_n$ 是 BAN 估计。而一般情况下矩估计不是 BAN 估计，从这个角度比较最大似然估计略胜一筹。

7.2　区间估计

频率派的区间估计就是利用样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathsf{T}}$ 构造两个统计量 $\overline{\theta}(\boldsymbol{X})$ 和 $\underline{\theta}(\boldsymbol{X})$ 满足 $\underline{\theta}(\boldsymbol{X})\le\overline{\theta}(\boldsymbol{X})$，它们分别充当闭区间的上下限，区间 $[\underline{\theta}(\boldsymbol{X}),\overline{\theta}(\boldsymbol{X})]$ 称为未知参数 θ 的一个区间估计。由于这样区间的上下限是随机变量，它覆盖住未知参数 θ 的概率 $\mathsf{P}_\theta\{\underline{\theta}(\boldsymbol{X})\le\theta\le\overline{\theta}(\boldsymbol{X})\}$ 一般是依赖于 θ 的。

定义 7.13 (置信度与置信系数)**.** 如果 $\forall\theta\in\Theta$ 皆有

$$\mathsf{P}_\theta\{\underline{\theta}(\boldsymbol{X})\le\theta\le\overline{\theta}(\boldsymbol{X})\}\ge 1-\alpha，\text{其中常数 }\alpha\in(0,1) \tag{7.15}$$

则称区间估计 $[\underline{\theta}(\boldsymbol{X}),\overline{\theta}(\boldsymbol{X})]$ 具有置信水平或置信度 (confidence level) $1-\alpha$，或称 $[\underline{\theta}(\boldsymbol{X}),\overline{\theta}(\boldsymbol{X})]$ 是 θ 的置信度为 $1-\alpha$ 的置信区间。通常 α 是一个接近 0 的正实数，如 $\alpha=0.05,0.01$ 等。显然，对于任何 $\beta>\alpha$ 皆有 $\mathsf{P}_\theta\{\underline{\theta}(\boldsymbol{X})\le\theta\le\overline{\theta}(\boldsymbol{X})\}\ge 1-\beta$，即 $1-\beta$ 也是置信度。我们把置信度中最大者，即 $\inf\limits_{\theta\in\Theta}\mathsf{P}_\theta\{\underline{\theta}(\boldsymbol{X})\le\theta\le\overline{\theta}(\boldsymbol{X})\}$，称为置信系数 (confidence coefficient)。

利用 Markov 不等式 (2.74) 和参数 θ 的点估计 $\hat{\theta}=\hat{\theta}(\boldsymbol{X})$（不要求 $\hat{\theta}$ 是无偏的，仅要求 $\mathsf{V}_\theta(\hat{\theta})<\infty$）可以给出 θ 的区间估计：令 $Y=(\hat{\theta}-\theta)^2, k=\epsilon^2\mathsf{E}(Y)$，代入到 $\mathsf{P}(Y\le k)\ge 1-\mathsf{E}(Y)/k$ 中得到

$$\mathsf{P}_\theta\left\{(\hat{\theta}-\theta)^2\le\epsilon^2\mathsf{E}_\theta(\hat{\theta}-\theta)^2\right\}\ge 1-\frac{1}{\epsilon^2} \tag{7.16}$$

$$\hat{\theta}-\epsilon\sqrt{\mathsf{E}_\theta(\hat{\theta}-\theta)^2}\le\theta\le\hat{\theta}+\epsilon\sqrt{\mathsf{E}_\theta(\hat{\theta}-\theta)^2} \tag{7.17}$$

对式 (7.17) 中的 $\mathsf{E}_\theta(\hat{\theta}-\theta)^2$ 做一个适当的放大使其不再含有 θ，这样得到的区间就是 θ 的置信度为 $1-1/\epsilon^2$ 的置信区间。

例 7.21. 已知样本 $X_1,\cdots,X_n\overset{iid}{\sim}p\langle 1\rangle+(1-p)\langle 0\rangle$，参数 p 未知。显然，$\mathsf{E}(\overline{X})=p,\mathsf{V}(\overline{X})=p(1-p)/n\le 1/(4n)$。由式 (7.17) 得到参数 p 的置信度

为 $1-1/\epsilon^2$ 的置信区间 $\overline{X}-\frac{\epsilon}{2\sqrt{n}} \le p \le \overline{X}+\frac{\epsilon}{2\sqrt{n}}$。这个区间还能继续得到改进：利用式 (7.16)，

$$(\overline{X}-p)^2 \le \frac{\epsilon^2 p(1-p)}{n} \quad \Leftrightarrow \quad \left(1+\frac{\epsilon^2}{n}\right)p^2 - \left(2\overline{X}+\frac{\epsilon^2}{n}\right)p + \overline{X}^2 \le 0$$

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 $p_1 < p_2$，则 $\mathsf{P}(p_1 \le p \le p_2) \ge 1-1/\epsilon^2$。当 n 足够大时，$\overline{X} \xrightarrow{P} p$，利用式 (7.17) 得到参数 p 的置信度为 $1-1/\epsilon^2$ 的置信区间为

$$\overline{X} - \epsilon\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \le p \le \overline{X} + \epsilon\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

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 $[\underline{\theta}(\boldsymbol{X}), \infty)$。

定义 7.14 (置信上限与置信下限)**.** 如果 $\forall\theta \in \Theta$ 皆有

$$\mathsf{P}_\theta\{\underline{\theta}(\boldsymbol{X}) \le \theta\} \ge 1-\alpha，\text{其中常数 } \alpha \in (0,1) \tag{7.18}$$

$$\mathsf{P}_\theta\{\theta \le \overline{\theta}(\boldsymbol{X})\} \ge 1-\beta，\text{其中常数 } \beta \in (0,1) \tag{7.19}$$

则称 $\underline{\theta}(\boldsymbol{X})$ 是 θ 的置信度为 $1-\alpha$ 的**置信下限**，称 $\underline{\theta}(\boldsymbol{X})$ 是 θ 的置信度为 $1-\beta$ 的**置信上限**。。

本节内容 第一小节是 Neyman 的置信区间理论，它与第八章即将介绍的 Neyman-Pearson 假设检验理论有着密切联系。通过试验读者可以了解置信区间的概率意义。第二小节是对 Fisher 信任区间估计的简介。

学习目标 (1) 掌握利用枢轴量求解置信区间的基本方法，并理解它的内在含义；(2) 了解大样本时的近似置信区间的估计；(3) 粗略了解信任推断和信任区间估计。

7.2.1　Neyman 的置信区间

波兰统计学家 Jerzy Neyman (1894-1981) 于 1934-1937 年间提出了置信区间的理论 [63,64]，其基本思想是用样本构造一个（随机）闭区间的上下限，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 $0<\alpha<1$。如果对区间长度不作要求，置信区间一般不唯一。

定义 7.15 (枢轴量)**.** 如果未知参数 θ 和简单随机样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^\mathsf{T}$ 的函数 $h(\boldsymbol{X},\theta)$ 的分布与 θ 无关，则称 $h(\boldsymbol{X},\theta)$ 为枢轴量 (pivot)。

有了枢轴量就可以构造置信区间（或置信限）：首先找到实数 $c_1<c_2$ 使得 $\mathsf{P}\{c_1\le h(\boldsymbol{X},\theta)\le c_2\}\ge 1-\alpha$，然后解不等式 $c_1\le h(\boldsymbol{X},\theta)\le c_2$ 得到 $\underline{\theta}(\boldsymbol{X})\le\theta\le\overline{\theta}(\boldsymbol{X})$ 即是 θ 的置信度为 $1-\alpha$ 的置信区间。

例 7.22. 已知样本 $X_1,X_2,\cdots,X_n\overset{iid}{\sim}\mathrm{N}(\mu,\sigma^2)$，下面分别在不同的情况之下，对参数 μ 和 σ^2 进行区间估计。

- 参数 μ 未知，但参数 σ^2 已知。因为枢轴量 $\sqrt{n}(\overline{X}-\mu)/\sigma\sim\mathrm{N}(0,1)$，于是得到 μ 的置信度为 $1-\alpha$ 的置信区间

$$\overline{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\le\mu\le\overline{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\tag{7.20}$$

  其中 $z_{1-\alpha/2}$ 是标准正态分布 $Z\sim\mathrm{N}(0,1)$ 的 $(1-\alpha/2)$-分位数（见第 107 页的定义 2.13），满足 $\mathsf{P}(|Z|\le z_{1-\alpha/2})=1-\alpha$。显然，$\mathsf{P}(|Z|>z_{1-\alpha/2})=\alpha$。

- 参数 μ,σ^2 都未知。因为枢轴量 $\sqrt{n}(\overline{X}-\mu)/S\sim t(n-1)$，于是得到 μ 的置信度为 $1-\alpha$ 的置信区间

$$\overline{X}-t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\le\mu\le\overline{X}+t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\tag{7.21}$$

其中 $t_{n-1,1-\alpha/2}$ 是 $T \sim t(n-1)$ 分布的 $(1-\alpha/2)$-分位数，满足 $\mathsf{P}(T \le t_{n-1,1-\alpha/2}) = 1-\alpha/2$。显然，$\mathsf{P}(|T| > t_{n-1,1-\alpha/2}) = \alpha$。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)。

- 参数 μ 已知，但参数 σ^2 未知。因为枢轴量 $\sum_{j=1}^n (X_j-\mu)^2/\sigma^2 \sim \chi_n^2$，于是得到 σ^2 的置信度为 $1-\alpha$ 的置信区间

$$\frac{\sum_{j=1}^n (X_j-\mu)^2}{\chi^2_{n,1-\alpha/2}} \le \sigma^2 \le \frac{\sum_{j=1}^n (X_j-\mu)^2}{\chi^2_{n,\alpha/2}} \tag{7.22}$$

其中 $\chi^2_{n,\alpha/2}$ 是 χ_n^2 分布的 $(\alpha/2)$-分位数，满足 $\mathsf{P}(\chi_n^2 \le \chi^2_{n,\alpha/2}) = \alpha/2$。

- 参数 μ, σ^2 都未知。因为枢轴量 $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$，于是得到 σ^2 的置信度为 $1-\alpha$ 的置信区间（μ 是冗余参数）

$$\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \tag{7.23}$$

本书所采用的分位数缺省地是下侧 α-分位数 (lower α-th quantile)，即 $q_\alpha \in \mathbb{R}$ 使得 $\mathsf{P}(X \le q_\alpha) = F_X(q_\alpha) = \alpha$。有的教科书采用上侧 α-分位数 (upper α-th quantile)，即 q'_α 使得 $\mathsf{P}(X > q'_\alpha) = \alpha$。读者在阅读文献的时候，注意上下文中对分位数的约定。如果密度函数关于 $X=0$ 对称，则有 $-q_\alpha = q_{1-\alpha} = q'_\alpha = -q'_{1-\alpha}$。

例 7.23. 设食品厂生产的某袋装食品的重量服从正态分布 $\mathrm{N}(\mu,\sigma^2)$，参数 μ, σ^2 都未知。现随机抽取 20 袋食品测得重量（单位：千克）分别为 0.11262844, 0.08596988, 0.09544452, 0.08610892, 0.09072335, 0.09706382, 0.10381781, 0.10115408, 0.10432509, 0.10224744, 0.09520061, 0.10722380, 0.09094787, 0.11757568, 0.10688860, 0.10324404, 0.10575931, 0.10432506, 0.11252202, 0.09518698，分别求 μ, σ^2 的置信度为 95% 的置信区间。

解. 分别利用用例 7.22 中第二和第四种情况求参数的置信区间。

```
> alpha <- 1-0.95
> n <- length(x)                              # 样本量
> mean(x)-sd(x)/sqrt(n)*qt(1-alpha/2,df=n-1) # 用函数 qt 求 t(n-1) 分布的分位数
[1] 0.09683864
> mean(x)+sd(x)/sqrt(n)*qt(1-alpha/2,df=n-1)
[1] 0.1049971
> (n-1)*var(x)/qchisq(1-alpha/2,df=n-1)       # 求 chi^2_{n-1} 分布的分位数
[1] 4.393633e-05
> (n-1)*var(x)/qchisq(alpha/2,df=n-1)
[1] 0.0001620623
```

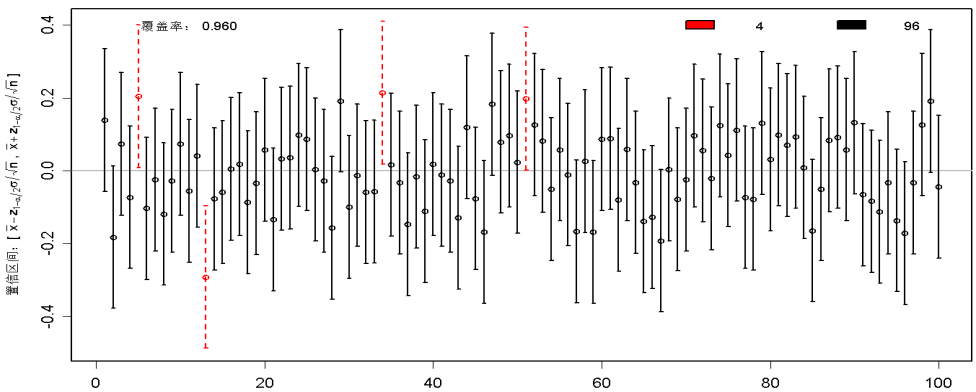


图 7.1: 置信度为 $1-\alpha$ 的置信区间的频率解释：例 7.22 中假设总体为 $N(0,1)$，在 100 次独立重复的随机试验中，第一种情况下 μ 的置信度为 95% 的置信区间 $[\underline{\mu}(\boldsymbol{x}),\overline{\mu}(\boldsymbol{x})]$ 有 96 次覆盖住 $\mu=0$（实线），4 次未覆盖住 $\mu=0$（虚线）。

注记 7.2. 参数 θ 的置信度为 $1-\alpha$ 的置信区间 $[\underline{\mu}(\boldsymbol{x}),\overline{\mu}(\boldsymbol{x})]$ 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 $[\underline{\mu}(\boldsymbol{x}),\overline{\mu}(\boldsymbol{x})]$ 要么覆盖住 θ，要么未覆盖住，并无随机性可言。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 $[\underline{\mu}(\boldsymbol{X}),\overline{\mu}(\boldsymbol{X})]$ 覆盖住未知参数 θ 的概率，而与具体观察到的样本值 $\boldsymbol{x}$ 并无多大关系。频率派拿尚未观察到的数据为当前的估计结果 $[\underline{\mu}(\boldsymbol{x}),\overline{\mu}(\boldsymbol{x})]$ “撑腰助威”的这一作法常被贝叶斯学派诟病。

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 $[\underline{\theta}, \overline{\theta}]$，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，如例 7.22 中第三、四种情况，并不奢求最短的区间。

例 7.24. 考虑例 7.22 中的第一种情况，下面验证它就是最短的置信区间。令 $a < b$ 满足

$$G(a) = \mathsf{P}\left\{a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le b\right\} = \int_a^b \phi(t)dt = 1 - \alpha$$

其中 b 是 a 的函数，必有 $dG/da = 0$。为了使得区间 $[\overline{X} - b\sigma/\sqrt{n}, \overline{X} - a\sigma/\sqrt{n}]$ 的长度 $L(a) = (b-a)\sigma/\sqrt{n}$ 取得最小，令 $dL/da = 0$ 得到

$$\left.\begin{aligned} \frac{dG}{da} &= \phi(b)\frac{db}{da} - \phi(a) = 0 \\ \frac{dL}{da} &= \frac{\sigma}{\sqrt{n}}\left(\frac{db}{da} - 1\right) = 0 \end{aligned}\right\} \Rightarrow \phi(a) = \phi(b) \Rightarrow a = b \text{ 或 } a = -b$$

$a = b$ 之解无意义，所以必有 $a = -b$，进而 $b = z_{1-\alpha/2}, a = -z_{1-\alpha/2}$。此外，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 $n \ge 4z_{1-\alpha/2}^2\sigma^2/d^2$。仿照此方法，请读者自行验证例 7.22 中第二种情况所给出的也是最短的置信区间。

以上的例子都是小样本下的置信区间估计。在大样本的情况之下，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 7.8 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。

例 7.25. 设简单随机样本 $X_1, X_2, \cdots, X_n$ 来自总体 X，已知总体具有有限期望和方差 $\mathsf{E}(X) = \mu, \mathsf{V}(X) = \sigma^2 > 0$，它们都是未知的。由 Lindeberg-Lévy 中心极限定理 5.14 知，$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{L} \mathrm{N}(0,1)$，进而 $\frac{\overline{X}-\mu}{S/\sqrt{n}} \xrightarrow{L} \mathrm{N}(0,1)$。所以近似地有 μ 的置信度为 $1-\alpha$ 的置信区间 $\overline{X} \pm z_{1-\alpha/2}S/\sqrt{n}$。

例 7.26 (Fisher-Behrens 问题)**.** 设样本 $X_1, \cdots, X_m \overset{iid}{\sim} \mathrm{N}(\mu_X, \sigma_X^2)$ 和样本 $Y_1, \cdots, Y_n \overset{iid}{\sim} \mathrm{N}(\mu_Y, \sigma_Y^2)$ 来自两个独立总体，若 $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ 都未知，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。在大样本情况下，令 m, n 充分大时利用正态逼近

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/m + S_Y^2/n}} \sim \mathrm{N}(0,1)$$

得到 $\mu_X - \mu_Y$ 的置信度为 $1-\alpha$ 的置信区间 $\overline{X} - \overline{Y} \pm z_{1-\alpha/2}\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}$。

例 7.27. 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 7.8 的条件，则近似地有 θ 的置信度为 $1-\alpha$ 的置信区间 $\hat{\theta}_n \pm z_{1-\alpha/2}/\sqrt{n\mathcal{I}(\theta)}$。

7.2.2* Fisher 的信任估计

1930 年，Fisher 提出从观察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 [65]。有别于贝叶斯学派通过参数的先验分布和观察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 也意识到它的局限性，该方法终因缺少系统理论的支持而未被广泛接受。

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*。譬如，考虑样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中 μ 未知而 σ^2 已知。从事实 $Y = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ 不难得到 $\mu = -\frac{\sigma}{\sqrt{n}}Y + \overline{X} \sim \mathrm{N}(\overline{X}, \sigma^2/n)$，即参数 μ 的信任分布。进而得到信任区间 $\overline{X} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$，与置信区间估计取得了相同的结果，见第 273 页的例 7.22 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果也是相同的。

例 7.28. 考虑例 7.26 的信任区间估计。显然，$\xi = \frac{\overline{X}-\mu_X}{S_X/\sqrt{m}} \sim t(m-1)$ 与 $\eta = \frac{\overline{Y}-\mu_Y}{S_Y/\sqrt{n}} \sim t(n-1)$ 独立，从 $\mu_X - \mu_Y - (\overline{X} - \overline{Y}) = \frac{S_X}{\sqrt{m}}\xi - \frac{S_Y}{\sqrt{n}}\eta$ 的信任分布可求得 δ 使得 $\mathsf{P}\left\{\left|\frac{S_X}{\sqrt{m}}\xi - \frac{S_Y}{\sqrt{n}}\eta\right| \le \delta\right\} = 1 - \alpha$。于是便得到信任系数为 $1 - \alpha$ 的信任区间 $[\overline{X} - \overline{Y} - \delta, \overline{X} - \overline{Y} + \delta]$。

注记 7.3. 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。信任推断法在 Fisher 的诸多成就中是颇受争议的，许多追随者企图发展 Fisher 的信任推断法也未能取得实质成效，所以应用并不广泛。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。

*Fisher 对使用 Bayes 公式持非常谨慎的态度，他是强烈反对贝叶斯学派的。

7.3　习题

7.1. 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的简单随机样本，设 $\mathsf{E}(X) = \mu$，$\mathsf{V}(X) = \sigma^2$。(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 $(\overline{X})^2 - cS^2$ 为 μ^2 的无偏估计。

7.2. 设 $\hat{\theta}$ 是参数 θ 的无偏估计并且 $\mathsf{V}(\hat{\theta}) > 0$，试证明：$\hat{\theta}^2$ 不是 θ^2 的无偏估计。

7.3. 设 X_1, X_2 是总体 X 的一个简单随机样本，已知 $\mathsf{E}(X) = \mu, \mathsf{V}(X) = \sigma^2$。试问：$\hat{\mu}_1 = \frac{1}{2}(X_1 + X_2)$ 和 $\hat{\mu}_2 = a_1X_1 + a_2X_2$（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？

☆ 7.4. 总体 X 的密度函数为 $f_\theta(x) = \begin{cases} e^{-(x-\theta)} & \text{当 } x \geq \theta \\ 0 & \text{当 } x < \theta \end{cases}$ 其中 $\theta \in \mathbb{R}$ 是未知参数。设 $X_1, \cdots, X_n$ 是来自 X 的一个简单随机样本，试证明：$\hat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n} X_j - 1$ 和 $\hat{\theta}_2 = \min(X_1, \cdots, X_n) - \frac{1}{n}$ 都是 θ 的无偏估计且 $\mathsf{V}(\hat{\theta}_2) \leq \mathsf{V}(\hat{\theta}_1)$。

7.5. 设样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathsf{N}(\mu, \sigma^2)$，其中 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。

7.6. 设简单随机样本 $X_1, X_2, \cdots, X_n$ 来自的总体 X 具有密度函数 $f_\theta(x) = \begin{cases} (\theta+1)x^\theta & \text{当 } 0 < x < 1 \\ 0 & \text{其他} \end{cases}$（其中参数 $\theta > 1$ 未知），试求：参数 θ 的矩估计和最大似然估计。

7.7. 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。

7.8. 设简单随机样本 $X_1, X_2, \cdots, X_n$ 来自总体 X，总体的概率分布为 $\theta^2\langle 1\rangle + 2\theta(1-\theta)\langle 2\rangle + (1-\theta)^2\langle 3\rangle$，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。

7.9. 设简单随机样本 $X_1, \cdots, X_n$ 是来自对数正态分布总体 X，其中 $\ln X \sim \mathrm{N}(\mu, \sigma^2)$，其中参数 $\mu \in \mathbb{R}, \sigma > 0$ 未知，试求参数 $\theta_1 = \mathsf{E}(X)$ 和 $\theta_2 = \mathsf{V}(X)$ 的最大似然估计。

☆ 7.10. 例 7.18 中参数空间为 $\Theta = (0, +\infty)$ 时，求 θ 的最大似然估计。

7.11. 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 $\overline{X}$ 和 $\overline{Y}$，确定常数 p 使 $T = p\overline{X} + (1-p)\overline{Y}$ 的方差达到最小。

☆ 7.12. 有 n 台测量光速的仪器，其中第 j 台的测量值 $X \sim \mathrm{N}(\theta, \sigma_j^2), j = 1, 2, \cdots, n$。用这些仪器独立地对光速 θ 各测量一次，得到样本 $X_1, X_2, \cdots, X_n$。问 $k_1, k_2, \cdots, k_n$ 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 $\mathsf{V}(\hat{\theta})$ 最小？

7.13. 设样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中参数 μ, σ 未知，求 $\ln \sigma^2$ 的置信度为 $1-\alpha$ 的置信区间。

☆ 7.14. 设样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{U}[0, \theta]$，证明：对于任意给定的 $1-\alpha$，其中 $0 < \alpha < 1$，存在常数 c_n 使 $[\max(X_1, \cdots, X_n), c_n \max(X_1, \cdots, X_n)]$ 为 θ 的一个置信度为 $1-\alpha$ 的置信区间。

7.15. 例 7.26 条件之下，给出 σ_Y^2/σ_X^2 的置信度为 $1-\alpha$ 的置信区间。

7.16. 设样本 $X_1, X_2, \cdots, X_m \overset{iid}{\sim} \mathrm{Expon}(\lambda_1)$，样本 $Y_1, Y_2, \cdots, Y_n \overset{iid}{\sim} \mathrm{Expon}(\lambda_2)$，试求 λ_2/λ_1 的置信度为 $1-\alpha$ 的置信区间。

第八章

假设检验

关于总体分布的假设称为统计假设。假设检验 (hypothesis testing) 是统计推断一个重要的组成部分，其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)。

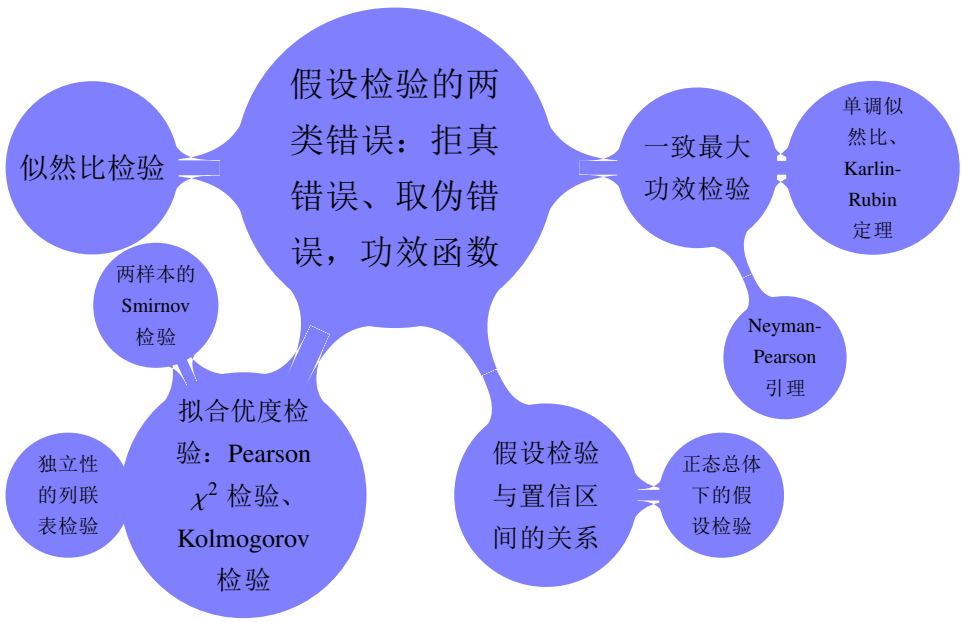
例 8.1. 工厂生产一批零件，其长度（单位：毫米）服从分布 $\mathrm{N}(\mu, 10^{-2})$，其中参数 μ 未知，$\mu_0 = 100$ 为合格零件的长度。随机抽取 15 个零件测得其长度分别为：100.095, 100.101, 100.248, 100.156, 99.946, 100.243, 100.041, 100.145, 100.054, 100.113, 100.055, 100.080, 99.895, 100.135, 100.056。零假设是这批零件的长度合格，即 $H_0 : \mu = \mu_0$。

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”意味着观察数据（即样本值）不支持零假设，“不拒绝 H_0”意味着观察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。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。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。本章后续正文将给出该例假设检验的细节。

从是否已知总体分布类型的角度，我们可以把统计假设分为参数假设和非参数假设两类：像例 8.1，总体分布类型已知且仅涉及未知参数的统计假设被称为参数假设；而总体分布类型未知时，我们把仅涉及总体分布类型的统计假设称为非参数假设，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，统计假设又可分为简单假设和复合假设两类：若一个统计假设能确定总体的分布，则称之为简单假设 (simple hypothesis)；否则就是复合假设 (composite hypothesis)。例 8.1 中的零假设 $H_0 : \mu = \mu_0$ 就是一个简单假设。再如这样的非参数假设，$H_0 : X \sim \mathrm{N}(100, 10^{-2})$，也是简单假设。对例 8.1 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。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。再如，考察两个样本是否来自同一个总体的非参数假设 $H_0 : F(x) = G(x)$ 也是复合假设，其中 $F(x), G(x)$ 分别代表两个样本的总体分布。



8.1　Neyman-Pearson 假设检验理论

已知简单随机样本 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 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。

以 $H_0 : \theta \leq \theta_0$ 为例，如果样本均值 $\overline{X}$ 是 θ 的点估计，则 $\overline{X}$ 越小 H_0 成立的可能就越大。不妨设定一个临界值 c，把样本空间划分为 $R = \{\boldsymbol{x} : \overline{x} > c\}$ 和 $R^c = \{\boldsymbol{x} : \overline{x} \leq c\}$ 两部分：当样本 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 落于 R 中时，就拒绝零假设 H_0 并接受备择假设 H_1；否则就接受 H_0 并拒绝 H_1。假设检验的目标就是寻找样本空间中这样的区域 R，称之为该检验的**拒绝域** (rejection region)*。我们把 R 的指示函数 $\delta(\boldsymbol{x}) = I_R(\boldsymbol{x})$ 称作**检验函数** (test function)，在下文中，选择拒绝域 R 与构造检验函数 $\delta(\boldsymbol{x})$ 被视为一回事。给定拒绝域后，凭借样本 $\boldsymbol{X}$ 是否落于其中来拒绝或接受 H_0 便带来了随机性，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**）

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，拒真错误比取伪错误更难以接受。再如，“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

*拒绝域 R 的补集 $A = R^c$ 称为**接受域**，当样本 $\boldsymbol{X}$ 落于 A 中时，就接受零假设。

将导致生产成本的增加。一般地，人们把力图否定的命题定为零假设 H_0，假设检验就像是在用数据“抬杠”，总想对零假设说“不”。

第一类错误的概率（称为*拒真概率*）是 $\alpha = \mathsf{P}\{\text{拒绝 } H_0 | \theta \in \Theta_0\}$，第二类错误的概率（称为*取伪概率*）是 $1 - \mathsf{P}\{\text{拒绝 } H_0 | \theta \in \Theta_1\}$。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，但当样本量一定时，同时无限制地减少二者是不可能的。通常情况下，人们把关注的统计假设当作零假设 H_0，拒真错误就显得更重要。Neyman 和 Pearson 提出一个挑选检验的原则：*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** $\beta = \mathsf{P}\{\text{拒绝 } H_0 | \theta \in \Theta_1\}$ *尽可能地大*。

行为 \ 真实情况	H_0 为真	H_0 为假
拒绝 H_0	α	β
接受 H_0	$1-\alpha$	$1-\beta$

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。

学习目标 (1) 理解两类错误、功效函数、UMP 检验、单调似然比等基本概念；(2) 会利用 Neyman-Pearson 引理和 Karlin-Rubin 定理构造 UMP 检验；(3) 掌握假设检验与置信区间估计的关系，熟悉正态总体下对参数的假设检验。

*概率 $\beta = \mathsf{P}\{\text{拒绝 } H_0 | \theta \in \Theta_1\}$ 称为检验对备择假设的*功效*或*势* (power)。在拒真概率不超过给定的一个很小的正实数 α 的时候，β 当然越大越好。

8.1.1　功效函数与一致最大功效检验

定义 8.1. 设定一个临界概率 α，譬如 $\alpha = 0.05$ 或 0.01，在零假设 H_0 成立（也记作 $\theta \in \Theta_0$）的前提下，如果观察到样本 $\boldsymbol{X}$ 的概率不超过 α，即如果 $\mathsf{P}\{\boldsymbol{X}|\theta \in \Theta_0\} \le \alpha$，则拒绝 H_0。我们称这个临界概率 α 为*显著水平* (significance level) 或*水平*，并称在水平 α 拒绝 H_0。特别地，如果拒绝域 R 由 $T(\boldsymbol{x}) \ge c$ 给出，其中 $T(\boldsymbol{X})$ 是一个统计量，c 为一待定常数，拒真概率的上确界记为 $\alpha(c)$，即

$$\alpha(c) = \sup_{\theta\in\Theta_0} \mathsf{P}_\theta\{T(\boldsymbol{X}) \ge c\} = \sup_{\theta\in\Theta_0} \mathsf{P}_\theta\{\text{拒绝 } H_0\} \tag{8.1}$$

我们称 T 为*检验统计量* (test statistic)，称 c 为*临界值* (critical value)。

定义 8.2. 概率 $\mathsf{P}_\theta\{\text{拒绝 } H_0\}$ 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 $\beta_\delta(\theta)$。功效函数显然满足：

$$\begin{aligned}\beta_\delta(\theta) &= \mathsf{E}_\theta\delta(\boldsymbol{X}) = \mathsf{P}_\theta\{\delta(\boldsymbol{X}) = 1\} = \mathsf{P}_\theta\{\text{拒绝 } H_0\} \\ &= \begin{cases} \text{拒真概率}, & \text{如果 } \theta \in \Theta_0 \\ 1 - \text{取伪概率}, & \text{如果 } \theta \in \Theta_1 \end{cases}\end{aligned} \tag{8.2}$$

拒真概率和取伪概率可由功效函数统一表示，形式上很方便。

假设检验的一般过程是：(1) 首先把整个参数空间 Θ 划分为 Θ_0 和 Θ_1，列出零假设和备择假设 $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$。给出显著水平 $0 < \alpha < 1$，譬如 $\alpha = 0.01$ 或 0.05 等。(2) 定义拒绝域为 $R = \{\boldsymbol{x} \in \mathbb{R}^n : T(\boldsymbol{x}) \ge c\}$，其中 T 为某一统计量，临界值 c 待定。(3) 由 $T(\boldsymbol{X}) \sim G_\theta(t)$，其中 $G_\theta(t)$ 为某一已知分布函数，得到拒绝 H_0 的概率 $\mathsf{P}_\theta\{T(\boldsymbol{X}) \ge c\}$ 的表达式。为使得拒真错误不超过 α，由 $\alpha(c) = \sup_{\theta\in\Theta_0} \mathsf{P}_\theta\{T(\boldsymbol{X}) \ge c\} = \alpha$ 解出临界值 c。当 $\boldsymbol{X} \in R$ 时，拒绝零假设 H_0；否则，接受 H_0。

例 8.2. 已知样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中参数 σ^2 已知，μ 未知。参数空间为 $\Theta = \{\mu_0, \mu_1\}$，其中 $\mu_0 < \mu_1$。为检验 $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu =$

μ_1，定义功效函数如下并得到拒真概率的上确界 $\alpha(c)$，

$$\beta_\delta(\mu) = \mathsf{P}_\mu\{\overline{X} \geq c\} = 1 - \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$$
$$\alpha(c) = \sup_{\mu=\mu_0} \beta_\delta(\mu) = \beta_\delta(\mu_0) = 1 - \Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right)$$

由 $\alpha(c) = \alpha$ 可得 $c = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$，进而

$$\text{取伪概率} = 1 - \beta_\delta(\mu_1) = \Phi\left[z_{1-\alpha} - \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}\right]$$

明显地，拒真概率 $\leq \alpha$。当 $\alpha \to 0$ 时，取伪概率 $\to 1$。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。

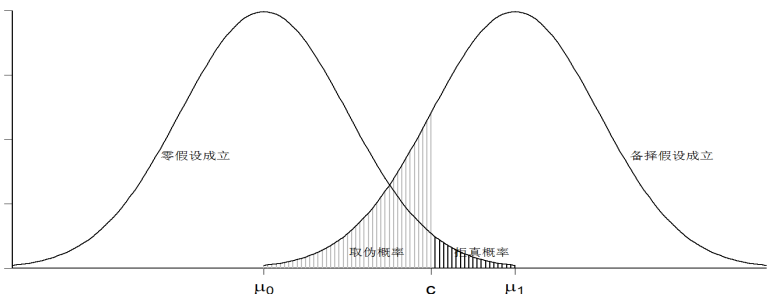


图 8.1: 拒真概率 $\mathsf{P}\{\overline{X} > c|H_0 \text{ 成立}\}$（取伪概率 $\mathsf{P}\{\overline{X} < c|H_1 \text{ 成立}\}$）为深色（浅色）阴影部分的面积：当拒真概率趋向于 0 时，取伪概率趋向于 1。

例 8.3. 条件与例 8.2 相同，为检验 $H_0 : \mu \leq \mu_0 \leftrightarrow H_1 : \mu > \mu_0$，定义功效函数如下并得到拒真概率的上确界 $\alpha(c)$，

$$\beta_\delta(\mu) = \mathsf{P}_\mu\{\overline{X} \geq c\} = 1 - \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$$
$$\alpha(c) = \sup_{\mu\leq\mu_0} \beta_\delta(\mu) = \beta_\delta(\mu_0) = 1 - \Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right)$$

由 $\alpha(c)=\alpha$ 可得 $c=\mu_0+z_{1-\alpha}\sigma/\sqrt{n}$。于是，

$$\beta_\delta(\mu)=\Phi\left[\frac{\sqrt{n}(\mu-\mu_0)}{\sigma}-z_{1-\alpha}\right]$$

显然有 $\lim\limits_{\mu\to\mu_0}\beta_\delta(\mu)=\alpha$ 且 $\lim\limits_{\mu\to\infty}\beta_\delta(\mu)=1$，它的含义也很明显：$\mu$ 较之 μ_0 越大，拒绝 H_0 的概率就越大。

定理 8.1. 已知充分统计量 $T(\boldsymbol{X})$，对于任意检验 $\delta(\boldsymbol{x})$，存在只依赖于 $T(\boldsymbol{X})$ 且与 δ 有相同功效函数的检验。

证明. 令 $\psi(t)=\mathsf{E}_\theta[\delta(\boldsymbol{X})|T(\boldsymbol{X})=t]=\mathsf{E}[\delta(\boldsymbol{X})|T(\boldsymbol{X})=t]$，则 $\beta_\psi(\theta)=\mathsf{E}_\theta\psi(T)=\mathsf{E}_\theta[\mathsf{E}\delta(\boldsymbol{X})|T]=\mathsf{E}_\theta\delta(\boldsymbol{X})$。 □

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 8.3. 基于观察到的样本值 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^\mathsf{T}$，利用检验统计量 T 定义 *p*-值为

$$p\text{-值}=\alpha[T(\boldsymbol{x})]=\sup_{\theta\in\Theta_0}\mathsf{P}_\theta\{T(\boldsymbol{X})\geq T(\boldsymbol{x})\}\tag{8.3}$$

在上面的两个例子中，*p*-值都为 $1-\Phi[\sqrt{n}(\bar{x}-\mu_0)/\sigma]$，其中 $\bar{x}=\frac{1}{n}\sum_{j=1}^n x_j$。

定义 8.4. 在拒真错误被控制在不超过 α 的前提之下，人们希望得到对备择假设 $\theta\in\Theta_1$ 而言*最大功效* (most powerful, MP) 检验 δ^*，即

$$\beta_{\delta^*}(\theta)\geq\beta_\delta(\theta)，\text{其中 }\delta\text{ 是任一检验函数}\tag{8.4}$$

更有甚者，如果 $\forall\theta\in\Theta_1$，某检验函数 δ^* 满足条件 (8.4)，则称之为*一致最大功效* (uniformly most powerful, UMP) 检验。

8.1.2 Neyman-Pearson 引理和单调似然比

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 [66]。

引理 8.1 (Neyman-Pearson, 1933)**.** 已知样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} f_\theta(x)$，以及简单假设 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$ 的检验函数 δ_k，满足

$$\mathsf{E}_{\theta_0}\left[\delta_k(\boldsymbol{X})\right] \leq \alpha，\text{其中 } k \geq 0 \text{ 且} \tag{8.5}$$

$$\delta_k(\boldsymbol{x}) = \begin{cases} 1, & \text{若 } \prod_{j=1}^{n} f_{\theta_1}(x_j) \geq k \prod_{j=1}^{n} f_{\theta_0}(x_j) \\ 0, & \text{若 } \prod_{j=1}^{n} f_{\theta_1}(x_j) < k \prod_{j=1}^{n} f_{\theta_0}(x_j) \end{cases} \tag{8.6}$$

令 δ 是任一满足条件 $\mathsf{E}_{\theta_0}\left[\delta(\boldsymbol{X})\right] = \beta_\delta(\theta_0) \leq \beta_{\delta_k}(\theta_0) = \mathsf{E}_{\theta_0}\left[\delta_k(\boldsymbol{X})\right]$ 的检验函数，则有 $\mathsf{E}_{\theta_1}\left[\delta(\boldsymbol{X})\right] = \beta_\delta(\theta_1) \leq \beta_{\delta_k}(\theta_1) = \mathsf{E}_{\theta_1}\left[\delta_k(\boldsymbol{X})\right]$。

证明. 令 $\mathscr{L}(\theta_0; \boldsymbol{x}) = \prod_{j=1}^{n} f_{\theta_0}(x_j)$ 和 $\mathscr{L}(\theta_1; \boldsymbol{x}) = \prod_{j=1}^{n} f_{\theta_1}(x_j)$ 分别是 H_0, H_1 成立时的似然函数。构造函数 $g(\boldsymbol{x}) = \left[\delta_k(\boldsymbol{x}) - \delta(\boldsymbol{x})\right]\left[\mathscr{L}(\theta_1; \boldsymbol{x}) - k\mathscr{L}(\theta_0; \boldsymbol{x})\right]$，记 $D_1 = \{\boldsymbol{x} : \mathscr{L}(\theta_1; \boldsymbol{x}) \geq k\mathscr{L}(\theta_0; \boldsymbol{x})\}, D_2 = \{\boldsymbol{x} : \mathscr{L}(\theta_1; \boldsymbol{x}) < k\mathscr{L}(\theta_0; \boldsymbol{x})\}$，于是

$$\begin{aligned} \int_{\mathbb{R}^n} g(\boldsymbol{x})d\boldsymbol{x} =& \mathsf{E}_{\theta_1}\left[\delta_k(\boldsymbol{X})\right] - \mathsf{E}_{\theta_1}\left[\delta(\boldsymbol{X})\right] - k\left\{\mathsf{E}_{\theta_0}\left[\delta_k(\boldsymbol{X})\right] - \mathsf{E}_{\theta_0}\left[\delta(\boldsymbol{X})\right]\right\} \\ =& \left\{\int_{D_1} + \int_{D_2}\right\} g(\boldsymbol{x})d\boldsymbol{x} \\ =& \int_{D_1} \left[1 - \delta(\boldsymbol{x})\right]\left[\mathscr{L}(\theta_1; \boldsymbol{x}) - k\mathscr{L}(\theta_0; \boldsymbol{x})\right] d\boldsymbol{x} + \\ & \int_{D_2} \left[-\delta(\boldsymbol{x})\right]\left[\mathscr{L}(\theta_1; \boldsymbol{x}) - k\mathscr{L}(\theta_0; \boldsymbol{x})\right] d\boldsymbol{x} \geq 0 \end{aligned}$$

由 $\mathsf{E}_{\theta_0}\left[\delta(\boldsymbol{X})\right] \leq \mathsf{E}_{\theta_0}\left[\delta_k(\boldsymbol{X})\right]$ 易得 $\mathsf{E}_{\theta_1}\left[\delta(\boldsymbol{X})\right] \leq \mathsf{E}_{\theta_1}\left[\delta_k(\boldsymbol{X})\right]$。 □

例 8.4. 考虑 $H_0 : X \sim \mathrm{N}(0,1) \leftrightarrow H_1 : X \sim \mathrm{Cauchy}(1,0)$，似然比为

$$\frac{f_1(x)}{f_0(x)} = \frac{1/(\pi + \pi x^2)}{1/\sqrt{2\pi}\exp\{-x^2/2\}} = \sqrt{\frac{2}{\pi}}\frac{\exp\{x^2/2\}}{1+x^2}$$

Neyman-Pearson 最大功效检验具有形式 $\delta(x) = \begin{cases} 1 & \text{如果 } \sqrt{\frac{2}{\pi}}\frac{\exp\{x^2/2\}}{1+x^2} \geq k \\ 0 & \text{其他} \end{cases}$

其中 k 由 $\mathsf{E}_0\delta(X) = \alpha$ 唯一确定，其中 $\mathsf{E}_0\delta(X)$ 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 $\Phi(k_1) - \Phi(-k_1) = 1 - \alpha$ 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 $k_1 > 1$。对备择假设的功效是 $\mathsf{E}_1\delta(X) = 1 - \int_{-k_1}^{k_1}(\pi + \pi x^2)^{-1}dx = 1 - \frac{2}{\pi}\arctan z_{1-\alpha/2}$。

例 8.5. 已知样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中 σ^2 已知而 μ 未知。考虑 $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu = \mu_1$，似然比为

$$\begin{aligned} \lambda(\boldsymbol{x}) &= \frac{\exp\{-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j - \mu_1)^2\}}{\exp\{-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j - \mu_0)^2\}} \\ &= \exp\left\{\sum_{j=1}^n x_j\left(\frac{\mu_1}{\sigma^2} - \frac{\mu_0}{\sigma^2}\right) + n\left(\frac{\mu_0^2}{2\sigma^2} - \frac{\mu_1^2}{2\sigma^2}\right)\right\} \end{aligned}$$

定义检验函数为 $\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{若 } \lambda(\boldsymbol{x}) \geq k \\ 0 & \text{若 } \lambda(\boldsymbol{x}) < k \end{cases}$ 如果 $\mu_1 \geq \mu_0$，则 $\lambda(\boldsymbol{x}) \geq k$ 当且仅当 $\sum_{j=1}^n x_j \geq k_1$，其中 $k_1 = z_{1-\alpha}\sigma\sqrt{n} + n\mu_0$ 是由下面的条件确定的：

$$\alpha = \mathsf{P}_{\mu_0}\left\{\sum_{j=1}^n X_j \geq k_1\right\} = \mathsf{P}\left\{\frac{\sum_{j=1}^n X_j - n\mu_0}{\sigma\sqrt{n}} \geq \frac{k_1 - n\mu_0}{\sigma\sqrt{n}}\right\}$$

$\mu_1 < \mu_0$ 的情况也是类似处理，留给读者练习。

定义 8.5. 检验 $\delta_k^*(\boldsymbol{x}), 0 \le k < \infty$ 被称为 N-P 检验 (N-P test)，如果

$$\delta_k^*(\boldsymbol{x}) = \begin{cases} \delta_k(\boldsymbol{x}), & \text{当 } \boldsymbol{x} \in \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \neq k\} \\ \text{任意}, & \text{当 } \boldsymbol{x} \in \{\boldsymbol{x} : \lambda(\boldsymbol{x}) = k\} \end{cases} \tag{8.7}$$

- 在水平 $\mathsf{P}_{\theta_0}[\delta_k^* = 1]$，N-P 检验 δ_k^* 是对简单假设 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$ 的最大功效检验。
- 如果检验函数 δ 的两类错误概率都不超过 δ_k 的，则 δ 是一个 N-P 检验，即仅在 $\lambda = k$ 时与 δ_k 不同。
- 如果统计量 T 对 θ 而言是充分的，则对简单假设 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$，N-P 检验是 T 的函数。

定义 8.6. 分布族 $\{f_\theta(\boldsymbol{x}) : \theta \in \Theta\}$ 称为对统计量 $T(\boldsymbol{X})$ 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 $\lambda(\boldsymbol{x}) = f_{\theta_1}(\boldsymbol{x})/f_{\theta_0}(\boldsymbol{x})$ 是关于 $T(\boldsymbol{x})$ 的非减函数。

例 8.6. 令 $X \sim \text{Cauchy}(1, \theta)$，则当 $x \to \pm\infty$

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = \frac{1 + (x - \theta_0)^2}{1 + (x - \theta_1)^2} \to 1$$

于是 Cauchy$(1, \theta)$ 没有单调似然比。

例 8.7. 单参数指数族 $f_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{q(\theta)T(\boldsymbol{x}) + g(\theta)\}$ 对统计量 $T(\boldsymbol{X})$ 具有单调似然比，其中 $q(\theta)$ 关于 θ 非减。

定理 8.2 (Karlin-Rubin)**.** 已知 $\boldsymbol{X} \sim f_\theta, \theta \in \Theta$，其中 $\{f_\theta\}$ 对统计量 $T(\boldsymbol{X})$ 具有单调似然比。为检验 $H_0 : \theta \le \theta_0 \leftrightarrow H_1 : \theta > \theta_0$，检验函数

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{若 } T(\boldsymbol{x}) > t_0 \\ 0 & \text{若 } T(\boldsymbol{x}) \le t_0 \end{cases} \tag{8.8}$$

具有非减的功效函数且是水平 $\alpha = \mathsf{P}\{T > t_0\}$ 的 UMP 检验。

证明. 见 V. K. Rohatgi 的《概率论及数理统计导论》 [78] 第九章第四节（第 420-421 页）或 G. Casella 和 R. L. Berger 的《统计推断》 [23] 第八章第三节。 □

定理 8.3. 令 $\theta_0 < \theta_1$，对于单参数指数族（见例 8.7）存在对 $H_0 : \theta \le \theta_0$ 或 $\theta \ge \theta_1 \leftrightarrow H_1 : \theta_0 < \theta < \theta_1$ 的 UMP 检验

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{若 } c_1 < T(\boldsymbol{x}) < c_2 \\ 0 & \text{若 } T(\boldsymbol{x}) \le c_1 \text{ 或 } T(\boldsymbol{x}) \ge c_2 \end{cases} \tag{8.9}$$

其中 c_1, c_2 由 $\mathsf{E}_{\theta_0}\delta(\boldsymbol{X}) = \mathsf{E}_{\theta_1}\delta(\boldsymbol{X}) = \alpha$ 解出，α 是给定的显著水平。

例 8.8. 已知样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, 1)$，对 $H_0 : \mu \le \mu_0$ or $\mu \ge \mu_1 \leftrightarrow H_1 : \mu_0 < \mu < \mu_1$，UMP 检验函数为

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{若 } c_1 < \sum_{j=1}^n x_j < c_2 \\ 0 & \text{若 } \sum_{j=1}^n x_j \le c_1 \text{ 或 } \sum_{j=1}^n x_j \ge c_2 \end{cases}$$

$$\begin{aligned} \alpha = \mathsf{P}_{\mu_0}\left\{c_1 < \sum_{j=1}^n X_j < c_2\right\} &= \mathsf{P}_{\mu_0}\left\{\frac{c_1 - n\mu_0}{\sqrt{n}} < \frac{\sum_{j=1}^n X_j - n\mu_0}{\sqrt{n}} < \frac{c_2 - n\mu_0}{\sqrt{n}}\right\} \\ &= \Phi\left(\frac{c_2 - n\mu_0}{\sqrt{n}}\right) - \Phi\left(\frac{c_1 - n\mu_0}{\sqrt{n}}\right) \end{aligned}$$

同理，$\Phi[(c_2 - n\mu_1)/\sqrt{n}] - \Phi[(c_1 - n\mu_1)/\sqrt{n}] = \alpha$，联立解出 c_1, c_2 即可。

8.1.3 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：令 $[\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$ 是参数 θ 的置信度为 $1-\alpha$ 的置信区间，定义检验函数 $\delta(\boldsymbol{x};\theta)$ 如下：

$$\delta(\boldsymbol{x};\theta) = \begin{cases} 0 & \text{若 } \theta \in [\underline{\theta}(\boldsymbol{x}), \overline{\theta}(\boldsymbol{x})] \\ 1 & \text{否则} \end{cases} \tag{8.10}$$

于是对于 $H_0 : \theta = \theta_0 \leftrightarrow \theta \neq \theta_0$，区间 $[\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$ 是水平 α 下的接受域，这是因为 $\mathsf{P}_{\theta_0}[\delta(\boldsymbol{X};\theta_0) = 1] = 1 - \mathsf{P}_{\theta_0}[\underline{\theta}(\boldsymbol{X}) \le \theta_0 \le \overline{\theta}(\boldsymbol{X})] \le \alpha$。反之，若 $\delta(\boldsymbol{x};\theta_0)$ 是 $H_0 : \theta = \theta_0 \leftrightarrow \theta \neq \theta_0$ 的一个水平为 α 的检验且 $\{\theta : \delta(\boldsymbol{X};\theta) = 0\}$ 是一个区间，该区间即是 θ 的一个置信度为 $1-\alpha$ 的置信区间。类似地，如果 $\underline{\theta}(\boldsymbol{X})$ 是 θ 的置信度为 $1-\alpha$ 的置信下限，把式 (8.10) 中 $\delta(\boldsymbol{x};\theta) = 0$ 的条件修改为 $\theta \in [\underline{\theta}(\boldsymbol{x}), \infty)$，则 $\{\boldsymbol{x} : \underline{\theta}(\boldsymbol{x}) \le \theta_0\}$ 是 $H_0 : \theta \le \theta_0 \leftrightarrow H_1 : \theta > \theta_0$ 的水平 α 下的接受域。反之，若 $\delta(\boldsymbol{x};\theta_0)$ 是 $H_0 : \theta \le \theta_0 \leftrightarrow H_1 : \theta > \theta_0$ 的一个水平为 α 的检验且 $\{\theta : \delta(\boldsymbol{X};\theta) = 0\}$ 是一个形如 $[\underline{\theta}(\boldsymbol{X}), \infty)$ 的区间，$\underline{\theta}(\boldsymbol{X})$ 即是 θ 的一个置信度为 $1-\alpha$ 的置信下限。

例 8.9. 已知样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，样本均值和方差分别为 $\overline{X}$ 和 S^2（它们的观察值分别记为 $\overline{x}$ 和 s^2）。在 σ^2 已知和未知两种情况之下，考虑双侧检验 (two-sided test) $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$，由例 7.22 中的第一、二种情况不难给出拒绝 H_0 的条件。对 μ 的单侧检验 (one-sided test)，请读者验证下面拒绝 H_0 的条件（下表中的最后两行）。

$H_0 \leftrightarrow H_1$	σ^2 已知	σ^2 未知
$\mu = \mu_0 \leftrightarrow \mu \neq \mu_0$	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{\sigma/\sqrt{n}} \ge z_{1-\alpha/2}$	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{s/\sqrt{n}} \ge t_{n-1,1-\alpha/2}$
$\mu \ge \mu_0 \leftrightarrow \mu < \mu_0$	$\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \le z_{\alpha}$	$\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} \le t_{n-1,\alpha}$
$\mu \le \mu_0 \leftrightarrow \mu > \mu_0$	$\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \ge z_{1-\alpha}$	$\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} \ge t_{n-1,1-\alpha}$

例 8.10. 已知样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中 σ^2 未知。在 μ 已知和未知两种情况之下，给出拒绝零假设 H_0 的条件如下。

$H_0 \leftrightarrow H_1$	μ 已知	μ 未知
$\sigma^2 \geq \sigma_0^2 \leftrightarrow \sigma^2 < \sigma_0^2$	$\dfrac{\sum_{j=1}^n (x_j-\mu)^2}{\sigma_0^2} \leq \chi^2_{n,\alpha}$	$\dfrac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{n-1,\alpha}$
$\sigma^2 \leq \sigma_0^2 \leftrightarrow \sigma^2 > \sigma_0^2$	$\dfrac{\sum_{j=1}^n (x_j-\mu)^2}{\sigma_0^2} \geq \chi^2_{n,1-\alpha}$	$\dfrac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{n-1,1-\alpha}$
$\sigma^2 = \sigma_0^2 \leftrightarrow \sigma^2 \neq \sigma_0^2$	$\dfrac{\sum_{j=1}^n (x_j-\mu)^2}{\sigma_0^2} \leq \chi^2_{n,\alpha/2}$ 或 $\dfrac{\sum_{j=1}^n (x_j-\mu)^2}{\sigma_0^2} \geq \chi^2_{n,1-\alpha/2}$	$\dfrac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{n-1,\alpha/2}$ 或 $\dfrac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{n-1,1-\alpha/2}$

例 8.11. 样本 $X_1, \cdots, X_m \overset{iid}{\sim} \mathrm{N}(\mu_X, \sigma_X^2)$ 和样本 $Y_1, \cdots, Y_n \overset{iid}{\sim} \mathrm{N}(\mu_Y, \sigma_Y^2)$ 来自两个独立的总体，它们的样本方差分别为 S_X^2 和 S_Y^2，两个样本的*合并样本方差* (pooled sample variance) 定义为 $S^2 = \frac{1}{m+n-2}[(m-1)S_X^2 + (n-1)S_Y^2]$。在总体方差已知和未知（但知道 $\sigma_X^2 = \sigma_Y^2$）两种情况之下，给出拒绝零假设 H_0 的条件如下。

$H_0 \leftrightarrow H_1$	σ_X^2, σ_Y^2 已知	$\sigma_X^2 = \sigma_Y^2 = \sigma^2$ 未知
$\mu_X \geq \mu_Y \leftrightarrow \mu_X < \mu_Y$	$\dfrac{\overline{x}-\overline{y}}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} \leq z_\alpha$	$\dfrac{\overline{x}-\overline{y}}{s\sqrt{1/m+1/n}} \leq t_{m+n-2,\alpha}$
$\mu_X \leq \mu_Y \leftrightarrow \mu_X > \mu_Y$	$\dfrac{\overline{x}-\overline{y}}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} \geq z_{1-\alpha}$	$\dfrac{\overline{x}-\overline{y}}{s\sqrt{1/m+1/n}} \geq t_{m+n-2,1-\alpha}$
$\mu_X = \mu_Y \leftrightarrow \mu_X \neq \mu_Y$	$\dfrac{\lvert\overline{x}-\overline{y}\rvert}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} \geq z_{1-\alpha/2}$	$\dfrac{\lvert\overline{x}-\overline{y}\rvert}{s\sqrt{1/m+1/n}} \geq t_{m+n-2,1-\alpha/2}$

右列结果根据的是性质 6.5，即 $\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{S\sqrt{1/m+1/n}} \sim t(m+n-2)$。

例 8.12. 已知样本 $(X_1,Y_1)^\mathsf{T},\cdots,(X_n,Y_n)^\mathsf{T} \overset{iid}{\sim} \mathrm{N}(\boldsymbol{\mu},\Sigma)$，其中期望 $\boldsymbol{\mu}=(\mu_X,\mu_Y)^\mathsf{T}$ 且协方差阵 $\Sigma=[\sigma_X^2,\rho\sigma_X\sigma_Y;\rho\sigma_X\sigma_Y,\sigma_Y^2]$ 未知。显然，$D_j=X_j-Y_j \overset{iid}{\sim} \mathrm{N}(\mu_X-\mu_Y,\sigma^2)$，其中 $\sigma^2=\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y$。定义 $\overline{D}=\frac{1}{n}\sum_{j=1}^n D_j$ 并且 $S^2=\frac{1}{n-1}\sum_{j=1}^n(D_j-\overline{D})^2$。

$H_0 \leftrightarrow H_1$	拒绝零假设 H_0 的条件
$\mu_X-\mu_Y \geq d_0 \leftrightarrow \mu_X-\mu_Y < d_0$	$\dfrac{\overline{d}-d_0}{s/\sqrt{n}} \leq t_{n-1,\alpha}$
$\mu_X-\mu_Y \leq d_0 \leftrightarrow \mu_X-\mu_Y > d_0$	$\dfrac{\overline{d}-d_0}{s/\sqrt{n}} \geq t_{n-1,1-\alpha}$
$\mu_X-\mu_Y = d_0 \leftrightarrow \mu_X-\mu_Y \neq d_0$	$\dfrac{\lvert\overline{d}-d_0\rvert}{s/\sqrt{n}} \geq t_{n-1,1-\alpha/2}$

例 8.13. 已知来自两个独立总体的样本 $X_1,\cdots,X_m \overset{iid}{\sim} \mathrm{N}(\mu_X,\sigma_X^2)$ 与 $Y_1,\cdots,Y_n \overset{iid}{\sim} \mathrm{N}(\mu_Y,\sigma_Y^2)$，样本方差分别为 S_X^2 和 S_Y^2，于是 $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1,n-1)$。在下面的假设检验中，拒绝零假设 H_0 的条件是

$H_0 \leftrightarrow H_1$	μ_X,μ_Y 已知	μ_X,μ_Y 未知
$\sigma_X^2 \geq \sigma_Y^2 \leftrightarrow \sigma_X^2 < \sigma_Y^2$	$\dfrac{\sum_{j=1}^n(y_j-\mu_Y)^2/n}{\sum_{j=1}^m(x_j-\mu_X)^2/m} \geq F_{n,m,1-\alpha}$	$\dfrac{s_Y^2}{s_X^2} \geq F_{n-1,m-1,1-\alpha}$
$\sigma_X^2 \leq \sigma_Y^2 \leftrightarrow \sigma_X^2 > \sigma_Y^2$	$\dfrac{\sum_{j=1}^m(x_j-\mu_X)^2/m}{\sum_{j=1}^n(y_j-\mu_Y)^2/n} \geq F_{m,n,1-\alpha}$	$\dfrac{s_X^2}{s_Y^2} \geq F_{m-1,n-1,1-\alpha}$
$\sigma_X^2 = \sigma_Y^2 \leftrightarrow \sigma_X^2 \neq \sigma_Y^2$	$\begin{cases}\dfrac{\sum_{j=1}^m(x_j-\mu_X)^2/m}{\sum_{j=1}^n(y_j-\mu_Y)^2/n} \geq F_{m,n,1-\alpha/2} \\ \text{或} \\ \dfrac{\sum_{j=1}^n(y_j-\mu_Y)^2/n}{\sum_{j=1}^m(x_j-\mu_X)^2/m} \geq F_{n,m,1-\alpha/2}\end{cases}$	$\begin{cases}\dfrac{s_X^2}{s_Y^2} \geq F_{m-1,n-1,1-\alpha/2} \\ \text{如果 } s_X^2 \geq s_Y^2 \text{ 或} \\ \dfrac{s_Y^2}{s_X^2} \geq F_{n-1,m-1,1-\alpha/2} \\ \text{如果 } s_X^2 < s_Y^2\end{cases}$

例 8.14. 考察 9 个人在新的饮食计划实施前后的体重，以确定该饮食计划是否有助于减轻体重。假定前后的体重满足 $(X,Y)^\mathsf{T} \overset{iid}{\sim} \mathrm{N}(\boldsymbol{\mu},\Sigma)$，在给定的显著水平 $\alpha=0.01$ 下考察 $H_0:\mu_X-\mu_Y\leq 0 \leftrightarrow H_1:\mu_X-\mu_Y>0$。

体重（千克）	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

$\overline{d} = (8-2+8-2+8+0-3-4+5)/9 = 2, d_0 = 0, n = 9, s = 5.17, t_{8,0.99} = 2.896 \Rightarrow \frac{\overline{d}-d_0}{s/\sqrt{n}} < t_{n-1,1-\alpha} \Rightarrow$ 接受 H_0，即该饮食计划无助于减轻体重。

8.2 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构造合理的检验，很多情况下也能带来形式上的简化。例如，当样本分布为指数族时，如下定义的似然比检验*（或称 Wald 检验）具有一定的可行性，但在一般情况下此方法有计算上的困难。

定义 8.7. 令 $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$ 为一个向量参数，样本 $\boldsymbol{X} = (X_1, \cdots, X_n)^\mathsf{T}$ 的似然函数为 $\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x})$，其中 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^\mathsf{T}$。考虑假设 $H_0 : \boldsymbol{\theta} \in \Theta_0 \leftrightarrow H_1 : \boldsymbol{\theta} \in \Theta_1$，定义似然比 $\lambda(\boldsymbol{x})$ 为

$$\lambda(\boldsymbol{x}) = \frac{\sup\limits_{\boldsymbol{\theta}\in\Theta_0} \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x})}{\sup\limits_{\boldsymbol{\theta}\in\Theta} \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x})} = \frac{\mathscr{L}(\hat{\boldsymbol{\theta}}_0; \boldsymbol{x})}{\mathscr{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{x})} \tag{8.11}$$

其中，$\hat{\boldsymbol{\theta}}_0$ 和 $\hat{\boldsymbol{\theta}}$ 分别是参数限定在 Θ_0 和 Θ 中的 MLE。显然 $0 \le \lambda(\boldsymbol{x}) \le 1$。我们把“拒绝零假设 H_0 当且仅当 $\lambda(\boldsymbol{x}) < c$”这样的检验称为似然比检验 (likelihood ratio test)。

直观上，如果 H_0 为真，则似然比 (8.11) 必然接近 1。换句话说，如果这个比值很小就应该否定 H_0。确定拒绝域 $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) < c\}$ 需要最大似然估计，有时还要定性地利用 $\lambda(\boldsymbol{x})$ 的单调性来求解。临界值 $c \in (0, 1)$ 由 $\sup\limits_{\boldsymbol{\theta}\in\Theta_0} \mathsf{P}_{\boldsymbol{\theta}}\{\boldsymbol{X} : \lambda(\boldsymbol{X}) < c\} = \alpha$ 决定，其中 α 是给定的显著水平。在一般情况下，由于难以求出零假设成立时 $\lambda(\boldsymbol{X})$ 的分布而变得复杂。幸运的是，当样本量足够地大，在一定的条件之下 $-2\ln\lambda(\boldsymbol{X})$ 渐近于 χ^2 分布，于是临界值 c 可近似求得。鉴于此，本书把似然比检验划归为大样本检验，虽然小样本情况下它有时也是可行的（见下面的例子）。

*似然比检验是 J. Neyman 和 E. S. Pearson 于 1928 年提出的，适用范围较广。虽然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。有的文献把式 (8.11) 定义的似然比称为“广义似然比”以区别于 §8.1.2 讨论的似然比。也有人把广义似然比定义为 $\sup\limits_{\boldsymbol{\theta}\in\Theta} \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x}) / \sup\limits_{\boldsymbol{\theta}\in\Theta_0} \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x})$，只是行文习惯不同而已。

例 8.15. 设样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} p\langle 1\rangle + (1-p)\langle 0\rangle$，其中参数 p 未知。令 $X = X_1 + X_2 + \cdots + X_n$，则 $X \sim \mathrm{B}(n, p)$。对假设 $H_0 : p \le p_0 \leftrightarrow H_1 : p > p_0$ 进行水平为 α 的似然比检验。因为 $p^x(1-p)^{n-x}$ 是关于 p 的单峰函数，在 x/n 处取得最大值，所以

$$\sup_{0\le p\le 1} p^x(1-p)^{n-x} = (x/n)^x(1-x/n)^{n-x}$$

$$\sup_{p\le p_0} p^x(1-p)^{n-x} = \begin{cases} p_0^x(1-p_0)^{n-x} & \text{如果 } p_0 < x/n \\ (x/n)^x(1-x/n)^{n-x} & \text{如果 } p_0 \ge x/n \end{cases}$$

$$\lambda(x) = \frac{\sup\limits_{p\le p_0} C_n^x p^x(1-p)^{n-x}}{\sup\limits_{0\le p\le 1} C_n^x p^x(1-p)^{n-x}} = \begin{cases} 1 & \text{如果 } x \le np_0 \\ \dfrac{p_0^x(1-p_0)^{n-x}}{(x/n)^x(1-x/n)^{n-x}} & \text{如果 } x > np_0 \end{cases}$$

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 $x > c'$，似然比检验否定 H_0。

$$\sup_{p\le p_0} \mathsf{P}_p\{X > c'\} = \mathsf{P}_{p_0}\{X > c'\} = 1 - \sum_{k=0}^{\lfloor c'\rfloor} C_n^k p_0^k (1-p_0)^{n-k}$$

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'：$\mathsf{P}_{p_0}\{X > c'\} \le \alpha$ 且 $\mathsf{P}_{p_0}\{X > c'-1\} > \alpha$。

本节内容　第一节继续列举了似然比检验的实例，并给出了大样本情况下的似然比检验。为检验总体服从是某一给定的分布，或属于某一分布族，第二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，它是第三节所讨论的内容。

学习目标　(1) 掌握似然比检验、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。

8.2.1 似然比检验

例 8.16. 总体的分布为 $\mathrm{N}(\mu,\sigma^2)$，其中参数 $\boldsymbol{\theta}=(\mu,\sigma^2)^\mathsf{T}$ 未知。用似然比的方法对 $H_0:\mu=\mu_0 \leftrightarrow H_1:\mu\neq\mu_0$ 进行检验。

- 零假设成立时的参数空间为 $\Theta_0=\{(\mu_0,\sigma^2)^\mathsf{T}:\sigma^2>0\}$，此时似然函数的上确界为

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta_0} f_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \sup_{\sigma^2>0}\left[\frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left\{-\frac{\sum_{j=1}^n(x_j-\mu_0)^2}{2\sigma^2}\right\}\right]\\ &= f_{\hat{\sigma}^2}(\boldsymbol{x})=\left[\frac{1}{\sqrt{2\pi e}\hat{\sigma}}\right]^n\end{aligned}$$

  其中，$\hat{\sigma}^2=\sum_{j=1}^n(x_j-\mu_0)^2/n$ 是参数 σ^2 的最大似然估计值。

- 在整个参数空间 $\Theta=\{(\mu,\sigma^2)^\mathsf{T}:\mu\in\mathbb{R},\sigma^2>0\}$ 上，参数 $\boldsymbol{\theta}=(\mu,\sigma^2)^\mathsf{T}$ 的最大似然估计值和似然函数的上确界如下，进而得到似然比。

$$\hat{\boldsymbol{\theta}}=\left(\frac{1}{n}\sum_{j=1}^n x_j,\frac{1}{n}\sum_{j=1}^n(x_j-\overline{x})^2\right)^\mathsf{T}=(a_1,b_2)^\mathsf{T},\sup_{\boldsymbol{\theta}\in\Theta} f_{\boldsymbol{\theta}}(\boldsymbol{x})=\left[\frac{1}{\sqrt{2\pi e b_2}}\right]^n$$

  于是，似然比为 $\lambda(\boldsymbol{x})=\left(\dfrac{b_2}{\hat{\sigma}^2}\right)^{n/2}=\left[1+\dfrac{n(\overline{x}-\mu_0)^2}{\sum_{j=1}^n(x_j-\overline{x})^2}\right]^{-n/2}$

如果 $\lambda(\boldsymbol{x})<c$，似然比检验拒绝零假设 H_0。因为 $\lambda(\boldsymbol{x})$ 是 $n(\overline{x}-\mu_0)^2/\sum_{j=1}^n(x_j-\overline{x})^2$ 的减函数，所以

$$\left|\frac{\sqrt{n}(\overline{x}-\mu_0)}{\sqrt{\sum_{j=1}^n(x_j-\overline{x})^2}}\right|>c' \text{ 或者等价地，} \left|\frac{\sqrt{n}(\overline{x}-\mu_0)}{s}\right|>c''$$

其中 $s^2=\frac{1}{n-1}\sum_{j=1}^n(x_j-\overline{x})^2$。因为 $\sqrt{n}(\overline{X}-\mu_0)/S\sim t(n-1)$，故选取 $c''=t_{n-1,1-\alpha/2}$（此处由似然比检验导出的双侧 t 检验是 UMP 的）。

例 8.17. 已知样本 $X_1,\cdots,X_m \overset{iid}{\sim} \mathrm{N}(\mu_X,\sigma_X^2)$ 和 $Y_1,\cdots,Y_n \overset{iid}{\sim} \mathrm{N}(\mu_Y,\sigma_Y^2)$，且两总体是独立的，其中参数 $\boldsymbol{\theta}=(\mu_X,\sigma_X^2,\mu_Y,\sigma_Y^2)^\mathsf{T}$ 未知。用似然比的方法对 $H_0:\sigma_X^2=\sigma_Y^2 \leftrightarrow H_1:\sigma_X^2\neq\sigma_Y^2$ 进行检验。

- 整个参数空间是 $\Theta=\{(\mu_X,\sigma_X^2,\mu_Y,\sigma_Y^2)^\mathsf{T}:\mu_X,\mu_Y\in\mathbb{R},\sigma_X^2,\sigma_Y^2>0\}$，样本 $(X_1,\cdots,X_m,Y_1,\cdots,Y_n)^\mathsf{T}$ 的密度函数为

$$f_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})=\frac{\exp\left\{-\frac{1}{2\sigma_X^2}\sum_{j=1}^m(x_j-\mu_X)^2-\frac{1}{2\sigma_Y^2}\sum_{j=1}^n(x_j-\mu_Y)^2\right\}}{(2\pi)^{(m+n)/2}\sigma_X^m\sigma_Y^n}$$

  参数 $\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2$ 的最大似然估计值分别为 $\hat{\mu}_1=\overline{x},\hat{\mu}_2=\overline{y},\hat{\sigma}_X^2=\frac{1}{m}\sum_{j=1}^m(x_j-\overline{x})^2,\hat{\sigma}_Y^2=\frac{1}{n}\sum_{j=1}^n(y_j-\overline{y})^2$。

- 零假设成立的时候，参数空间为 $\Theta_0=\{(\mu_X,\sigma_X^2,\mu_Y,\sigma_Y^2)^\mathsf{T}:\mu_X,\mu_Y\in\mathbb{R},\sigma_X^2=\sigma_Y^2>0\}$。设 $\sigma_X^2=\sigma_Y^2=\sigma^2$，其最大似然估计值 $\hat{\sigma}^2=\frac{1}{m+n}[\sum_{j=1}^m(x_j-\overline{x})^2+\sum_{j=1}^n(y_j-\overline{y})^2]$。

经过简单的计算，得到似然比如下：

$$\begin{aligned}\lambda(\boldsymbol{x},\boldsymbol{y})&=\sqrt{\frac{m^m n^n\left[\sum_{j=1}^m(x_j-\overline{x})^2\right]^m\left[\sum_{j=1}^n(y_j-\overline{y})^2\right]^n}{(m+n)^{m+n}\left[\sum_{j=1}^m(x_j-\overline{x})^2+\sum_{j=1}^n(y_j-\overline{y})^2\right]^{m+n}}}\\&=\sqrt{\frac{m^m n^n}{(m+n)^{m+n}\left[1+\frac{m-1}{n-1}f\right]^n\left[1+\frac{n-1}{m-1}\frac{1}{f}\right]^m}}\\&\text{其中，}\ f=\frac{\sum_{j=1}^m(x_j-\overline{x})^2/(m-1)}{\sum_{j=1}^n(y_j-\overline{y})^2/(n-1)}\end{aligned}$$

请读者自行验证：$\lambda(\boldsymbol{x},\boldsymbol{y})<c$ 等价于 $f<c_1$ 或 $f>c_2$。根据 $F=\frac{\sum_{j=1}^m(X_j-\overline{X})^2/(m-1)}{\sum_{j=1}^n(Y_j-\overline{Y})^2/(n-1)}\sim F(m-1,n-1)$，选取 $c_1=F_{m-1,n-1,\alpha/2},c_2=F_{m-1.n-1,1-\alpha/2}$（该问题的似然比检验是 UMP 的）。

定理 8.4. 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，似然比具有如下的渐近分布：

$$-2\ln\lambda(\boldsymbol{X}) \sim \chi^2_m \tag{8.12}$$

如例 8.16 中，$-2\ln\lambda(\boldsymbol{X}) = n\ln[1 + n(\overline{X} - \mu_0)^2 / \sum_{j=1}^{n}(X_j - \overline{X})^2] \sim \chi^2_1$。

8.2.2　拟合优度检验

已知简单随机样本 $X_1, X_2, \cdots, X_n$ 来自总体 $X \sim F(x)$，对假设 $H_0 : F = F_0 \leftrightarrow F \neq F_0$ 进行的检验，或者更一般地，对假设 $H_0 : F \in \mathscr{F} \leftrightarrow F \notin \mathscr{F}$ 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），$\mathscr{F} = \{F_{\boldsymbol{\theta}}(x) : \boldsymbol{\theta} \in \Theta\}$ 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 $H_0 : F = \frac{1}{6}\langle 1\rangle + \cdots \frac{1}{6}\langle 6\rangle$。大样本检验的第一个重要结果是 K. Pearson 于 1900 年给出的下述引理，它也是统计学最重要的成果之一。

引理 8.2 (K. Pearson, 1900)**.** 已知多项分布 $\boldsymbol{Y} \sim \text{Multin}(n; p_1, p_2, \cdots, p_k)$，定义 Pearson χ^2 统计量为

$$\chi^2(\boldsymbol{Y}) = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} = \frac{1}{n}\sum_{j=1}^{k} \frac{Y_j^2}{p_j} - n, \text{ 其中 } \sum_{j=1}^{k} Y_j = n \tag{8.13}$$

当 $n \to \infty$ 时，渐近地有 $\chi^2 \sim \chi^2_{k-1}$。

证明. 见陈希孺的《高等概率统计学》 [9] 第六章第二节。 □

按照分点 $a_0 < \cdots < a_k$ 把实数轴 $\mathbb{R}$ 划分成 k 个两两不交的区间：$A_1 = (a_0, a_1), A_2 = [a_1, a_2), \cdots, A_k = [a_{k-1}, a_k]$，其中 $a_0 = -\infty, a_k = \infty$。

定理 8.5 (Pearson χ^2 检验)**.** 设 $p_j = \mathsf{P}(X \in A_j) > 0, j = 1, 2, \cdots, k$，显然 $\sum_{j=1}^{k} p_j = 1$。定义随机变量 Y_j 为 $X_1, \cdots, X_n$ 落于区间 A_j 内的个数，则 $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_k)^{\mathsf{T}} \sim \text{Multin}(n; p_1, p_2, \cdots, p_k)$。由引理 8.2，给定显著水平 α，当 Pearson χ^2 统计量的观察结果 $\chi^2(\boldsymbol{y}) > \chi^2_{k-1,1-\alpha}$ 时拒绝零假设；当 $\chi^2(\boldsymbol{y}) \leq \chi^2_{k-1,1-\alpha}$ 时接受零假设。

定理 8.6 (Fisher, 1924)**.** 已知样本 $X_1, \cdots, X_n \overset{iid}{\sim} F_{\boldsymbol{\theta}}(x)$，假设参数 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_r)^{\mathsf{T}}$ 的最大似然估计存在，设为 $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_r)^{\mathsf{T}}$。设 $\hat{p}_j = \mathsf{P}_{\hat{\boldsymbol{\theta}}}\{X \in A_j\} > 0, j = 1, \cdots, k$，定义随机变量 Y_j 为样本 $X_1, \cdots, X_n$ 落于 A_j 内的个

数。当 $n \to \infty$ 时，渐近地有

$$\chi^2(\boldsymbol{Y}) = \sum_{j=1}^{k} \frac{(Y_j - n\hat{p}_j)^2}{n\hat{p}_j} \sim \chi^2_{k-1-r} \tag{8.14}$$

对 $H_0 : F \in \mathscr{F} = \{F_{\boldsymbol{\theta}}(x)\}$ 的检验可转换为对 $H_0 : F = F_{\hat{\boldsymbol{\theta}}}(x)$ 的 Pearson χ^2 检验：$\chi^2(\boldsymbol{y}) > \chi^2_{k-1-r,1-\alpha}$ 时，在水平 α 下否定 H_0。虽然历史上 K. Pearson 由于疏忽了有参数情况下自由度应减小的事实，曾与 Fisher 有过不愉快的激烈争论，但为纪念 K. Pearson 发现引理 8.2 的学术功绩，习惯上仍然把基于定理 8.6 的检验也称作拟合优度的 Pearson χ^2 检验。式 (8.13) 和式 (8.14) 中的 Y_j 称为经验频次，np_j 和 $n\hat{p}_j$ 称为理论频次。

例 8.18. 72小时之内全国发生了 306 起交通事故，每小时事故数的观察结果见下表左边两列。问每小时的事故数 X 是否服从 Poisson 分布？

每小时事故数	y_j	$n\hat{p}_j$
0 或 1	4	5.38
2	10	9.28
3	15	13.14
4	12	13.96
5	12	11.87
6	6	8.41
7	5	5.10
8 或更多	7	4.86

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 $\hat{\lambda} = \overline{X} = 306/72 = 4.25$。根据递归关系

$$\frac{\mathsf{P}_{\hat{\lambda}}(X = j+1)}{\mathsf{P}_{\hat{\lambda}}(X = j)} = \frac{\hat{\lambda}}{j+1}$$

以及初始值 $\hat{p}_0 = \mathsf{P}_{\hat{\lambda}}(X = 0) = \exp\{-\hat{\lambda}\} = 0.0143$，可以得到 $\hat{p}_j = \mathsf{P}_{\hat{\lambda}}(X = j)$，进而求得 $n\hat{p}_j$（表中最右列），其中 $j = 0, 1, 2, \cdots$。

由 $k - 1 - r = 8 - 1 - 1 = 6$ 以及式 (8.14)，在水平 $\alpha = 0.05$ 之下，

$$\chi^2(\boldsymbol{y}) = \sum_{j=1}^{8} \frac{(y_j - n\hat{p}_j)^2}{n\hat{p}_j} = 2.58 < \chi^2_{6,0.95} = 12.59159 \Rightarrow \text{接受 } H_0$$

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 233 页的定理 6.2 的 Kolmogorov 检验。令 $F_n^*(x)$ 是由样本 $X_1, X_2, \cdots, X_n$ 构造的经验分布函

数，根据 Glivenko 定理 6.1，当 n 很大时统计量 $D_n = \sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)|$ 接近 0。对于零假设 $H_0 : F(x) = F_0(x)$，一个合理的检验是当 $D_n(\boldsymbol{x}) \geq c$ 时拒绝零假设，利用定理 6.2 可得 Kolmogorov 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 (6.12) 所定义的 Kolmogorov 分布 $K(z)$ 的 $(1-\alpha)$-分位数；在水平 $\alpha = 0.01$ 之下，取 $c = 1.628/\sqrt{n}$。

如何计算 D_n 呢？从图 6.2 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 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ 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

$$D_n^+ = \max_{1\leq j\leq n}\left\{\frac{j}{n} - F(X_{(j)})\right\}, \qquad D_n^- = \max_{1\leq j\leq n}\left\{F(X_{(j)}) - \frac{j-1}{n}\right\} \tag{8.15}$$

1944 年，苏联数学家 Vladimir Ivanovich Smirnov (1887-1974) 在定理 6.2 的基础上证明了统计量 D_n^+ 具有下面的极限性质。

$$\lim_{n\to\infty} \mathsf{P}\{\sqrt{n}D_n^+ \leq z\} = \begin{cases} 1 - \exp(-2z^2) & \text{若 } z > 0 \\ 0 & \text{若 } z \leq 0 \end{cases} \tag{8.16}$$

统计量 D_n^- 也有相同的结果。此外，Smirnov 还证明了下面的结论。

定理 8.7 (Smirnov, 1944)**.** 设简单随机样本 $X_{j1}, \cdots, X_{jn_j}$ 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 $F_1(x) = F_2(x)$，统计量 $D_{n_1,n_2} = \sup\{|F_{1n_1}^*(x) - F_{2n_2}^*(x)|\}$ 和 $D_{n_1,n_2}^+ = \sup\{F_{1n_1}^*(x) - F_{2n_2}^*(x)\}$ 具有如下极限性质。

$$\lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \mathsf{P}\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2} \leq z\right\} = K(z) \tag{8.17}$$

$$\lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \mathsf{P}\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2}^+ \leq z\right\} = \begin{cases} 1 - \exp(-2z^2) & \text{若 } z > 0 \\ 0 & \text{若 } z \leq 0 \end{cases} \tag{8.18}$$

统计量 D_{n_1,n_2} 和 $D^+_{n_1,n_2}$ 分别被称为 Smirnov 统计量和单侧 Smirnov 统计量。基于定理 8.7 可给出两个总体是否具有相同的连续分布函数的 Smirnov 检验：当 $D_{n_1,n_2}(\boldsymbol{x}_1,\boldsymbol{x}_2) \geq K_{1-\alpha}/\sqrt{n}$ 时拒绝零假设 $H_0: F_1(x) = F_2(x)$，其中 $n = n_1n_2/(n_1+n_2)$。

性质 8.1. 由式 (8.16) 可得到渐近关系：当 $n \to \infty$ 时，$4n(D^+_n)^2 \sim \chi^2_2$，这是因为 $\lim\limits_{n\to\infty} \mathsf{P}\{4n(D^+_n)^2 \leq x\} = 1 - \exp(-x/2)$ 即是 χ^2_2 的分布函数。类似地也有渐近关系，$\frac{4n_1n_2}{n_1+n_2}(D^+_{n_1,n_2})^2 \sim \chi^2_2$。

例 8.19. 设样本 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} F(x)$，其中 $F(x)$ 为一维连续分布函数。对于零假设 $H_0: F(x) \geq F_0(x)$ 的单侧检验是当 $D^+_n(\boldsymbol{x}) \geq c$ 时拒绝零假设。根据性质 8.1，在水平 $\alpha = 0.05$ 之下，取 $c = \frac{\sqrt{\chi^2_{2,0.95}}}{2\sqrt{n}} = 1.224/\sqrt{n}$；类似地，在水平 $\alpha = 0.01$ 之下，取 $c = 1.517/\sqrt{n}$。而对零假设 $H_0: F(x) \leq F_0(x)$，当 $D^-_n(\boldsymbol{x}) \leq c$ 时拒绝零假设，c 的选取同上。

8.2.3　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 = 受教育程度，B = 收入）是否相互关联感兴趣时，常把属性 A 分为 r 个等级 $A_1, A_2, \cdots, A_r$，把属性 B 分为 s 个等级 $B_1, B_2, \cdots, B_s$，这样共产生 rs 个组合子类。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，如下构造的 $r \times s$ 数据表被称为列联表 (contingency table)，其中，$n_{i\cdot} = \sum_{j=1}^{s} n_{ij}, n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$。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。

$_A{}^B$	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 $(X, Y)^{\mathsf{T}}$ 满足 $\mathsf{P}(X = i, Y = j) = p_{ij}$，其中 p_{ij} 表示 $(X, Y)^{\mathsf{T}}$ 属于 (A_i, B_j) 子类的概率，$i = 1, 2, \cdots, r$ 且 $j = 1, 2, \cdots, s$。对零假设“H_0 : X, Y 相互独立”的检验即验证存在非负常数 $p_{1\cdot}, \cdots, p_{r\cdot}$ 和 $p_{\cdot 1}, \cdots, p_{\cdot s}$ 满足 $\sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{s} p_{\cdot j} = 1$ 使得 $\mathsf{P}(X = i, Y = j) = p_{i\cdot} p_{\cdot j}$。若 H_0 成立，为确定 $(X, Y)^{\mathsf{T}}$ 的分布，必须把未知参数 $p_{1\cdot}, \cdots, p_{r\cdot}$ 和 $p_{\cdot 1}, \cdots, p_{\cdot s}$ 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 $\hat{p}_{i\cdot} = n_{i\cdot}/n, \hat{p}_{\cdot j} = n_{\cdot j}/n$，进而得到经验频次 n_{ij} 和理论频次 $n\hat{p}_{ij} = n_{i\cdot} n_{\cdot j}/n$。利用定理 8.6 可得 $n \to \infty$ 时，渐近地有

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n)^2}{n_{i\cdot} n_{\cdot j}/n} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n n_{ij} - n_{i\cdot} n_{\cdot j})^2}{n n_{i\cdot} n_{\cdot j}} \sim \chi^2_m \tag{8.19}$$

其中，$m = rs - 1 - (r + s - 2) = (r - 1)(s - 1)$。若 $\chi^2 > \chi^2_{m,1-\alpha}$，则在水平 α 下拒绝 $H_0 : X, Y$ 相互独立。

例 8.20. 为检验假设“H_0 = 在句子中单词 w_1 和 w_2 相互独立”，随机选取 n 个句子，得到如下 2×2 列联表。

频次	出现 w_2	不出现 w_2	求和
出现 w_1	a	b	$a + b$
不出现 w_1	c	d	$c + d$
求和	$a + c$	$b + d$	$a + b + c + d$

根据列联表检验和式 (8.19)，如果 $\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(b+d)(c+d)} > \chi^2_{1,1-\alpha}$，则在水平 α 下拒绝 H_0。例如，$\chi^2_{1,0.95} = 3.841459$ 和 $\chi^2_{1,0.99} = 6.634897$。

8.3　习题

8.1. 考虑第 281 页的例 8.1，在显著水平 $\alpha = 0.05$ 下，问这批零件的长度是否合格？$z_{0.975} = 1.959964$

8.2. 已知正态总体的方差 σ^2，期望 μ 只可能取 μ_0 或 μ_1 二值之一，$\overline{X}$ 为总体容量为 n 的样本均值。在显著水平 α 下，对 $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu = \mu_1 > \mu_0$ 的检验犯两类错误的概率为 α, γ，求样本容量 n。

8.3. 已知样本 $X_1, \cdots, X_{25} \overset{iid}{\sim} \mathrm{N}(\mu, 9)$，其中 μ 未知。设 $\overline{X}$ 为样本均值，在显著水平 $\alpha = 0.05$ 下，若 $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$ 的否定域为 $R = \{\boldsymbol{x} : |\overline{x} - \mu_0| \geq c\}$，试确定常数 c。

8.4. 为了比较甲、乙两个品种农作物的优劣，各种 10 亩，假设亩产量服从正态分布。收获后测得甲品种的亩产量（单位：千克）的均值为 32，标准差为 23；乙品种的亩产量的均值为 21，标准差为 12。问：在显著水平 $\alpha = 0.01$ 下，这两个品种有无差别？$t_{18,0.995} = 2.878440, F_{9,9,0.995} = 6.54109$

☆ 8.5. 某测量值 $(X, Y)^{\mathsf{T}}$ 或服从 $\mathrm{N}(0, 0, 1, 1, 0.6)$ 分布（总体 1）或服从 $\mathrm{N}(1, 1, 1, 1, 0.6)$ 分布（总体 2）。请给出一个统计量 $T(X, Y)$ 和一个临界值 c，使得如下分类规则的两个错误分类的概率 $\mathsf{P}_2(T \geq c), \mathsf{P}_1(T < c)$ 尽可能地小：若 $T \geq c$，则 $(X, Y)^{\mathsf{T}}$ 来自总体 1；如果 $T < c$，则 $(X, Y)^{\mathsf{T}}$ 来自总体 2。

8.6. 已知样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$，其中 μ, σ^2 未知。请给出对假设 $H_0 : \sigma^2 \leq \sigma_0^2 \leftrightarrow H_1 : \sigma^2 > \sigma_0^2$ 的似然比检验。

☆ 8.7. 设样本 $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$。请给出对假设 $H_0 : \sigma^2 = \sigma_0^2 \leftrightarrow H_1 : \sigma^2 \neq \sigma_0^2$ 的似然比检验。

☆ 8.8. 已知样本 $X_1, \cdots, X_m \overset{iid}{\sim} \mathrm{N}(\mu_1, \sigma^2)$ 和 $Y_1, \cdots, Y_n \overset{iid}{\sim} \mathrm{N}(\mu_2, \sigma^2)$，请给出对假设 $H_0 : \mu_1 = \mu_2 \leftrightarrow H_1 : \mu_1 \neq \mu_2$ 的似然比检验。

8.9. 按 Mendel 的遗传定律，让开粉红花的豌豆随机交配，子代可分为红花、粉红花和白花三类，其比例为 1 : 2 : 1。为检验这个理论设计试验：100 株豌豆中开红花 30 株，开粉红花 48 株，开白花 22 株。问 Mendel 遗传定律是否成立（置信水平 $\alpha = 0.05$）？

8.10. 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（置信水平 $\alpha = 0.05$）？

抛次 k	1	2	3	4	5	6	≥ 7
频次 n_k	533	233	121	53	29	15	16

8.11. 当 $k = 2$ 时，证明引理 8.2。

第九章

线性模型的回归分析与方差分析

在经典数学、物理学的理论中，为揭示了自然的本质规律，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。变量之间除了这样的函数关系外，还有一种非确定性的关系，即所谓的*相关关系*：自变量的取值确定时，因变量的取值虽不能完全确定，但与前者有着某种联系。譬如，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）。当多个自变量 $a_1, \cdots, a_k$ 的取值确定时，因变量 X 是一个随机变量，它与 $a_1, \cdots, a_k$ 之间的相关关系可描述为：

$$\mathsf{E}X = f(a_1, \cdots, a_k)，\text{或者等价地，} X = f(a_1, \cdots, a_k) + \epsilon \tag{9.1}$$

其中，$\epsilon \sim \mathrm{N}(0, \sigma^2)$ 是一个*随机误差项*[*]，函数 f 称为*回归函数*。*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 [3,92]，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 $a_1, \cdots, a_n$ 来预测 X，或通过控制 $a_1, \cdots, a_n$ 来控制 X（譬如，通过控制

[*]该随机误差项的期望之所以设为 0，是因为 $X = f(a_1, \cdots, a_k) + \mathsf{E}\epsilon + (\epsilon - \mathsf{E}\epsilon)$ 中总可以把 $\eta = \epsilon - \mathsf{E}\epsilon \sim \mathrm{N}(0, \sigma^2)$ 当做随机误差项。

土质、烧结温度、烧制时间等来控制砖的硬度）。最简单的回归函数是线性函数，即 $\mathsf{E}X = \beta_0 + \beta_1 a_1 + \cdots + \beta_k a_k$，其中 $\beta_0, \beta_1, \cdots, \beta_k$ 是待定的参数，称为**回归系数**，要利用 X 的样本把它们都估计出来。

例 9.1. 观察 Iris 数据中 setosa 类的萼片长度（自变量，横轴）与宽度（纵轴）的散点图，有直线可以近似地描绘这两个变量之间的关系。软件 R 自带了 Iris 数据，列举部分数据如下：

```
1                1   2   3   4   5   6   7   8 ...  44  45  46  47  48  49  50
2  Sepal.Length 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 ... 5.0 5.1 4.8 5.1 4.6 5.3 5.0
3  Sepal.Width  3.5 3.0 3.2 3.1 3.6 3.9 3.4 3.4 ... 3.5 3.8 3.0 3.8 3.2 3.7 3.3
```

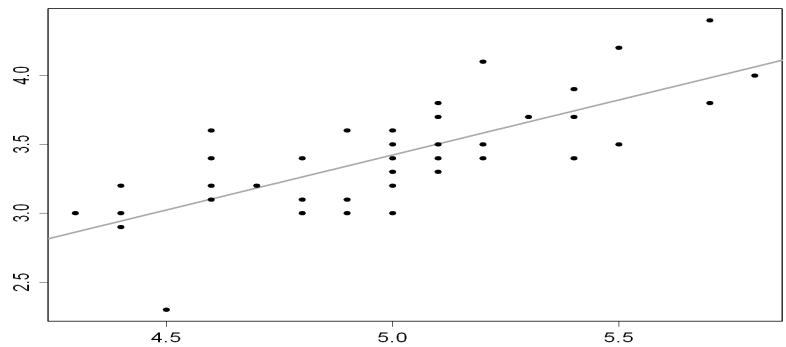


图 9.1: 萼片长度 a 确定时，宽度 X 是不确定的，但基本围绕在 $\beta_0 + \beta_1 a$ 的周围，其中 β_0, β_1 是待定的参数。

对人的身高和体重这两个变量，身高 a 的取值确定时，体重 X 是一个随机变量，按照经验有线性的数量关系表达式（单位：千克、米）

$$X = \beta_0 + \beta_1 a + \epsilon \tag{9.2}$$

其中，$\epsilon \sim \mathrm{N}(0, \sigma^2)$ 是一个随机误差项，σ^2, β_0 和 β_1 都是待定的参数。从式 (9.2) 不难看出，随机变量 X 由两部分组成：确定的 $\beta_0 + \beta_1 a$ 和不确定的 ϵ。根据样本值 $(a_1, x_1)^\mathsf{T}, \cdots, (a_n, x_n)^\mathsf{T}$，利用参数估计的方法（即最小二乘法，详见 §9.1.1）可以分别得到 $\sigma^2, \beta_0, \beta_1$ 的点估计值 $\hat{\sigma}^2, \hat{\beta}_0, \hat{\beta}_1$。我们把 $\hat{x}_i = \hat{\beta}_0 + \hat{\beta}_1 a_i$ 称作**理论值**（在不同场合下也称作**回归值**、**预测值**、**拟合值**等），把相应的 x_i 称作观测值，把 $x_i - \hat{x}_i$ 称作**残差** (residual)，把 $\boldsymbol{x} - \hat{\boldsymbol{x}}$ 称作残差向量，其中 $\hat{\boldsymbol{x}} = (\hat{x}_1, \cdots, \hat{x}_n)^\mathsf{T}$。

9.1　线性回归模型

因变量 X 是一个随机变量，它与（可精确测量或可控制的）一般变量 $a_1,\cdots,a_k$ 之间所有可能的关系中最简单的就是下面的线性关系，其中 $\beta_0,\beta_1,\cdots,\beta_k$ 是待定的参数，$\epsilon\sim \mathrm{N}(0,\sigma^2)$ 是随机误差。

$$X=\beta_0+\beta_1 a_1+\cdots+\beta_k a_k+\epsilon \tag{9.3}$$

定义 9.1 (线性回归模型)**.** 设式 (9.3) 中自变量 $a_1,\cdots,a_k$ 的取值分别为 $a_{i1},\cdots,a_{ik}$ 的时候观察到样本点 X_i，$i=1,\cdots,n$ 且 $n>k$，于是便得到 n 个线性方程 $X_i=\beta_0+\beta_1 a_{i1}+\cdots+\beta_k a_{ik}+\epsilon_i$ 构成的方程组，其中假定 $\epsilon_1,\cdots,\epsilon_n \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$，该方程组称为 k 元线性回归模型。为了记述和推导的方便，把线性回归模型整理为矩阵的形式。

$$\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}=\begin{pmatrix}1&a_{11}&\cdots&a_{1k}\\1&a_{21}&\cdots&a_{2k}\\\vdots&\vdots&\cdots&\vdots\\1&a_{n2}&\cdots&a_{nk}\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{pmatrix}+\begin{pmatrix}\epsilon_1\\\epsilon_2\\\vdots\\\epsilon_n\end{pmatrix} \tag{9.4}$$

记参数向量 $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_k)^{\mathsf{T}}\in\mathbb{R}^{k+1}$，令随机向量 $\boldsymbol{\epsilon}=(\epsilon_1,\cdots,\epsilon_n)^{\mathsf{T}}\sim \mathrm{N}_n(\mathbf{0},\sigma^2 I)$，矩阵 $A=(a_{ij})_{n\times(k+1)}$ 是确定的，线性回归模型在形式上进一步简化为

$$\boldsymbol{X}=A\boldsymbol{\beta}+\boldsymbol{\epsilon}，\text{或者等价地，}\mathsf{E}\boldsymbol{X}=A\boldsymbol{\beta} \tag{9.5}$$

如果样本 $\boldsymbol{X}=(X_1,\cdots,X_n)^{\mathsf{T}}$ 使得式 (9.5) 成立，则称它满足一个线性模型 (linear model)。如果变量之间的关系不是线性的，某些情况下我们可以通过变换使之成为线性的。譬如，

$$\frac{1}{y}=a+\frac{b}{x}\quad\overset{\substack{y'=1/y\\x'=1/x}}{\rightsquigarrow}\quad y'=a+bx'$$

我们把这类经过变换后得到的线性模型称为广义线性模型，有关内容详见 [62]。

练习 9.1. 请读者验证下面的函数关系都能够通过变换成为线性的。

$$y = ax^b,\ \ y = a\exp(bx),\ \ y = a + b\ln x$$

$$y = \frac{1}{a + b\exp(-x)},\ \text{其中 } a, b > 0$$

$$y = b_0 + b_1 x + \cdots + b_n x^n,\ \ z = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2$$

9.1.1　最小二乘估计

既然回归模型是用来做预测或控制的，人们当然希望误差越小越好，即待定的参数 $\boldsymbol{\beta}$ 要使得误差平方和*$\sum_{i=1}^{n}\epsilon_i^2=\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon}=\|\boldsymbol{x}-A\boldsymbol{\beta}\|^2$ 达到最小，其中 $\boldsymbol{x}$ 为 $\boldsymbol{X}$ 的观察结果（即样本值）。在此标准之下，对未知参数 $\boldsymbol{\beta}$ 的估计就归结为一个最优化的问题——这种求估计值的方法被称为*最小二乘法*。最小二乘法深刻地影响了统计学的发展，曾是十九世纪的热点研究，它的重要性“犹如微积分之于数学”[10]。法国数学家 Legendre 在 1805 年发表的著作《计算彗星轨道的新方法》的附录中明确提出了最小二乘法，他与德国数学家 Gauss 之间曾有过最小二乘法优先权之争†。

最小二乘法的思想是朴素的。例如，对某物体长度的多次测量得到了观察结果 $x_1,x_2,\cdots,x_n$，测量误差分别为 $\epsilon_i=x_i-\theta,i=1,2,\cdots,n$，其中 θ 为真实长度。为了使 $\sum_{i=1}^{n}\epsilon_i^2=\sum_{i=1}^{n}(x_i-\theta)^2$ 达到最小，对未知参数 θ 的估计值是 $\hat{\theta}=\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，即样本值的算术平均。

定义 9.2. 对于线性回归模型 (9.5)，如果 $\forall\boldsymbol{\beta}\in\mathbb{R}^{k+1}$ 都有 $\|\boldsymbol{x}-A\hat{\boldsymbol{\beta}}\|^2\le\|\boldsymbol{x}-A\boldsymbol{\beta}\|^2$，则称 $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}(A,\boldsymbol{x})$ 为 $\boldsymbol{\beta}$ 的*最小二乘估计* (least square estimate, LSE)，有时也称统计量 $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}(A,\boldsymbol{X})$ 为 $\boldsymbol{\beta}$ 的最小二乘估计。

线性回归模型 (9.5) 中参数的最小二乘估计总是存在的（其存在性的证明和几何意义见下一节），该问题所要最小化的目标函数是一个关于 $\boldsymbol{\beta}$ 的函数：$\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon}=(\boldsymbol{x}-A\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{x}-A\boldsymbol{\beta})=\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}-\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathsf{T}}A^{\mathsf{T}}\boldsymbol{x}+\boldsymbol{\beta}^{\mathsf{T}}A^{\mathsf{T}}A\boldsymbol{\beta}$，根据附录 G 中定理 G.4，不难从 $\partial\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon}/\partial\boldsymbol{\beta}=0$ 得到所谓的*正则方程*：

*采用误差平方和比采用误差绝对值之和在理论推导和计算上更简捷些，所以习惯上用误差平方和来构造最优化问题中的目标函数。

†Gauss 声称他对最小二乘法的研究在先，鉴于他在数学界的权威，数学史把最小二乘法的优先权归功于 Gauss。另外，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。此外，为解出最小二乘估计，Gauss 还提出了线性方程组的“Gauss 消去法”。

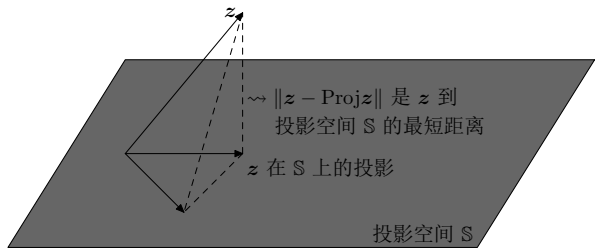
$A^\top A\boldsymbol{\beta} = A^\top \boldsymbol{x}$。若 $A^\top A$ 非奇异，则可解出 $\boldsymbol{\beta}$ 的最小二乘估计。

$$\hat{\boldsymbol{\beta}} = (A^\top A)^{-1} A^\top \boldsymbol{x} \tag{9.6}$$

进而得到回归方程 $\hat{\boldsymbol{x}} = A\hat{\boldsymbol{\beta}} = A(A^\top A)^{-1}A^\top \boldsymbol{x}$。

练习 9.2. 请读者验证残差的平方和为 $\|\boldsymbol{x} - A\hat{\boldsymbol{\beta}}\|^2 = \boldsymbol{x}^\top[I - A(A^\top A)^{-1}A^\top]\boldsymbol{x}$。

向量 $\boldsymbol{z} \in \mathbb{R}^n$ 在线性空间 $\mathbb{S} \subseteq \mathbb{R}^n$ 里的最佳近似是 $\boldsymbol{z}$ 在 $\mathbb{S}$ 上的投影向量 $\operatorname{Proj}\boldsymbol{z}$，其几何直观源自直和分解 $\boldsymbol{z} = \operatorname{Proj}\boldsymbol{z} \oplus (\boldsymbol{z} - \operatorname{Proj}\boldsymbol{z})$，其中向量 $(\boldsymbol{z} - \operatorname{Proj}\boldsymbol{z}) \perp \mathbb{S}$。该分解满足勾股定理 $\|\boldsymbol{z}\|^2 = \|\operatorname{Proj}\boldsymbol{z}\|^2 + \|\boldsymbol{z} - \operatorname{Proj}\boldsymbol{z}\|^2$。



定理 9.1. 线性回归模型 (9.5) 参数 $\boldsymbol{\beta}$ 的最小二乘估计总是存在的，且 $\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的最小二乘估计当且仅当 $\hat{\boldsymbol{\beta}}$ 满足正则方程。

证明. 由矩阵 A 的所有列向量张成的线性空间即为 $\mathbb{S} = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} = A\boldsymbol{\beta}$，其中 $\boldsymbol{\beta} \in \mathbb{R}^{k+1}\}$。令 $A\tilde{\boldsymbol{\beta}}$ 是向量 $\boldsymbol{x}$ 在线性空间 $\mathbb{S}$ 上的投影，显然 $\|\boldsymbol{x} - A\tilde{\boldsymbol{\beta}}\| \le \|\boldsymbol{x} - A\boldsymbol{\beta}\|$，即 $\tilde{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的最小二乘估计，存在性得证。

如果 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的最小二乘估计，则必有 $A\hat{\boldsymbol{\beta}} = \operatorname{Proj}\boldsymbol{x}$，如若不然，将导致 $\|\boldsymbol{x} - A\tilde{\boldsymbol{\beta}}\| < \|\boldsymbol{x} - A\hat{\boldsymbol{\beta}}\|$，矛盾！于是条件 $A\hat{\boldsymbol{\beta}} = \operatorname{Proj}\boldsymbol{x}$ 是 $\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的最小二乘估计的充要条件，它等价于 $\boldsymbol{x} - A\hat{\boldsymbol{\beta}} \perp \mathbb{S}$，即 $\boldsymbol{x} - A\hat{\boldsymbol{\beta}}$ 垂直于 A 的每个列向量，也就是说 $A^\top(\boldsymbol{x} - A\hat{\boldsymbol{\beta}}) = \boldsymbol{0}_{k+1}$，故 $\hat{\boldsymbol{\beta}}$ 满足正则方程，得证。 □

如果 A 是满秩的，线性回归模型 (9.5) 参数的最小二乘估计是唯一的。如果 A 不是满秩的，A 的列向量就是线性相关的，用这些列向量来线性表示 $\operatorname{Proj}\boldsymbol{x}$ 就可能不唯一，即最小二乘估计可能不唯一。对模型 (9.5) 中的未知参数 $\boldsymbol{\beta}$ 如果有约束条件，如 $\boldsymbol{\beta}^\top\boldsymbol{\beta} \le 2$ 等，$\boldsymbol{\beta}$ 的最小二乘估计可以通过 Lagrange 乘子法得到。

例 9.2. 接着考虑例 9.1 的一元线性回归模型 $\boldsymbol{x} = A\boldsymbol{\beta} + \boldsymbol{\epsilon}$，其中 $A^{\mathsf{T}} = \begin{pmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{pmatrix}, \boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathsf{T}}, \boldsymbol{x} = (x_1, \cdots, x_n)^{\mathsf{T}}$。为了最小化 $\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon} = \sum_{i=1}^n (x_i - \beta_0 - \beta_1 a_i)^2$，解得 $\hat{\beta}_0 = \overline{x} - \hat{\beta}_1 \overline{a}$ 且 $\hat{\beta}_1 = \sum_{i=1}^n (a_i - \overline{a})(x_i - \overline{x}) / \sum_{i=1}^n (x_i - \overline{x})^2$。也可以直接从式 (9.6) 求出 $\boldsymbol{\beta}$ 的最小二乘估计。

$$\begin{aligned} \hat{\boldsymbol{\beta}} = (A^{\mathsf{T}}A)^{-1}(A^{\mathsf{T}}\boldsymbol{x}) &= \begin{pmatrix} n & \sum_{i=1}^n a_i \\ \sum_{i=1}^n a_i & \sum_{i=1}^n a_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n a_i x_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n a_i^2 - (\sum_{i=1}^n a_i)^2} \begin{pmatrix} \sum_{i=1}^n a_i^2 \sum_{i=1}^n x_i - \sum_{i=1}^n a_i \sum_{i=1}^n a_i x_i \\ n\sum_{i=1}^n a_i x_i - \sum_{i=1}^n a_i \sum_{i=1}^n x_i \end{pmatrix} \end{aligned}$$

```
(%i1) M : matrix([n,sum(a[i],i,1,n)],[sum(a[i],i,1,n),sum((a[i])^2,i,1,n)]) $
(%i2) N : matrix([sum(x[i],i,1,n)],[sum(a[i]*x[i],i,1,n)]) $
(%i3) ratsimp(invert(M) . N) ;
```

定理 9.2. 对于线性回归模型 (9.5)，若 $A_{n\times(k+1)}$ 为满秩（即秩为 $k+1$），则最小二乘估计 $\hat{\boldsymbol{\beta}} = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}\boldsymbol{X}$ 和 $\hat{\sigma}^2 = \|\boldsymbol{X} - A\hat{\boldsymbol{\beta}}\|^2/(n-k-1)$ 分别是 $\boldsymbol{\beta}$ 和 σ^2 的无偏估计，并且 $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(A^{\mathsf{T}}A)^{-1})$ 以及 $(n-k-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-k-1}$。

证明. 若 $A_{n\times(k+1)}$ 为满秩，则 $A^{\mathsf{T}}A$ 正定，存在逆矩阵。

$$\begin{aligned} \mathsf{E}\hat{\boldsymbol{\beta}} &= (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}(\mathsf{E}\boldsymbol{X}) = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}A\boldsymbol{\beta} = \boldsymbol{\beta} \\ \mathsf{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) &= \mathsf{Cov}[(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}(A\boldsymbol{\beta} + \boldsymbol{\epsilon}), (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}(A\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &= \mathsf{Cov}[(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}\boldsymbol{\epsilon}, (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}\boldsymbol{\epsilon}] \\ &= (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}\mathsf{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})A(A^{\mathsf{T}}A)^{-1} = \sigma^2(A^{\mathsf{T}}A)^{-1} \end{aligned}$$

已知 $\boldsymbol{X}$ 服从正态分布，经过线性变换后 $\hat{\boldsymbol{\beta}} = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}\boldsymbol{X}$ 依然服从正态分布。由上述结果可得 $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(A^{\mathsf{T}}A)^{-1})$。令 $B = I - A(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$，

根据附录 G 中定理 G.2 有

$$\begin{aligned}\mathsf{E}\|\boldsymbol{X}-A\hat{\boldsymbol{\beta}}\|^2 &= \mathsf{E}(\boldsymbol{X}^{\mathsf{T}}B\boldsymbol{X}) = \mathsf{E}[\mathrm{tr}(\boldsymbol{X}^{\mathsf{T}}B\boldsymbol{X})] = \mathsf{E}[\mathrm{tr}(B\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})]\\ &= \mathrm{tr}[B\cdot\mathsf{E}(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})] = \mathrm{tr}[B\cdot(\sigma^2 I + A\boldsymbol{\beta}\boldsymbol{\beta}^{\mathsf{T}}A)]\\ &= \sigma^2\mathrm{tr}(B)，\text{因为 } BA = O_{n\times(k+1)}\\ &= \sigma^2\{\mathrm{tr}(I) - \mathrm{tr}[A(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}]\} = \sigma^2\{n - \mathrm{tr}[(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}A]\}\\ &= \sigma^2(n-k-1)\end{aligned}$$ □

定理 9.3 (Gauss-Markov)**.** 设 $\boldsymbol{c}\in\mathbb{R}^{k+1}$，在 $\boldsymbol{c}^{\mathsf{T}}\boldsymbol{\beta}$ 的所有线性无偏估计（即形为 $\boldsymbol{d}^{\mathsf{T}}\boldsymbol{X}$ 的无偏估计，其中 $\boldsymbol{d}\in\mathbb{R}^{k+1}$）当中，最小二乘估计 $\boldsymbol{c}^{\mathsf{T}}\hat{\boldsymbol{\beta}}$ 是方差最小的无偏估计。

注记 9.1. 稳健性 (robustness) 是衡量统计方法优劣的一个标准，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。最小二乘法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。

9.1.2　回归模型的假设检验

对线性回归模型 (9.5) 中未知参数 $\beta_0,\beta_1,\cdots,\beta_k$ 之间线性关系的假设都可以统一地表示为

$$H_0: H\boldsymbol{\beta}=0，\text{其中 } H_{r\times(k+1)} \text{ 是一个秩为 } r\le k+1 \text{ 的矩阵} \tag{9.7}$$

这类假设被称为线性假设 (linear hypothesis)，它陈述的是参数 $\beta_0,\beta_1,\cdots,\beta_k$ 满足 r 个独立的线性约束。譬如，对模型 (9.2)，零假设可以是 $H_0:\beta_1=0$，此时它等同于 $H_0: H\boldsymbol{\beta}=0$, 其中 $H=(0,1)$。

定理 9.4. 考虑线性回归模型 (9.5)，线性假设为 $H_0: H\boldsymbol{\beta}=\mathbf{0}$，其中 H 是一个秩为 $r\le k+1$ 的 $r\times(k+1)$ 矩阵。令 $\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}$ 分别是 $\boldsymbol{\beta}$ 在参数空间 Θ 和 Θ_0 里的最大似然估计，构造统计量

$$F=\frac{(\boldsymbol{X}-A\hat{\boldsymbol{\beta}})^{\mathsf{T}}(\boldsymbol{X}-A\hat{\boldsymbol{\beta}})-(\boldsymbol{X}-A\tilde{\boldsymbol{\beta}})^{\mathsf{T}}(\boldsymbol{X}-A\tilde{\boldsymbol{\beta}})}{(\boldsymbol{X}-A\hat{\boldsymbol{\beta}})^{\mathsf{T}}(\boldsymbol{X}-A\hat{\boldsymbol{\beta}})} \tag{9.8}$$

则有以下结果

$$\frac{n-k-1}{r}F \quad\sim\quad F(r,n-k-1) \tag{9.9}$$

在给定的显著水平 α 似然比检验拒绝零假设 H_0 的条件是

$$\frac{n-k-1}{r}F>F_{r,n-k-1,1-\alpha} \tag{9.10}$$

例 9.3. 下面我们考虑线性回归模型 $X_i=\beta_0+\beta_1 a_i+\epsilon_i, i=1,\cdots,n$，并假定 $\epsilon_1,\cdots,\epsilon_n \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$。该模型写成式 (9.5) 的形式，其中 $\boldsymbol{\beta}=(\beta_0,\beta_1)^{\mathsf{T}}, \boldsymbol{\epsilon}=(\epsilon_1,\cdots,\epsilon_n)^{\mathsf{T}}$，并且 $A=\begin{pmatrix}1 & 1 & \cdots & 1\\ a_1 & a_2 & \cdots & a_n\end{pmatrix}^{\mathsf{T}}$。

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 $\boldsymbol{X}=(X_1,\cdots,X_n)^{\mathsf{T}}$ 的

密度函数为

$$f(\boldsymbol{x};\beta_0,\beta_1,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\beta_0-\beta_1 a_i)^2\right\}$$

利用最大似然估计的方法，我们得到 Θ 上参数点估计的结果：

$$\begin{aligned}\hat{\beta}_0 &= \overline{X}-\hat{\beta}_1\overline{a}\\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(a_i-\overline{a})^2(X_i-\overline{X})}{\sum_{i=1}^{n}(a_i-\overline{a})^2}\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\beta}_0-\hat{\beta}_1 a_i)^2\end{aligned}$$

其中，$\overline{a}=\sum_{i=1}^{n}a_i/n$ 且 $\overline{X}=\sum_{i=1}^{n}X_i/n$。在零假设 H_0 成立时，参数在 Θ_0 上最大似然估计的结果是

$$\tilde{\beta}_0=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X},\qquad \tilde{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

根据式 (9.8)，构造统计量如下

$$\begin{aligned}F &= \frac{\sum_{i=1}^{n}(X_i-\overline{X}+\hat{\beta}_1\overline{a}-\hat{\beta}_1 a_i)^2-\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X}+\hat{\beta}_1\overline{a}-\hat{\beta}_1 a_i)^2}\\ &= \frac{\hat{\beta}_1^2\sum_{i=1}^{n}(a_i-\overline{a})^2}{\sum_{i=1}^{n}(X_i-\overline{X}+\hat{\beta}_1\overline{a}-\hat{\beta}_1 a_i)^2}\end{aligned}$$

根据式 (9.9)，我们有 $(n-2)F\sim F(1,n-2)$。由 t 分布的定义知：$\sqrt{(n-2)F}\sim t(n-2)$，这是因为

$$\frac{\chi_1^2}{\chi_{n-2}^2/(n-2)}\sim F(1,n-2)$$

如果 $\sqrt{(n-2)F}\geq t_{n-2,1-\alpha}$，则在水平 α 拒绝 H_0 假设。

例 9.4. 还是上例中的线性回归模型。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

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n a_i X_i}{\sum_{i=1}^n a_i^2} = \hat{\beta}_1 + \frac{n\hat{\beta}_0 \overline{a}}{\sum_{i=1}^n a_i^2}, \qquad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \tilde{\beta}_1 a_i)^2$$

根据式 (9.8)，构造统计量如下

$$\begin{aligned} F &= \frac{\sum_{i=1}^n (X_i - \overline{X} + \hat{\beta}_1 \overline{a} - \hat{\beta}_1 a_i)^2 - \sum_{i=1}^n (X_i - \tilde{\beta}_1 a_i)^2}{\sum_{i=1}^n (X_i - \overline{X} + \hat{\beta}_1 \overline{a} - \hat{\beta}_1 a_i)^2} \\ &= \frac{\hat{\beta}_0^2 n \sum_{i=1}^n (a_i - \overline{a})^2 / \sum_{i=1}^n a_i^2}{\sum_{i=1}^n (X_i - \overline{X} + \hat{\beta}_1 \overline{a} - \hat{\beta}_1 a_i)^2} \end{aligned}$$

与上例类似，如果 $\sqrt{(n-2)F} \geq t_{n-2,1-\alpha}$，则在水平 α 拒绝 H_0 假设。

9.1.3 置信区间与预测

9.2　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。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。
- 某人受教育的程度对其经济收入的影响。
- 土壤、肥料、日照时间等因素对某农作物产量的影响。
- 不同饲料对牲畜体重增长的效果。

影响指标的因素之间互相制约同时又互相依存，它们的共同作用决定着事物的表象— 指标值。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。这类试验的目的无外乎以下几点：

9.1. 通过数据分析找出显著影响指标的因素。

9.2. 对某个因素而言，哪个水平使得指标值最大或最小？

9.3. 各因素之间的相互作用：所有因素以什么样的水平搭配使得指标最优？

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*。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。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费用，我们不应简单地进行全面试验，而是要通过*试验设计*的方法选取一些有代表性的水平组合来进行试验。

Fisher 于 1918-1925 年间提出*方差分析* (analysis of variance, ANOVA) 的理论。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；*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。

9.2.1　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 个观察值 $X_{i1},\cdots,X_{in_i}$。考虑下面的线性模型

$$X_{ij}=\mu_i+\epsilon_{ij},\quad j=1,2,\cdots,n_i,\quad i=1,2,\cdots,k \tag{9.11}$$

或者简单地，

$$\boldsymbol{X}=A\boldsymbol{\mu}+\boldsymbol{\epsilon} \tag{9.12}$$

其中，随机向量 $\boldsymbol{X}=(X_{11},\cdots,X_{1n_1},X_{21},\cdots,X_{2n_2},\cdots,X_{k1},\cdots,X_{kn_k})^{\mathsf{T}}$ 满足 $\sum_{i=1}^{k}n_i=n$, 即共有 n 个观察样本。

$\boldsymbol{\mu}=(\mu_1,\cdots,\mu_k)^{\mathsf{T}},\boldsymbol{\epsilon}=(\epsilon_{11},\cdots,\epsilon_{1n_1},\epsilon_{21},\cdots,\epsilon_{2n_2},\cdots,\epsilon_{k1},\cdots,\epsilon_{kn_k})^{\mathsf{T}}$ 满足 $\epsilon_{ij}\overset{iid}{\sim}\mathrm{N}(0,\sigma^2)$，矩阵 $A_{n\times k}=\begin{pmatrix}\mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}\\ \vdots & \vdots & \cdots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_k}\end{pmatrix}^{\mathsf{T}}$，其中列向量 $\mathbf{1}_{n_1}=(\overbrace{1,1,\cdots,1}^{n_1\text{个}})^{\mathsf{T}}$。

如果我们认为不同水平对指标的影响没有差异，零假设可设置为 $H_0:\mu_1=\mu_2=\cdots=\mu_k$，或者等价地用线性假设的形式表示，$H_0:H\boldsymbol{\mu}=0$，其中矩阵 $H_{(k-1)\times k}$ 的秩为 $k-1$，具体为

$$H=\begin{pmatrix}1 & -1 & 0 & \cdots & 0\\ 1 & 0 & -1 & \cdots & 0\\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & -1\end{pmatrix}$$

样本的密度函数为

$$f(\boldsymbol{x};\boldsymbol{\mu},\sigma^2)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2\right\} \tag{9.13}$$

参数最大似然估计的结果为

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i}X_{ij}}{n_i}=\overline{X}_{i\cdot},\quad i=1,2,\cdots,k \tag{9.14}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2}{n} \tag{9.15}$$

在零假设 H_0 成立的情况之下，我们设 $\mu_1=\mu_2=\cdots=\mu_k=\mu$，则样本的密度函数 (9.13) 简化为

$$f(\boldsymbol{x};\boldsymbol{\mu},\sigma^2)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\mu)^2\right\} \tag{9.16}$$

参数最大似然估计的结果为

$$\tilde{\mu} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}}{n}=\overline{X} \tag{9.17}$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2}{n} \tag{9.18}$$

按照定理 9.4，我们构造统计量

$$F=\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2-\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2} \tag{9.19}$$

由于 $\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})=0$，我们对总的偏差平方和 (total sum of squares)

$\mathrm{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2$ 有如下的分解

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot}+\overline{X}_{i\cdot}-\overline{X})^2\\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2+\sum_{i=1}^{k}n_i(\overline{X}_{i\cdot}-\overline{X})^2\\ \text{简记作 TSS} &= \mathrm{WSS}+\mathrm{BSS}\end{aligned} \tag{9.20}$$

将式 (9.20) 代入式 (9.19)，我们得到

$$F=\frac{\mathrm{BSS}}{\mathrm{WSS}} \tag{9.21}$$

其中，$\mathrm{BSS}=\sum_{i=1}^{k}n_i(\overline{X}_{i\cdot}-\overline{X})^2$ 称为*组间偏差平方和* (between sum of squares)，$\mathrm{WSS}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2$ 称为*组内偏差平方和* (within sum of squares)。进一步由定理 9.4 可得

$$\begin{aligned}\frac{n-k}{k-1}F &= \frac{\mathrm{BSS}/(k-1)}{\mathrm{WSS}/(n-k)}=\frac{\sum_{i=1}^{k}n_i(\overline{X}_{i\cdot}-\overline{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2/(n-k)}\\ &\sim F(k-1,n-k)\end{aligned} \tag{9.22}$$

当 $\frac{n-k}{k-1}F \geq F_{k-1,n-k,1-\alpha}$ 时，似然比检验拒绝零假设 H_0。

来源	偏差平方和	自由度	平均偏差平方和
组间	$\mathrm{BSS}=\sum_{i=1}^{k}n_i(\overline{X}_{i\cdot}-\overline{X})^2$	$k-1$	$\mathrm{BSS}/(k-1)$
组内	$\mathrm{WSS}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2$	$n-k$	$\mathrm{WSS}/(n-k)$
总的	$\mathrm{TSS}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2$	$n-1$	$\mathrm{TSS}/(n-1)$

例 9.5. 我们考察三种牌子电池的寿命，零假设是三者之间没有差异。我们把三个牌子定为三个水平，下面是观察到的 $n=15$ 个样本，每个因素对应的样本数为 $n_1=5, n_2=4, n_3=6$。

	X_1	X_2	X_3
	40	60	60
	30	40	50
	50	55	70
	50	65	65
	30		75
			40
均值	40	55	60

$$\bar{x} = \frac{200 + 220 + 360}{15} = 52, \qquad \sum_{j=1}^{5}(x_{1j} - \bar{x}_{1\cdot})^2 = 400,$$

$$\sum_{j=1}^{4}(x_{2j} - \bar{x}_{2\cdot})^2 = 350, \qquad \sum_{j=1}^{6}(x_{3j} - \bar{x}_{3\cdot})^2 = 850$$

组内偏差平方和、组间偏差平方和分别为

$$\begin{aligned} \text{WSS} &= 400 + 350 + 850 = 1600, \\ \text{BSS} &= 5(40 - 52)^2 + 4(55 - 52)^2 + 6(60 - 52)^2 = 1140 \end{aligned}$$

偏差来源	偏差平方和	自由度	F-比
组间	1140	2	4.28
组内	1600	12	

因为 F-比 $> F_{2,12,0.05} = 3.89$，所以在水平 $\alpha = 0.05$ 拒绝零假设 $H_0 : \mu_1 = \mu_2 = \mu_3$。

9.2.2　两因素方差分析

在实际应用中，我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，我们把它抽象为线性模型

$$X_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \tag{9.23}$$
$$\text{其中} i = 1, \cdots, a, j = 1, \cdots, b$$

γ_{ij} 是 α_i 和 β_j 相互作用的结果。不失一般性，我们假定

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0 \tag{9.24}$$

我们感兴趣的假设，譬如

$$H_\gamma : \gamma_{ij} = 0, \forall i, j$$

为什么会有约束 (9.24) 呢？假设 $\mu', \alpha_i', \beta_j', \gamma_{ij}'$ 不满足 (9.24)，

$$X_{ij} = \mu' + \alpha_i' + \beta_j' + \gamma_{ij}' + \epsilon_{ij}$$

我们可以如下构造 $\mu, \alpha_i, \beta_j, \gamma_{ij}$ 使之满足 (9.24)：

$$\begin{aligned}
\mu &= \mu' + \overline{\alpha}' + \overline{\beta}' + \overline{\gamma}' \\
\alpha_i &= \alpha_i' - \overline{\alpha}' + \overline{\gamma}_{i\cdot}' - \overline{\gamma}' \\
\beta_j &= \beta_j' - \overline{\beta}' + \overline{\gamma}_{\cdot j}' - \overline{\gamma}' \\
\gamma_{ij} &= \gamma_{ij}' - \overline{\gamma}_{i\cdot}' - \overline{\gamma}_{\cdot j}' + \overline{\gamma}'
\end{aligned}$$

其中

$$\overline{\alpha}' = \frac{1}{a}\sum_{i=1}^{a}\alpha_i', \quad \overline{\beta}' = \tfrac{1}{b}\textstyle\sum_{j=1}^{b}\beta_j' \quad \text{并且}$$

$$\overline{\gamma}_{i\cdot}' = \frac{1}{b}\sum_{j=1}^{b}\gamma_{ij}', \quad \overline{\gamma}_{\cdot j}' = \tfrac{1}{a}\textstyle\sum_{i=1}^{a}\gamma_{ij}', \quad \overline{\gamma}' = \dfrac{1}{ab}\displaystyle\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}'$$

我们把观察样本整理成如下列表。

因素 1	因素 2 的水平				
的水平	1	2	$\cdots$	b	均值
1	X_{111}	X_{121}	$\cdots$	X_{1b1}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
1	X_{11m}	X_{12m}	$\cdots$	X_{1bm}	$\overline{X}_{1\cdot\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	X_{a11}	X_{a21}	$\cdots$	X_{ab1}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
a	X_{a1m}	X_{a2m}	$\cdots$	X_{abm}	$\overline{X}_{a\cdot\cdot}$
均值	$\overline{X}_{\cdot 1\cdot}$	$\overline{X}_{\cdot 2\cdot}$	$\cdots$	$\overline{X}_{\cdot b\cdot}$	$\overline{X}$

均值与偏差平方和

我们考虑线性模型

$$X_{ijs} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijs}, \quad \epsilon_{ijs} \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2) \tag{9.25}$$

其中 $i = 1, \cdots, a, \quad j = 1, \cdots, b, \quad s = 1, \cdots, m$

满足

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = \sum_{i=1}^{a}\gamma_{ij} = \sum_{j=1}^{b}\gamma_{ij} = 0 \tag{9.26}$$

定义 9.3. 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：

$$\overline{X} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{s=1}^{m}X_{ijs}}{n} \tag{9.27}$$

$$\overline{X}_{ij\cdot} = \frac{\sum_{s=1}^{m}X_{ijs}}{m} \tag{9.28}$$

$$\overline{X}_{i\cdot\cdot} = \frac{\sum_{j=1}^{b}\sum_{s=1}^{m}X_{ijs}}{mb} \tag{9.29}$$

$$\overline{X}_{\cdot j\cdot} = \frac{\sum_{i=1}^{a}\sum_{s=1}^{m}X_{ijs}}{ma} \tag{9.30}$$

定义 9.4. 下面定义的偏差平方和分别是：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。

$$\mathrm{SS}_1 = bm\sum_{i}(\overline{X}_{i\cdot\cdot}-\overline{X})^2 \tag{9.31}$$

$$\mathrm{SS}_2 = am\sum_{j}(\overline{X}_{\cdot j\cdot}-\overline{X})^2 \tag{9.32}$$

$$\mathrm{SSI} = m\sum_{i,j}(\overline{X}_{ij\cdot}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot j\cdot}+\overline{X})^2 \tag{9.33}$$

$$\mathrm{SSE} = \sum_{i,j,s}(\overline{X}_{ijs}-\overline{X}_{ij\cdot})^2 \tag{9.34}$$

两因素方差分析的假设检验

9.1. 对模型 (9.25)，我们考虑零假设 $H_\alpha : \alpha_1 = \cdots = \alpha_a = 0$，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：

$$\hat{\mu} = \tilde{\mu} = \overline{X},\quad \hat{\alpha}_i = \overline{X}_{i\cdot\cdot}-\overline{X},\quad \hat{\beta}_j = \tilde{\beta}_j = \overline{X}_{\cdot j\cdot}-\overline{X} \tag{9.35}$$

$$\hat{\gamma}_{ij} = \tilde{\gamma}_{ij} = \overline{X}_{ij\cdot}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot j\cdot}+\overline{X} \tag{9.36}$$

构造统计量如下

$$\begin{aligned} F &= \frac{\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{ij\cdot}+\overline{X}_{i\cdot\cdot}-\overline{X})^2-\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{ij\cdot})^2}{\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{ij\cdot})^2} \\ &= \frac{bm\sum\limits_{i}(\overline{X}_{i\cdot\cdot}-\overline{X})^2}{\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{ij\cdot})^2}=\frac{\mathrm{SS}_1}{\mathrm{SSE}} \end{aligned} \tag{9.37}$$

由式 (9.8) 可得

$$\frac{ab(m-1)}{a-1}F \sim F[a-1,ab(m-1)] \tag{9.38}$$

9.2. 对模型 (9.25)，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：

$$\tilde{\mu}=\overline{X},\quad \tilde{\alpha}_i=\overline{X}_{i\cdot\cdot}-\overline{X},\quad \tilde{\beta}_j=\overline{X}_{\cdot j\cdot}-\overline{X} \tag{9.39}$$

构造统计量如下

$$\begin{aligned} F &= \frac{\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot j\cdot}+\overline{X})^2-\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{ij\cdot})^2}{\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{ij\cdot})^2} \\ &= \frac{m\sum\limits_{i,j,s}(\overline{X}_{ij\cdot}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot j\cdot}+\overline{X})^2}{\sum\limits_{i,j,s}(X_{ijs}-\overline{X}_{ij\cdot})^2}=\frac{\mathrm{SSI}}{\mathrm{SSE}} \end{aligned} \tag{9.40}$$

由式 (9.8) 可得

$$\frac{ab(m-1)}{(a-1)(b-1)}F \sim F[a-1,ab(m-1)] \tag{9.41}$$

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	$ab/(m-1)$	$MSE = SSE/[ab(m-1)]$	

例 9.6. 下面的例子选自 [78] pp528-529：请三位老师分别使用三种不同教学方法，学生的成绩作为指标，我们得到一些观察样本如下：

教学	教师		
方法	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

求得几类均值如下：

		$\overline{X}_{ij\cdot}$		$\overline{X}_{i\cdot\cdot}$
	82	75	77	78.0
	86	89	75	83.3
	80	78	85	81.0
$\overline{X}_{\cdot j\cdot}$	82.7	80.7	79.0	$\overline{X} = 80.8$

$$\mathrm{SS}_1 = bm\sum_{i=1}^{a}(\overline{X}_{i\cdot\cdot} - \overline{X})^2 = 3 \times 4 \times 14.13 = 169.56$$

$$\mathrm{SS}_2 = am\sum_{j=1}^{b}(\overline{X}_{\cdot j\cdot} - \overline{X})^2 = 82.32, \mathrm{SSI} = 561.80, \mathrm{SSE} = 1830.00$$

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。

9.3　习题

9.1. 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 $\epsilon_1, \epsilon_2, \epsilon_3 \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$。试求：(1) β_1 和 β_2 的最小二乘法估计 $\hat{\beta}_1, \hat{\beta}_2$；(2) 给出 $\hat{\beta}_1, \hat{\beta}_2$ 的分布并证明 $\hat{\beta}_1, \hat{\beta}_2$ 相互独立。

9.2. 设有一元线性回归模型 $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$，$i = 1, 2, \cdots, n$，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。

9.3. 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

9.4. 设线性模型：$y_i = a + bx_i + \varepsilon_i,\ i = 1, 2, \cdots, n$, 其中各 ε_i 相互独立，且均 $\sim \mathrm{N}(0, \sigma^2)$. 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。

9.5. 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。

9.6. 一工厂用三种不同的工艺生产某类型电池。从各种工艺生产的电池中分别抽取样本并测得样本的寿命（使用时间）如下：（单位：小时）

工艺 1：40, 46, 38, 42, 44

工艺 2：26, 34, 30, 28, 32

工艺 3：39, 40, 43, 48, 50

取 $\alpha = 0.05$，对此进行方差分析，检验三种工艺的电池寿命有无显著差异。

9.7. 在上题中，已拒绝了原假设 H_0，现要对三种工艺作比较（μ 越大越好）。由于 $\overline{X_3} > \overline{X_1} > \overline{X_2}$，故我们认为 $\mu_3 > \mu_1 > \mu_2$，即第三种工艺最优，第一种次之，第二中最差。求第二种和第三种工艺所生产的电池平均寿命的差别？并求出区间估计（$\alpha = 0.05$）。

9.8. 为了测试新试制的汽船的性能，在规定里程内测量三种不同的风浪条件下汽船航行的时间（单位：分钟）。数据如下（都符合正态分布）：

无风浪：26, 19, 16, 22

风稍有浪：25, 27, 25, 20, 18, 23

大风大浪：23, 25, 28, 31, 16

用这些数据检验航行条件对航行时间有无影响？（$\alpha = 0.05$）

9.9. 有两种品牌的软饮料拟在三个地区进行销售，为了分析饮料的品牌（“品牌”因素）和销售地区（“地区”因素）对销售量的影响，对每种品牌在三个地区的销售量取得以下数据：

品牌 1：558, 627, 484

品牌 2；464, 528, 616

试分析品牌和销售地区对饮料的销售量是否有显著影响？（取$\alpha = 0.05$）

9.10. 证明

$$\sum_{j,k}(x_{jk} - \overline{x})^2 = \sum_{j,k}(x_{jk} - \overline{x}_{j.})^2 + \sum_{j,k}(\overline{x}_{j.} - \overline{x})^2$$

9.11. 希望确定汽油 A, B, C, D 间在每加仑行驶里数上是否存在差异。使用 4 个不同的驾驶员，4 辆汽车和 4 条不同的道路进行实验。设计该试验。

9.12. 描述如何进行三方式分组或三因素试验的方差分析技术（每一试验有单一数据）。

第十章

非参数统计学简介

10.1 次序统计量

第十一章

统计决策与贝叶斯分析概要

贝叶斯学派认为概率描述的是信念度，而非大量重复性试验中随机事件出现频率的极限，所以概率反映的是个体对不确定性的主观认识，不是随机现象本身固有的属性。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。频率派不承认先验知识，只会根据历史数据做推断。为了臆造出可重复的随机试验，频率派不得不把“相似性”这一主观认识强加给数据。在日常生活中常会遇到无据可查的窘况，即便是频率派最顽固的坚信者也不会反对用 0,1 之间的某个数字表达一下自己的遗憾程度，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？所有科学研究的目标都是为了揭示客观规律，在向客观目标前进的过程中，主观选择了路线，离开了主观认识连规律的内容都无法交待清楚。

11.1 先验分布

11.1.1 无信息先验

Let $\boldsymbol{X}$ be a random vector and Y any rv, for any function $g(\boldsymbol{x})$,

$$\begin{aligned} \mathsf{E}\left\{[Y-g(\boldsymbol{X})]^2|\boldsymbol{X}=\boldsymbol{x}\right\} &= \mathsf{E}\left\{[Y-g(\boldsymbol{x})]^2|\boldsymbol{X}=\boldsymbol{x}\right\} \\ &\geq \mathsf{E}\left\{[Y-\mathsf{E}(Y|\boldsymbol{X}=\boldsymbol{x})]^2|\boldsymbol{X}=\boldsymbol{x}\right\} \end{aligned} \tag{11.1}$$

By the fact that the equality in $V(X) \leq \mathsf{E}(X-c)^2$ holds iff $c = \mathsf{E}(X)$. Therefore,

$$\mathsf{E}\left\{[Y-g(\boldsymbol{X})]^2|\boldsymbol{X}\right\} \geq \mathsf{E}\left\{[Y-\mathsf{E}(Y|\boldsymbol{X})]^2|\boldsymbol{X}\right\} \tag{11.2}$$

$$\text{we have, } \mathsf{E}[Y-\mathsf{E}(Y|\boldsymbol{X})]^2 \leq \mathsf{E}[Y-g(\boldsymbol{X})]^2 \tag{11.3}$$

where the equality holds iff $g(\boldsymbol{X}) = \mathsf{E}(Y|\boldsymbol{X})$. We call $\mathsf{E}(Y|\boldsymbol{X})$ the **mean square error** (MSE) predictor of Y, given $\boldsymbol{X}$.

Examples of unbiasness Let $X_1, X_2, \cdots, X_n$ be random sample from a population of X.

11.1. If $\mathsf{E}(X) < \infty$, $\overline{X}$ is unbiased for $\mathsf{E}(X)$.

11.2. If $V(X) < \infty$, S^2 is unbiased for $V(X)$.

11.3. If $m_k = \mathsf{E}(X^k)$ exists, a_k is unbiased for m_k.

Unbiased vs biased: which one is better?

From UMVUE to most efficient estimate

Theorems about UMVUE The unbiased estimates with finite variances are of concern.

11.1. There exists at most one UMVUE for θ.

11.2. If UMVUE exists, say $T = T(X_1, X_2, \cdots, X_n)$, it must be a symmetric function of $X_1, X_2, \cdots, X_n$.

11.3. (Rao-Blackwell Theorem, 1949) Let X, Y be random variables with $\mathsf{E}(Y) = \mu, V(Y) = \sigma_Y^2$. Let $\mathsf{E}(Y|x) = \eta(x)$, then

$$\mathsf{E}[\eta(X)] = \mu, \quad \sigma_{\eta(X)}^2 \le \sigma_Y^2 \tag{11.4}$$

Proof. $\mathsf{E}[\eta(X)] = \mathsf{E}[\mathsf{E}(Y|X)] = \mathsf{E}(Y) = \mu$ and $\eta(X)$ is the MSE predictor of Y.

定理 11.1 (推广了的 Rao-Blackwell 定理)**.** 如果损失函数 L 是凸函数，则

$$\mathsf{E}[L(\delta_1(X))] \le \mathsf{E}[L(\delta(X))] \tag{11.5}$$

其中 $\delta_1(X) = \mathsf{E}[\delta(X)|T(X)]$，$T(X)$ 为充分统计量。

11.2 后验分布

第三部分

概率统计中的一些实用算法

在本书的第三部分，我们将介绍在实际应用中常用的一些算法。

第十二章

Markov 链和隐 Markov 模型

12.1 Markov 链

12.1.1 随机过程简介

Stochastic process

定义 12.1 (Stochastic Process)**.** A stochastic process is a collection of random variables $\{X(t)|t \in T\}$. Typically, there are two cases.

12.1. T is continuous: $\{X(t)|t \geq 0\}$.

12.2. T is discrete: $\{X_n|n = 0, 1, 2, \cdots\}$.

The set of all possible values of $X(t)$ is called the state space, denoted by $\mathbb{S}$. Each element in $\mathbb{S}$ is called a state.

➠Further reading: Rosenblatt, M. (1974) Random Processes, Spring-Verlag New York Inc.

Homogeneous Markov chains A Markov chain is a stochastic process $\{X_n|n = 0, 1, 2, \cdots\}$ with continuous or discrete $\mathbb{S}$, satisfying the following

Markov property.

$$\begin{aligned} & P(X_n = y|X_{n-1} = x) \\ = \quad & P(X_n = y|X_{n-1} = x, X_{n-2} = *, \cdots, X_0 = *) \end{aligned} \tag{12.1}$$

where $x, y \in \mathbb{S}$. A Markov chain is called time homogeneous if the transition probability $P(X_n = y|X_{n-1} = x)$ satisfies that

$$P(X_n = y|X_{n-1} = x) = A(x, y) \tag{12.2}$$

$$\text{where } \sum_{y\in\mathbb{S}} A(x, y) = 1 \text{ for all } x \in \mathbb{S} \tag{12.3}$$

is called the transition function.

12.1.2 转换矩阵与转换函数

Matrix of transition probabilities ✂We will consider homogeneous Markov chains from now on.

If the state space is countable, the transition function can be represented by the matrix of transition probabilities.

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{12.4}$$

where $p_{ij} = P(X_n = j|X_{n-1} = i) = A(i, j)$. By (12.3), it is obvious that

$$\sum_j p_{ij} = 1 \tag{12.5}$$

Examples of finite Markov chains

12.1. Bernoulli trials: there are two states $\mathbb{S} = \{1, 2\}$ with transition matrix $P = \begin{pmatrix} p & 1-p \\ p & 1-p \end{pmatrix}$.

12.2. The random walk of particle: $p_{11} = p_{ss} = 1$ (i.e., states $1, s$ are absorbing) and for $2 \le i \le s-1$

$$p_{ij} = \begin{cases} p & \text{when } j = i+1 \\ 1-p & \text{when } j = i-1 \\ 0 & \text{otherwise} \end{cases} \tag{12.6}$$

Markov model of Mendel genetics

Gene types: AA, Aa, aa

Possible states: For N individuals in a generation, there are $2N$ genes. There are $2N+1$ states for the number of gene A — $0, 1, \cdots, 2N$.

Inheritance: The new generation inherits genes from parents in Bernoulli scheme.

Transition probabilities: From state i to state j,

$$p_{ij} = C_{2N}^{j} \left(\frac{i}{2N}\right)^{j} \left(1 - \frac{i}{2N}\right)^{2N-j}$$

Obviously, states $0, 2N$ are absorbing.

Chapman-Kolmogorov equation

n-step transition probabilities Let $p_{ij}^{(n)}$ denote the probability from state i to state j after n transitions. The matrix $P_n = \left(p_{ij}^{(n)}\right)$ is called the matrix of

n-step transition probabilities. Obviously,

$$p_{ij}^{(n)} = \sum_{k \in \mathbb{S}} p_{ik}^{(m)} p_{kj}^{(n-m)} \tag{12.7}$$

$$\text{or equivalently, } P_n = P_m P_{n-m} \tag{12.8}$$

(12.8) is called Chapman-Kolmogorov equation, where $1 \leq m < n$. It is easy to verify that

$$P_n = P^n \tag{12.9}$$

n-step transition function Let $A^{(n)}(x, y)$ denote the *n*-step transition function, defined by

$$A^{(n)}(x, y) = P(X_n = y | X_0 = x) \tag{12.10}$$

where $A^{(1)}(x, y) = A(x, y)$.

The Chapman-Kolmogorov equation is

$$A^{(n)}(x, y) = \int_{\mathbb{S}} A^{(m)}(x, z) A^{(n-m)}(z, y) dz \tag{12.11}$$

where $1 \leq m < n$.

12.1.3 遍历定理

Irreducible Markov chain

Irreducible Markov chain State j is accessible from state i, denoted by

$i \to j$, if $\exists m \geq 0$ such that $p_{ij}^{(m)} > 0$, where

$$p_{ij}^{(0)} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{12.12}$$

States i and j communicate, denoted by $i \leftrightarrow j$, if $i \to j$ and $j \to i$. Communication is an equivalence relation, since

12.1. $i \leftrightarrow i$

12.2. $i \leftrightarrow j \Rightarrow j \leftrightarrow i$

12.3. $i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k$

The state space $\mathbb{S}$ is divided into mutually exclusive classes by "$\leftrightarrow$". A Markov chain is irreducible if $\forall i, j \in \mathbb{S}, i \leftrightarrow j$.

Illustration of periodicity of states State i has period $d = d(i)$ if

12.1. $p_{ii}^{(n)} > 0$ only for values of n divisible by d (i.e., $d|n$).

12.2. d is the largest number satisfying (1).

We set $d(i) = 0$ if $p_{ii}^{(n)} = 0$ for all $n \geq 1$ — "gone for ever".
State j is called aperiodic if $d(j) = 1$.

$$P = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix}$$

States $1, 2$ are aperiodic, and $d(3) = 2$.

Classification of states: recurrence vs transience

Recurrence vs transience

$$f_{ij}^{(k)} = P(X_0 = i, X_k = j, X_s \neq j, 1 \leq s \leq k-1) \tag{12.13}$$

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} \tag{12.14}$$

$$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)} \text{ is the probability of returning back to } i \tag{12.15}$$

$$\text{State } i \text{ is } \begin{cases} \text{recurrent} & \text{if } f_{ii} = 1 \begin{cases} \text{positive} & \text{if } \sum_{n=1}^{\infty} n f_{ii}^{(n)} < \infty \\ \text{null} & \text{if } \sum_{n=1}^{\infty} n f_{ii}^{(n)} = \infty \end{cases} \\ \text{transient} & \text{if } f_{ii} < 1 \end{cases}$$

Illustration of recurrence

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{pmatrix}_{s \times s}$$

- For $2 \leq i \leq s-1, i \to 1$ but $1 \nrightarrow i$.
- $1 \leftrightarrow s$.
- States $1, s$ are recurrent, but not periodic.

Properties of recurrence and transience

12.1. State i is

recurrent: iff $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$. Starting in state i, the process will reenter i infinitely often. Especially when $p_{ii} = 1$, state i is called absorbing.

transient: iff $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$.

12.2. If j is recurrent (transient) and $i \leftrightarrow j$, then i is recurrent (transient).

12.3. If j is transient, then $p_{ij}^{(n)} \to 0$ as $n \to \infty$ for all i.

12.4. Not all states in a finite Markov chain can be transient.

12.5. All states of a finite irreducible Markov chain are recurrent.

Proof of properties

❑ With $p_{ii}^{(0)} = 1$,

$$\begin{aligned}\sum_{n=1}^{\infty} p_{ii}^{(n)} &= \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \sum_{n=k}^{\infty} p_{ii}^{(n-k)} \\ &= f_{ii}\sum_{n=0}^{\infty} p_{ii}^{(n)} = f_{ii}\left(1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}\right)\end{aligned}$$

❑ Let $p_{ij}^{(a)} > 0, p_{ji}^{(b)} > 0$, we have

$$p_{ii}^{(n+a+b)} \geq p_{ij}^{(a)} p_{jj}^{(n)} p_{ji}^{(b)}$$

If $\sum_{n=1}^{\infty} p_{jj}^{(n)} = \infty$, then $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$.

Continued: proof of properties

❑ Similar to the first proof.

$$\begin{aligned}\sum_{n=1}^{\infty} p_{ij}^{(n)} &= \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=k}^{\infty} p_{jj}^{(n-k)} \\ &= f_{ij}\sum_{n=0}^{\infty} p_{jj}^{(n)} \le \sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty\end{aligned}$$

By the fact of $f_{ij} \le 1$.

❑ Homework: Prove the left properties.

Ergodicity and ergodic theorems for Markov chains

Ergodic theorems for Markov chains A state i is said to be ergodic if it is aperiodic and positive recurrent. If all states in a Markov chain are ergodic, then the chain is said to be ergodic.

❑ If a Markov chain is ergodic, then $\forall i, j$

$$\lim_{n\to\infty} p_{ij}^{(n)} = p_j \tag{12.16}$$

It means that, as $n \to \infty$, the probability of reaching at state j is independent of the starting state.

❑ A finite Markov chain (with s states) is ergodic if there exists $n \in \mathbb{N}$ such that $\forall i, j = 1, 2, \cdots, s$

$$p_{ij}^{(n)} > 0 \tag{12.17}$$

Stationary distribution Moreover, $\{p_j\}$ in (12.16) is the unique solution

of the equations

$$p_j = \sum_{i=1}^{s} p_i p_{ij} \tag{12.18}$$

$$\text{or equivalently, } \pi = P^{\mathsf{T}}\pi \tag{12.19}$$

where $\pi(X = j) = p_j$ is the stationary distribution of states, satisfying $p_j > 0$ and $\sum_{j=1}^{s} p_j = 1$.

In other words, the stationary distribution $\pi = (p_1, \cdots, p_s)^{\mathsf{T}}$ is a normalized eigenvector of P^{T} associated with the eigenvalue 1. We call π invariant under P^{T}.

Reversible Markov chains If the time is reversed, we have

$$P(X_n = i|X_{n+1} = j) = \frac{P(X_n = i)P(X_{n+1} = j|X_n = i)}{P(X_{n+1} = j)}$$

A finite Markov chain is called reversible, if it satisfies the detailed balance condition as follows.

$$p_i p_{ij} = p_j p_{ji} \tag{12.20}$$

Moreover, we have the property of invariance.

$$\sum_{i=1}^{s} p_i p_{ij} = \sum_{i=1}^{s} p_j p_{ji} = p_j \sum_{i=1}^{s} p_{ji} = p_j$$

Reversible Markov chains Generally, a Markov chain is called reversible with respect to $\pi(x)$, if it satisfies the detailed balance condition as follows.

$$\pi(x)A(x, y) = \pi(y)A(y, x) \tag{12.21}$$

Similarly, we have the property of invariance.

$$\begin{aligned}\int \pi(x)A(x,y)dx &= \int \pi(y)A(y,x)dx \\ &= \pi(y)\int A(y,x)dx = \pi(y)\end{aligned}$$

✍The detailed balance ensures the $\pi(x)$ as the stationary distribution, but not vice versa.

Further readings

12.1. Kemeny, J. G. and Snell, J. L. (1976) Finite Markov Chains, Spring-Verlag New York Inc.

12.2. Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4): 711-732.

12.3. Liu, J.S. (2001) Monte Carlo Strategies in Scientific Computing, Spring-Verlag New York Inc.

12.4. Robert, C. P. and Casella, G. (2004) Monte Carlo Statistical Methods (second edition), Spring-Verlag New York Inc.

12.5. Rubinstein, R. Y. and Kroese, D. P. (2007) Simulation and the Monte Carlo Method (second edition), John Wiley & Sons.

12.2 隐 Markov 模型及其算法

Urn model analogy of hidden Markov model

12.1. Props: n urns labeled with $1, 2, \cdots, n$ and in each urn there are balls with m distinct colors.

12.2. Game rule: God chooses an urn, picks up a ball randomly and tells the blind Abing its color, then puts it back. God repeats the process T times.

12.3. Game: Abing is requested to give the sequence of urns that God chooses.

For instance, suppose that an urn is a part-of-speech (POS) and a color is a word, then from color sequence to an urn sequence is nothing but from a word sequence to a POS sequence.

Hidden Markov model (HMM)

12.1. Finite Urns: $U = \{1, 2, \cdots, n\}$

12.2. Transition Matrix: $A = (a_{ij})_{n\times n}$

12.3. Initial Distribution: $\boldsymbol{\pi} = (\pi_i)_{n\times 1}$

12.4. Colors: $C = \{c_1, c_2, \cdots, c_m\}$

12.5. Observation of Color Sequence: $O = o_1 o_2 \cdots o_T$

12.6. Color Distribution: $B = (b_{ik})_{n\times m}$ satisfying that $b_{ik} = P(O_t = c_k | X_t = i)$, denoted by $b_i(k)$, and $\sum_{k=1}^{m} b_i(k) = 1$, where X_t is the state and O_t is the observation at time t.

Hidden Markov model (HMM) is $\mathcal{M} = \langle U, C, A, B, \boldsymbol{\pi}\rangle$, in which $\Theta = \langle A, B, \boldsymbol{\pi}\rangle$ is called the parameter set of HMM.

Three problems of hidden Markov model

The probability of observations: Given Θ and O, $P(O|\Theta)$ =?

The hidden sequence of states: Given Θ and O, finding the urn sequence X which maximizes $P(X|O, \Theta)$.

Parameter training: Given O, training $\Theta_{best} = \underset{\Theta}{\operatorname{argmax}}\, P(O|\Theta)$.

Correspondingly, we will introduce the following algorithms.

12.1. Forward-backward algorithm

12.2. Viterbi algorithm

12.3. Baum-Welch algorithm

12.2.1 向前-向后算法与 Viterbi 算法

A cursory solution to problem 1

12.1. Let $S = s_1 s_2 \cdots s_T$ be a state sequence;

12.2. Suppose that the observations are independent:
$P(O|S,\Theta) = \prod_{i=1}^{T} P(o_t|s_t,\Theta) = b_{s_1}(o_1)b_{s_2}(o_2)\cdots b_{s_T}(o_T)$

12.3. $P(S|\Theta) = \pi_{s_1} a_{s_1 s_2} a_{s_2 s_3} \cdots a_{s_{T-1} s_T}$

12.4. $P(O|\Theta) = \sum_{S} P(O|S,\Theta)P(S|\Theta)$

The complexity of the upper cursory algorithm is $O(Tn^T)$. We can improve it by dynamic programming.

Forward algorithm by dynamic programming Forward variable is defined by

$$\alpha_t(i) = P(o_1 o_2 \cdots o_t, X_t = i|\Theta) \tag{12.22}$$

Forward Algorithm

12.1. initialization: $\alpha_1(i) = \pi_i b_i(o_1), 1 \le i \le n$, the probability of getting o_1 from the i-th urn.

12.2. recursion: $\alpha_{t+1}(j) = \left[\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right] b_j(o_{t+1})$, the probability of observing $o_1 o_2 \cdots o_t$ at time t with next urn j.

12.3. result: $P(O|\Theta) = \sum_{i=1}^{n} \alpha_T(i)$, the probability of observing $o_1 o_2 \cdots o_T$ at time T.

Complexity of forward algorithm

12.1. The complexity of $\alpha_t(j)$ is $O(n)$, since there are at most n transitions at time $t-1$.

$$\alpha_t(j) = \left[\sum_{i=1}^{n} \alpha_{t-1}(i) a_{ij}\right] b_j(o_t)$$

12.2. For a fixed $t, \{\alpha_t(i) : 1 \le i \le n\}$ costs $O(n^2)$ steps.

12.3. Because $t = 1, 2, \cdots, T$, the complexity of forward algorithm is $O(Tn^2)$.

$$P(O|\Theta) = \sum_{i=1}^{n} \alpha_T(i)$$

Backward algorithm Backward variable is defined by

$$\beta_t(i) = P(o_{t+1} o_{t+2} \cdots o_T | X_t = i, \Theta) \tag{12.23}$$

Backward Algorithm

12.1. initialization: $\beta_T(i) = 1, 1 \le i \le n$

12.2. recursion: $\beta_t(i) = \sum_{j=1}^{n} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$, where $1 \le i \le n$ and $t = T - 1, \cdots, 1$

12.3. result: $P(O|\Theta) = \sum_{i=1}^{n} \pi_i \beta_1(i)$

Forward-backward algorithm

$$\begin{aligned} P(O|\Theta) &= \sum_{i=1}^{n} P(o_1 \cdots o_T, X_t = i|\Theta) \\ &= \sum_{i=1}^{n} P(o_1 \cdots o_{t-1}, X_t = i, o_t \cdots o_T|\Theta) \\ &= \sum_{i=1}^{n} P(o_1 \cdots o_{t-1}, X_t = i|\Theta) P(o_t \cdots o_T|X_t = i, \Theta) \\ &= \sum_{i=1}^{n} \alpha_t(i)\beta_t(i) \text{ where } 1 \le t \le T \end{aligned} \tag{12.24}$$

Viterbi variable The task of Problem 2 is to find X satisfying that

$$\hat{X} = \underset{X}{\operatorname{argmax}}\, P(X, O|\Theta) = \underset{X}{\operatorname{argmax}}\, P(X|O, \Theta) \tag{12.25}$$

Viterbi variable:

$$\delta_t(i) = \max_{X_1, \cdots, X_{t-1}} P(X_1, \cdots, X_{t-1}, X_t = i, o_1 \cdots o_t|\Theta) \tag{12.26}$$

We have the following recursion.

$$\delta_{t+1}(j) = \left[\max_{1 \le i \le n} \delta_t(i) a_{ij}\right] b_j(o_{t+1}) \tag{12.27}$$

Let $\Delta_t(i)$ denote the former state of i, at time $t - 1$, in the path which has probability $\delta_t(i)$.

Viterbi algorithm by dynamic programming

$$
\begin{aligned}
\text{initialization}: \quad & \delta_1(i) = \pi_i b_i(o_1) && 1 \le i \le n \\
& \Delta_1(i) = 0 \\
\text{recursion}: \quad & \delta_t(j) = \left[\max_{1\le i\le n} \delta_{t-1}(i) a_{ij}\right] b_j(o_t) && 2 \le t \le T, 1 \le j \le n \\
& \Delta_t(j) = \operatorname*{argmax}_{1\le i\le n} \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t) \\
\text{calculation}: \quad & P(\hat{X}, O|\Theta) = \max_{1\le i\le n} [\delta_T(i)] \\
\text{traceback}: \quad & \hat{X}_T = \operatorname*{argmax}_{1\le i\le n} [\delta_T(i)] \\
& \hat{X}_t = \Delta_{t+1}(\hat{X}_{t+1}) && t = T-1, \cdots, 1
\end{aligned}
$$

The complexity of Viterbi Algorithm is $O(Tn^2)$.

12.2.2　模型参数的训练：Baum-Welch 算法

Training of parameters ☞The third problem is to optimize the parameters $\Theta = \langle A, B, \boldsymbol{\pi} \rangle$ such that $P(O|\Theta)$ maximal. Details can be found in the slides of introducing Expectation Maximization (EM) algorithm (available at http://icl.pku.edu.cn/yujs/lecture.htm).

Given the observation O and parameters Θ, the conditional probability, at time t, from state i to state j is:

$$
\begin{aligned}
\xi_t(i,j) &= P(X_t = i, X_{t+1} = j|O, \Theta) \\
&= \frac{P(X_t = i, X_{t+1} = j, O|\Theta)}{P(O|\Theta)} \\
&= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}
\end{aligned} \tag{12.28}
$$

Some values used in Baum-Welch algorithm The conditional probability

of state i at time t is

$$\gamma_t(i) = P(X_t = i|O,\Theta) = \sum_{j=1}^{n} \xi_t(i,j) \tag{12.29}$$

$$\begin{aligned} \sum_{t=1}^{T-1} \gamma_t(i) &= \text{probability of state } i \text{ in } O \\ &= P(i|O,\Theta) \end{aligned} \tag{12.30}$$

$$\begin{aligned} \sum_{t=1}^{T-1} \xi_t(i,j) &= \text{probability of } i \to j \text{ in } O \\ &= P(i \to j|O,\Theta) \end{aligned} \tag{12.31}$$

Baum-Welch algorithm

initialization:	Θ, ϵ	ϵ is experiential
calculation:	$\bar{\Theta} = \langle \bar{A}, \bar{B}, \bar{\pi} \rangle$, in which $\bar{a}_{ij} = \frac{P(i \to j|O,\Theta)}{P(i|O,\Theta)}$ $\bar{b}_i(k) = \frac{\sum_{t=1}^{T-1} \gamma_t(i) 1_{o_t=k}}{P(i|O,\Theta)}$ $\bar{\pi}_i = \gamma_1(i)$ where $1_{o_t=k} = \begin{cases} 1 & \text{if } o_t = k \\ 0 & \text{otherwise} \end{cases}$	updating parameters
condition:	if $\lvert \log P(O|\bar{\Theta}) - \log P(O|\Theta) \rvert < \epsilon$	end
goto:	otherwise, let $\Theta = \bar{\Theta}$	goto calculation

12.3　习题

12.1.

第十三章

期望最大化算法与最大熵算法

不完全数据下如何使用最大似然法对未知参数进行估计呢？例如，已知样本 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathsf{T}}$ 来自高斯混合总体 $w\phi(x|\mu_1, \sigma_1^2) + (1-w)\phi(x|\mu_2, \sigma_2^2)$，其中 $\mu_j, \sigma_j^2 (j = 1, 2)$ 和 $0 < w < 1$ 都未知，问它们的最大似然估计？

定义 13.1. 给定 $\boldsymbol{Y} = \boldsymbol{y}$ 后随机变量 X 的条件期望定义为

$$\mathsf{E}(X|\boldsymbol{Y} = \boldsymbol{y}) = \begin{cases} \sum\limits_{x} xp(x|\boldsymbol{y}) & \text{离散情形} \\ \int\limits_{\mathbb{R}} xf(x|\boldsymbol{y})dx & \text{连续情形} \end{cases} \tag{13.1}$$

性质 13.1.

$$\mathsf{E}[h(\boldsymbol{Y})|\boldsymbol{Y}] = h(\boldsymbol{Y}) \tag{13.2}$$

$$\mathsf{E}[r(\boldsymbol{X})h(\boldsymbol{Y})|\boldsymbol{Y}] = h(\boldsymbol{Y})E[r(\boldsymbol{X})|\boldsymbol{Y}] \tag{13.3}$$

$$\mathsf{E}[\mathsf{E}(Z|\boldsymbol{Y})] = \mathsf{E}(Z) \tag{13.4}$$

$$\mathsf{E}[r(\boldsymbol{X})h(\boldsymbol{Y})] = E\left\{h(\boldsymbol{Y})E[r(\boldsymbol{X})|\boldsymbol{Y}]\right\} \tag{13.5}$$

$$\mathsf{E}[Z - r(\boldsymbol{X})]^2 = E[Z - \mathsf{E}(Z|\boldsymbol{X})]^2 + E[r(\boldsymbol{X}) - \mathsf{E}(Z|\boldsymbol{X})]^2 \tag{13.6}$$

定理 13.1 (均方误差预测)**.** Mean square error (MSE) predicator Let $\boldsymbol{X}$ be a random vector and Z any rv, for any function $g(\boldsymbol{x})$,

$$\mathsf{E}[Z - \mathsf{E}(Z|\boldsymbol{X})]^2 \leq \mathsf{E}[Z - g(\boldsymbol{X})]^2 \tag{13.7}$$

当 $g(\boldsymbol{X}) = \mathsf{E}(Z|\boldsymbol{X})$ 时等号成立。$\mathsf{E}(Z|\boldsymbol{X})$ 称为条件 $\boldsymbol{X}$ 下 Z 的均方误差 (mean square error, MSE)。

定理 13.2 (Rao-Blackwell, 1949)**.**

$$\mathsf{V}(Z) = \mathsf{E}[Z - \mathsf{E}(Z|\boldsymbol{X})]^2 + \mathsf{V}[\mathsf{E}(Z|\boldsymbol{X})] \tag{13.8}$$

随机变量 Z 可由随机变量 $\mathsf{E}(Z|\boldsymbol{X})$ 来近似，并且 $\mathsf{V}[\mathsf{E}(Z|\boldsymbol{X})] \leq \mathsf{V}(Z)$。

13.1　期望最大化算法

13.1.1　完整数据与最大似然估计

Problems of MLE in case of missing data

13.1. (Gaussian mixtures) Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ be a sample from the population $(1 - w)\phi(x|\mu_1, \sigma_1^2) + w\phi(x|\mu_2, \sigma_2^2)$, where $0 < w < 1$, what about the MLEs of μ_i, σ_i^2 and w?

13.2. (Censored data) Let $X_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$ be the response rv of the jth element among those receiving the ith treatment. If some X_{ij} are unknown, what about the MLEs of $\mu_1, \cdots, \mu_k, \sigma^2$?

Problems of MLE in case of missing data

13.3. (Mixture-density problem) Given a sample $\boldsymbol{X} = (X_1, \cdots, X_n)^\mathsf{T}$ from the

表 13.1: One-way layout with missing data

Treatments	Results			
1	X_{11}	X_{12}	$\cdots$	X_{1n_1}
2	X_{21}	X_{22}	$\cdots$	X_{2n_2}
$\vdots$				
k	X_{k1}	X_{k2}	$\cdots$	X_{kn_k}

population with the mixture density

$$f_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{m} w_i p_i(x|\theta_i) \tag{13.9}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^\mathsf{T}$ and $w_i > 0$ are the prior probabilities of each mixture component satisfying $\sum_{i=1}^{m} w_i = 1$. What about the MLEs of $\boldsymbol{\theta}$ and w_i?

13.4. The parameter estimation of hidden Markov model (HMM), in which the latent data is the sequence of hidden states. In fact, Baum-Welch algorithm is a special implementation of EM algorithm.

Complete data and complete MLE 样本 $\boldsymbol{X} = (X_1, \cdots, X_n)^\mathsf{T}$ 连同缺失数据（或隐藏数据）$\boldsymbol{Y}$ 称为完全数据 (complete data)。

完全似然: 令 $f_\theta(\boldsymbol{x}, \boldsymbol{y})$ 是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的联合密度函数，其中参数 θ 未知（该参数也可以为参数向量），完全似然函数定义为 $\mathscr{L}(\theta|\boldsymbol{x}, \boldsymbol{y}) = f_\theta(\boldsymbol{x}, \boldsymbol{y})$。很多时候完全对数似然函数 $\ell(\theta|\boldsymbol{x}, \boldsymbol{y}) = \ln \mathscr{L}(\theta|\boldsymbol{x}, \boldsymbol{y})$ 更常用。

完全最大似然估计: 令 $f_\theta(\boldsymbol{y}|\boldsymbol{x})$ 为给定 $\boldsymbol{X} = \boldsymbol{x}$ 下的 $\boldsymbol{Y}$ 的条件密度函数。完全最大似然估计定义为

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ln f_\theta(\boldsymbol{x}) = \underset{\theta}{\operatorname{argmax}} [\ln \mathscr{L}(\theta|\boldsymbol{x}, \boldsymbol{y}) - \ln f_\theta(\boldsymbol{y}|\boldsymbol{x})] \tag{13.10}$$

13.1.2　期望最大化算法及其变种

Why EM algorithms?

- Likelihood-based inference is of central importance in statistical theory and data analysis.
- MLE is the most frequently-used estimation technique in the frequentist framework, it seems ubiquitous.
- In practice, there is a challenge of complicated likelihood function resulting in difficult-to-compute maximization problems. This difficulty could be analytical or computational or even both.
- **For many problems, it is possible to formulate the model with "augmented data" and work out the MLEs in an analytically and computationally simpler manner.**

Motivation of EM algorithm

Motivation of EM algorithm Some of the data drawn from $N_2(\boldsymbol{\mu}, \Sigma)$ are missing.

X	0	2	1	-1	$*$	3	1
Y	1	0	3	1	0	$*$	$*$

We get the initialization $\hat{\mu}_1 = (0 + 2 + 1 - 1 + 3 + 1)/6 = 1$ and $\hat{\mu}_2 = (1 + 0 + 3 + 1 + 0)/5 = 1$. Replace the $*$'s by the means.

表 13.2: Initialization of the missing data

X	0	2	1	-1	1	3	1
Y	1	0	3	1	0	1	1

Motivation of EM algorithm The MLE of Σ is

$$\begin{aligned}
\hat{\sigma}_{11} &= [(0-1)^2+(2-1)^2+(1-1)^2 \\
&\quad +(-1-1)^2+(1-1)^2+(3-1)^2+(1-1)^2]/7 = 10/7 \\
\hat{\sigma}_{22} &= [(1-1)^2+(0-1)^2+(3-1)^2 \\
&\quad +(1-1)^2+(0-1)^2+(1-1)^2+(1-1)^2]/7 = 6/7 \\
\hat{\sigma}_{12} &= [(0-1)(1-1)+(2-1)(0-1)+(1-1)(3-1) \\
&\quad +(-1-1)(1-1)+(1-1)(0-1)+(3-1)(1-1) \\
&\quad +(1-1)(1-1)]/7 = -1/7 \\
X|Y &\sim \mathrm{N}\left(\mu_2+\sigma_{12}\sigma_{22}^{-1}(Y-\mu_2), \sigma_{11}-\sigma_{12}^2\sigma_{22}^{-1}\right) \qquad (13.11) \\
Y|X &\sim \mathrm{N}\left(\mu_1+\sigma_{12}\sigma_{11}^{-1}(X-\mu_1), \sigma_{22}-\sigma_{12}^2\sigma_{11}^{-1}\right) \qquad (13.12)
\end{aligned}$$

Motivation of EM algorithm

13.1. By (13.11) and (13.12), we update the missing data by their conditional expectations.

表 13.3: Updating the missing data by their conditional expectations

X	0	2	1	−1	7/6	3	1
Y	1	0	3	1	0	4/5	1

13.2. The MLE of $\boldsymbol{\mu}$: $\hat{\mu}_1 = 43/42, \hat{\mu}_2 = 34/35$, the MLE of Σ.

13.3. Repeat steps 1-2 many times, until no improvement could be done.

给定 $\boldsymbol{X}=\boldsymbol{x}$ 和对未知参数的当前估计为 $\hat{\theta}=\theta_{t-1}$，对数似然函数 $\ell(\theta|\boldsymbol{x})$ 具有如下分解。

$$\begin{aligned}\ell(\theta|\boldsymbol{x}) &= \ln f_\theta(\boldsymbol{x}) = \ln f_\theta(\boldsymbol{x},\boldsymbol{Y}) - \ln f_\theta(\boldsymbol{Y}|\boldsymbol{x}) \\ &= \mathsf{E}_{\theta_{t-1}}[\ln f_\theta(\boldsymbol{x},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{x}] - \mathsf{E}_{\theta_{t-1}}[\ln f_\theta(\boldsymbol{Y}|\boldsymbol{x})|\boldsymbol{X}=\boldsymbol{x}] \\ &= Q(\theta,\theta_{t-1}) - H(\theta,\theta_{t-1}) \end{aligned} \tag{13.13}$$

$$\begin{aligned}Q(\theta,\theta_{t-1}) &= \mathsf{E}_{\theta_{t-1}}\left[\ln f_\theta(\boldsymbol{x},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{x}\right] \\ &= \int f_{\theta_{t-1}}(\boldsymbol{y}|\boldsymbol{x})\ln f_\theta(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y} = \int \frac{f_{\theta_{t-1}}(\boldsymbol{x},\boldsymbol{y})}{f_{\theta_{t-1}}(\boldsymbol{x})}\ell(\theta|\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y}\end{aligned} \tag{13.14}$$

$$\theta_t \leftarrow \underset{\theta}{\operatorname{argmax}}\, Q(\theta,\theta_{t-1}) = \underset{\theta}{\operatorname{argmax}} \int f_{\theta_{t-1}}(\boldsymbol{x},\boldsymbol{y})\ell(\theta|\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y} \tag{13.15}$$

保证了 $\ell(\theta_t|\boldsymbol{x}) \geq \ell(\theta_{t-1}|\boldsymbol{x})$，等号成立当且仅当 $Q(\theta_t,\theta_{t-1}) = Q(\theta_{t-1},\theta_{t-1})$。事实上，$\ell(\theta_t|\boldsymbol{x}) - \ell(\theta_{t-1}|\boldsymbol{x}) = [Q(\theta_t,\theta_{t-1}) - Q(\theta_{t-1},\theta_{t-1})] + [H(\theta_{t-1},\theta_{t-1}) - H(\theta_t,\theta_{t-1})] \geq 0$，这是因为（见附录 H 的定理 H.4）

$$H(\theta_{t-1},\theta_{t-1}) - H(\theta_t,\theta_{t-1}) = \int f_{\theta_{t-1}}(\boldsymbol{y}|\boldsymbol{x})\ln\frac{f_{\theta_{t-1}}(\boldsymbol{y}|\boldsymbol{x})}{f_{\theta_t}(\boldsymbol{y}|\boldsymbol{x})}d\boldsymbol{y} \geq 0 \tag{13.16}$$

Missing information principle Let θ^* be the maximum of $\ell(\theta|\boldsymbol{x})$. By normal-based inference, $\theta|\boldsymbol{X}=\boldsymbol{x} \sim \mathrm{N}_d(\theta^*,\Sigma_{\mathrm{obv}})$, where

$$\begin{aligned}\Sigma_{\mathrm{obv}}^{-1} &= -\left.\frac{\partial^2 \ln f_\theta(\boldsymbol{x})}{\partial\theta^2}\right|_{\theta^*} = -\left.\frac{\partial^2 \ell(\theta|\boldsymbol{x})}{\partial\theta^2}\right|_{\theta^*} \\ &= -\left.\frac{\partial^2 Q(\theta,\theta^*)}{\partial\theta^2}\right|_{\theta^*} + \left.\frac{\partial^2 H(\theta,\theta^*)}{\partial\theta^2}\right|_{\theta^*} \\ &= -\int \left.\frac{\partial^2 \ln \ell(\theta|\boldsymbol{x},\boldsymbol{y})}{\partial\theta^2}\right|_{\theta^*} f_{\theta^*}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} \\ &\quad + \int \left.\frac{\partial^2 \ln f_\theta(\boldsymbol{y}|\boldsymbol{x})}{\partial\theta^2}\right|_{\theta^*} f_{\theta^*}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}\end{aligned} \tag{13.17}$$

We have missing information principle — "observed information = complete information - missing information".

Monte Carlo EM (MCEM) algorithm

Monte Carlo implementation of the E-step The E-step (13.14) can be approximated by

13.1. Draw $y_1, \cdots, y_m$ from $f_{\theta_{t-1}}(\boldsymbol{y}|\boldsymbol{x})$.

13.2. Maximize the following approximation of $Q(\theta, \theta_{t-1})$.

$$Q(\theta, \theta_{t-1}) \approx \frac{1}{m} \sum_{j=1}^{m} \ell(\theta|\boldsymbol{x}, y_j) \tag{13.18}$$

For instance, in the genetic linkage model (see the next subsection), sample from $B(x_1, p_{t-1})$, we have $E_{\theta_{t-1}}(Y_2|\boldsymbol{X} = \boldsymbol{x}) \approx \sum_{j=1}^{m} y_j/m$. Let $m = 10, \theta_0 = 0.4$, after 9-12 iterations, we get the estimate result 0.627.

Convergence rate of EM algorithm

Convergence rate of EM algorithm The EM algorithm defines a self-map $M : \Theta \to \Theta$ such that $\theta_t = M(\theta_{t-1})$. Then $\theta^* = M(\theta^*)$. Dempster, Laird and Rubin (1977) showed that

$$\left[\left. \frac{\partial M(\theta)}{\partial \theta} \right|_{\theta^*} \right] \left[\left. \frac{\partial^2 Q(\theta, \theta^*)}{\partial \theta^2} \right|_{\theta^*} \right] = \left[\left. \frac{\partial^2 H(\theta, \theta^*)}{\partial \theta^2} \right|_{\theta^*} \right] \tag{13.19}$$

and the authors argued that in a neighborhood of θ^*, the rate of convergence is given by

$$\left(\frac{\partial^2 H}{\partial \theta^2} \right) \left(\frac{\partial^2 Q}{\partial \theta^2} \right)^{-1} \tag{13.20}$$

In Louis (1982) proved that

$$-\frac{\partial^2 H}{\partial \theta^2} = V\left[\frac{\partial \ell(\theta|\boldsymbol{x}, \boldsymbol{Y})}{\partial \theta}\right] \tag{13.21}$$

GEM algorithms

GEM algorithms GEM algorithm: Instead of maximizing $Q(\theta, \theta_{t-1})$, we find θ_t such that $Q(\theta_t, \theta_{t-1}) > Q(\theta_{t-1}, \theta_{t-1})$.

GEM: Generalized Expectation/Maximization
Dempster, Laird, Rubin, 1977
⇓
ECM: Expectation/Conditional Maximization
Meng, Rubin, 1993
⇓
ECME: Expectation/Conditional Maximization Either
Liu, Rubin, 1994
⇓
AECM: Alternating ECM
Meng, van Dyk, 1995

Further readings

13.1. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B, 39, 1-38.

13.2. Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society B, 44(2), pp226-233.

13.3. Tanner, M. A. (1996) Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Spring-Verlag New York, Inc.

13.4. McLachlan, G. and Krishnan, T. (1997) The EM Algorithm and Extensions. John Wiley & Sons, Inc.

13.5. Little, R. J. A. and Rubin, D. B. (2002) Statistical Analysis with Missing Data (2nd Ed.). John Wiley & Sons, Inc.

13.1.3 期望最大化算法的应用

EM algorithm for genetic linkage model

假设动物分为四个类 $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^\mathsf{T}$，概率为with probabilities $(1/2 + \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)^\mathsf{T}$. Among 197 animals, we observed $\boldsymbol{x} = (125, 18, 20, 34)^\mathsf{T}$.

Method of MLE: The likelihood function is

$$L(\theta|\boldsymbol{x}) = (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}$$

By $dL(\theta|\boldsymbol{x})/d\theta = 0$ we get the MLE $\hat{\theta} \approx 0.6268$.

EM algorithm: Let $X_1 = Y_1 + Y_2$ satisfying $p(y_1) = 1/2$ and $p(y_2) = \theta/4$. The complete data are $(Y_1, Y_2, X_2, X_3, X_4)^\mathsf{T}$.

EM algorithm for genetic linkage model The (complete) likelihood is simplified (in structure) to be

$$\begin{aligned} L(\theta|\boldsymbol{x}, \boldsymbol{y}) &= \theta^{y_2+x_4}(1-\theta)^{x_2+x_3}, \text{ and} \\ Q(\theta, \theta_{t-1}) &= E_{\theta_{t-1}}[(Y_2 + x_4)\ln\theta + (x_3 + x_4)\ln(1-\theta)|\boldsymbol{X} = \boldsymbol{x}] \\ &= [E_{\theta_{t-1}}(Y_2|\boldsymbol{X} = \boldsymbol{x}) + x_4]\ln\theta + (x_3 + x_4)\ln(1-\theta) \end{aligned}$$

$Y_2|\theta_{t-1}, \boldsymbol{X} = \boldsymbol{x} \sim B(x_1, p_{t-1})$, where $p_{t-1} = \theta_{t-1}/(\theta_{t-1}+2)$. By $dQ(\theta, \theta_{t-1})/d\theta = 0$, we have

$$\theta_t = \frac{E_{\theta_{t-1}}(Y_2|\boldsymbol{X}=\boldsymbol{x}) + x_4}{E_{\theta_{t-1}}(Y_2|\boldsymbol{X}=\boldsymbol{x}) + x_2 + x_3 + x_4} = \frac{\frac{x_1\theta_{t-1}}{\theta_{t-1}+2} + x_4}{\frac{x_1\theta_{t-1}}{\theta_{t-1}+2} + x_2 + x_3 + x_4}$$

Starting from $\theta_0 = 0.5$, we get $\hat{\theta} \approx \theta_5 \approx 0.6268$.

EM algorithm for genetic linkage model We will give the variance of $\hat{\theta}$ by (13.17) and (13.21).

$$\begin{aligned}
-\left.\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta^2}\right|_{\hat{\theta}} &= \frac{E_{\theta_{t-1}}(Y_2|\boldsymbol{X}=\boldsymbol{x}) + x_4}{\hat{\theta}^2} + \frac{x_3 + x_4}{(1-\hat{\theta})^2} \\
&= \frac{29.83 + 34}{0.6268^2} + \frac{38}{(1-0.6268)^2} = 435.2 \\
V\left[\left.\frac{\partial \ell(\theta|\boldsymbol{x}, \boldsymbol{Y})}{\partial \theta}\right|_{\hat{\theta}}\right] &= \frac{V_{\hat{\theta}}(Y_2|\boldsymbol{X}=\boldsymbol{x})}{\hat{\theta}^2} \\
&= \left(\frac{125}{\theta^2}\right)\left(\frac{\hat{\theta}}{\hat{\theta}+2}\right)\left(\frac{2}{\hat{\theta}+2}\right) = 57.8 \\
-\left.\frac{\partial^2 \ell(\theta|\boldsymbol{x})}{\partial \theta^2}\right|_{\hat{\theta}} &= 435.3 - 57.8 = 377.5 \text{ by (13.17) and (13.21)}
\end{aligned}$$

The standard error of $\hat{\theta}$ is $\sqrt{1/377.5} = 0.05$.

EM algorithm approach to Gaussian mixtures

EM algorithm approach to Gaussian mixtures Let $D = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ be from $(1-w)\phi(\boldsymbol{x}|\boldsymbol{\mu}_1, \Sigma) + w\phi(\boldsymbol{x}|\boldsymbol{\mu}_2, \Sigma)$, the problem is to estimate $\boldsymbol{\theta} = (w, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$.

The complete data are $\left\{\left.\begin{pmatrix}\boldsymbol{x}_i \\ y_i\end{pmatrix}\right| i = 1, 2, \cdots, n\right\}$, where $y_i \in \{0, 1\}$ is the class label of $\boldsymbol{x}_i$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^{\mathsf{T}}$ are the missing data. The complete log-

likelihood function is

$$\ell(\boldsymbol{\theta}|D,\boldsymbol{y}) = \sum_{i=1}^{n}\left[(1-y_i)\ln\phi(\boldsymbol{x}_i|\boldsymbol{\mu}_1,\Sigma) + y_i\ln\phi(\boldsymbol{x}_i|\boldsymbol{\mu}_2,\Sigma)\right] \tag{13.22}$$

Let $\partial\ell(\boldsymbol{\theta}|D,\boldsymbol{y})/\partial\boldsymbol{\theta} = \mathbf{0}$, we get the MLEs of $\boldsymbol{\theta}$. $\partial f(A)/\partial A$ is defined as $(\partial f(A)/\partial a_{ij})$, where $f(A)$ is a scale function of matrix A.

EM algorithm approach to Gaussian mixtures We have

$$\frac{\partial \boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x}}{\partial A} = \begin{cases} \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} & \text{if } A \text{ is asymmetric} \\ 2\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} - \text{diag}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} & \text{if } A \text{ is symmetric} \end{cases} \tag{13.23}$$

$$\frac{\partial |A|}{\partial A} = \begin{cases} |A|(A^{-1})^{\mathsf{T}} & \text{if } A \text{ is asymmetric} \\ |A|(2A^{-1} - \text{diag}A^{-1}) & \text{if } A \text{ is symmetric} \end{cases} \tag{13.24}$$

Homework: The MLEs of $\boldsymbol{\theta}$ in (13.22) are

$$\hat{w} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\boldsymbol{\mu}}_1 = \sum_{i=1}^{n}(1-y_i)\boldsymbol{x}_i \Big/ \left(n - \sum_{i=1}^{n} y_i\right), \quad \hat{\boldsymbol{\mu}}_2 = \sum_{i=1}^{n} y_i\boldsymbol{x}_i \Big/ \sum_{i=1}^{n} y_i$$

$$\hat{\Sigma}^{-1} = \frac{1}{n}\sum_{i=1}^{n}\left[(1-y_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_1)(\boldsymbol{x}_i-\boldsymbol{\mu}_1)^{\mathsf{T}} + y_i(\boldsymbol{x}_i-\boldsymbol{\mu}_2)(\boldsymbol{x}_i-\boldsymbol{\mu}_2)^{\mathsf{T}}\right]$$

EM algorithm approach to Gaussian mixtures

E-step: Starting with some initial $\boldsymbol{\theta}_0$. We find (13.22) is linear of $\boldsymbol{y}$, The E-step is determined by

$$E_{\boldsymbol{\theta}_{t-1}}(Y_i|D) = \frac{w^{(t-1)}\phi(\boldsymbol{x}_i|\boldsymbol{\mu}_2^{(t-1)},\Sigma^{(t-1)})}{(1-w)\phi(\boldsymbol{x}_i|\boldsymbol{\mu}_1^{(t-1)},\Sigma^{(t-1)}) + w\phi(\boldsymbol{x}_i|\boldsymbol{\mu}_2^{(t-1)},\Sigma^{(t-1)})}$$

where $\boldsymbol{\theta}_{t-1} = \left(w^{(t-1)},\boldsymbol{\mu}_1^{(t-1)},\boldsymbol{\mu}_2^{(t-1)},\Sigma^{(t-1)}\right)$.

M-step: Let $y_i \leftarrow E_{\boldsymbol{\theta}_{t-1}}(Y_i|D)$ and updates $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_t$ in the way of last slide.

EM algorithms are desirable to make the procedure of MLE easier by iterative computations.

EM algorithm for censored data

EM algorithm for censored data Let $\boldsymbol{x} = (x_1, \cdots, x_m)^\mathsf{T}$ be from $\mathrm{N}(\theta, 1)$, with censored data at a. The complete loglikelihood is

$$\begin{aligned} \ell(\theta|\boldsymbol{x}, \boldsymbol{y}) &= \sum_{i=1}^{m}(x_i - \theta)^2 + \sum_{i=m+1}^{n}(y_i - \theta)^2 \\ \text{thus, } Q(\theta, \theta_{t-1}) &= \sum_{i=1}^{m}(x_i - \theta)^2 + (n-m)E_{\theta_{t-1}}(Y_1 - \theta)^2 \end{aligned}$$

where the missing data $\boldsymbol{y} = (y_{m+1}, \cdots, y_n)^\mathsf{T}$ are from a truncated normal. The updating of θ is

$$\begin{aligned} \theta_t &= \frac{m\bar{x} + (n-m)E_{\theta_{t-1}}(Y_1)}{n} \\ &= \frac{m}{n}\bar{x} + \frac{n-m}{n}\theta_{t-1} + \frac{\phi(a - \theta_{t-1})}{n[1 - \Phi(a - \theta_{t-1})]} \end{aligned} \tag{13.25}$$

Truncated normal distribution Let $X \sim \mathrm{N}(0, 1), A$ be the event of $X \in [c, d]$, then $P(A) = \Phi(d) - \Phi(c)$ and the conditional density of $X|A$ is

$$f(x|A) = \frac{\phi(x)}{\Phi(d) - \Phi(c)} \tag{13.26}$$

Moreover, the moment generating function, expectation and variance of $X|A$ are

$$M(t) = \mathsf{E}(e^{tX}|A) = \frac{\int_c^d e^{tx}\phi(x)dx}{\Phi(d) - \Phi(c)} = e^{t^2/2}\frac{\Phi(d-t) - \Phi(c-t)}{\Phi(d) - \Phi(c)} \tag{13.27}$$

$$\mathsf{E}(X|A) = M'(t)|_{t=0} = -\frac{\phi(d) - \phi(c)}{\Phi(d) - \Phi(c)} \tag{13.28}$$

$$\mathsf{V}(X|A) = 1 - \frac{d\phi(d) - c\phi(c)}{\Phi(d) - \Phi(c)} - \left[\frac{\phi(d) - \phi(c)}{\Phi(d) - \Phi(c)}\right]^2 \tag{13.29}$$

13.2 最大熵算法

第十四章

随机模拟技术

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。

14.1 Markov 链 Monte Carlo (MCMC) 方法

14.1.1 Metropolis-Hastings 算法

Motivation and history of MCMC methods

14.1. MCMC methods includes many algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. The more steps, the better performance.

14.2. MCMC methods are widely applied in the numerical computation of multi-dimensional integrals, which often arise in Bayesian statistics,

physics and computational biology.

In 1953, N. Metropolis published the algorithm for Boltzmann distribution, and W. K. Hastings generalized it in 1970. The Gibbs sampling, a special case of the Metropolis-Hastings algorithm, is faster but less applicable in general.

Metropolis 算法与 Hastings 算法

Metropolis algorithm Boltzmann distribution is frequently used in physics.

$$\pi(x) = \frac{1}{Z(T)} \exp\left\{-\frac{U(x)}{kT}\right\} \tag{14.1}$$

where x is a pattern of physical system (may be a vector), $U(x)$ is its potential energy, T is its temperature, k is the Boltzmann constant, and $Z(T)$ is the normalizing constant.

In 1953, Metropolis proposed a method to sample from arbitrary distribution in the following universal form.

$$\pi(x) = \frac{1}{Z} \exp\{-h(x)\} \tag{14.2}$$

Metropolis algorithm We define the probability **transition function** of states $T(x, y)$ as a non-negative function satisfying $\sum_y T(x, y) = 1$ for all x.

Metropolis assumed that $T(x, y)$ is symmetric. The current state is x_n, now we consider the next possible state.

14.1. Generate the proposal y from $T(x_n, y)$.

14.2. Generate $U \sim \mathrm{U}[0, 1]$,

$$x_{n+1} = \begin{cases} y & \text{if } U \le \pi(y)/\pi(x_n) = \exp\{h(y) - h(x_n)\} \\ x_n & \text{otherwise} \end{cases} \tag{14.3}$$

Examples of Metropolis algorithm Let $\pi(x) \propto \exp\{-|x-4|\}$ (Laplace distribution). A new state is proposed by randomly drawing from the normal distribution centered at the current state, with standard deviation given by the stepsize.

Examples of Metropolis algorithm Let $-\ln \pi(x,y,z) = x^2/2 + y^2/2 + z^2/2 + (x+y+z)^2 + 10^4/(1+x^2+y^2+z^2)$.

Metropolis-Hastings algorithm Hastings generalized the symmetric transition function to be "$T(x,y) > 0 \Leftrightarrow T(y,x) > 0$", and proposed that

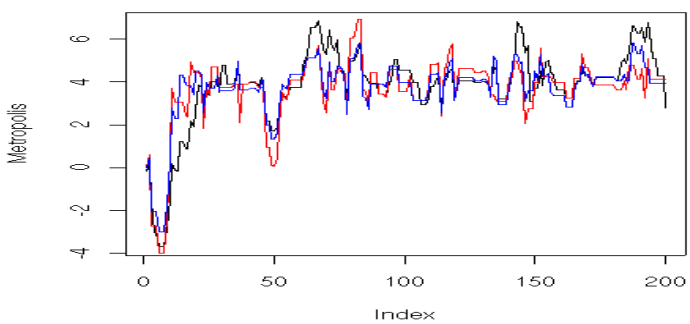
14.1. Generate the proposal y from $T(x_n, y)$.

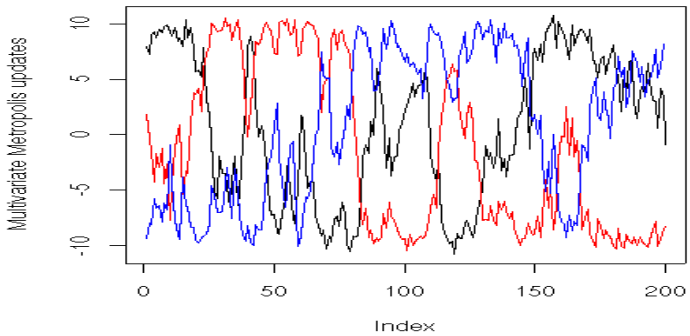
14.2. Generate $U \sim \mathrm{U}[0,1]$,

$$x_{n+1} = \begin{cases} y & \text{if } U \le r(x_n, y) \\ x_n & \text{otherwise} \end{cases} \tag{14.4}$$

where the acceptance function $r(x,y)$ is suggested to be

$$r(x,y) = \min\left\{1, \frac{\pi(y)T(y,x)}{\pi(x)T(x,y)}\right\} \tag{14.5}$$





Why does Metropolis-Hastings algorithm work? For M-H algorithm, $\pi(x)A(x,y)$ is symmetric function of x, y, where

$$A(x,y) = T(x,y)r(x,y) = T(x,y)\min\left\{1, \frac{\pi(y)T(y,x)}{\pi(x)T(x,y)}\right\}$$

therefore the detailed balance condition of (12.21) is satisfied, i.e., the Markov chain induced by M-H algorithm is reversible. For the following acceptance functions, (12.21) is satisfied too.

$$r(x,y) = \frac{\pi(y)T(y,x)}{\pi(y)T(y,x) + \pi(x)T(x,y)} \tag{14.6}$$

$$r(x,y) = \frac{\delta(x,y)}{\pi(x)T(x,y)} \leq 1 \tag{14.7}$$

where $\delta(x,y)$ is a symmetric function of x, y.

Mutilple-try Metropolis (MTM) algorithm

Mutilple-try Metropolis (MTM) algorithm Liu, J. S., Liang, F. and Wong, W. H. (2000) The use of multiple-try method and local optimization in Metropolis sampling, Journal of the American Statistical Association, 96(454): 561-573.

Let $T(x,y)$ be an arbitrary proposal function satisfying $T(x,y)>0 \Leftrightarrow T(y,x)>0$. Define

$$w(x,y)=\pi(x)T(x,y)\lambda(x,y) \tag{14.8}$$

where $\lambda(x,y)$ is a non-negative symmetric function in x and y, defined by users. Suppose the current state is x.

Mutilple-try Metropolis (MTM) algorithm

14.1. Draw m independent trial proposals $y_1,\cdots,y_m$ from $T(x,\cdot)$. Compute the weights $w(y_i,x)$.

14.2. Select y from the y_i with probability proportional to the weights.

14.3. Produce a reference set by drawing $x_1,\cdots,x_{m-1}$ from the distribution $T(y,\cdot)$. Set $x_m=x$.

14.4. Accept y with probability

$$r(x,y)=\min\left\{1,\frac{w(y_1,x)+\cdots+w(y_m,x)}{w(x_1,y)+\cdots,w(x_m,y)}\right\} \tag{14.9}$$

It is proved that MTM algorithm satisfies the detailed balance requirement (12.21).

14.1.2　Gibbs 抽样与切片抽样

Gibbs sampling — a special MCMC method $\boldsymbol{X}=(X_1,\cdots,X_d)^{\mathsf{T}}\sim\pi(\boldsymbol{x})$ is the random vector of concern. Let $\boldsymbol{X}_{-i}$ denote $(X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_d)$ and $\boldsymbol{x}^{(n)}=(x_1^{(n)},\cdots,x_d^{(n)})$ for the n-th iteration of Gibbs sampling.

Systematic-scan Gibbs sampler: For $i = 1, \cdots, d$, draw $x_i^{(n+1)}$ for the conditional distribution

$$X_i^{(n+1)} \sim \pi(x_i | x_1^{(n+1)}, \cdots, x_{i-1}^{(n+1)}, x_{i+1}^{(n)}, \cdots, x_d^{(n)}) \tag{14.10}$$

Random-scan Gibbs sampler: Two steps from $\boldsymbol{x}^{(n)}$ to $\boldsymbol{x}^{(n+1)}$.

14.1. Select a coordinate i from $\{1, \cdots, d\}$ according to a given probability distribution $p(i) = p_i$.

14.2. Draw $x_i^{(n+1)}$ from $\pi(x_i | \boldsymbol{x}_{-i}^{(n)})$ and update $\boldsymbol{x}_{-i}^{(n+1)} = \boldsymbol{x}_{-i}^{(n)}$.

Example of Gibbs sampling Let $\boldsymbol{X} \sim \mathrm{N}_2(\boldsymbol{0}, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The systematic-scan Gibbs sampler is designed by

$$X_1^{n+1} | X_2^{(n)} = x_2^{(n)} \sim \mathrm{N}(\rho x_2^{(n)}, 1 - \rho^2) \tag{14.11}$$

$$X_2^{n+1} | X_1^{(n+1)} = x_1^{(n+1)} \sim \mathrm{N}(\rho x_1^{(n+1)}, 1 - \rho^2) \tag{14.12}$$

We have

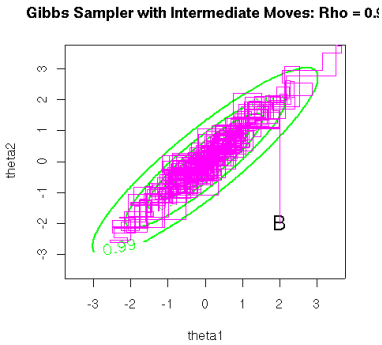
$$\begin{pmatrix} X_1^{(n)} \\ X_2^{(n)} \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} \rho^{2n-1} x_2^{(0)} \\ \rho^{2n} x_2^{(0)} \end{pmatrix}, \begin{pmatrix} 1 - \rho^{4n-2} & \rho - \rho^{4n-1} \\ \rho - \rho^{4n-1} & 1 - \rho^{4n} \end{pmatrix} \right) \tag{14.13}$$

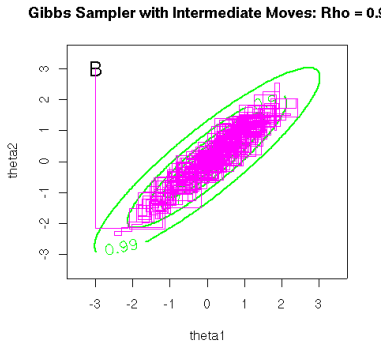
Example of Gibbs sampling Let $\boldsymbol{\theta} \sim \mathrm{N}_2(\boldsymbol{0}, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$. The simulations start from $(2, -2)$ and $(-3, 3)$ respectively.

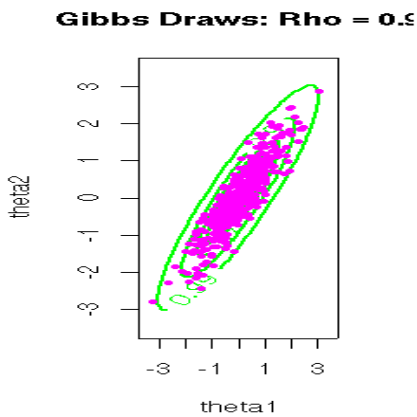
Example of Gibbs sampling

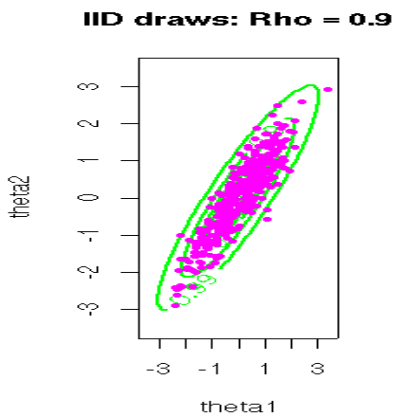
Slice sampling — a special Gibbs sampling 📎Basic idea: Let $X \sim \pi(x)$, then

$$\pi(x) = \int_0^{\pi(x)} dy \tag{14.14}$$


Gibbs Sampler with Intermediate Moves: Rho = 0.
theta2
theta1
B
0.99


Gibbs Sampler with Intermediate Moves: Rho = 0.
theta2
theta1
B
0.99


Gibbs Draws: Rho = 0.
theta2
theta1
0.99


IID draws: Rho = 0.9
theta2
theta1
0.99

$\pi(x)$ is the marginal density of the joint distribution $(X, Y) \sim \mathrm{U}\{(x, y) : 0 < y < \pi(x)\}$. A Gibbs sampler can be designed as follows.

14.1. Draw $y^{(n+1)} \sim \mathrm{U}[0, \pi(x^{(n)})]$.

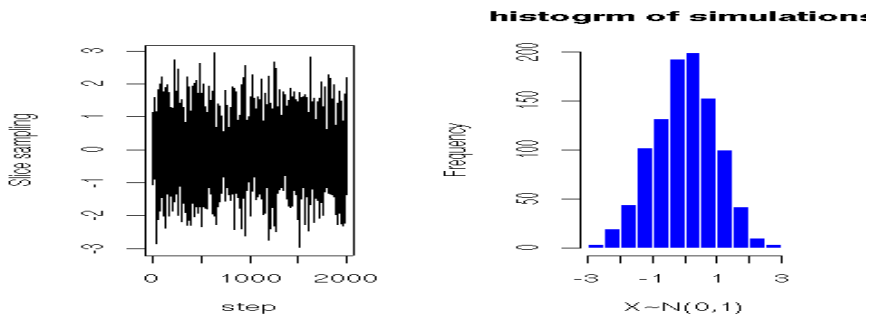
14.2. Draw $x^{(n+1)}$ uniformly from the region $S^{(n+1)} = \{x|\pi(x) \geq y^{(n+1)}\}$. However, step 2 is often as difficult as the original Monte Carlo simulation.

Further reading: Radford M. Neal, "Slice Sampling". The Annals of Statistics, 31(3):705-767, 2003.

Example of slice sampling To sample $X \sim \mathrm{N}(0, 1)$, we start from $x = 1.14$.

14.1. $Y \sim \mathrm{U}\left[0, \exp\{-x^2/2\}/\sqrt{2\pi}\right]$.

14.2. $X \sim \mathrm{U}[-a, a]$, where $a = \sqrt{-2\ln(y\sqrt{2\pi})}$.



14.1.3 混合 Monte Carlo 方法

Hybrid Monte Carlo (HMC) Some sophisticated algorithms prevent the walker from doubling back, so that the convergence becomes faster. Hybrid Monte Carlo, also called **Hamiltonian Monte Carlo**, is such a method.

Hybrid Monte Carlo method implements Hamiltonian dynamics, in which the

potential function is the target density. The momentum samples are discarded after sampling. The result of Hybrid MCMC is that proposals move across the sample space in larger steps (therefore less correlated) and converge to the target distribution more rapidly.

➠Further readings: Liu, J.S. (2001) Monte Carlo Strategies in Scientific Computing, Spring-Verlag New York Inc. pp183-203.

14.1.4　可逆跳 MCMC 方法

Gaussian mixture with unknown number of components Problem: Let the sample be drawn from $M_k = \sum_{j=1}^{k} w_j \mathrm{N}(\mu_j, \sigma_j^2)$, where the number of components k, the mixture proportions w_j (satisfying $0 \le w_j \le 1$ and $\sum_{j=1}^{k} w_j = 1$), and $\rho_j = (\mu_j, \sigma_j^2)$ are unknown for $j = 1, 2, \cdots, k$.

✎The difficulty of statistical model learning of heterogeneous population M_k lies in the varying dimension of k-component state $\boldsymbol{\theta}^{(k)} = (\boldsymbol{w}^{(k)}, \boldsymbol{\rho}^{(k)}) = (w_1, \cdots, w_k, \rho_1, \cdots, \rho_k)$, which unavoidably leads to the moves between parameter spaces with different dimensions. Fortunately, this problem can be worked out by at least two MCMC simulation methods developed recently.

14.1. reversible jump MCMC (RJMCMC, Green 1995, Richardson & Green 1997)

14.2. birth-death MCMC (Stephens 2000)

Green's reversible jump MCMC (RJMCMC) For $k \neq k'$, the key idea of Green's reversible jump MCMC is to supplement two suitable simulations $\boldsymbol{\vartheta}^{(k)} \sim f_k(\boldsymbol{\vartheta}^{(k)})$ and $\boldsymbol{\vartheta}^{(k')} \sim f_{k'}(\boldsymbol{\vartheta}^{(k')})$ to $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\theta}^{(k')}$ respectively, such that

$$(\boldsymbol{\theta}^{(k')}, \boldsymbol{\vartheta}^{(k')}) = T_{k \to k'}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\vartheta}^{(k)}) \tag{14.15}$$

is a bijection. The acceptance probability of moving from model M_k to model $M_{k'}$, in the Metropolis-Hastings form, is $\min(1, A_{k\to k'})$, where $A_{k\to k'}$ is

$$\underbrace{\frac{\pi(k', \boldsymbol{\theta}^{(k')})}{\pi(k, \boldsymbol{\theta}^{(k)})}}_{\text{model ratio}} \times \underbrace{\frac{\pi_{k'\to k} f_{k'}(\boldsymbol{\vartheta}^{(k')})}{\pi_{k\to k'} f_k(\boldsymbol{\vartheta}^{(k)})}}_{\text{proposal ratio}} \times \underbrace{\left|\frac{\partial T_{k\to k'}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\vartheta}^{(k)})}{\partial(\boldsymbol{\theta}^{(k)}, \boldsymbol{\vartheta}^{(k)})}\right|}_{\text{Jacobian } |J|} \tag{14.16}$$

Further readings of RJMCMC

14.1. Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4), 711-732.

14.2. Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion), Journal of the Royal Statistical Society Series B, 59, 731-792.

14.3. Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components: an alternative to reversible jump methods, Annals of Statistics, 28, 40-74.

第四部分

附录

附录 A

软件 R、Maxima 和 GnuPlot 简介

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*。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。

A.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 Gibbs 抽样工具 BUGS 或 JAGS、Python、C 等。“众人拾柴

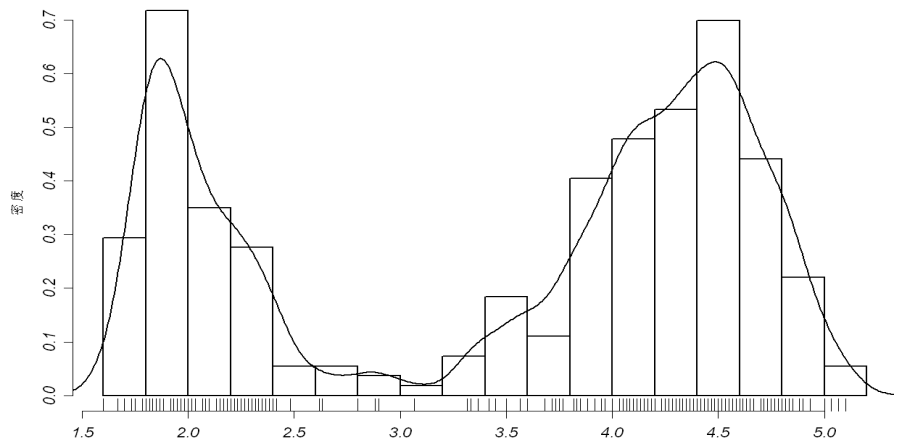
*R、Maxima 和 GnuPlot 都是跨平台的，既支持命令行交互模式，也支持脚本。

†1998 年，John Chambers 因对 S 语言的杰出贡献获得了 ACM 软件系统奖 (ACM Software Systems Award)。

火焰高”，开源为 R 的普及铺平了道路，并使得 R 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。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。

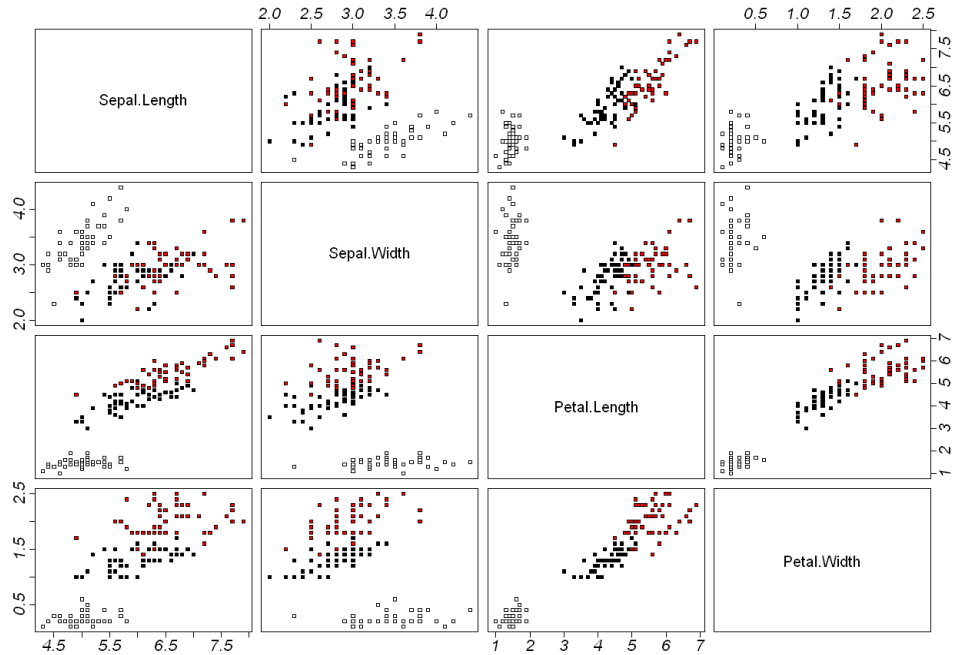
例 A.1 (交互式，> 是命令行提示符)**.** 利用 summary 函数考察一维数据的分布情况；利用 hist 绘出直方图，并叠加上密度估计曲线。

```
> attach(faithful)        # 导入数据
> summary(eruptions)      # 对数据的大致描述
   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)    # 画直方图
> lines(density(eruptions, bw=0.1))                 # 叠加密度曲线
> rug(eruptions)          # 显示数据点
```



例 A.2 (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。

```
pairs(iris[1:4], pch=22, font.axis = "3", cex.axis = 2,
      bg=c("white", "black", "red")[unclass(iris$Species)])
```



A.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

*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。

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）。

例 A.3 (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]);
                                                    2                 n
                      %k  (2 n + 1)          %k  (2 n  + 2 n - 1) (- 1)
                        2                      1
(%o2)     T  = ------------------------ - --------------------------
           n    (n - 1) n (n + 1) (n + 2)    (n - 1) n (n + 1) (n + 2)
```

其中，提示符 (%o2) 表示第 2 个输入 (%i2) 的输出结果。

例 A.4. 用组合数学的方法证明李善兰恒等式 (1.50) 并非易事，可仅用 Maxima 的寥寥数行代码就能验证它。

```
load("simplify_sum") $   /* 导入“简化求和”包 */
assume(m > n) $          /* 假设 m>n */
sum ((binomial(n,k))^2 * binomial(m+2*n-k,2*n), k, 0, n);
simplify_sum(%);         /* 简化上述结果 */
```

例 A.5. 级数求和 $\sum_{n=1}^{\infty} n^{-2}$，不定积分 $\int (1+x^3)^{-1} dx$ 等都通过 Maxima 计算。另外，Maxima 还可调用外部绘图程序 GnuPlot 实现绘图功能。

```
(%i1) sum (1/n^2, n, 1, inf), simpsum;
                                        2
                                     %pi
(%o1)                                ----
                                      6
(%i2) integrate(1/(1+x^3),x);
                                         2 x - 1
                      2             atan(-------)
                  log(x  - x + 1)        sqrt(3)    log(x + 1)
(%o2)           - --------------- + ------------- + ----------
                         6             sqrt(3)          3
(%i3) load(draw) $
(%i4) 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)),
```

```
15        gr3d(surface_hide = true, parametric_surface(cos(a)*(10+b*cos(a/2)),
16        sin(a)*(10+b*cos(a/2)), b*sin(a/2), a,-%pi,%pi,b,-1,1))) $
```

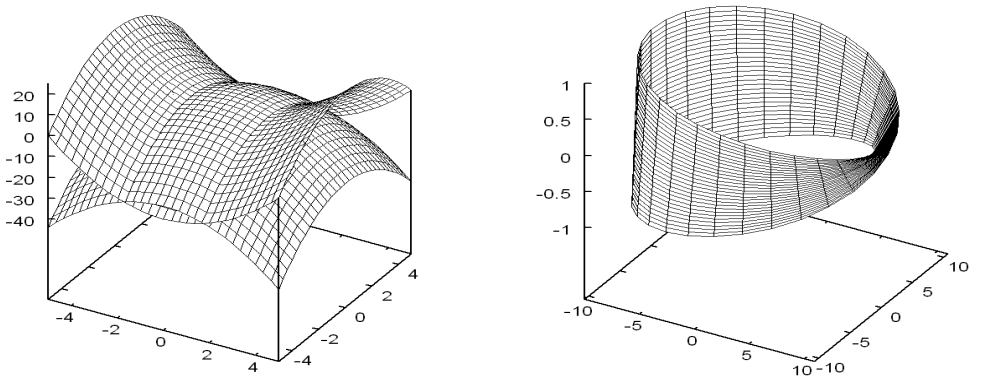


图 A.1: Maxima 的 draw 函数调用外部绘图程序 GnuPlot 完成绘图：左图是曲面 $x^2 - y^2$ 与 $6 - x^2 - y^2$ 之交，右图是不可定向曲面 Möbius 带。

例 A.6. Maxima 环境下可以不断调用 draw 以产生模拟效果。

```
1  load(draw) $
2  block([history:[[0,0,0]], lst, pos],
3  for k:1 thru 10000 do
4    (lst: copylist(last(history)),
5    pos: random(3) + 1,
6    lst[pos]: lst[pos] + random(2)*2-1,
7    history: endcons(lst, history)),
8  draw3d(point_type = 0, points_joined = true, points(history))) $
```

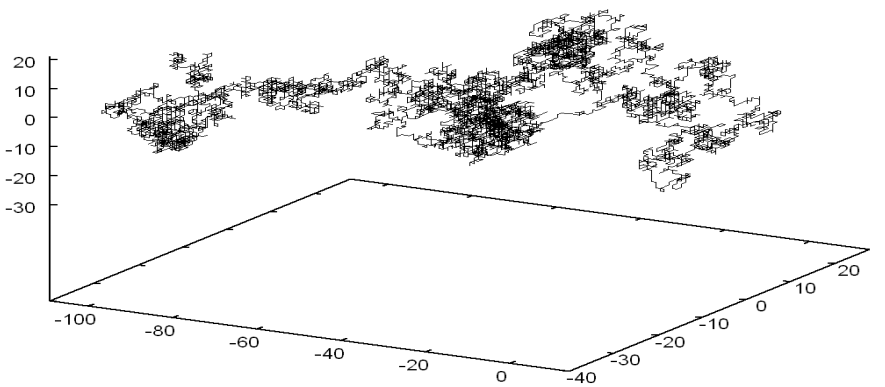
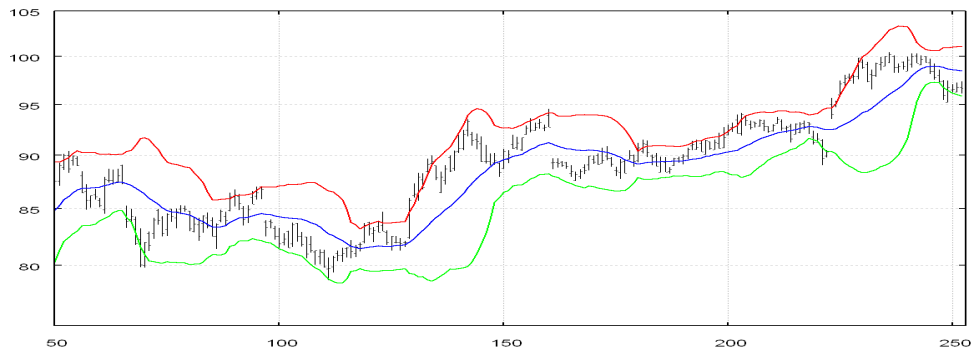


图 A.2: 三维空间中，粒子 Brown 运动的模拟。

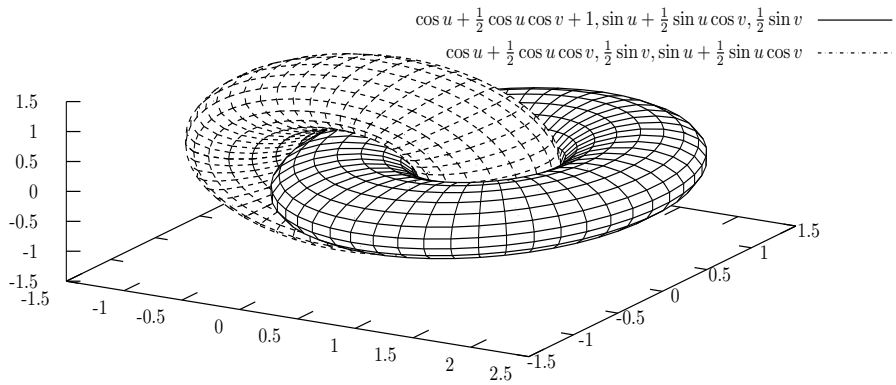
A.3　GnuPlot —— 强大的函数绘图工具

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。其特点如下：

- 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。如股票 K 线图和 Bollinger 带。



- GnuPlot 支持多种输出格式，如 eps, fig, jpeg, LaTeX, MetaPost, pbm, pdf, png, postscript, svg 等。输出的 MetaPost 文档可插入数学公式，再经 `mpost` 编译生成精美的矢量图直接嵌入 LaTeX 文件。



*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。

附录 B

一些模拟试验的 R 演示源码

❑ Bertrand 悖论（见例 1.16）的 R 演示源码。

```
## 目的: 实现 Bertrand 几何概率问题的三种取弦方式
## 输出: n 条随机弦的中点, 并显示前 m 条随机弦 (m < n)
rm(list=ls())                # 清空当前数据
n  <- 5000                   # 随机弦的条数
m  <- 500                    # 画出的随机弦的条数
op <- par(mar=c(0,0,0,0))    # 输出图形的边白厚度为 0

## 第一种选弦方式
rho   <- runif (n)           # 极径: (0,1] 上均匀分布的随机数
theta <- runif(n) * 2 * pi   # 角度: [0, 2*pi) 上均匀分布的随机数

## 画出随机弦的中点并保存输出图形
x <- rho * cos(theta)        # 随机弦的中点 (x,y)
y <- rho * sin(theta)
plot(x, y, pch ='.', cex = 2, axes = FALSE, xlab = "", ylab = "")
dev2bitmap (file = 'Bertrand1-scatterplot.png', type = 'pnggray')

## 定义函数 chord(theta, rho), 其中 theta 为角度, rho 为极径
## 目的: 计算两个端点的坐标 (x1,y1) 和 (x2,y2) 并画出连线
chord <- function(theta, rho){
    eta1  <- theta - acos(rho)
    eta2  <- theta + acos(rho)
    x1 <- cos(eta1)
    y1 <- sin(eta1)
    x2 <- cos(eta2)
```

```
    y2 <- sin(eta2)
    plot(x1, y1, pch = '.', axes = FALSE, xlab = "", ylab = "")
    segments(x1[1:m], y1[1:m], x2[1:m], y2[1:m])      # 画出前 m 条随机弦
}                           # 函数 chord 的定义结束

## 按照第一种选弦方式画出随机弦并保存输出图形
chord(theta, rho)
dev2bitmap (file = 'Bertrand1.png', type = 'pnggray')

## 第二种选弦方式
r1 <- runif(n) * 2 * pi     # 产生 [0, 2*pi) 上均匀分布的随机数
r2 <- runif(n) * 2 * pi
x1 <- cos(r1)               # 计算两个端点 (x1,y1) 和 (x2,y2)
y1 <- sin(r1)
x2 <- cos(r2)
y2 <- sin(r2)
x  <- (x1 + x2)/2           # 随机弦的中点 (x,y)
y  <- (y1 + y2)/2

## 画出随机弦的中点并保存输出图形
plot(x, y, pch = '.', cex = 2, axes = FALSE, xlab = "", ylab = "")
dev2bitmap(file = 'Bertrand2-scatterplot.png', type = 'pnggray')

## 按照第二种选弦方式画出随机弦并保存输出图形
plot(x1, y1, pch = '.', axes = FALSE, xlab = "", ylab = "")
segments(x1[1:m], y1[1:m], x2[1:m], y2[1:m])
dev2bitmap (file = 'Bertrand2.png', type = 'pnggray')

## 第三种选弦方式
n <- floor (4 * n/pi)       # 随机样本的个数
x <- runif(n) * 2 - 1       # 正方形 [-1,1]x[-1,1] 上的均匀分布
y <- runif(n) * 2 - 1
z <- sqrt(x^2 + y^2)        # 向量的长度
idx <- z < 1                # 落于单位圆内的 (x,y) 的指标
x <- x[idx]                 # 单位圆上均匀分布的随机抽样
y <- y[idx]
rho   <- sqrt(x^2 + y^2)    # 计算两个端点 (x1,y1) 和 (x2,y2)
theta <- atan2(y,x)

## 画出随机弦的中点并保存输出图形
plot(x, y, pch = '.', cex = 2, axes = FALSE, xlab = "", ylab = "")
dev2bitmap (file = 'Bertrand3-scatterplot.png', type = 'pnggray')

## 按照第三种选弦方式画出随机弦并保存输出图形
```

```
chord(theta, rho)
dev2bitmap (file = 'Bertrand3.png', type = 'pnggray')
```

❑ 求圆周率近似值（例 1.19）的 R 源码。

```
## 目的: 通过投钉试验模拟计算圆周率
## 输出: n 次投钉试验得到的圆周率的近似值
rm(list=ls())                          # 清空当前数据
Num        <- 10000                    # N 为投钉的次数
Experiment <- 3                        # 模拟试验的次数
expand = 0.05                          # 估计值的变化范围

## 定义函数 SimulatePi(n): n 次投钉所得到的 n 个圆周率的近似值
SimulatePi <- function(n){             # n 为投钉次数
  x <- runif(n, min=-0.5, max=0.5)     # 分别产生 n 个 [-0.5,0.5] 区间上
  y <- runif(n, min=-0.5, max=0.5)     # 均匀分布的随机数作为横坐标和纵坐标
  z <- x^2 + y^2 <= 0.25               # 判断钉是否投在圆盘上
  first <- sum(z[1])                   # 第一个钉是否投在原盘上?
  init  <- 4*first                     # 圆周率的第一个近似值
  apx   <- rep(init,n)                 # 圆周率 n 次估计结果的初始化
  SUM   <- rep(first,n)                # 前 n 次投钉结果的初始化

## 计算 1,2,...,n 次投钉所得圆周率的近似值
  for (i in 2:n){
    SUM[i] <- SUM[i-1] + z[i]          # 前 n 次投钉落于圆内的钉数
    apx[i] <- 4 * SUM[i]/i             # 计算第 i 个近似值
  }
  return(apx)                          # 返回近似值序列给函数 SimulatePi(n)
}                                      # 函数 SimulatePi 定义结束

## 绘图参数的设定
op  <- par(lwd = "2", font = "3", font.axis = "3", font.lab = "1", # 线宽与字体
           cex = 1.2, cex.axis = 1.2, mar = par("mar")+c(0,0,0,0)) # 边白宽度

## 画出第一轮投钉试验求得的圆周率的近似值
plot(SimulatePi(Num), type = "l",
     ylim = c((1-expand) * pi,(1+expand) * pi),     # 纵坐标的变化范围
     xlab = paste("投 钉 次 数"),                   # x 轴标记
     ylab = paste("圆 周 率 的 估 计 值"))          # y 轴标记

## 画出其余 Experiment-1 轮模拟试验的结果
for (j in 2:Experiment){
  points(SimulatePi(Num), type = "l", col = j)      # col=2 为红色, col=3 为绿色
}
```

```
40
41 ## 画出圆周率精确值的参考线
42 abline(a=pi, b=0, lty = 2)                    # lty=2 表示虚线
```

❑ 例 1.18 中用于随机模拟的函数 SimulatePi(n)。

```
1  ## 定义函数 SimulatePi(n): n 次投针所得到的 n 个圆周率的近似值
2  SimulatePi <- function(n){          # n 为投针次数
3    L <- 0.75                         # Buffon 试验中的针长，区域宽度设为单位 1
4    x <- runif(n, min=0, max=pi)      # 分别产生 n 个 [-0.5,0.5] 区间上
5    y <- runif(n, min=0, max=0.5)     # 均匀分布的随机数作为横坐标和纵坐标
6    temp <- sin(x)*L/2
7    z <- as.numeric(y <= temp)        # 判断针是否与区域边界相交
8    first <- sum(z[1])                # 第一个针是否投在正弦取弦之下?
9    apx   <- rep(0,n)                 # 圆周率 n 次估计结果的初始化
10   SUM   <- rep(first,n)             # 前 n 次投针结果的初始化
11   if (SUM[1]>0) apx[1] <- 2 * L
12   else apx[1] <- 10^2
13
14 ## 计算 1,2,...,n 次投针所得圆周率的近似值
15   for (i in 2:n){
16     SUM[i] <- SUM[i-1] + z[i]       # 前 n 次投针落于正弦取弦之下的针数
17     if (SUM[i]>0)
18     apx[i] <- 2*L*i/SUM[i]          # 计算第 i 个近似值
19     else apx[i] <- 10^2
20   }
21   return(apx)                       # 返回近似值序列给函数 SimulatePi(n)
22 }                                   # 函数 SimulatePi 定义结束
```

附录 C

正态分布的由来

正态分布的密度函数 $\phi(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$ 是如何得到的？C. F. Gauss (1809) 和 P. Laplace (1812) 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。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 $\mathrm{B}(n,1/2)$ 可用正态分布 $\mathrm{N}(n/2,n/4)$ 来近似（见例 4.1），具体说来，$C_n^k 2^{-n}\approx\phi(k|n/2,n/4)$。遗憾的是，de Moivre 没有明确定义正态分布，这件工作是 Gauss 于 1809 年完成的。

1812 年，Laplace 将 de Moivre 的结果推广到 $\mathrm{B}(n,p)$ 的情形：对于满足条件 $|k-np|=o(npq)^{2/3}$ 的所有 k 皆有

$$P_n(k)=C_n^k p^k q^{n-k}\sim\frac{1}{\sqrt{2\pi npq}}\exp\left\{-\frac{(k-np)^2}{2npq}\right\} \tag{C.1}$$

其中 $q=1-p,0<p<1$。这就是历史上著名的 de Moivre-Laplace 中心极限定理。下面用数学分析的方法往证 (C.1) 并借此了解 $\phi(x|\mu,\sigma^2)$ 的由来：若 k 使得 $|k-np|=o(npq)^{2/3}$，则有 $\lim\limits_{n\to\infty}|\tilde{p}-p|=0$。利用 Stirling

公式* $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$，不难验证

$$C_n^k = \frac{n!}{k!(n-k)!} \sim \frac{1}{\sqrt{2\pi n \frac{k}{n}\left(1-\frac{k}{n}\right)}\left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}}$$

令 $\tilde{p} = k/n$，则有

$$\begin{aligned} C_n^k p^k q^{n-k} &\sim \frac{1}{\sqrt{2\pi n \tilde{p}(1-\tilde{p})}} \left(\frac{p}{\tilde{p}}\right)^k \left(\frac{1-p}{1-\tilde{p}}\right)^{n-k} \\ &= \frac{1}{\sqrt{2\pi n \tilde{p}(1-\tilde{p})}} \exp\left\{-n\left[\frac{k}{n}\ln\frac{\tilde{p}}{p} + \left(1-\frac{k}{n}\right)\ln\frac{1-\tilde{p}}{1-p}\right]\right\} \\ &= \frac{1}{\sqrt{2\pi n \tilde{p}(1-\tilde{p})}} \exp\{-nh(\tilde{p})\} \end{aligned} \tag{C.2}$$

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

$$h(x) = h(p) + h'(p)(x-p) + \frac{1}{2}h''(p)(x-p)^2 + o(|x-p|^2) = \frac{(x-p)^2}{2pq} + o(|x-p|^2)$$

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 (C.2) 中便得到 (C.1)。

性质 C.1. 二项分布 $\mathrm{B}(n,p)$ 的分布函数 $F(x)$ 与标准正态分布的分布函数 $\Phi(x)$ 有如下渐近关系：

$$F(x) = \Phi\left[\frac{x-np+0.5}{\sqrt{np(1-p)}}\right] + O\left(\frac{1}{\sqrt{n}}\right) \tag{C.3}$$

*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 D

函数项级数的一致收敛性

定义 D.1. 一个函数序列 $g_1(t), g_2(t), \cdots, g_n(t), \cdots$ 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。

基于函数序列的一致收敛性，我们可以定义函数项级数 $\sum_{k=1}^{\infty} u_k(t)$ 的一致收敛性。记 $\sum_{k=1}^{\infty} u_k(t)$ 的前 n 项之和为

$$S_n(t) = \sum_{k=1}^{n} u_k(t), \text{ 其中 } t \in T$$

定义 D.2. 函数项级数 $\sum_{k=1}^{\infty} u_k(t)$ 在集合 T 上一致收敛于 $S(t)$ 当且仅当函数序列 $S_1(t), S_2(t), \cdots, S_n(t), \cdots$ 在集合 T 上一致收敛于 $S(t)$。

性质 D.1. 下面列举几个一致收敛的函数项级数常用的性质。

- 已知函数项级数 $\sum_{k=1}^{\infty} u_k(t)$ 一致收敛，如果 $u_1(t), u_2(t), \cdots$ 在 $t = t_0$ 处连续，则 $\sum_{k=1}^{\infty} u_k(t)$ 在 t_0 处也连续。

- 如果 $u_1(t), u_2(t), \cdots$ 在 $[a, b]$ 上连续，且 $\sum_{k=1}^{\infty} u_k(t)$ 在 $[a, b]$ 上一致

收敛，则 $\forall x_0, x \in [a,b]$

$$\int_{x_0}^{x} \sum_{k=1}^{\infty} u_k(t)dt = \sum_{k=1}^{\infty} \int_{x_0}^{x} u_k(t)dt \tag{D.1}$$

并且上式右端的级数在 $[a,b]$ 上也一致收敛。

❑ 已知函数项级数 $\sum_{k=1}^{\infty} u_k(t)$ 收敛，如果 $u_1'(t), u_2'(t), \cdots$ 连续且 $\sum_{k=1}^{\infty} u_k'(t)$ 一致收敛，则有 $\sum_{k=1}^{\infty} u_k(t)$ 也一致收敛，并且

$$\frac{d}{dt} \sum_{k=1}^{\infty} u_k(t) = \sum_{k=1}^{\infty} \frac{d}{dt} u_k(t) \tag{D.2}$$

判定函数项级数一致收敛常用到如下的方法。

Cauchy 判别法： 函数项级数 $\sum_{k=1}^{\infty} u_k(t)$ 在 t 的某个集合 T 上一致收敛当且仅当 $\forall t \in T$，对任意的 $\epsilon > 0$ 总能找到 $N \in \mathbb{N}$ 使得 $n, m > N$ 时，$|S_n(t) - S_m(t)| < \epsilon$ 成立。

Weierstrass 判别法： 对于一个函数项级数 $\sum_{k=1}^{\infty} u_k(t)$，如果存在一个收敛的正常数项级数 $\sum_{k=1}^{\infty} C_k$ 使得 $|u_k(t)| \leq C_k$，则级数 $\sum_{k=1}^{\infty} u_k(t)$ 一致收敛且绝对收敛。

Abel 判别法： 已知级数 $\sum_{k=1}^{\infty} u_k(t)$ 在 T 上一致收敛且 $C_1, C_2, \cdots$ 是一个递减的正实数序列，则级数 $\sum_{k=1}^{\infty} C_k u_k(t)$ 在 T 上也一致收敛。

Dirichlet 判别法： 已知 $C_1, C_2, \cdots$ 是一个递减的正实数序列且 $\lim_{n\to\infty} C_n = 0$，如果 $\forall t \in T, \forall n \in \mathbb{N}$ 能找到某一常数 C 使得 $|\sum_{k=1}^{n} u_k(t)| \leq C$ 成立，则级数 $\sum_{k=1}^{\infty} C_k u_k(t)$ 在 T 上一致收敛。

附录 E

Riemann-Stieltjes 积分

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，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。

定义 E.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 $a = x_0 < x_1 < x_2 < \cdots < x_n = b$，令 $\xi_j \in [x_j, x_{j+1}]$，如果下面的极限存在

$$\lim_{\max\{x_{j+1}-x_j\}\to 0} \sum_{j=0}^{n-1} g(\xi_j)[u(x_{j+1}) - u(x_j)] \tag{E.1}$$

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

$$S = \int_a^b g(x)du(x) \quad \text{或简记为} \quad S = \int_a^b g du \tag{E.2}$$

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 $u(x) = x$。

性质 E.1. 在有限区间 $[a, b]$ 上，Riemann-Stieltjes 积分具有以下性质。

❶ 如果 g_1, g_2 关于 u 都是 R-S 可积的，则 g_1, g_2 的线性组合关于 u 也是 R-S 可积的，且

$$\int_a^b (cg_1 + dg_2)du = c\int_a^b g_1 du + d\int_a^b g_2 du \tag{E.3}$$

❷ 如果 g 关于 u_1, u_2 都是 R-S 可积的，则 g 关于 u_1, u_2 的线性组合也是 R-S 可积的，且

$$\int_a^b gd(cu_1 + du_2) = c\int_a^b gdu_1 + d\int_a^b gdu_2 \tag{E.4}$$

❸ 如果 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 R-S 可积，且

$$\int_a^b gdu = \int_a^c gdu + \int_c^b gdu \tag{E.5}$$

但其逆命题不成立。

❹ 如果 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

$$\int_a^b gdu = g(b)u(b) - g(a)u(a) - \int_a^b udg \tag{E.6}$$

❺ 如果 g 在 $[a, b]$ 上有界，$m \le g(x) \le M$，并且 u 在 $[a, b]$ 上单调增，则存在 $w \in [m, M]$ 使得

$$\int_a^b gdu = w[u(b) - u(a)] \tag{E.7}$$

这就是 R-S 积分中值定理。特别地，如果 g 在 $[a,b]$ 上连续，则存在 $\xi \in [a,b]$ 使得 $w = f(\xi)$。

❻ 如果 $u(x)$ 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）

$$\int_a^b g(x)du(x) = \int_a^b g(x)u'(x)dx \tag{E.8}$$

❼ 如果 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。

❽ 如果 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

$$\left|\int_a^b g du\right| \le \sup_{x\in[a,b]} |f(x)| \bigvee_a^b (u) \tag{E.9}$$

❾ 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（“一致收敛”见附录 D），则

$$\lim_{n\to\infty} \int_a^b g_n du = \int_a^b g du \tag{E.10}$$

---

*1881 年，法国数学家 Camille Jordan (1838-1922) 定义了闭区间上实函数 $u(x)$ 的一个重要数字特征——变差，具体描述如下。

$$\bigvee_a^b (u) = \sup_{x_0,\cdots,x_n} \sum_{j=1}^n \left|u(x_{j+1}) - u(x_j)\right|$$

其中，$a = x_0 < x_1 < \cdots < x_n = b$ 是对闭区间 $[a,b]$ 的任意分割。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 $\lim\limits_{n\to\infty} \bigvee_{-n}^n(u) < \infty$，则称 u 为 $\mathbb{R}$ 上的有界变差函数。Jordan 分解定理说，u 是有界变差函数的充要条件是存在两个增函数 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 $u \in \bigvee[c,b]$。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 $u(0) = 0$。有界变差函数的不连续点至多可数，并且都是第一类的。

附录 F

可测函数与 Lebesgue 积分

定理 F.1. 以下三种说法与定义 1.7 等价，都可作为可测空间 $(\Omega, \mathscr{S})$ 上可测函数 g 的定义：$\forall r \in \mathbb{R}$，(1) $\{\omega : g(\omega) > r\} \in \mathscr{S}$；(2) $\{\omega : g(\omega) \geq r\} \in \mathscr{S}$；(3) $\{\omega : g(\omega) < r\} \in \mathscr{S}$。

定理 F.2. 已知 g, h 都是可测空间 $(\Omega, \mathscr{S})$ 上可测函数，

❶ 则函数 $|g|, \max(f, g), \min(g, h), g^+ = \max(g, 0), g^- = -\min(g, 0), rg$ 都是可测的，其中 $r \in \mathbb{R}$。

❷ 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，$g + h, g - h, gh$ 是可测的；当 $h \neq 0$ 时，g/h 也是可测的。

❸ 设 $g_n, n = 1, 2, \cdots$ 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*(almost everywhere, *a.e.*) 收敛于 g，则 g 是可测的。

❹ 已知 f 是可测空间 $(\Omega, \mathscr{S})$ 上的可测函数，g 是 $\mathbb{R}$ 上的 Borel 函数，则 $g(f)$ 是 $(\Omega, \mathscr{S})$ 上的可测函数。

1902 年，法国数学家 Henri Léon Lebesgue (1875-1941) 发表了论文《积分、长度和面积》。论文中定义的 Lebesgue 测度和 Lebesgue

*一个性质在集合 S 上“几乎处处”成立，意味着使得此性质不成立的点集是零测集。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。

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 [14,19]。Lebesgue 的论文标志着古典分析过渡到现代分析 [47,79,80]。

简而言之，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，$\mathbb{R}$ 上开集 $\bigcup_{j=1}^{\infty}(a_j,b_j)\subset[a,b]$ 的 Lebesgue 测度等于 $\sum_{j=1}^{\infty}(b_j-a_j)$，闭集 $[a,b]-\bigcup_{j=1}^{\infty}(a_j,b_j)$ 的 Lebesgue 测度为 $(b-a)-\sum_{j=1}^{\infty}(b_j-a_j)$。而为了定义任意集合 $S\subseteq[a,b]$ 的 Lebesgue 测度，我们需要一个辅助概念——外测度。

定义 F.1. 集合 $S\subseteq[a,b]$ 的所有开覆盖，即包含 S 的所有开集，都具有确定的测度，其下确界定义为 S 的外测度，记作 $m^*(S)$。如果下面的关系式成立，则称 S 是 Lebesgue 可测的，其测度定义为 $m(S)=m^*(S)$。

$$m^*(S)+m^*(S^c)=b-a \tag{F.1}$$

该定义可自然推广到 $\mathbb{R}^n$ 上。可以证明，刚体运动不改变 Lebesgue 可测集的测度。$\mathbb{R}$ 上零测集的典型例子有：有理数集 $\mathbb{Q}$（在 $\mathbb{R}$ 上处处稠密）、Cantor 集（具有连续统）等。

定义 F.2 (简单函数)**.** 定义在可测空间 $(\Omega,\mathscr{S})$ 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 $r_1,r_2,\cdots$（两两不等），r_j 的逆像的全体记为 Ω_j，显然 $\bigcup_{j=1}^{\infty}\Omega_j=\Omega$。

定义 F.3 (可积的简单函数)**.** 已知测度空间 $(\Omega,\mathscr{S},\mu)$（见定义 1.12）和定义在 $(\Omega,\mathscr{S})$ 上的简单函数 g，如果级数 $\sum_{j=1}^{\infty}r_j\mu(\Omega_j)$ 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*

$$\int_{\Omega}g\,d\mu=\sum_{j=1}^{\infty}r_j\mu(\Omega_j) \tag{F.2}$$

例 F.1. 考虑定义在闭区间 $[0,1]$ 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，$D(x)$ 的 Riemann 积分不存在。因为 $[0,1]$ 上无理数集合的 Lebesgue 测度为 1 且 $D(x)$ 是可积的简单函数，所以 $D(x)$ 在 $[0,1]$ 上的 Lebesgue 积分等于 1。

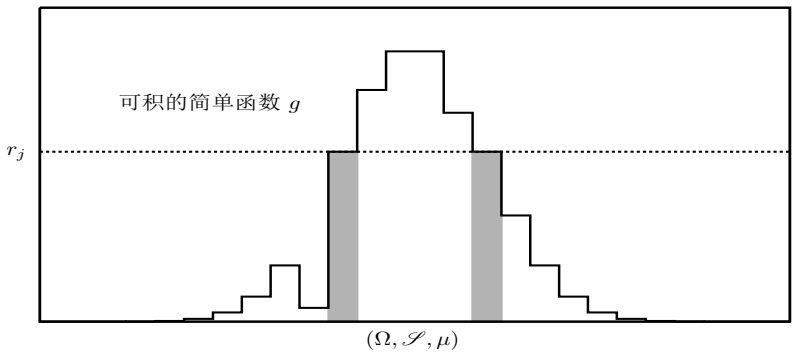


图 F.1: 可积的简单函数 g：把取值为 r_j 那些 ω 所组成的可测集记为 Ω_j，图中阴影部分表示 $r_j\mu(\Omega_j)$。Lebesgue 积分的过程好比计算大量各种面值硬币的总值，先把硬币按面值（如 $r_1, r_2, \cdots$）分堆，相同面值（取值 r_j）的放在一起，然后分别计算它们的个数 $\mu(\Omega_j)$ 并将面值乘以个数得到每堆的值 $r_j\mu(\Omega_j)$，最后将各堆的值加起来得到总值 $\sum_{j=1}^{\infty} r_j\mu(\Omega_j)$。而 Riemann 积分的过程则好比把不同面值的硬币都混在一起用依次累加的方法得到总值。

定义 F.4 (可积函数)**.** 对于函数 $f : \Omega \to \mathbb{R}$，若能找到一个可积的简单函数序列 $g_n, n = 1, 2, \cdots$ 在 Ω 上（可以除去一个零测集）一致收敛于 f 且 $\lim\limits_{n\to\infty} \int_\Omega g_n d\mu < \infty$，则称 f 是可积的，记作 $f \in L^1(\Omega, \mu)$。定义 f 在 Ω 上的 Lebesgue 积分为

$$\int_\Omega f d\mu = \lim_{n\to\infty} \int_\Omega g_n d\mu \tag{F.3}$$

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 $h_n, n = 1, 2, \cdots$ 一致收敛于 f，结果还是一样。

性质 F.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，它具有以下性质。

*泛函 (functional) 即把函数映为实数或复数的映射。

❶ 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。

❷ 如果 $f,g \in L^1(\Omega,\mu)$ 只在一个零测集上不相等，则 $\int_\Omega f d\mu = \int_\Omega g d\mu$。

$$\int_A f d\mu = 0 \Rightarrow \begin{cases} \mu(A) = 0 & \text{若 } f \text{ 在 } A \text{ 上几乎处处为正} \\ f \overset{a.e.}{=} 0 & \text{若 } f \text{ 在 } A \text{ 上几乎处处非负} \end{cases}$$

❸ 如果 $f \in L^1(\Omega,\mu)$，则 $f \in L^1(\Omega,\mu)$ 且

$$\left|\int_\Omega f d\mu\right| \le \int_\Omega |f| d\mu \tag{F.4}$$

❹ 如果 f 有界（设 $m \le f \le M$）且可测，则 $f \in L^1(\Omega,\mu)$ 且

$$m\mu(\Omega) \le \int_\Omega f d\mu \le M\mu(\Omega) \tag{F.5}$$

❺ 如果 $f \in L^1(\Omega,\mu), |g| \le f$ 且 g 可测，则 $g \in L^1(\Omega,\mu)$ 且

$$\left|\int_\Omega g d\mu\right| \le \int_\Omega f d\mu \tag{F.6}$$

❻ 令 I_A 是集合 A 的指示函数（见例 2.3），如果 $A \subseteq \Omega$ 是可测的，则

$$\int_A f d\mu = \int_\Omega f I_A d\mu \tag{F.7}$$

❼ 如果 $f \in L^1(A,\mu)$，则 $f \in L^1(B,\mu)$，其中 $\forall B \subseteq A$。

❽ 如果 $f \in L^1(A,\mu)$，可测集 $A_1, A_2, \cdots, A_n, \cdots$ 是 A 的划分，则

$$\int_A f d\mu = \sum_{n=1}^{\infty} \int_{A_n} f d\mu \tag{F.8}$$

引理 F.1 (Fatou, 1906)**.** 若 $\{f_n : n = 1,2,\cdots\}$ 是一个非负可测函数序列，则 $\int_A (\liminf_{n\to\infty} f_n) d\mu \le \liminf_{n\to\infty} \int_A f_n d\mu$。该式中 $\liminf$ 换成 $\limsup$ 时，不等号 $\le$ 换成 $\ge$。

定义 F.5. 对于式子 (F.3)，当 $\Omega = \mathbb{R}^n$ 时，(1) 若 μ 是 Lebesgue 测度，Lebesgue 积分记作 $\int_{\mathbb{R}^n} f(x)dx$，可积函数的全体记为 $L^1(\mathbb{R}^n)$。(2) 若 μ 不是 Lebesgue 测度，式子 (F.3) 称作 Lebesgue-Stieltjes 积分。

定理 F.3 (Lebesgue 控制收敛定理, 1909)**.** 若在可测集 $A \subseteq \mathbb{R}$ 上，有一列可测函数 $\{f_n(x) : n = 1, 2, \cdots\}$ 几乎处处收敛于 $f(x)$，并且存在可积函数 $g(x)$ 使得 $|f_n(x)| \leq g(x)$，则 $f_n(x)$ 和 $f(x)$ 在 A 上都是可积的，并且还有

$$\lim_{n\to\infty} \int_A f_n(x)dx = \int_A f(x)dx \tag{F.9}$$

附录 G

矩阵计算的若干基本结果

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，尤其对多元统计分析它更是一件不可缺少的工具。该附录中我们仅列出本书所需要的矩阵计算知识，更多的内容见专著 [11,41,51]。在本书中我们约定：零矩阵记为 O，单位阵记为 (identity matrix) 为 I。向量缺省地是列向量，如 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^\mathsf{T} \in \mathbb{R}^n$。向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的内积定义为 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\mathsf{T}\boldsymbol{y} = \sum_{j=1}^n x_j y_j$。$\forall \boldsymbol{x} \neq \boldsymbol{0}$，若方阵 $A_{n\times n}$ 满足 $\boldsymbol{x}^\mathsf{T}A\boldsymbol{x} > 0$ 或等价地，$\langle A\boldsymbol{x}, \boldsymbol{x} \rangle > 0$，则称 A 为正定矩阵；若 $\boldsymbol{x}^\mathsf{T}A\boldsymbol{x} \geq 0$，则称 A 为半正定。

定理 G.1. 方阵 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) 存在非退化的上（下）三角矩阵 Q 使得 $A = Q^\mathsf{T}Q$。(3) 存在可逆矩阵 C 使得 $A = C^\mathsf{T}C$。

定义 G.1. 一个复值函数 $\varphi : \mathbb{R}^n \to \mathbb{C}$ 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 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m \in \mathbb{R}^n, c_1, \cdots, c_m \in \mathbb{C}$ 皆有 $\sum_{i,j=1}^m c_i \bar{c}_j \varphi(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq 0$，或等价地，$A = [\varphi(\boldsymbol{x}_i - \boldsymbol{x}_j)]_{m\times m}$ 为半正定矩阵。

定理 G.2. 对于任意的 $\boldsymbol{x} \in \mathbb{R}^n$ 和 n 阶方阵 A 有 $\boldsymbol{x}^\mathsf{T}A\boldsymbol{x} = \mathrm{tr}(A\boldsymbol{x}\boldsymbol{x}^\mathsf{T})$。用 Maxima 验证 $n = 2$ 时，定理成立。

```
(%i1) load ("functs") $
(%i2) A : genmatrix (a, 2, 2) $
(%i3) x : matrix ([u], [v]) $
(%i4) is(equal(tracematrix (A . x . transpose(x)), transpose(x) . A . x));
(%o4)                                true
```

定义 G.2. 令 $f(\boldsymbol{x})$ 是关于 $\boldsymbol{x}=(x_1,\cdots,x_n)^{\mathsf{T}}$ 的一个纯量函数，在 $\Omega\subseteq\mathbb{R}^n$ 上有定义且可微，f 的*梯度*定义为

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}=\left(\frac{\partial f}{\partial x_1},\cdots,\frac{\partial f}{\partial x_n}\right)^{\mathsf{T}} \tag{G.1}$$

也常记作 $\mathrm{grad}f$ 或 $\nabla_{\boldsymbol{x}}f$ 或 ∇f。

定义 G.3. 令 $\boldsymbol{y}=(y_1,\cdots,y_m)^{\mathsf{T}}$ 中的每个分量都是 $\boldsymbol{x}\in\mathbb{R}^n$ 的纯量函数，在 $\Omega\subseteq\mathbb{R}^n$ 上有定义且可微，定义函数 $F:\boldsymbol{x}\mapsto\boldsymbol{y}$ 的*雅可比矩阵* J_F 为

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}=\left(\frac{\partial y_i}{\partial x_j}\right)_{m\times n} \tag{G.2}$$

其中，$\partial y_i/\partial x_j$ 是 $m\times n$ 方阵 $\partial\boldsymbol{y}/\partial\boldsymbol{x}$ 的 (i,j) 元素。例如，

```
(%i1) jacobian ([sin (u + v), u * sin (v)], [u, v]);
                          [ cos(v + u)  cos(v + u) ]
(%o1)                     [                        ]
                          [   sin(v)     u cos(v)  ]
```

性质 G.1. 对于任意 n 阶方阵 A 有 $\partial(A\boldsymbol{x})/\partial\boldsymbol{x}=A$。

证明. 列向量 $\boldsymbol{y}=A\boldsymbol{x}$ 的第 i 个元素为 $a_{i1}x_1+\cdots+a_{ij}x_j+\cdots+a_{in}x_n$，所以 $\partial\boldsymbol{y}/\partial\boldsymbol{x}$ 的 (i,j) 元素为 a_{ij}，得证。 □

定理 G.3. 已知函数 $F:\boldsymbol{x}\mapsto\boldsymbol{y}$ 在点 $\boldsymbol{p}\in\mathbb{R}^n$ 可微，则在 $\boldsymbol{p}$ 的足够小的邻域里 $F(\boldsymbol{x})$ 可由线性函数来近似，即 $F(\boldsymbol{x})\approx F(\boldsymbol{p})+J_F(\boldsymbol{p})(\boldsymbol{x}-\boldsymbol{p})$。

定理 G.4. 令 $f(\boldsymbol{x})$ 如下表左列所定义，其中 $\boldsymbol{c}=(c_1,\cdots,c_n)^{\mathsf{T}}$ 为一个 n 维向量，令 $S_{n\times n}$ 为任意 n 阶对称阵，则有下面的结果：

$f(\boldsymbol{x})$	$\partial f(\boldsymbol{x})/\partial \boldsymbol{x}$	$\partial^2 f(\boldsymbol{x})/\partial \boldsymbol{x}^2$
$\boldsymbol{x}^\mathsf{T}\boldsymbol{c}$ or $\boldsymbol{c}^\mathsf{T}\boldsymbol{x}$	$\boldsymbol{c}$	O
$\boldsymbol{x}^\mathsf{T}\boldsymbol{x}$	$2\boldsymbol{x}$	$2I$
$\boldsymbol{x}^\mathsf{T}S\boldsymbol{x}$	$2S\boldsymbol{x}$	$2S$

证明. 记 S 的 (i,j) 元素为 s_{ij}，往证 $\partial(\boldsymbol{x}^\mathsf{T}S\boldsymbol{x})/\partial\boldsymbol{x} = 2S\boldsymbol{x}$ 如下：

$$\boldsymbol{x}^\mathsf{T}S\boldsymbol{x} = \sum_{i,j=1}^{n} x_i s_{ij} x_j \Rightarrow \frac{\partial(\boldsymbol{x}^\mathsf{T}S\boldsymbol{x})}{\partial x_k} = \sum_{j=1}^{n} s_{kj}x_j + \sum_{i=1}^{n} x_i s_{ik} = 2\sum_{j=1}^{n} s_{kj}x_j$$

其中 $k = 1, 2, \cdots, n$。其他结论请读者作为练习自己证明。 □

定义 G.4. 下面的方阵称为 *海赛矩阵* (Hessian matrix)：

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}^2} = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{n\times n} \tag{G.3}$$

```
(%i1) hessian (u * sin (v), [u, v]);
                              [   0       cos(v)   ]
(%o1)                         [                    ]
                              [ cos(v)  - u sin(v) ]
```

海赛矩阵是以德国数学家 Ludwig Otto Hesse (1811-1874) 命名的，有时也简记作 $H(f)$ 或 $\nabla_{\boldsymbol{x}}^2 f$ 或 $\nabla^2 f$。海赛矩阵常用于近似计算

$$f(\boldsymbol{x}+\Delta\boldsymbol{x}) \approx f(\boldsymbol{x}) + \Delta\boldsymbol{x}^\mathsf{T}\nabla f + \frac{1}{2}\Delta\boldsymbol{x}^\mathsf{T}\nabla^2 f \Delta\boldsymbol{x} \tag{G.4}$$

例如，当非负的可积函数 $g(\boldsymbol{x})$ 的积分 $\int_{\mathbb{R}^n} g(\boldsymbol{x})d\boldsymbol{x}$ 难以精确求解时，可以应用 Laplace *近似求解法*：若在 $\boldsymbol{x} = \boldsymbol{x}_0$ 处 $\nabla g(\boldsymbol{x}_0) = \boldsymbol{0}$，则 $\ln g(\boldsymbol{x}) \approx \ln g(\boldsymbol{x}_0) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\mathsf{T}A(\boldsymbol{x}-\boldsymbol{x}_0)$，其中 $A = -\nabla^2 \ln g(\boldsymbol{x}_0)$。进而近似地有，非负 n 元实值函数 $\sqrt{|A|}g(\boldsymbol{x})[(2\pi)^{n/2}g(\boldsymbol{x}_0)]^{-1}$ 是多元正态分布 $\mathrm{N}_n(\boldsymbol{0}, A^{-1})$ 的密

度函数，于是

$$\int_{\mathbb{R}^n} g(\boldsymbol{x})d\boldsymbol{x} \approx g(\boldsymbol{x}_0)\frac{(2\pi)^{n/2}}{\sqrt{|A|}} \tag{G.5}$$

定理 G.5. 令 $D \subseteq \mathbb{R}^n$ 为开凸集（见附录 H），若实值函数 $f(\boldsymbol{x})$ 在 D 上存在一阶和二阶偏导数且 $\forall \boldsymbol{x} \in D$ 皆有 $H(\boldsymbol{x}) = -\nabla^2 f$ 是正定矩阵，则方程组 $\nabla f = \boldsymbol{0}$ 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 $\nabla f = \boldsymbol{0}$ 在 D 内有两个解 $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$，则函数 $g(t) = f[t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2]$ 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 $g''(t_0) = -(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathsf{T}} H[t_0\boldsymbol{x}_1 + (1-t_0)\boldsymbol{x}_2](\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$。由于 $\forall \boldsymbol{x} \in D, H(\boldsymbol{x})$ 是正定矩阵，所以 $(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{T}} H[t_0\boldsymbol{x}_1+(1-t_0)\boldsymbol{x}_2](\boldsymbol{x}_1-\boldsymbol{x}_2) > 0$，矛盾！于是至多有一个解。若 $\boldsymbol{x}_0$ 是 $\nabla f = \boldsymbol{0}$ 的解，$\forall \boldsymbol{x} \in D$，函数 $h(t) = f[t\boldsymbol{x} + (1-t)\boldsymbol{x}_0]$ 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，$f(\boldsymbol{x}) = h(1) < h(0) = f(\boldsymbol{x}_0)$。 □

定理 G.6 (奇异值分解)**.** 对称阵 $S_{n\times n}$ 具有以下的分解：

$$S = U\Lambda U^{\mathsf{T}} \tag{G.6}$$

其中，U 是一个正交方阵（即 $U^{\mathsf{T}}U = I$），$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ 是一个对角阵。显然，该分解等价于 $SU = U\Lambda$，进而等价于

$$S\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \tag{G.7}$$

其中，$\boldsymbol{u}_i$ 是 U 的第 i 列向量。即，λ_i 是 S 的特征值，$\boldsymbol{u}_i$ 是对应的特征向量。

附录 H

凸性与 Jensen 不等式

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形式，在应用中非常实用。例如，证明推广了的 Rao-Blackwell 定理、Kullback-Leibler 信息量非负等都需要用到 Jensen 不等式。

定义 H.1 (凸集)**.** 集合 $S \subset \mathbb{R}^d$ 中的任意两点 $\boldsymbol{x}, \boldsymbol{y} \in S$ 若满足 $\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y} \in S, \forall \alpha \in [0,1]$，则称 S 为凸集 (convex set)。它的几何含义是：以 S 上任意两点为端点的线段仍在 S 上。

定义 H.2 (凸函数)**.** 定义在凸集 S 上的实值函数 $g : S \to \mathbb{R}$ 若满足 $\forall \boldsymbol{x}, \boldsymbol{y} \in S, \forall \alpha \in [0,1]$ 皆有 $g[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \leq \alpha g(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{y})$，则称 g 为 S 上的凸函数。它的几何含义是：线段 $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}$ 在 g 下的像在线段 $\alpha g(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{y})$ 的下方。

性质 H.1. 凸函数具有如下的一些性质 [74]：

❶ 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。

❷ 若 $d > 1$，$g(\boldsymbol{x})$ 是凸函数当且仅当 $\forall \boldsymbol{u} \in \mathbb{R}^d, \forall \boldsymbol{x} \in S$ 皆有 $\boldsymbol{u}^{\mathsf{T}} \frac{\partial^2 g(\boldsymbol{x})}{\partial \boldsymbol{x}^2} \boldsymbol{u} > 0$。

❸ 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。

❹ 凸性在仿射变换下不变：若 $f(\boldsymbol{x})$ 是凸函数，其中 $\boldsymbol{x} \in \mathbb{R}^m$，则 $g(\boldsymbol{y}) = f(A_{m\times n}\boldsymbol{y} + \boldsymbol{c})$ 也是凸函数，其中 $\boldsymbol{y} \in \mathbb{R}^n$ 且 $\boldsymbol{c} \in \mathbb{R}^m$。

定理 H.1 (Jensen 不等式的离散版)**.** 若 $g(x)$ 是凸函数，则

$$g\left(\sum_{j=1}^{n} \alpha_j x_j\right) \le \sum_{j=1}^{n} \alpha_j g(x_j) \tag{H.1}$$

其中 $\alpha_j \ge 0$ 满足 $\sum_{j=1}^{n} \alpha_j = 1$。等号成立当且仅当 $x_1 = x_2 = \cdots = x_n$ 或 $g(x)$ 是线性函数。

证明. 显然 $n = 2$ 时成立。设 $n \le k$ 时都成立，不妨设 $\alpha_1 \ne 1$

$$\begin{aligned} g\left(\sum_{j=1}^{k+1} \alpha_j x_j\right) &= g\left[\alpha_1 x_1 + (1-\alpha_1)\sum_{j=2}^{k+1} \frac{\alpha_j}{1-\alpha_1} x_j\right] \\ &\le \alpha_1 g(x_1) + (1-\alpha_1) g\left[\sum_{j=2}^{k+1} \frac{\alpha_j}{1-\alpha_1} x_j\right] \\ &\le \alpha_1 g(x_1) + \alpha_2 g(x_2) + \cdots + \alpha_{k+1} g(x_{k+1}) \end{aligned}$$ □

练习 H.1. 由 Jensen 不等式证明 Hölder 不等式 (2.64)。

由 Jensen 不等式 (H.1) 可直接推出“算术与几何平均不等式”：$\sqrt[n]{z_1 z_2 \cdots z_n} \le \frac{1}{n}(z_1 + z_2 + \cdots + z_n)$，其中 $z_1, z_2, \cdots, z_n$ 非负。公式 (H.1) 首先由 O. Hölder 于 1889 年提出 [48]，下面的结果才是 Jensen 于 1906 年提出并证明了的。

定理 H.2 (Jensen 不等式的连续版, 1906)**.** 若 $g(x)$ 是 $S \subset \mathbb{R}$ 上凸函数，则

$$g\left[\int_D \lambda(t)x(t)dt\right] \le \int_D \lambda(t) g[x(t)]dt \tag{H.2}$$

其中 $x(D) \subseteq S$ 且 $\forall t \in D, \lambda(t) \ge 0$ 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。

定理 H.3 (Jensen 不等式的数学期望版)**.** 已知 $g(\boldsymbol{x})$ 是 $S \subset \mathbb{R}^d$ 上的凸函数，d 维随机向量 $\boldsymbol{X}$ 有有限期望 $\mathsf{E}\boldsymbol{X}$ 且 $\mathsf{P}(\boldsymbol{X} \in S) = 1$，则 $\mathsf{E}\boldsymbol{X} \in S$，随机变量 $g(\boldsymbol{X})$ 的期望存在并且 $g(\mathsf{E}\boldsymbol{X}) \leq \mathsf{E}[g(\boldsymbol{X})]$。等号成立当且仅当存在 $c \in \mathbb{R}$ 和 $\boldsymbol{w} \in \mathbb{R}^d$ 使得 $\mathsf{P}\{g(\boldsymbol{X}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X} + c\} = 1$。例如，$g(x) = x^2$ 在 $\mathbb{R}$ 上是凸函数，则 $(\mathsf{E}X)^2 \leq \mathsf{E}(X^2)$。

推论 H.1. 如果非常数的随机变量 $X > 0$ 具有有限期望，则 $[\mathsf{E}(X)]^{-1} \leq \mathsf{E}(X^{-1})$ 且 $\mathsf{E}(\ln X) \leq \ln[\mathsf{E}(X)]$。

定义 H.3 (Kullback-Leibler 信息量)**.** 为刻画两个密度函数 $f(\boldsymbol{x})$ 和 $g(\boldsymbol{x})$ 之间的相似程度，1968 年美国密码专家兼数学家 Solomon Kullback (1907-1994) 定义了 Kullback-Leibler 信息量 (information divergence) 或相对熵 (relative entropy) 或判别信息量如下：

$$K(f,g) = \mathsf{E}_f\left[\ln\frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right] = \int_{\mathbb{R}^d} f(\boldsymbol{x})\ln\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}d\boldsymbol{x} \tag{H.3}$$

有的文献中将 Kullback-Leibler 信息量称为“Kullback-Leibler 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 信息量不满足对称性，即 $K(f,g) \neq K(g,f)$。下述结果对第十三章的 EM 算法很重要。

定理 H.4. 利用推论 H.1 可证 Kullback-Leibler 信息量非负，即 $K(f,g) \geq 0$。等号成立当且仅当 $f(\boldsymbol{x}) = g(\boldsymbol{x})$。

证明. $K(f,g) = -\mathsf{E}_f\{\ln[g(\boldsymbol{X})/f(\boldsymbol{X})]\} \geq -\ln\mathsf{E}_f[g(\boldsymbol{X})/f(\boldsymbol{X})] = 0$。 □

附录 I

习题答案或提示

1.1 $1-(5/6)^3=91/216$。

1.2 令 A 表示事件“点数之和为 7 或 11”，则 $\mathsf{P}(A)=8/36=2/9$。所求的概率为 $[1-\mathsf{P}(A)]^2=49/81$。

1.3 $0.2^3+3\times 0.8\times 0.2^2=0.104$。

1.4 令 A_m 表示“取到的 k 个球中最大编号是 m”，k 个球的最大编号不超过 m 共有 m^k 种可能，所以事件 A_m 包含的基本事件个数为 $m^k-(m-1)^k$，基本事件的总数为 n^k，故 $\mathsf{P}(A_m)=m^k-(m-1)^k/n^k$。

1.5 $1-(1-P)^3=0.875$，解得 $P=0.5$。

1.6 $1-C_{10}^4 2^4/C_{20}^4=99/323$。

1.7 令 A 表示“该夫妇坐在一起”，$\mathsf{P}(A)=C_{10}^1C_2^1A_8^8/A_{10}^{10}=2/9$。

1.8 $C_n^2 n!/n^n$。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。

1.9 令 A_i 表示“第 i 个人拿到第 i 号球”，$i=1,2,\cdots,n$，则有 $\mathsf{P}(A_i)=(n-1)!/n!=1/n,\mathsf{P}(A_iA_j)=(n-2)!/n!=1/A_n^2$（其中 $i\neq j$），

$\cdots, \mathsf{P}(A_1A_2\cdots A_n) = 1/n!$。所求概率为

$$\begin{aligned}\mathsf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) &= \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{1\le i\le j\le n} \mathsf{P}(A_iA_j) + \cdots \\ &+(-1)^{n-1}\mathsf{P}(A_1A_2\cdots A_n) = C_n^1 1/n - C_n^2 1/A_n^2 + \cdots + (-1)^{n-1}C_n^n 1/A_n^n \\ &= 1 - 1/2! + \cdots + (-1)^{n-1}1/n! = \sum_{k=1}^{n}(-1)^{k-1}1/k!\end{aligned}$$

1.10 由式 (1.10)，$\mathsf{P}(A_k) = C_m^k A_{Np}^k A_{Nq}^{m-k}/A_N^m$，其中 $p = n/N, q = 1 - p$。

1.11 令 B 表示“第 k 次摸出的是白球”。(1) $\mathsf{P}(B) = w(w + b - 1)!/(w + b)! = w/(w + b)$；(2) $\mathsf{P}(B) = C_{w+b-1}^{w-1}/C_{w+b}^{w} = w/(w + b)$。

1.12 盒子装有可辨的 m 个黑球和 n 个白球，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。

1.13 见附录 A 中的例 A.4。

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“$x + y \le 6/5$”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 $i = 1, 2, \cdots, n$。由 $\mathsf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - \mathsf{P}(A_1^c A_2^c \cdots A_n^c) = 1 - 0.4^n \ge 0.95$ 解得 $n \ge 3.27$，故至少要使用 4 位情报员。

1.17 $\mathsf{P}(ABC) = 0, \mathsf{P}(A + B + C) = 5/8, \mathsf{P}(A^cB^cC) = 1/8, \mathsf{P}(ABC + A^cBC + AB^cC + ABC^c) = 1/8$。

1.18 解法 1：n 次试验中 A 至少出现一次的概率为 $1-[1-P(A)]^n$，随着 n 的增加而趋于 1。解法 2：试验无穷多次，每个基本事件都是 A 和 A^c 构成的无限字符串，基本事件集合与 $[0,1]$ 区间上的二进制小数可一一对应，事件“A 总不发生”仅对应着一个点。解法 3：前 $n-1$ 次试验 A 不发生，第 n 次试验 A 发生的概率是 $\mathsf{P}(n)=[1-\mathsf{P}(A)]^{n-1}\mathsf{P}(A)$。显然 $\sum_{n=1}^{\infty}\mathsf{P}(n)=1$。

1.19 令 A 表示“期中考试及格”，B 表示“期末考试及格”，则 $\mathsf{P}(B)=\mathsf{P}(A)\mathsf{P}(B|A)+\mathsf{P}(A^c)\mathsf{P}(B|A^c)=p^2+(1-p)p/2=p/2+p^2/2$。(1) $\mathsf{P}(A\cup B)=\mathsf{P}(A)+\mathsf{P}(B)-\mathsf{P}(AB)=p+p/2+p^2/2-p^2=3p/2-p^2/2$；(2) $\mathsf{P}(A|B)=\mathsf{P}(AB)/\mathsf{P}(B)=2p/(1+p)$。

1.20 当 $AB=\emptyset$ 时，$\mathsf{P}(AB)-\mathsf{P}(A)\mathsf{P}(B)$ 最小，此时有 $\mathsf{P}(AB)-\mathsf{P}(A)\mathsf{P}(B)=-\mathsf{P}(A)\mathsf{P}(B)\geq-[\mathsf{P}(A)+\mathsf{P}(B)]^2/4=-[\mathsf{P}(A+B)]^2/4\geq-1/4$。

1.21 提示：利用 Borel-Cantelli 引理 1.1。

1.22 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 $\mathsf{P}(B_2|B_1)=\mathsf{P}(B_1B_2)/\mathsf{P}(B_1)$。

1.23 由 $\mathsf{P}(B|A)=\mathsf{P}(B|A^c)$ 可得 $\mathsf{P}(AB)/\mathsf{P}(A)=\mathsf{P}(A^cB)/\mathsf{P}(A^c)=[\mathsf{P}(B)-\mathsf{P}(AB)]/[1-\mathsf{P}(A)]$，整理后即证得充分性。

1.24 若 $A_j\in\mathscr{S}$ 两两互斥，则 $\bigcup_{j=1}^{\infty}A_j+(\bigcup_{j=1}^{\infty}A_j)^c=\Omega=(\bigcup_{j=1}^{\infty}A_j)^c+A_1+A_2+\cdots$，于是 $\mathsf{P}(\bigcup_{j=1}^{\infty}A_j)+\mathsf{P}[(\bigcup_{j=1}^{\infty}A_j)^c]=1=\mathsf{P}[(\bigcup_{j=1}^{\infty}A_j)^c]+\mathsf{P}(A_1)+\mathsf{P}(A_2)+\cdots$，得证。

1.25 利用 $\sum_{k=1}^{n}A_k$ 的非交并分解、推论 1.3 和 Boole 不等式 (1.32)。

$$\mathsf{P}\left(\sum_{k=1}^{n}A_k\right)=\sum_{k=1}^{n}\mathsf{P}\left(A_k-\sum_{k<j}A_kA_j\right)=\sum_{k=1}^{n}\mathsf{P}(A_k)-\mathsf{P}\left(\sum_{k<j}A_kA_j\right)$$

1.26 $\mathsf{P}(A_k)=1/6\cdot(5/6)^{k-1}$ 且 $A=\sum_{k=1}^{\infty}A_k$，于是 $\mathsf{P}(A)=1$。

1.27 $P_k = C_{2n-k}^{n-k} p^{n+1} q^{n-k} + C_{2n-k}^{n-k} q^{n+1} p^{n-k}$

1.28 假设两场比赛为 1 轮，则甲在任意一轮比赛中获得 1 分和 2 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 $\mathsf{P}(A_k) = \alpha^2(2\alpha\beta)^{k-1}$，其中 $k = 1, 2, \cdots$。所以甲获得奖牌的概率为 $\sum_{k=1}^{\infty} \mathsf{P}(A_k) = \sum_{k=1}^{\infty} \alpha^2(2\alpha\beta)^{k-1} = \alpha^2/(1 - 2\alpha\beta)$。乙获得奖牌的概率为 $1 - \alpha^2/(1 - 2\alpha\beta) = \beta^2/(1 - 2\alpha\beta)$。

1.29 令 $A_t^{t+\Delta t}$ 表示“在 $(t, t + \Delta t]$ 内不与其它分子碰撞”，则 $A_{t+\Delta t} = A_t A_t^{t+\Delta t}$。由乘法公式得到 $\mathsf{P}(A_{t+\Delta t}) = \mathsf{P}(A_t)[1 - \lambda\Delta t - o(\Delta t)]$，将 $\mathsf{P}(A_t)$ 作为变量 t 的函数，令 $\Delta t \to 0$ 可得到微分方程 $d\mathsf{P}(A_t)/dt = -\lambda\mathsf{P}(A_t)$，解此方程即得 $\mathsf{P}(A_t) = e^{-\lambda t}$。

1.30 令 A 表示“第 k 次摸到黑球”，则 $\mathsf{P}(A) = 1 - \mathsf{P}(A^c) = 1 - [(n - 1)/n]^{k-1}/n$。

1.31 令 A 表示“盒子里原装的是白球”，令 B_1 表示“取出的是白球”，B_2 表示“盒子里剩下的球也是白球”，则 $\mathsf{P}(A) = 1/2, \mathsf{P}(B_1) = \mathsf{P}(A)\mathsf{P}(B_1|A) + \mathsf{P}(A^c)\mathsf{P}(B_1|A^c) = 3/4, \mathsf{P}(B_1B_2) = 1/2$，故所求概率为 $\mathsf{P}(B_2|B_1) = \mathsf{P}(B_1B_2)/\mathsf{P}(B_1) = 2/3$。

1.32 $m/m + n2^r$。提示：利用贝叶斯公式。

1.33 令 D 表示“放回后仍为 MAXIMA”，令 H_1 表示“脱落的两字母相同”，H_2 表示“脱落的两字母不同”。所求概率为 $\mathsf{P}(D) = \mathsf{P}(H_1)\mathsf{P}(D|H_1) + \mathsf{P}(H_2)\mathsf{P}(D|H_2) = C_2^1/C_6^2 \times 1 + (1 - C_2^1/C_6^2) \times 1/2 = 17/30$。

1.34 记 B_k 表示事件“两次故障间共生产 k 件正品”，A_n表示“两次故障间共生产 n 件产品”，其中 $n = 0, 1, 2, \cdots$。则事件 $A_0, A_1, A_2, \cdots$ 是基本事件集合的一个划分，利用全概率公式求出 $\mathsf{P}(B_k) = \sum_{n=k}^{\infty} C_n^k p^k(1-p)^{n-k}\lambda^n e^{-\lambda}/n! = (\lambda p)^k e^{-\lambda p}/k!$（当 $n < k$ 时，$\mathsf{P}(B_k|A_n) =$

0）。在求解程中用到了 e^x 在 $x=0$ 处的幂级数展开 $e^x = 1 + x + x^2/2! + \cdots + x^n/n! + \cdots$。当 $m < k$ 时，$\mathsf{P}(A_m|B_k) = 0$；当 $m \geq k$ 时，$\mathsf{P}(A_m|B_k) = (\lambda q)^{m-k} e^{-\lambda q}/(m-k)!$，其中 $q = 1 - p$。

1.35 (1) 4/15，(2) 5/13。

1.36 令 A_k 表示“取到盒子 A_k”，$k = 0, 1, \cdots, N$，由题意知 $\mathsf{P}(A_k) = 1/(N+1)$。令 B_n 表示“任取一个盒子，连续 n 次有放回抽取均为黑球”，令 A_{kn} 表示“在盒子 A_k 中进行了 n 次有放回地抽取均为黑球”。于是，$\mathsf{P}(B_n|A_k) = \mathsf{P}(A_{kn}) = (k/N)^n$。由全概率公式可得 $\mathsf{P}(B_n) = \sum_{k=0}^{N} \mathsf{P}(A_k)\mathsf{P}(B_n|A_k) \approx 1/(n+1)$。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，$\mathsf{P}(B_{n+1}) \approx 1/(n+2)$。故所求概率为 $\mathsf{P}(B_{n+1}|B_n) = \mathsf{P}(B_n B_{n+1})/\mathsf{P}(B_n) \approx (n+1)/(n+2)$。

1.37 下面只给出第一个不等式的证明，第二个雷同。根据性质 1.7，

$$\begin{aligned}
\mathsf{P}\left\{\bigcup_{j=1}^{\infty} A_j B_j\right\} &= \mathsf{P}(A_1B_1) + \mathsf{P}[(A_1B_1)^c A_2B_2] + \mathsf{P}[(A_1B_1)^c (A_2B_2)^c A_3B_3] + \cdots \\
&\geq \mathsf{P}(A_1B_1) + \mathsf{P}(A_1^c A_2B_2) + \mathsf{P}(A_1^c A_2^c A_3B_3) + \cdots \\
&\geq \mathsf{P}(A_1)\mathsf{P}(B_1) + \mathsf{P}(A_1^c A_2)\mathsf{P}(B_2) + \mathsf{P}(A_1^c A_2^c A_3)\mathsf{P}(B_3) + \cdots \geq \alpha \mathsf{P}\left\{\bigcup_{j=1}^{\infty} A_j\right\}
\end{aligned}$$

1.38 $\mathsf{P}(\bigcap_{j=1}^{\infty} A_j)^c = \mathsf{P}(\bigcup_{j=1}^{\infty} A_j^c) = \lim\limits_{n\to\infty} \mathsf{P}(\bigcup_{j=1}^{n} A_j^c) \leq \lim\limits_{n\to\infty} \sum_{j=1}^{n} \mathsf{P}(A_j^c) = 0$。

2.1 所求分布列为 $\mathsf{P}\{X = k\} = (1-p)^{k-1} p$，其中 $k = 1, 2, \cdots$。

2.2 所求分布列为 $X \sim \frac{1}{64}\langle 1\rangle + \frac{7}{64}\langle 2\rangle + \frac{19}{64}\langle 3\rangle + \frac{37}{64}\langle 4\rangle$。

2.3 $\mathsf{P}\{X = k\} = A_4^{k-1} C_4^1 / A_8^k$，其中 $k = 1, 2, 3, 4, 5$。

2.4 当 $k < (n+1)p$ 时，$\mathsf{P}\{X = k\}/\mathsf{P}\{X = k-1\} > 1$；当 $k = (n+1)p$ 时，$\mathsf{P}\{X = k\}/\mathsf{P}\{X = k-1\} = 1$；当 $k > (n+1)p$ 时，$\mathsf{P}\{X = k\}/\mathsf{P}\{X = k-1\} < 1$。故若 $(n+1)p$ 为整数，则当 $k = (n+1)p$ 或

$k=(n+1)p-1$ 时，$\mathsf{P}\{X=k\}$ 都取到最大值；若 $(n+1)p$ 不是整数，则当 $k=[(n+1)p]$ 时，$\mathsf{P}\{X=k\}$ 取到最大值。

2.5 $\mathsf{P}\{X=0\}=(1-p)^2=1-\mathsf{P}\{X\geq 1\}=1-5/9=4/9$，从而 $p=1/3$。$\mathsf{P}\{Y\geq 1\}=1-\mathsf{P}\{Y=0\}=1-(1-1/3)^3=19/27$。

2.6 (1) $p=(1+e^{-2\lambda})/2$；(2) $p=(1+e^{-6})/2$。

2.7 分布函数为 $F(x)=\begin{cases}0 & \text{当 } x<0\\ 1-(2+x)\sqrt{1-x}/2 & \text{当 } 0\leq x<1\\ 1 & \text{当 } x\geq 1\end{cases}$

2.8 验证 $F(x)$ 满足分布函数的三条性质（见定理 2.4）：(1) 对于任意实数 $x_1<x_2$ 皆有 $F(x_1)=aF_1(x_1)+bF_2(x_1)\leq aF_1(x_2)+bF_2(x_2)=F(x_2)$，即 $F(x)$ 单调不减；(2) 由于 $F_1(x)$ 和 $F_2(x)$ 右连续性，所以 $F(x)$ 也右连续；(3) $F(-\infty)=0, F(+\infty)=1$ 也是显然的。

2.9 验证 $G(x)$ 满足分布函数的三条性质。

2.10 $X\sim \mathrm{U}[0,1]$。

2.11 (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) $1-(e^{-2}+e^{-1})/2$。

2.12 由 $\sum_{k=1}^{+\infty}1/(ck!)=1/c[(\sum_{k=0}^{+\infty}1/k!)-1]=1/c(e-1)=1$，得 $c=e-1$。

2.13 (1) $a=1,b=-1$; (2) $f(x)=\begin{cases}xe^{-x^2/2} & \text{当 } x>0\\ 0 & \text{当 } x\leq 0\end{cases}$ (3) $\mathsf{P}(1<X<2)=F(2)-F(1)=e^{-1/2}-e^{-2}\approx 0.4712$。

2.14 $f_Y(y)=\begin{cases}|y-1|/10 & \text{当 } -3<y<3\\ 0 & \text{其它}\end{cases}$

2.15 (1) 当 $y>0$ 时，$f_Y(y)=f_X(\ln y)|(\ln y)'|=\exp\{-(\ln y)^2/2\}/(y\sqrt{2\pi})$；当 $y\leq 0$ 时，$f_Y(y)=0$。(2) 当 $z>0$ 时，$f_Z(z)=f_X(z^2)|(z^2)'|+f_X(-z^2)|(-z^2)'|=4z\exp\{-z^4/2\}/\sqrt{2\pi}$；当 $z\leq 0$ 时，$f_Z(z)=0$。

2.16 $f(x)=k\exp\{-(x+1)^2/2+1/2\}$ 在 $\mathbb{R}$ 上积分为 1，所以 $k=1/\sqrt{2\pi e}$。

2.17 首先 $f(x)\geq 0$。当 $c>0$时，令 $u=x^2/(2c)$，则 $\int_{-\infty}^{+\infty}f(x)dx=\int_0^{+\infty}e^{-u}du=1$。

2.18 当 $y\in[0,1]$ 时，$F_Y(y)=\mathsf{P}\{Y\leq y\}=\mathsf{P}\{1-e^{-2X}\leq y\}=\mathsf{P}\{X\leq -1/2\ln(1-y)\}=F_X\{-1/2\ln(1-y)\}=y$。

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 $X\sim \mathrm{U}[-2,2]$。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，只有 $1<r\leq 2$ 或 $r>2$ 可考虑，无论哪种情况都有 $r=2$。

2.21 所求密度函数为 $f_Z(z)=\begin{cases}1-z/2 & \text{当 } 0<z<2\\ 0 & \text{其他}\end{cases}$

2.22 (4) $Z=\max(X,Y)\sim \frac{1}{10}\langle -1\rangle+\frac{1}{5}\langle 1\rangle+\frac{7}{10}\langle 2\rangle$

2.23 $\mathsf{P}\{X>0,Y<0\}=1/3$。

2.24 (1) $a=1/2,b=1/\pi$；(2) $\mathsf{P}\{X\geq 0,Y\geq 0\}=9/32$。

2.25 (1) 利用 $f(x,y)$ 在 $\mathbb{R}^2$ 上积分等于 1 得到 $k=1$；(2) $\mathsf{P}\{X<2,Y<2\}=(1-e^{-2})^2$。

2.26 (1) 求得边缘分布 $f_X(x)$ 和 $f_Y(y)$，从 $f_X(x)f_Y(y)=f(x,y)$ 判定 X 与

Y 相互独立。(2) 计算 $F_Z(z) = \mathsf{P}\{Z \le z\} = \mathsf{P}\{X + Y \le z\}$ 得到

$$F_Z(z) = \iint\limits_{x+y\le z} f(x,y)dxdy = \begin{cases} 0 & \text{当 } z < 0 \\ z - 1 + e^{-z} & \text{当 } 0 \le z \le 1 \\ 1 + (1-e)e^{-z} & \text{当 } z > 1 \end{cases}$$

(3) $\mathsf{P}\{Z > 3\} = 1 - \mathsf{P}\{Z \le 3\} = (e-1)e^{-3}$。

2.27 提示：随机变量 $X_1/\sum_{j=1}^n X_j, \cdots, X_n/\sum_{j=1}^n X_j$ 独立同分布。

2.28 提示：由 $0 \le \mathsf{E}(X) \le 1$ 和 $X^2 \le X$ 得出 $\mathsf{V}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 \le \mathsf{E}(X) - [\mathsf{E}(X)]^2 \le 1/4$ 。若要等号成立当且仅当 $\mathsf{P}(X = 0) = \mathsf{P}(X = 1) = 1/2$。

2.29 $\mathsf{P}\{X \ge a\} = \mathsf{P}\{Y \ge e^{\lambda a}\} = \int_{e^{\lambda a}}^{\infty} dF_Y(y) \le \int_{e^{\lambda a}}^{\infty} ye^{-\lambda a}dF_Y(y) \le e^{-\lambda a}\mathsf{E}(Y)$。

2.30 令 $Y = X - \mu$，则 $\mathsf{E}(|Y|) = 2\int_0^{\infty} y\phi(y|0,\sigma^2)dy = \sigma\sqrt{2/\pi}$。

2.31 提示：利用附录 H 的 Jensen 不等式。

2.32 往证“$\Leftarrow$”：单点分布 $X \sim \langle 0\rangle$ 的期望和方差都等于 0，所以 $\mathsf{E}(X^2) = \mathsf{V}(X) + [\mathsf{E}(X)]^2 = 0$。往证“$\Rightarrow$”：由 $\mathsf{E}(X^2) = \mathsf{V}(X) + [\mathsf{E}(X)]^2$ 推出 $\mathsf{E}(X) = 0, \mathsf{V}(X) = 0$。对任意 $n \in \mathbb{N}$，利用 Chebyshev 不等式 $\mathsf{P}\{|X - \mathsf{E}(X)| \ge 1/n\} \le n^2\mathsf{V}(X)$ 进而得到 $\mathsf{P}\{|X| < 1/n\} = 1$。而事件 $\{X = 0\}$ 即事件 $\bigcap_{n=1}^{\infty}\{|X| < 1/n\}$。由第一章最后一道习题的结果，$\mathsf{P}(\bigcap_{n=1}^{\infty}\{|X| < 1/n\}) = 1$，得证。

2.33 若 $x < 0$，则 $-x = \int_{-\infty}^{\infty}(t-x)dF_X(t) \le \int_x^{\infty}(t-x)dF_X(t)$。进而，$x^2 \le \left[\int_x^{\infty}(t-x)dF_X(t)\right]^2 \le \int_x^{\infty} dF_X(t)\int_x^{\infty}(t-x)^2dF_X(t) \le \mathsf{P}(X \ge x)(\sigma^2+x^2)$。

2.34 由独立性得到 $\mathsf{E}(Z) = 2\mathsf{E}(X) - \mathsf{E}(Y) + 3 = 5, \mathsf{V}(Z) = 2^2\mathsf{V}(X) + \mathsf{V}(Y) = 9$，所以 $Z \sim \mathrm{N}(5,9)$。

2.35 (1) 由 $\mathsf{V}(Z_1) = \mathsf{V}(Z_2) = (a^2 + b^2)\sigma^2, \mathsf{E}(Z_1) = \mathsf{E}(Z_2) = 0, \mathsf{E}(Z_1Z_2) = \mathsf{E}(a^2X^2 - b^2Y^2) = (a^2 - b^2)\sigma^2$ 得到 $\rho(Z_1Z_2) = (a^2 - b^2)/(a^2 + b^2)$。(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（见练习 2.8）。

2.36 求得 $(U, V)^{\mathsf{T}}$ 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

$$f_U(u) = \begin{cases} ue^{-u} & \text{当 } u > 0 \\ 0 & \text{当 } u \le 0 \end{cases} \text{ 且 } f_V(v) = \begin{cases} 1/(1 + v)^2 & \text{当 } v > 0 \\ 0 & \text{当 } v \le 0 \end{cases}$$

由 $f(u, v) = f_U(u)f_V(v)$ 推得 U 与 V 相互独立。

2.37 提示：利用 Chebyshev 不等式。

2.38 提示：计算出 $\mathsf{E}(X_n) = 0, \mathsf{V}(X_n) = 2$。令 $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$，进而得出 $\mathsf{E}(\overline{X}) = 0, \mathsf{V}(\overline{X}) = 2/n$，由 Chebyshev 不等式推得 $1 \ge \mathsf{P}(|\overline{X}| < \epsilon) \ge 1 - \mathsf{V}(\overline{X})/\epsilon^2 = 1 - 2/(n\epsilon^2)$。

2.39 利用 Chebyshev 不等式，$\mathsf{P}\{|X - \mathsf{E}(X)| \ge \sqrt{2\mathsf{V}(X)}\} \le 1/2$，于是 $\mathsf{E}(X) - \sqrt{2\mathsf{V}(x)} \le \mathsf{M}(X) \le \mathsf{E}(X) + \sqrt{2\mathsf{V}(x)}$。

2.40 (1) $r = 1$ 的情形是显然的。设 $r > 1$，则 $|X + Y|^r \le |X| \cdot |X + Y|^{r-1} + |Y| \cdot |X + Y|^{r-1}$，根据 Hölder 不等式，$\mathsf{E}|X + Y|^r \le \{\mathsf{E}|X|^r\}^{1/r}\{\mathsf{E}|X + Y|^{(r-1)s}\}^{1/s} + \{\mathsf{E}|Y|^r\}^{1/r}\{\mathsf{E}|X + Y|^{(r-1)s}\}^{1/s} = [\{\mathsf{E}|X|^r\}^{1/r} + \{\mathsf{E}|Y|^r\}^{1/r}]\{\mathsf{E}|X + Y|^{(r-1)s}\}^{1/s}$。注意到 $(r - 1)s = r$，Minkowski 不等式得证。(2) Schwarz 不等式是 Hölder 不等式在 $r = 1/2$ 时的特例。

2.41 利用 Markov 不等式 (2.74) 和 Lyapunov 不等式 (2.92) 可证。

2.42 (1) $\rho = 0$；(2) X 与 Y 不独立。

2.43 $\gamma_{Y|X} = 1, \gamma_{X|Y} = 1/2, \rho(X,Y) = \sqrt{1/2}$。

2.44 $\rho(Y,Z) = 0.9$。

2.45 提示：按照 X, Y 的定义有 $\mathsf{Cov}(X,Y) = \mathsf{P}(AB) - \mathsf{P}(A)\mathsf{P}(B)$。

2.46 $\mathsf{E}[\max(X^2,Y^2)] = \mathsf{E}[\frac{1}{2}(X^2+Y^2+|X^2-Y^2|)] = \frac{1}{2}[\mathsf{E}(X^2)+\mathsf{E}(Y^2)+\mathsf{E}|X^2-Y^2|] \le \frac{1}{2}[\mathsf{V}(X)+\mathsf{V}(Y)+\sqrt{\mathsf{E}(X+Y)^2\mathsf{E}(X-Y)^2}] = 1+\sqrt{1-\rho^2}$。

2.47 (1) $y = \mathsf{E}(Y|X=x) = (1+x)/2$；(2) $x = \mathsf{E}(X|Y=y) = y/2$。

2.48 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，

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \frac{\frac{\partial f_\theta(x)}{\partial\theta}}{f_\theta(x)} f_\theta(x)dx = 0 \Rightarrow \int_{-\infty}^{+\infty} \frac{\partial \ln f_\theta(x)}{\partial\theta} f_\theta(x)dx = 0\\
&\Rightarrow \int_{-\infty}^{+\infty} \left\{ \frac{\partial^2 \ln f_\theta(x)}{\partial\theta^2} f_\theta(x) + \frac{\partial \ln f_\theta(x)}{\partial\theta}\frac{\partial f_\theta(x)}{\partial\theta} \right\} dx = 0\\
&\Rightarrow \int_{-\infty}^{+\infty} \left\{ \frac{\partial^2 \ln f_\theta(x)}{\partial\theta^2} + \left[\frac{\partial \ln f_\theta(x)}{\partial\theta}\right]^2 \right\} f_\theta(x)dx = 0
\end{aligned}$$

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 都不是。

3.3 (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) $\mathrm{U}[-1,1]$。

3.4 提示：利用概率密度函数积分为 1 的性质求出 $c = a/2$ 再求特征函数，结果为 $a^2/(a^2+t^2)$。

3.5 提示：令 $X_k = \begin{cases} 1 & \text{第 } k \text{ 次试验 } A \text{ 发生} \\ 0 & \text{第 } k \text{ 次试验 } A \text{ 不发生} \end{cases}$
于是 $X = \sum_{k=1}^n X_k$，进而 $\varphi_x(t) = \prod_{k=1}^n (q_k + p_k e^{it})$。

3.6 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 $Z\sim \mathrm{B}(m+n,p)$。

3.7 提示：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 $\mathsf{P}(X=k)=(n+k-1)!p^nq^k/[k!(n-1)!]$，其中 $k=0,1,2,\cdots$，即负二项分布 $\mathrm{NegB}(n,p)$。

3.8 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，$1-\varphi(2t)=\int_{-\infty}^{\infty}[1-\cos(2tx)]dF(x)=2\int_{-\infty}^{\infty}sin^2(tx)dF(x)=2\int_{-\infty}^{\infty}[1-\cos(tx)][1+\cos(tx)]dF(x)\le 4\int_{-\infty}^{\infty}[1-\cos(tx)]dF(x)=4[1-\varphi(t)]$。

3.9 提示：欲证 $\varphi(t)\equiv 1$，只需验证 $X\sim\langle 0\rangle$ 或 $\mathsf{E}(X^2)=0$ 即可（参见性质 2.8）。

3.10 经验证，$\sum_{k=1}^n\sum_{j=1}^n\varphi(t_k-t_j)z_k\bar{z}_j=\int_{-\infty}^{+\infty}|\sum_{k=1}^n e^{it_kx}z_k|^2dF(x)\ge 0$。

4.1 约为 0.8887。

4.2 $\mathsf{E}(X)=(n+1)/2,\mathsf{V}(X)=(n-1)(n+1)/12,\gamma_1=0,\gamma_2=-6(n^2+1)/[5(n-1)(n+1)],c_v=sqrt(n-1)/[3(n+1)]$。

4.3 定义 Y 为 4 个随机数中不超过 a 的个数，则由 $\mathsf{P}(0<X\le a)=a$ 知 $Y\sim\mathrm{B}(4,a)$。从 $\mathsf{P}\{Y=4\}=C_4^4a^4(1-a)^0=0.1$ 解得 $a\approx 0.5623$。

4.4 $\mathsf{E}(Y)=-(1+\ln 2)/2,\mathsf{V}(Y)=\ln^2 2/4+\ln 2/2+3/4$ 。

4.5 利用分布函数 $1-\exp\{-\beta Y\}\sim\mathrm{U}[0,1]$ 可得 $h(x)=-\beta^{-1}\ln(1-x)$。

4.6 提示：$2(X_1+X_2+\cdots+X_n)$ 与 χ^2_{2n} 的特征函数都为 $(1-2it)^{-n}$。或参考表 6.1 的最后一行。

4.7 只需往证 $Y=-\ln X\sim\mathrm{Expon}(1)$，由上一题的结论便可得证。事实上，由定理 2.10 知 Y 的密度函数为 $f(y)=\begin{cases}0 & \text{当 } y\le 0\\ \exp(-y) & \text{当 } y>0\end{cases}$

4.8 提示：参考例 3.10。当 n 为奇数时，$\mathsf{E}(X^n)=0,\mathsf{V}(X^n)=(2n-1)!!$；当 n 为偶数时，$\mathsf{E}(X^n)=(n-1)!!,\mathsf{V}(X^n)=(2n-1)!!-[(n-1)!!]^2$。

4.9 提示：$\max(X,Y)+\min(X,Y)=X+Y,\max(X,Y)-\min(X,Y)=|X-Y|$。根据 $X+Y\sim \mathrm{N}(2\mu,2\sigma^2),X-Y\sim \mathrm{N}(0,2\sigma^2)$ 可求出 $\mathsf{E}(X+Y)=2\mu,\mathsf{E}(|X-Y|)=2\sigma/\sqrt{\pi}$，于是 $\mathsf{E}[\max(X,Y)]=\mu+\sigma/\sqrt{\pi},\mathsf{E}[\min(X,Y)]=\mu-\sigma/\sqrt{\pi}$。

4.10 提示：$\sin x\sin y\exp[-(x^2+y^2)/2]$ 关于 x,y 都是奇函数。

4.11 $\mathsf{E}(|X|)=1,\mathsf{V}(|X|)=1,\mathsf{Cov}(X,|X|)=0$。$X$ 与 $|X|$ 不独立。

4.12 $f_Y(y)=\beta^\alpha(\beta+1)^{-\alpha}\Gamma(y+\alpha)[\Gamma(\alpha)y!(\beta+1)^y]^{-1},y=0,1,2,\cdots$

4.13 $f_{X/Y}(x)=\begin{cases}\frac{x^{p-1}\Gamma(p+q)}{(x+1)^{p+q}\Gamma(p)\Gamma(q)} & \text{当 } x>0\\ 0 & \text{当 } x\le 0\end{cases}$

$f_{X/(X+Y)}(x)=\begin{cases}\frac{x^{p-1}(1-x)^{q-1}\Gamma(p+q)}{\Gamma(p)\Gamma(q)} & \text{当 } 0<x<1\\ 0 & \text{其他}\end{cases}$

4.14 提示：令 $g(x)=\mathsf{P}(X>x)$，其中 $x\ge 0$，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，往证有 $g(m/n)=[g(1/n)]^m=[g(1)]^{m/n}$。再说明对于实数 $x\ge 0$ 有 $g(x)=[g(1)]^x$。令 $\beta=-\ln g(1)$，则有 $\mathsf{P}(X\le x)=1-\exp\{-\beta x\}$，这是指数分布 $\mathrm{Expon}(\beta)$ 的分布函数。

4.15 (1) $F_U(u)=\mathsf{P}\{\max(X_1,\cdots,X_n)\le u\}=\mathsf{P}(X_1\le u,X_2\le u,\cdots,X_n\le u)=\mathsf{P}(X_1\le u)\cdots\mathsf{P}(X_n\le u)=\begin{cases}0 & \text{当 } u\le 0\\ (1-e^{-\beta u})^n & \text{当 } u>0\end{cases}$，进而 U 的密度函数为 $f_U(u)=F_U'(u)=\begin{cases}0 & \text{当 } u\le 0\\ \beta n(1-e^{-\beta u})^{n-1}e^{-\beta u} & \text{当 } u>0\end{cases}$

(2) $F_V(v)=\mathsf{P}(\min(X_1,\cdots,X_n)\le v)=1-\mathsf{P}(\min(X_1,\cdots,X_n)>v)=$

$1-\mathsf{P}(X_1>v)\cdots\mathsf{P}(X_n>v)=1-[1-F_{X_1}(v)]\cdots[1-F_{X_n}(v)]=\begin{cases}0 & \text{当 } v<0\\ 1-e^{-\beta nv} & \text{当 } v\geq 0\end{cases}$，进而 $f_V(v)=F_V'(v)=\begin{cases}\beta n e^{-\beta nv} & \text{当 } v>0\\ 0 & \text{当 } v\leq 0\end{cases}$

4.16 由 $\mathsf{P}\{X\geq 1\}=\mathsf{P}\{X\leq 1\}$ 计算出 $\beta=\ln 2$，所以 $\mathsf{P}\{X\geq k\}=(1/2)^k$，进而 $\sum_{k=1}^{\infty}\mathsf{P}\{X\geq k\}=1$。

4.17 $f_Z(z)=\begin{cases}\dfrac{\Gamma\left(\frac{m+n}{2}\right)z^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)(z+1)^{\frac{m+n}{2}}} & \text{当 } z>0\\ 0 & \text{当 } z\leq 0\end{cases}$

4.18 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。而 $T^2=X^2/(Y/n)$ 且 $X^2\sim\chi_1^2$ 与 Y 相互独立，所以 $T^2\sim F(1,n)$。

4.19 $\mathsf{E}(1/X)=\sqrt{2\pi}/2\sigma$。

4.20 提示：参考练习 2.8。$(W,V)^{\mathsf{T}}$ 服从正态分布且 $\mathsf{Cov}(W,Y)=0$。

4.21 $X+Y\sim\mathrm{N}(8,18)$。

4.22 $\mathrm{N}(A\boldsymbol{\mu}+\boldsymbol{\alpha},A\Sigma A^{\mathsf{T}})$。

4.23 $X_1,X_2,\cdots,X_n$ 相互独立当且仅当 $\boldsymbol{X}$ 的特征函数为

$$\varphi(\boldsymbol{t})=\exp\left\{i\sum_{j=1}^{n}t_j\mu_j-\frac{1}{2}\sum_{j=1}^{n}\sigma_j^2t_j^2\right\}=\varphi(t_1)\varphi(t_2)\cdots\varphi(t_n)$$

4.24 利用性质 4.17 和 Beta 分布的数字特征可证。

4.25 $\mathsf{E}(\boldsymbol{W})=\sum_{j=1}^{n}\mathsf{E}(\boldsymbol{X}_j\boldsymbol{X}_j^{\mathsf{T}})=n\Sigma$，因为 $\mathsf{E}(\boldsymbol{X}_j\boldsymbol{X}_j^{\mathsf{T}})=\Sigma$。

5.1 $\mathsf{E}(X_k)=0,\mathsf{V}(X_k)=2,k=1,2,\cdots$，由 Chebyshev 弱大数律可证。

5.2 算得 $\mathsf{E}(X_k)=0,\mathsf{V}(X_k)=k^{2s},k=1,2,\cdots,n,\cdots$ 并且 $\{X_k\}$ 相互独立，有 $\frac{1}{n^2}\mathsf{V}(\sum_{k=1}^{n}X_k)=\frac{1}{n^2}\sum_{k=1}^{n}\mathsf{V}(X_k)=\frac{1}{n^2}\sum_{k=1}^{n}k^{2s}\leq n\cdot n^{2s}/n^2=n^{2s-1}$，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 1.35。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 $j = 1, 2, \cdots, n$，则 $S_n = \sum_{j=1}^{n} X_j$。由 $\mathsf{P}\{X_j = 1\} = 1/n$ 求得 $\mathsf{V}(X_j) = (n-1)/n^2, \mathsf{Cov}(X_j, X_k) = 1/[n^2(n-1)]$，利用式 (2.101) 算出 $\mathsf{V}(S_n) = 1$。由 Markov 弱大数律可证。

5.4 提示：利用式 (2.101) 往证 $\lim\limits_{n\to\infty} \frac{1}{n^2}\mathsf{V}(\sum_{i=1}^{n} X_i) = 0$，再利用 Markov 弱大数律即可证得。这个结果被称为 Bernstein 定理。

$$\begin{aligned} \frac{1}{n^2}\mathsf{V}\left(\sum_{i=1}^{n} X_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{V}(X_i) + \frac{2}{n^2}\sum_{1\le j<k\le n}\rho_{jk}\sqrt{\mathsf{V}(X_j)}\sqrt{\mathsf{V}(X_k)} \\ &\le \frac{c}{n} + \frac{2c}{n^2}\sum_{1\le j<k\le n}|\rho_{jk}| \end{aligned}$$

因为当 $|k-j| \to \infty$ 时，$\rho_{kj} \to 0$，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 $k - j \le N$ 的 ρ_{kj} 至多有 N 个，从而满足 $0 < k - j \le N$ 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 $|\rho_{kj}| \le 1$，

$$\begin{aligned} \frac{1}{n^2}\mathsf{V}\left(\sum_{i=1}^{n} X_i\right) &\le \frac{c}{n} + \frac{2c}{n^2}\left(\sum_{0<k-j\le N}|\rho_{kj}| + \sum_{k-j>N}|\rho_{kj}|\right) \\ &\le \frac{c}{n} + \frac{2c}{n^2}\left[Nn + \frac{n(n-1)}{2}\cdot\frac{\epsilon}{c}\right] = \frac{(2N+1)c}{n} + \left(1 - \frac{1}{n}\right)\epsilon \end{aligned}$$

由于 N 由 ϵ 确定，故当 $n \to \infty$ 时，有 $\lim\limits_{n\to\infty} \frac{1}{n^2}\mathsf{V}(\sum_{i=1}^{n} X_i) \le \epsilon$。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 $\sum_{k=1}^{\infty}\mathsf{V}(X_k)/k^2 < \infty$，所以 $\forall \epsilon > 0$，存在 N_1 使得 $m > N_1, n > m > N_1$ 时，有 $\sum_{k=m+1}^{n}\mathsf{V}(X_k)/k^2 < \epsilon/2$。又因为 $\mathsf{V}(X_k)$ 有限，存在 N_2 使得 $n > N_2$ 时，有 $\frac{1}{n^2}\sum_{k=1}^{m}\mathsf{V}(X_k) < \epsilon/2$。取 $N = \max(N_1, N_2)$，当 $n > N$ 时，有 $\frac{1}{n^2}\mathsf{V}(\sum_{k=1}^{n} X_k) = \frac{1}{n^2}\sum_{k=1}^{n}\mathsf{V}(X_k) \le \frac{1}{n^2}\sum_{k=1}^{m}\mathsf{V}(X_k) +$

$\sum_{k=m+1}^{n} \mathsf{V}(X_k)/k^2 < \epsilon/2 + \epsilon/2 = \epsilon$，再利用 Markov 弱大数律可证。

5.6 都成立。

5.7 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 $\mathsf{E}(\ln X_j) = -1$。利用 Khinchin 弱大数律可证，并得到 $c = e^{-1}$。

5.8 必要性：设 X_n 的分布函数为 $F_n(x)$，则有

$$\begin{aligned}\mathsf{E}[h(|X_n|)] &= \int_{|x|>\delta} h(|x|)dF_n(x) + \int_{|x|\le\delta} h(|x|)dF_n(x) \\ &\le \sup_{x\ge 0} h(x) \int_{|x|>\delta} dF_n(x) + h(\delta) \int_{|x|\le\delta} dF_n(x) \le c\mathsf{P}(|X_n| > \delta) + h(\delta)\end{aligned}$$

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 $\mathsf{P}(|X_n| > \delta) < \epsilon/(2c)$，于是 $\mathsf{E}[h(|X_n|)] < \epsilon$。

充分性：说明 $\mathsf{E}[h(|X_n|)] \ge h(\delta)\mathsf{P}(|X_n| > \delta)$，从而当 $n \to \infty$ 时 $\mathsf{P}(|X_n| > \delta) \to 0$，即 $X_n \xrightarrow{P} 0$。

5.9 提示：令 $Y_n = \frac{2}{n(n+1)}\sum_{k=1}^{n} kX_k$，则 $\mathsf{E}(Y_n) = \mu, \mathsf{V}(Y_n) \le 4\sigma^2/(n+1)$。于是 $\forall \epsilon > 0$，当 $n \to \infty$ 时有 $\mathsf{P}\{|Y_n - \mu| \le \epsilon\} \ge 1 - \mathsf{V}(Y_n)/\epsilon^2 \to 1$。

5.10 满足中心极限定理。提示：仿照例 5.5。

5.11 由 Lindeberg-Lévy 中心极限定理，$\mathsf{P}\{\sum_{i=1}^{100} X_i \ge 90\} \approx 1 - \Phi(-0.65) = \Phi(0.65) \approx 0.7422$。

5.12 由 $p = 1/4, n = 400$ 得到 $\sqrt{np(1-p)} = 5\sqrt{3}$，利用 de Moivre-Laplace 中心极限定理，$\mathsf{P}\{50 \le X \le 150\} = \mathsf{P}\left\{\frac{50-100}{5\sqrt{3}} \le \frac{X-100}{5\sqrt{3}} \le \frac{150-100}{5\sqrt{3}}\right\} \approx 2\Phi(10/\sqrt{3}) - 1 \approx 1$。

5.13 优等品个数服从二项分布 $\mathrm{B}(100, 0.2)$，利用 de Moivre-Laplace 中心极限定理，$\mathsf{P}\{18 \le X \le 25\} \approx \Phi(1.25) + \Phi(0.5) - 1$。

5.14 由题意知随机变量序列 $F(X_1), F(X_2), \cdots, F(X_n), \cdots \overset{iid}{\sim} \mathrm{U}[0,1]$，由 Lindeberg-Lévy 中心极限定理知

$$\lim_{n\to\infty} \mathsf{P}\left\{\frac{\sum_{i=1}^n F(X_i) - \mathsf{E}[\sum_{i=1}^n F(X_i)]}{\sqrt{\mathsf{V}[\sum_{i=1}^n F(X_i)]}} \le x\right\} = \lim_{n\to\infty} \mathsf{P}\left\{\frac{\sum_{i=1}^n F(X_i) - n/2}{\sqrt{n/12}} \le x\right\} = \Phi(x)$$

取 $x = 0$ 即可证得。

5.15 (1) 随机变量序列 $\{Y_n\}$ 独立同分布，再利用大数律即可证。(2) $\mathrm{N}(2\lambda^{-2}, 20n^{-1}\lambda^{-4})$。

5.16 $Y_n \sim \mathrm{N}\left(m_2, \frac{m_4 - m_2^2}{n}\right)$。

5.17 令 $Y_j = X_{2j} - X_{2j-1}, j = 1, 2, \cdots$，则 $\{Y_j\}$ 独立同分布，由中心极限定理求得 $c = 1/\sqrt{2}$。

5.18 令 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathrm{Poisson}(1)$，则 $\mathsf{E}(X_n) = \mathsf{V}(X_n) = 1$。由 Lindeberg-Lévy 中心极限定理，$\mathsf{P}\{\sum_{k=1}^n X_k \le n\} = \mathsf{P}\{\sum_{k=1}^n (X_k - 1)/\sqrt{n} \le 0\} \to \Phi(0) = 1/2$。并且，$\sum_{k=1}^n X_k \sim \mathrm{Poisson}(n)$。于是，$\mathsf{P}\{\sum_{k=1}^n X_k \le n\} = e^{-n}\sum_{k=0}^n n^k/k! \to 1/2$ 得证。

6.1 提示：$\sum_{j=1}^n (X_j - c)^2 = \sum_{j=1}^n (X_j - \overline{X} + \overline{X} - c)^2 = \sum_{j=1}^n (X_j - \overline{X})^2 + n(\overline{X} - c)^2$，类似于式 (2.72)。

6.2 $\overline{Y} = (\overline{X} - a)/b$ 且 $S_Y^2 = S_X^2/b^2$。

6.3 设 $\mathsf{E}(X) = \mu, \mathsf{V}(X) = \sigma^2$，计算得 $\mathsf{V}(X_i - \overline{X}) = \sigma^2(n-1)/n, \mathsf{E}[(X_i - \overline{X})(X_j - \overline{X})] = -\sigma^2/n$，因而 $\rho = \mathsf{Cov}(X_i - \overline{X}, X_j - \overline{X})/\mathsf{V}(X_i - \overline{X}) = \mathsf{E}[(X_i - \overline{X})(X_j - \overline{X})]/\mathsf{V}(X_i - \overline{X}) = -(n-1)^{-1}$。

6.4 利用性质 6.3：(1) $\mathsf{E}(\overline{X}) = p$ 并且 $\mathsf{V}(\overline{X}) = p(1-p)/n$。(2) $\mathsf{E}(S^2) = p(1-p)$。(3) 当 $x < 0$ 时，$F_n^*(x) = 0$；当 $0 \le x < 1$ 时，$F_n^*(x) = 1 - m/n$；当 $x \ge 1$ 时，$F_n^*(x) = 1$。

6.5 由定理 6.6 知 $9S^2/4^2 \sim \chi^2_9$，$\mathsf{P}\{S^2 > a\} = \mathsf{P}\{9S^2/4^2 > 9a/4^2\} = 0.1$，所以 $9a/4^2 \approx \chi^2_9(0.9) \approx 14.684$，进而 $a \approx 26.105$。利用 R 语言中的函数 `qchisq(p,df)` 可求得 χ^2_n 分布的 p-分位数。

6.6 $X_{(1)}$ 的分布函数为 $F^*_1(y) = 1 - [1 - F(y)]^n = \begin{cases} 1 - e^{-n\lambda y} & \text{当 } y \geq 0 \\ 0 & \text{当 } y < 0 \end{cases}$

于是，$\mathsf{E}X_{(1)} = 1/(n\lambda), \mathsf{V}X_{(1)} = 1/(n\lambda)^2$。

6.7 因为 $\overline{X}_1, \overline{X}_2 \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2/n)$，所以 $\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}(0, 2\sigma^2/n)$。从 $\mathsf{P}\{|\overline{X}_1 - \overline{X}_2| > \sigma\} = \mathsf{P}\{|(\overline{X}_1 - \overline{X}_2)/\sqrt{2\sigma^2/n}| > \sigma/\sqrt{2\sigma^2/n}\} = 2[1 - \Phi(\sqrt{n/2})] = 0.01$ 得 $n = 14$。

6.8 提示：由 $\mathsf{Cov}(X_1 + X_2, X_1 - X_2) = 0$ 和练习 2.8 先证明 $X_1 + X_2 \sim \mathrm{N}(0, 2\sigma^2)$ 与 $X_1 - X_2 \sim \mathrm{N}(0, 2\sigma^2)$ 相互独立，再说明 $(X_1 + X_2)^2/(X_1 - X_2)^2 \sim F(1, 1)$，故 $\mathsf{P}\{(X_1 + X_2)^2/(X_1 - X_2)^2 < 4\} = \int_0^4 1/[\pi(1 + y)y^{1/2}]dy = 2\arctan(2)/\pi \approx 0.70$。

6.9 提示：$\frac{Y_1 - Y_2}{\sigma/\sqrt{2}} \sim \mathrm{N}(0, 1)$ 且 $2S^2/\sigma^2 \sim \chi^2_2$。

6.10 由 $\mathsf{E}(\overline{X} - \mu)^2 = \mathsf{V}(\overline{X}) = \frac{1}{n}\mathsf{V}(X) = \frac{4}{n} \leq 0.1$ 得 $n \geq 40$。

6.11 $\overline{X} = \frac{1}{16}\sum_{i=1}^{16} X_i \sim \mathrm{N}(0, 1)$，由 $\frac{1}{4}Y_1, \cdots, \frac{1}{4}Y_{16} \overset{iid}{\sim} \mathrm{N}(0, 1)$ 知 $\sum_{i=1}^{16}(\frac{1}{4}Y_i)^2 \sim \chi^2_{16}$。由 t 分布的定义知 $V \sim t(16)$。

6.12 由 $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) \sim \mathrm{N}(0, 1)$ 和 $Y_1^2 + \cdots + Y_n^2 \sim \chi^2_n$ 得到 $W \sim t(n)$。

6.13 由 $\mathsf{V}(X_1 - 2X_2) = 20, [(X_1 - 2X_2)/\sqrt{20}]^2 \sim \chi^2_1$ 得 $a = 1/20$。由 $\mathsf{V}(3X_3 - 4X_4) = 100, [(3X_3 - 4X_4)/10]^2 \sim \chi^2_1$ 得 $b = 1/100$。

6.14 因 $\frac{1}{\sqrt{2}\sigma}(X_1 + X_2) \sim \mathrm{N}(0, 1)$ 与 $\frac{1}{\sigma^2}(X_3^2 + X_4^2 + X_5^2) \sim \chi^2_3$ 独立，故 $Y \sim t(3)$，得 $a = \sqrt{3/2}$。

6.15 由 $\sqrt{2}(\overline{X} - \overline{Y}) \sim \mathrm{N}(0, 1)$，$\mathsf{P}\{|\overline{X} - \overline{Y}| > 0.3\} = \mathsf{P}\{\sqrt{2}|\overline{X} - \overline{Y}| > 0.3\sqrt{2}\} = 2[1 - \Phi(0.3\sqrt{2})] \approx 0.6714$。

6.16 $Y_1 \sim F(1,1), Y_2 \sim F(2,1), Y_3 \sim t(1)$。

6.17 $Y_j \sim \mathrm{N}((1+a)\mu, (n+2a+a^2)\sigma^2/n)$。

6.18 仿照例 6.8 的证法，或利用定理 6.9。

6.19 利用定理 6.9 可证统计量 $[\overline{X}, \sum_{j=1}^n (X_j-\overline{X})^2, \overline{Y}, \sum_{j=1}^n (Y_j-\overline{Y})^2, \sum_{j=1}^n (X_j-\overline{X})(Y_j-\overline{Y})]^{\mathsf{T}}$ 是充分的。

7.1 (1) $c = 1/2(n-1)$；(2) $c = 1/n$。

7.2 证明：$\mathsf{E}(\hat{\theta}^2) = \mathsf{V}(\hat{\theta}) + \theta^2 > \theta^2$。

7.3 $\hat{\mu}_1, \hat{\mu}_2$ 都是 μ 的无偏估计量且 $\mathsf{V}(\hat{\mu}_1) \le \mathsf{V}(\hat{\mu}_2)$。

7.4 $\mathsf{E}(X) = \theta + 1 = \mathsf{V}(X)$，于是 $\mathsf{E}(\hat{\theta}_1) = \theta, \mathsf{E}(\hat{\theta}_2) = \mathsf{E}(X_{(1)}) - 1/n$，其中 $X_{(1)} = \min(X_1, \cdots, X_n)$，它的密度函数为 $f_1(x) = \begin{cases} ne^{-n(x-\theta)} & \text{当 } x \ge \theta \\ 0 & \text{当 } x < \theta \end{cases}$ 于是，$\mathsf{E}(X_{(1)}) = \theta + 1/n$。因此，$\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是 θ 的无偏估计量。计算得 $\mathsf{V}(\hat{\theta}_2) = 1/n^2 \le \mathsf{V}(\hat{\theta}_1) = \frac{1}{n^2} \cdot \sum_{j=1}^n \mathsf{V}(X_j) = 1/n$。

7.5 提示：$\mathsf{E}(X-\mu) = \sigma\sqrt{2/\pi}$。

7.6 $\mathsf{E}(X) = \int_0^1 (\theta+1)x^{\theta+1}dx = \frac{\theta+1}{\theta+2}$，矩估计为 $\hat{\theta} = \frac{2\overline{X}-1}{1-\overline{X}}$。似然函数为 $\mathscr{L}(\theta; x_1, x_2, \cdots, x_n) = (\theta+1)^n (\prod_{j=1}^n x_j)^\theta$，解方程 $\frac{\partial \ell(\theta)}{\partial \theta} = \frac{n}{\theta+1} + \sum_{j=1}^n \ln x_j = 0$ 得 θ 的最大似然估计 $\hat{\theta} = -\frac{n}{\sum_{j=1}^n \ln X_j} - 1$。

7.7 设盒中有白球 w 个，有黑球 θw 个。从盒中有放回地抽一次球，抽得白球的概率为 $1/(1+\theta)$，抽得黑球的概率为 $\theta/(1+\theta)$。似然函数为 $\mathscr{L}(\theta) = (\frac{1}{1+\theta})^k (\frac{\theta}{1+\theta})^{n-k}$，求得 θ 的最大似然估计为 $\hat{\theta} = n/k - 1$。

7.8 矩估计 $\hat{\theta} = \frac{3-\overline{X}}{2}$。最大似然估计 $\hat{\theta} = \frac{2n_1+n_2}{2n}$，其中 n_1, n_2 分别是样本中 1, 2 的个数。

7.9 $\hat{\theta}_1 = \exp\left\{\frac{1}{n}\sum_{j=1}^n \ln X_j + \frac{1}{2n}\sum_{j=1}^n \left(\ln X_j - \frac{1}{n}\sum_{j=1}^n \ln X_j\right)^2\right\}$
$\hat{\theta}_2 = \hat{\theta}_1^2\left[\exp\left\{\frac{1}{n}\sum_{j=1}^n \left(\ln X_j - \frac{1}{n}\sum_{j=1}^n \ln X_j\right)^2\right\} - 1\right]$

7.10 似然函数 $\mathscr{L}(\theta; x_1, \cdots, x_n) = \begin{cases} \theta^{-n} & \text{当 } \theta \le x_1, \cdots, x_n \le 2\theta \\ 0 & \text{其他} \end{cases}$ 因为 $\theta \le \min\limits_{1\le j\le n} x_j \le \max\limits_{1\le j\le n} x_j \le 2\theta$，所以 $\frac{\theta}{2} \le \frac{1}{2}\min\limits_{1\le j\le n} x_j \le \frac{1}{2}\max\limits_{1\le j\le n} x_j \le \theta \le \min\limits_{1\le j\le n} x_j$。又 $\mathscr{L}$ 是 θ 的递减函数，因此 θ 的极大似然估计为 $\hat{\theta} = \frac{1}{2}\max\limits_{1\le j\le n} X_j$。

7.11 要使得 $\mathsf{V}(T) = p^2\mathsf{V}(\overline{X}) + (1-p)^2\mathsf{V}(\overline{Y}) = [p^2/n_1 + (1-p)^2/n_2]\sigma^2$ 达到最小，解得 $p = n_1/(n_1+n_2)$。

7.12 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 $\mathsf{V}(\hat{\theta}) = \sum_{j=1}^n k_j^2\sigma_j^2$ 达到最小，利用 Lagrange 乘子法构造函数 $f(k_1^2, \cdots, k_n^2) = \sum_{j=1}^n k_j^2\sigma_j^2 - \lambda(\sum_{j=1}^n k_j - 1)$，由 $\partial f/\partial k_j = 0$ 得到方程组 $2k_j\sigma_j^2 - \lambda = 0, j = 1, 2, \cdots, n$，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。

7.13 参见例 7.22 第四种情况，两边取对数即可。

7.14 $X_{(n)} = \max(X_1, \cdots, X_n)$ 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 $\mathsf{P}\{X_{(n)} \le \theta \le c_n X_{(n)}\} = \mathsf{P}(\theta/c_n \le X_{(n)} \le \theta) = \int_{\theta/c_n}^{\theta} nx^{n-1}\theta^{-n}dx = 1 - c_n^{-n}$，要使此值等于 $1-\alpha$ 只需取 $c_n = \alpha^{-1/n}$。

7.15 利用第 242 页的性质 6.5 得 $\left[F_{m-1,n-1,\alpha/2}\frac{S_Y^2}{S_X^2}, F_{m-1,n-1,1-\alpha/2}\frac{S_Y^2}{S_X^2}\right]$，其中 $F_{m-1,n-1,\alpha}$ 表示 $F(m-1, n-1)$ 分布的 α-分位数。

7.16 由第 179 页的定义 4.7 和性质 4.10，$2\lambda_1\sum_{j=1}^m X_j = 2m\lambda_1\overline{X} \sim \chi_{2m}^2$，同理 $2n\lambda_2\overline{Y} \sim \chi_{2n}^2$。取枢轴量 $\lambda_2\overline{Y}/(\lambda_1\overline{X}) \sim F_{2n,2m}$，可求得 λ_2/λ_1 的置信度为 $1-\alpha$ 的置信区间为 $[\overline{X}/(\overline{Y}F_{2m,2n,\alpha/2}), \overline{X}/(\overline{Y}F_{2m,2n,1-\alpha/2})]$。

8.1 参考第 292 页的例 8.9，$H_0: \mu = 100$ 在显著水平 $\alpha = 0.05$ 下被拒绝，即这批零件长度不合格。

8.2 $\gamma = \mathrm{P}\{\sqrt{n}(\overline{X}-\mu_0)/\sigma < z_{1-\alpha}|H_1 \text{ 成立}\} = \mathrm{P}\{\sqrt{n}(\overline{X}-\mu_1)/\sigma < z_{1-\alpha} - \sqrt{n}(\mu_1-\mu_0)/\sigma|H_1 \text{ 成立}\} = \Phi[z_{1-\alpha} - \sqrt{n}(\mu_1-\mu_0)/\sigma]$，由分位点的性质，$z_\gamma = -z_{1-\gamma} = z_{1-\alpha} - \sqrt{n}(\mu_1-\mu_0)/\sigma$，得 $n = (z_{1-\alpha}+z_{1-\gamma})^2\sigma^2/(\mu_1-\mu_0)^2$。

8.3 若零假设 $H_0 : \mu = \mu_0$ 成立，则 $\overline{X} \sim \mathrm{N}(\mu_0, 9/25)$，根据例 8.9 可得出 $c = z_{1-\alpha/2}\sigma/\sqrt{n} \approx 1.176$。

8.4 用 X, Y 表示甲、乙两个品种的亩产量，按题设 $X \sim \mathrm{N}(\mu_X, \sigma_X^2), Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$，有容量为 $m = n = 10$ 的两个样本，其均值和标准差为别为 $\overline{X} = 32, S_X = 23, \overline{Y} = 21, S_Y = 12$。依题意要检验的假设为：(1) $H_0 : \mu_X = \mu_Y \leftrightarrow H_1 : \mu_X \neq \mu_Y$，参考例 8.11；(2) $H_0 : \sigma_X^2 = \sigma_Y^2 \leftrightarrow H_1 : \sigma_X^2 \neq \sigma_Y^2$，参考例 8.13。经计算两个品种的亩产量是没有差别。

8.5 令 $\theta = 1, 2$ 表示来自哪个总体，视 θ 为随机向量 $(X, Y)^{\mathsf{T}}$ 所依赖的一个参数，此问题相当于检验假设 $H_0 : \theta = 1 \leftrightarrow H_1 : \theta = 2$。由 Neyman-Pearson 引理，N-P 检验是给定水平下的最优势检验，对数似然比函数为 $\ln \lambda(x, y, \theta) = (x + y - 1)/(\rho + 1)$，得出 N-P 检验拒绝零假设 H_0 的充要条件是 $X + Y > c$，令 $T(X, Y) = X + Y$，对于给定水平 α，又令 $\mathrm{P}_2\{T(X, Y) \geq c\} = \alpha$，则有 $c = 4\sqrt{5}z_{1-\alpha}/5$。

8.6 当且仅当 $n\hat{\sigma}^2/\sigma_0^2 = \frac{1}{\sigma_0^2}\sum_{j=1}^n (x_j - \overline{x})^2 \geq c$ 时拒绝 H_0。

8.7 参考例 8.16，当零假设 H_0 成立时，似然比为

$$\lambda(\boldsymbol{x}) = \frac{(2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{j=1}^n (x_j - \overline{x})^2\right\}}{(2\pi s_n^2)^{-n/2}\exp\{-n/2\}} = \left[\frac{s_n^2}{\sigma_0^2}\exp\left\{-\frac{s_n^2}{\sigma_0^2}\right\}\right]^{n/2} e^{n/2}$$

似然比检验具有形式 $\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{当 } s_n^2/\sigma_0^2 \leq c_1 \text{ 或 } s_n^2/\sigma_0^2 \geq c_2 \\ 0 & \text{当 } c_1 < s_n^2/\sigma_0^2 < c_2 \end{cases}$

8.8 当 H_0 成立时，似然比 $\lambda(\boldsymbol{x}, \boldsymbol{y}) = \left[1 + \frac{mn}{(m+n)^2}\frac{(\overline{x}-\overline{y})^2}{\hat{\sigma}^2}\right]^{-(m+n)/2}$ 是 $|\overline{x} - \overline{y}|/\hat{\sigma}$

的减函数。因为 $t(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sqrt{mn(m+n-2)}|\overline{X}-\overline{Y}|}{(m+n)\hat{\sigma}} \sim t(m+n-2)$，因此似然比检验具有形式

$$\delta(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} 1 & \text{当 } t(\boldsymbol{x},\boldsymbol{y}) \geq t_{m+n-2,1-\alpha/2} \\ 0 & \text{当 } t(\boldsymbol{x},\boldsymbol{y}) < t_{m+n-2,1-\alpha/2} \end{cases}$$

8.9 零假设为“H_0 : Mendel 遗传定律成立”。设 H_0 成立，利用 Pearson χ^2 检验，将 $n = 100, Y_1 = 30, Y_2 = 48, Y_3 = 22$ 代入式 (8.13)，得到 $\chi^2 = (30-25)^2/25 + (48-50)^2/50 + (22-25)^2/25 = 1.44 < \chi^2_{2,0.95} = 5.991465$。故在置信水平 $\alpha = 0.05$ 下接受 H_0，即 Mendel 遗传定律是正确的。

8.10 设零假设“H_0 : 硬币是均匀的”成立，则 $p_k = \mathsf{P}(X = k) = 0.5^k, k = 1, 2, \cdots, 6$ 且 $\mathsf{P}(X \geq 7) = 1 - \mathsf{P}(X \leq 6) = p_6$。利用 Pearson χ^2 检验，将 $n = 1000, Y_1 = 533, Y_2 = 233, \cdots, Y_7 = 16$ 代入式 (8.13)，得到 $\chi^2 = \frac{1}{n}(Y_7^2/p_6 + \sum_{j=1}^{6} Y_j^2/p_j) - n = 5.102 < \chi^2_{6,0.95} = 12.59159$，故在置信水平 $\alpha = 0.05$ 下接受 H_0，即此硬币是均匀的。

8.11 当 $k = 2$ 时，

$$\begin{aligned} P &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{[n - X_1 - n(1-p_1)]^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \xrightarrow{L} \chi_1^2, \text{ 当 } n \to \infty \text{ 时} \end{aligned}$$

9.1 (1) $\hat{\beta}_1 = (X_1+2X_2+X_3)/6$ 且 $\hat{\beta}_2 = (-X_2+2X_3)/5$；(2) $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2/6), \hat{\beta}_2 \sim \mathrm{N}(\beta_2, \sigma^2/5)$。由 $A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix}$ 得 $A^TA = \begin{pmatrix} 6 & 0 \\ 0 & 5 \end{pmatrix}$，所以 $\hat{\beta}_1, \hat{\beta}_2$ 相互独立。

9.2 $T = \frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{l_{xx}}}}$，$t(n-2)$，$|T| \geq t_{\frac{\alpha}{2}}(n-2)$。

9.3 百分比的均值有显著差异。

9.4 (1)$\left(\hat{a}-S_e\frac{t_{\alpha,n-1}\sqrt{\overline{x}+s_x^2}}{s_x\sqrt{n}},\ \hat{a}+S_e\frac{t_{\alpha,n-1}\sqrt{\overline{x}+s_x^2}}{s_x\sqrt{n}}\right)$
(2)$\left(\hat{b}-S_e\frac{t_{\alpha,n-1}}{s_x\sqrt{n}},\ \hat{b}+S_e\frac{t_{\alpha,n-1}}{s_x\sqrt{n}}\right)$
(3)$\left(\frac{(n-1)S_e^2}{\chi^2_{\alpha/2,n-1}},\ \frac{(n-1)S_e^2}{\chi^2_{1-\alpha/2,n-1}}\right)$
其中 S_e 是标准残差，$t_{\alpha,n-2}$ 是自由度为 $n-2$ 的 t 分布水平 α 双侧分位数，而 $\overline{x}=\frac{1}{n}\sum_{j=1}^n x_j$，$s_x^2=\frac{1}{n}\sum_{j=1}^n(x_j-\overline{x})^2$。

9.5 (1)$\hat{y}=6.28+0.18x$; (2)$(9.02,9.30)$。

9.6 $H_0:\mu_1=\mu_2=\mu_3$，$k=3,n=5,T_1=210,T_2=150,T_3=220,T=580,H=T_1^2+T_2^2+T_3^2=115000$，15 个数据平方和 $G=23174$，故有：
$SS_A=1/5\cdot H-1/15\cdot T^2=573.33,SS_e=G-1/5\cdot H=174,SS_T=G-1/15\cdot T^2=747.33$。
$F_\alpha(k,k(n-1))=F_{0.05}(2,12)=3.89<19.77=F_1$，故拒绝 H_0，认为有显著差异。

9.7 $\overline{x_3}=\overline{x_2}=44-30=14$，即我们估计：用工艺三生产的电池，平均来说其寿命比用工艺二生产的长 14 小时。
取 $\alpha=0.05$，作区间估计：$\sqrt{14.5}=3.808,n=5,\sqrt{2/n}=0.632,k=3,t_{0.025}(12)=2.179,0.63\times3.808\times2.179=5.242$。
故置信水平为 0.95 的区间估计为：$14\pm5.244=(8.765,19.244)$。

9.8 $H_0:\mu_1=\mu_2=\mu_3$，$k=3,N=4+6+5=15,T_1=83,T_2=138,T_3=133,T=354$，15 个数据平方和 $G=8584$，故有：
$SS_A=T_1^2/4+T_2^2/6+T_3^2/5-354^2/15=79.65,SS_e=G-(T_1^2/4+T_2^2/6+T_3^2/5)=149.95,SS_T=G-T^2/15=229.6$。
$F_\alpha(k-1,N-k)=F_{0.05}(2,12)=3.88>3.187=F_1$，故不能拒绝 H_0，认为无显著影响。

9.9 对品牌因素提出假设：$H_0:\mu_1=\mu_2=\mu_3$，品牌因素对销售量无显著影响。

对地区因素提出假设：$H_0 : \mu_1 = \mu_2 = \mu_3$，地区因素对销售量无显著影响。

$SS_T = (558-546.1667)^2+\cdots+(616-546.1667)^2 = 22496.8333, SS_A = 2\times((511-546.1667)^2+\cdots+(550-546.1667)^2) = 4466.333, SS_B = 620.1667, SS_e = SS_T - SS_A - SS_B = 17410.33$

$MS_A = SS_A/(k-1) = 2233.1667, MS_B = SS_B/(r-1) = 620.1667, MS_e = SS_e/(k-1)(r-1) = 8705.1667$

$F_A = MS_A/MS_e = 0.256533, F_B = MS_B/MS_e = 0.071241$

因$F_{0.05}(2,2) = 0.79584$，故不拒绝地区因素的假设，认为地区因素对销售量无显著影响。

因$F_{0.05}(1,2) = 0.81454$，故不拒绝品牌因素的假设，认为品牌因素对销售量无显著影响。

9.10 $x_{jk} - \overline{x} = (x_{jk} - \overline{x}_{j.}) + (\overline{x}_{j.} - \overline{x})$，两边平方再对 j 和 k 求和，即可得

$$\sum_{j,k}(x_{jk} - \overline{x})^2 = \sum_{j,k}(x_{jk} - \overline{x}_{j.})^2 + \sum_{j,k}(\overline{x}_{j.} - \overline{x})^2 + 2\sum_{j,k}(x_{jk} - \overline{x}_{j.})(\overline{x}_{j.} - \overline{x})$$

因为

$$\overline{x}_{j.} = \frac{1}{b}\sum_{k=1}^{b} x_{jk.}$$

所以最后一个和项为 0，得证。

9.11 假定不同的汽车用行表示，不同的驾驶员用列表示。现在将不同的汽油 A, B, C, D 随机的安排到行和列中，要每一行或每一列每一字母仅出现一次。因此每一驾驶员有机会驾驶每一辆汽车和使用每一型号的汽油（没有一辆汽车用同一种汽油被驾驶两次）。

现在再随机安排 4 条道路的使用，记为 $\alpha, \beta, \gamma, \delta$。限定它们有与对拉丁字母同样的要求。因此，每个驾驶员将有机会沿每一条道路驾驶。一个可行的安排如下表：（4 列表示 4 个驾驶员）

汽车1：$B_\gamma, A_\beta, D_\delta, C_\alpha$

汽车2：$A_\delta, B_\alpha, C_\gamma, D_\beta$

汽车3：$D_\alpha, C_\delta, B_\beta, A_\gamma$

汽车4：$C_\beta, D_\gamma, A_\alpha, B_\delta$

9.12 假设分组有 A 分群记为 $A_1, \cdots, A_a$，B 分群记为 $B_1, \cdots, B_b$，C 分群记为 $C_1, \cdots, C_c$。在 A_j, B_k, C_l处的值记为 x_{jkl}。值 $\overline{x}_{jk.}$ 为 A_j 和 B_k 保持固定时对 C 分群的平均值，类似地有 $\overline{x}_{j.l}$ 和 $\overline{x}_{.kl}$。值 $\overline{x}_{j..}$ 是 A_j 固定时对 B 和 C 分群的平均，等等。最后，总平均记为 $\overline{x}$。

总方差给定为

$$v = \sum_{j,k,l} (x_{jkl} - \overline{x})^2$$

它可以分解成以下 7 个方差的和：$v_A, v_B, v_C, v_{AB}, v_{BC}, v_{CA}, v_{ABC}$。这些方差是同型分群间的方差和不同型分群间的方差（交互项）。全体分群间的交互项称为剩余或随机方差。

12.1

索引

参考文献

[1] 华罗庚. 《高等数学引论》. 科学出版社, 1963.

[2] 华罗庚. 《华罗庚科普著作选集》. 上海教育出版社, 1984.

[3] 陈希孺，陈桂景等. 《线性模型参数的估计理论》. 科学出版社, 1985.

[4] 《中国大百科全书数学卷》. 中国大百科全书出版社, 1988.

[5] 陈家鼎，孙山泽，李东风. 《数理统计学讲义》. 高等教育出版社, 1993.

[6] 王梓坤. 《概率论基础及其应用》. 北京师范大学出版社, 1996.

[7] 《英汉数学词汇》. 科学出版社, 1997.

[8] 《现代数学手册》. 华中科技大学出版社, 1999.

[9] 陈希孺. 《高等数理统计学》. 中国科技大学出版社, 1999.

[10] 陈希孺. 《数理统计学简史》. 湖南教育出版社, 2000.

[11] 张贤达. 《矩阵分析与应用》. 清华大学出版社, 2004.

[12] 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.

[13] S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 1985.

[14] R. B. Ash and C. A. Doléans-Dade. *Probability & Measure Theory*. Elsevier, second edition, 2000.

[15] D. R. Bellhouse. The Reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth. *Statistical Science*, 19(1):3–43, 2004.

[16] 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.

[17] 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.

[18] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Prentice-Hall, Inc., second edition, 2001.

[19] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., 1995.

[20] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[21] C. M. Bishop. *Pattern Recognition and Machine Learning*. Spring Science+Business Media, LLC, 2006.

[22] E. A. Bishop. *Foundations of Constructive Analysis*. McGraw-Hill, Inc., 1967.

[23] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, second edition, 2002.

[24] Y. S. Chow, H. Robbins, and D. Siegmund. *The Theory of Optimal Stopping*. Dover Publications, 1991.

[25] K. L. Chung. *A Course in Probability Theory,*《概率论教程》. Academic Press, third edition, 2001.

[26] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. McGraw-Hill Companies, second edition, 2001.

[27] 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.

[28] H. Cramér. *Mathematical Methods of Statistics,*《统计学数学方法》. Princeton University Press, 1946.

[29] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[30] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41:1–31, 1979.

[31] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[32] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[33] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[34] 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.

[35] W. Feller. *An Introduction to Probability Theory and Its Applications,* 《概率论及其应用》第一卷*)*, volume 1. John Wiley & Son, Inc., 1968.

[36] W. Feller. *An Introduction to Probability Theory and Its Applications,* 《概率论及其应用》第二卷*)*, volume 2. John Wiley & Son, Inc., 1971.

[37] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society London, Series A*, 222A:309–368, 1922.

[38] D. Freedman, R. Pisani, R. Purves, and A. Adhikari. *Statistics*. W. W. Norton & Company, Inc., 1991.

[39] 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.

[40] B. V. Gnedenko. *The Theory of Probability,*《概率论教程》. Mir Publishers, Moscow, third edition, 1978.

[41] G. H. Golub and C. F. van Loan. *Matrix Computations,*《矩阵计算》. John Hopkins University Press, 1996.

[42] W. S. Goseet. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[43] Z. Govindarajulu. *Elements of Sampling Theory and Methods*. Prentice Hall, 1999.

[44] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison Wesley Publishing Company, Inc., second edition, 2002.

[45] A. Hald. *A History of Probability and Statistics and Their Applications before 1750*. John Wiley & Son, Inc., Hoboken, New Jersey, 2003.

[46] A. Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. Springer Science+Business Media, LLC, 2007.

[47] P. R. Halmos. *Measure Theory,*《测度论》. Springer Verlag, 1974.

[48] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition, 1952.

[49] A. Heyting. *Intuitionism: An Introduction*. North-Holland, 1971.

[50] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[51] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[52] E. T. Jaynes. The well-posed problem. *Foundations of Physics*, 3:477–493, 1973.

[53] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education, Inc., fifth edition, 2003.

[54] M. Kline. *Mathematical Thought From Ancient to Modern Times,*《古今数学思想》. Oxford University Press, 1972.

[55] D. E. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley Publishing Company, Inc., third edition, 1998.

[56] A. N. Kolmogorov. *Foundations of The Theory of Probability*. Chelsea Publishing Company, 1956.

[57] A. N. Kolmogorov and S. V. Fomin, Halmos. *Elements of the Theory of Functions and Functional Analysis*. Graylock Press, 1961.

[58] S. Kotz and N. L. Johnson. *Breakthroughs in Statistics*, volume 1. Spring-Verlag New York, Inc., 1992.

[59] E. L. Lehmann and G. Casella. *Theory of Point Estimation,*《点估计理论》. Spring-Verlag New York, Inc., second edition, 1998.

[60] D. V. Lindley. The philosophy of statistics. *The Statistician*, 49(3):293–337, 2000.

[61] M Loève. *Probability Theory*. Springer-Verlag Inc., fourth edition, 1977.

[62] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[63] J. Neyman. On the problem of confidence intervals. *Annals of Mathematical Statistics*, 6(3):111–116, 1935.

[64] 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.

[65] J. Neyman. Fiducial argument and the theory of confidence intervals. *Biometrika*, 32(2):128 - 150, 1941.

[66] 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.

[67] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[68] K. Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.

[69] V. V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, 1975.

[70] 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.

[71] I. Prigogine. *The End of Certainty: Time, Chaos, and the New Laws of Nature*. The Free Press, 1997.

[72] M. H. Quenouille. Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11:18–44, 1949.

[73] M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 61:353–360, 1956.

[74] C. R. Rao. *Linear Statistical Inference and Its Applications*. John Wiley & Son, Inc., second edition, 1973.

[75] C. R. Rao. R. A. Fisher: The founder of modern statistics. *Statistical Science*, 7(1):34–48, 1992.

[76] A. Rényi. *Probability Theory*. American Elsevier Publishing Company, Inc., New York, 1970.

[77] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[78] V. K. Rohatgi. *An introduction to Probability Theory and Mathematical Statistics*. John Wiley & Son, Inc., 1976.

[79] W. Rudin. *Principles of Mathematical Analysis*. The McGraw-Hill Companies, Inc., third edition, 1976.

[80] W. Rudin. *Real and Complex Analysis*. The McGraw-Hill Companies, Inc., third edition, 1987.

[81] W. Rudin. *Functional Analysis*. The McGraw-Hill Companies, Inc., second edition, 1991.

[82] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[83] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Spring-Verlag New York, Inc., 1995.

[84] A. N. Shiryayev. *Probability*. Spring-Verlag New York Inc., 1984.

[85] B. W. Silverman. *Density Estimation*. London: Chapman and Hall, 1986.

[86] D. J. Spiegelhalter, A. Thomas, N. G. Best, and W. R. Gilks. *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.50*. MRC Biostatistics Unit, Cambridge, 1995.

[87] S. M. Stigler. Thomas Bayes' Bayesian inference. *Journal of the Royal Statistical Society, Series A*, 145:250–258, 1982.

[88] J. W. Tukey. Bias and confidence in not-quite large samples. *The Annals of Statistics*, 29(2):614–623, 1958.

[89] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. The McGraw-Hill Companies, Inc., 1997.

[90] A. Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947.

[91] L. Wasserman. *All of Nonparametric Statistics,*《现代非参数统计》. Spring-Verlag New York, Inc., 2005.

[92] S. Weisberg. *Applied Linear Regression,*《应用线性回归》. John Wiley & Sons, Inc., second edition, 1985.

[93] K. M. Wolter. *Introduction to Variance Estimation,*《方差估计引论》. Spring-Verlag New York, Inc., 1985.