

概率论引论

谢践生

2021 年 11 月中旬草稿
(目前: 2021/12/08 版本)

目录

| | |
|---------------------------------|-----------|
| 前言与导读 | vi |
| 1 引言：随机现象与概率论 | 1 |
| 1.1 随机现象与随机事件 | 1 |
| 1.2 频率与概率 | 2 |
| 1.3 概率论简史 | 4 |
| 习题 1 | 13 |
| 2 从古典概率模型、几何概率模型到概率论的公理 | 17 |
| 2.1 事件与集合 | 17 |
| 2.2 古典概率模型 | 21 |
| 2.2.1 古典概率模型简介 | 21 |
| 2.2.2 计数方法简介 | 22 |
| 2.2.3 古典概率模型应用实例：(1) Polyá 坛子模型 | 23 |
| 2.2.4 古典概率模型应用实例：(2) 同生日问题 | 24 |
| 2.2.5 古典概率模型应用实例：(3) 唱票问题与反射原理 | 25 |
| 2.2.6 古典概率模型应用实例：(4) 图论中的染色问题 | 26 |
| 2.3 几何概率模型 | 27 |
| 2.3.1 几何概率模型简介 | 27 |
| 2.3.2 几何概率模型应用实例 | 28 |
| 2.3.3 Bertrand 悖论 | 29 |
| 2.4 概率的公理化及其性质 | 31 |
| 2.4.1 概率的公理 | 31 |
| 2.4.2 概率的基本性质 | 32 |
| 2.4.3 Jordan 公式 | 33 |
| 习题 2 | 34 |
| 3 经典条件概率与事件独立性 | 39 |
| 3.1 经典条件概率 | 39 |
| 3.1.1 条件概率空间与经典条件概率的定义 | 39 |
| 3.1.2 乘法公式 | 42 |
| 3.1.3 全概率公式 | 43 |
| 3.1.4 Bayes 公式 | 47 |
| 3.2 乘积概率空间与事件独立性 | 52 |

| | | |
|----------|--|------------|
| 3.2.1 | 乘积概率空间 | 53 |
| 3.2.2 | 事件独立性的定义 | 54 |
| 3.2.3 | 重复实验与条件实验 | 55 |
| 习题 3 | | 58 |
| 4 | Lebesgue 积分理论简介 | 60 |
| 4.1 | 可测集与可测函数 | 60 |
| 4.1.1 | Borel 可测集与一般可测集 | 60 |
| 4.1.2 | Borel 可测函数与一般的可测映射 | 61 |
| 4.2 | Lebesgue 测度与一般可测空间中的非负测度 | 62 |
| 4.2.1 | 测度、预测度、外测度的定义 | 62 |
| 4.2.2 | Carathéodory 扩张与 Lebesgue 测度 | 67 |
| 4.3 | 欧氏空间中的 Lebesgue 积分 | 72 |
| 4.3.1 | Lebesgue 积分的定义 | 72 |
| 4.3.2 | Lebesgue 积分与极限的交换 | 74 |
| 4.3.3 | Lebesgue 积分与 Riemann 积分 | 77 |
| 4.3.4 | 微积分基本定理的探讨 | 80 |
| 4.3.5 | 重积分与累次积分—Fubini 定理 | 82 |
| 4.4 | 抽象测度的 Lebesgue 积分 | 83 |
| 4.4.1 | 抽象测度的 Lebesgue 积分的定义 | 83 |
| 4.4.2 | 抽象测度的 Lebesgue 积分的极限交换问题 | 84 |
| 4.4.3 | 微积分基本定理的推广：Radon-Nikodym 定理 | 85 |
| 习题 4 | | 89 |
| 5 | 随机变量 (I) | 90 |
| 5.1 | 随机变量、随机向量及其分布律 | 90 |
| 5.1.1 | 随机变量与随机向量的定义 | 90 |
| 5.1.2 | 随机变量/向量的相互独立性 | 94 |
| 5.2 | 离散型分布 | 95 |
| 5.2.1 | 离散型随机变量的定义 | 95 |
| 5.2.2 | 常见离散型分布 | 97 |
| 5.2.3 | 两个极大似然估计的例子 | 99 |
| 习题 5 | | 101 |
| 6 | 数学期望 | 102 |
| 6.1 | 数学期望的定义 | 102 |
| 6.2 | 数学期望的性质 | 104 |
| 6.3 | 数学期望的计算公式 | 105 |
| 6.4 | 方差、协方差与独立性 | 108 |
| 习题 6 | | 110 |
| 7 | 条件数学期望与条件分布律 | 112 |
| 7.1 | 条件数学期望的定义 | 112 |
| 7.1.1 | 数学期望的一个性质 | 113 |
| 7.1.2 | 经典条件概率到相应的“经典”条件数学期望 | 113 |

| | | |
|-----------|---|------------|
| 7.1.3 | 抽象的条件数学期望的定义 | 114 |
| 7.1.4 | 由条件数学期望 $\mathbb{E}[\cdot \mathcal{G}]$ 到条件概率 $\mathbb{P}(\cdot \mathcal{G})$ | 117 |
| 7.1.5 | 抽象条件数学期望 $\mathbb{E}[Y \mathcal{G}]$ 的计算 | 117 |
| 7.2 | 条件数学期望的性质 | 118 |
| 7.3 | 条件数学期望的积分变换公式 | 119 |
| 7.4 | 条件分布律及条件数学期望的计算公式 | 120 |
| 7.5 | 条件数学期望与独立性 | 123 |
| 习题 7 | | 124 |
| 8 | 随机变量 (II) | 126 |
| 8.1 | 连续型分布与密度函数的定义 | 126 |
| 8.2 | 连续型随机变量的数学期望计算公式 | 127 |
| 8.3 | 边缘密度、条件密度与条件分布函数 | 128 |
| 8.4 | 概率微元法 | 129 |
| 8.5 | 常见连续型分布 | 132 |
| 习题 8 | | 144 |
| 9 | 随机变量 (III) | 145 |
| 9.1 | 一个奇异型分布的例子及随机变量的分类 | 145 |
| 9.2 | 分布函数的性质与随机变量的实现 | 147 |
| 9.2.1 | 分布函数的性质 | 147 |
| 9.2.2 | 次序统计量 | 148 |
| 9.2.3 | 随机数发生器的构造 | 151 |
| 习题 9 | | 152 |
| 10 | 随机变量列的收敛与大数律 | 155 |
| 10.1 | Chebyshev 不等式 | 155 |
| 10.2 | 各种收敛性的定义 | 157 |
| 10.3 | 几种收敛之间的关系 | 158 |
| 10.3.1 | L^p -收敛、依概率收敛与几乎处处收敛之间的关系 | 158 |
| 10.3.2 | 阅读材料：随机序与随机控制收敛定理 | 161 |
| 10.3.3 | 依概率收敛与依分布收敛之间的关系 | 163 |
| 10.4 | 大数律简介 | 163 |
| 10.4.1 | Borel-Cantelli 引理 | 163 |
| 10.4.2 | 从 Bernoulli 弱大数律到 Borel 强大数律 | 167 |
| 10.4.3 | 从 Khintchine 弱大数律到 Kolmogorov 强大数律 | 169 |
| 10.5 | 应用举例 | 173 |
| 习题 10 | | 175 |
| 11 | 随机变量的特征函数与中心极限定理 | 182 |
| 11.1 | 特征函数与分布函数 | 182 |
| 11.2 | 特征函数的唯一性定理 | 184 |
| 11.3 | 特征函数的连续性定理 | 187 |
| 11.4 | 特征函数的等价刻画 | 190 |
| 11.5 | 中心极限定理 | 191 |

| | |
|--|------------|
| 习题 11 | 194 |
| 12 检验我们的概率建模：步入《统计学》! | 201 |
| 12.1 对总体分布的分布律的检验 | 202 |
| 12.1.1 定性检验：P-P 图或 Q-Q 图 | 202 |
| 12.1.2 定量检验：分布函数 F 连续的情形 | 203 |
| 12.1.3 定量检验：分布函数 F 离散的情形 | 204 |
| 12.2 对独立性的检验 | 205 |
| 12.2.1 技术准备：关于卡方分布的讨论 | 205 |
| 12.2.2 Pearson-Fisher 定理 | 206 |
| 习题 12 | 207 |
| A 单调类定理及其在概率论中的应用 | 209 |
| A.1 集合类简介 | 209 |
| A.2 单调类定理 | 212 |
| A.3 乘积概率测度的存在唯一性 | 213 |
| A.4 定理 5.1.1 的证明：(5.3) \Rightarrow (5.5) 部分 | 214 |
| 习题 A | 215 |
| B 数学期望、条件数学期望的一些性质的证明 | 216 |
| B.1 数学期望的一些性质的证明 | 216 |
| B.1.1 数学期望的线性性质 | 216 |
| B.1.2 Jensen 不等式与矩不等式的证明 | 217 |
| B.1.3 Hölder 不等式与 Minkowski 不等式的证明 | 218 |
| B.2 条件数学期望的一些性质的证明 | 218 |
| B.2.1 条件数学期望 $\mathbb{E}[\cdot \mathcal{G}]$ 的单调性与 \mathcal{G} -线性性质 | 219 |
| B.2.2 条件数学期望 $\mathbb{E}[\cdot \mathcal{G}]$ 的性质 (6) 的证明 | 219 |
| B.2.3 条件 Jensen 不等式的证明 | 220 |
| 习题 B | 220 |
| C 矩问题与 Laplace 变换 | 221 |
| C.1 矩问题 | 221 |
| C.1.1 Hausdorff 矩问题 | 222 |
| C.1.2 Hamburger 矩问题：可解的条件 | 222 |
| C.1.3 Hamburger 矩问题：解的唯一性与不唯一性 | 223 |
| C.2 Laplace 变换 | 225 |
| C.2.1 Laplace 变换的定义与基本性质 | 226 |
| C.2.2 唯一性定理与 Laplace 变换的反演公式 | 228 |
| C.2.3 Laplace 变换的连续性定理 | 229 |
| C.2.4 Laplace 变换的刻画定理 | 229 |
| C.2.5 Laplace 变换的 Tauber 定理 | 230 |
| 习题 C | 233 |
| 参考文献 | 237 |

前言与导读

本讲义是我在 2021 年秋季学期为复旦大学数学科学学院的本科生主讲概率论荣誉课程的授课过程中逐渐整理而来。在此之前，我常年从事学院的多门概率论专业课程的教学工作，因此讲义中毫无疑问有众多内容来源于那些教学工作中积累的素材。特别是在开设本科生概率论荣誉课程之前，我刚好差不多完成了自己编撰的《现代概率论基础》课程的电子讲义的修订工作，这为本讲义的出现做了比较充分的准备工作。

从我本科阶段开始概率论方向的相关课程学习，直到自己成为教师来教授本科生概率论，所使用的教材总是或多或少让我感觉不舒服，有众多迷思：几乎所有的概率论教材对于为何（及为何如此模式地）引入相关的概念缺乏必要且令人信服的解说。比如说，为何把经典条件概率定义为 $\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$ ？为何把两事件的独立性定义为 $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$ ？为何大多数教材讲完概率论的几条公理后，后续介绍随机变量时不再明确说明随机变量赖以生存的样本空间了？一个取定了的概率空间怎么可能容纳源源不断在讨论的各式各样的随机变量？在实务工作中能用频率来代替概率的道理是什么？其他数学领域中的数学结构如线性空间、群、环、域、流形等有同构的概念，为啥概率论中概率空间作为一种概率结构不好好讲讲它的同构问题？为什么数学期望、条件数学期望按照大多数初等概率论教材中的模式定义？明明到了概率论的高级课程中数学期望、条件数学期望又有另一套抽象定义，这两套定义怎么联系起来？另外，数学期望、条件数学期望定义出来到底是为了解决什么实际问题？总不能说成是概率论学者们自娱自乐的数字计算游戏吧？类似的问题还有很多。所有这些问题在国内外的为初学者而准备的概率论教材中几乎都没有给出答案。当然这些“为何”及“如何”的问题我最终从概率论的高级课程的教材等材料的字里行间揣摩到了我自己认可的答案，这也是本讲义的写作动机。

带着这些“为何”与“如何”的问题（起初一大部分问题已在前期的学习和科研路上摸索出自己的回答），我在复旦多年的本科生概率论教学实践中，几乎每一年的讲法都与上一年有所差别，其目的就是在探索一种自己满意的（初等）概率论教学内容安排体系，使得学生们再次遇到我当初的困惑场景时，在这个教学体系下能够自然而然地如冬雪春融。这是一个教学的迭代、优化过程；这个过程反映在电子讲义上，我的《现代概率论基础》电子讲义草稿始于 2012 年，直到 2021 年才达到我基本满意的状态，但仍需少许修订，近乎十年磨一剑；如果加上前期的教学磨练，则远超十年。这是一个把过往大多数初等概率论教材的教学内容次序推倒之后逐步重新构建的过

程；当然概率论的理论还是那些理论，改变的是它们的出场次序以及介绍的模式；无疑，这需要勇气、决心与大量的付出。这也是一个痛并快乐的过程。痛，除了来源于这个探索过程中自身的精神折磨，还来源于外界。众所周知，当前的学界生态状况是重视科研远远重于教学的情况；一位对我常年关爱照顾的年长同事劝诫我，当前的状况是好的科研论文才是“硬通货”，教学中的投入对得起自己的良心就足够，应该花更大气力在科研上。这位我敬重于心犹如兄长的同事其实自己在教学上也是富有思想、兢兢业业的，但他知道教材的编写费时费力，即使出版后得到好评也就那样；特别是教学上投入的时间和心血非常难以量化比较。他说的这些我都能理解，因为我即是这个体系中的感同身受者；因此我也并没有说就放弃科研：融入科研体悟的教学和通过教学促进科研这两种快乐的体会我都真真切切地感受过。真正的痛是在教学过程中的不被学生理解，尽管这种状况不多。在我多年的教学过程中，有两次被学生投诉的历史（其中一次投诉原因让人哭笑不得，在此也不细说），还好两次投诉院系领导通融理解，没有最终升级成教学事故而进一步造成悲剧；也曾有同学询问，“为啥课程中要讲那么难的内容，反正我们也学不懂，最后考试也不考”云云，让我略感愤慨而联想到如今的教学管理又有点无言以对。我由于主要为本院的学生讲授专业课程，总体来说师生对数学的价值取向与品味趋同，我的授课方式还是能被大多数同学理解；身边一些为外院系开设数学课程的同事、朋友在教学中发生的类似故事更多。这里把这些事情写出来，只为呼吁同学们理解那些在教学一线兢兢业业服务的老师们，特别是青年教师。教学经验是在教学实践的磨练中逐步提高的；我清清楚楚记得自己在职业生涯的早期有过一次挂黑板的经历；至于授课过程中的口误或笔误，应该说只要继续从事教学我就无法向各位保证能避免。同时，我也呼吁管理层对于学生们的教学投诉审慎应对与处理。

除了中学阶段的初步积累，我的一点点物理学理论底子始于北大的理科实验班的“攀登计划”，现如今似乎已改名为“元培实验班”或“元培计划”。后来在跟从钱敏先生、并在刘培东教授实际指导下从事随机动力系统研究的过程中，逐步积攒了点统计力学、热力学的理论基础，也因受钱敏先生影响对物理学发生更多的兴趣。到复旦大学开始概率论课程的教学工作后，这才又慢慢开始去追寻概率论发展的历史。在反复阅读、思考的过程中注意到Hilbert在他倡导的公理化运动中，把概率和力学的公理化置于物理学的公理化的范畴内，大受震撼；但直到今年秋季的某天，才突然感觉自己摸索到了完成前述重建我心目中的初等概率论教学体系任务的钥匙：把概率论中所有的概念和理论置于概率空间是对现实中的随机现象进行的概率建模这一构想下来进行论述。当然有关论述最好能与学科发展的史实交叉融合。如今的讲义就是在这样的指导精神下组织相关材料，逐步展开有关理论论述的。个人感觉这种组织材料的方式可以给读者呈现概率理论是如何从模糊的直观中萌芽、生长的。Newton本人发展他的力学理论体系似乎也是如此；有关史料给我的感觉是：他对微积分理论的创建只是在力学理论构建过程中，出于解决有关问题的需要，在前人（如Archimedes、Galileo、Pascal等）工作基础上顺手发展的工具，在职业生涯的早期他自己也并没觉得有多了不起，相反他更看重基于直观和简单物理现象来发展力学理论体系的哲学思想及

与之对应的数学原理^{*}；到了职业生涯的后期他才被迫卷入微积分发明的优先权之争。

在此我们介绍一下全书的框架结构。在前述“概率建模”的观点和指导精神下，考虑到教学的具体安排，我们把全书正文内容分成了12章。

第1章用于引出概率论关心的随机现象与随机事件，并简单回顾概率论的发展历史。第2章论述了初学者最容易理解和接受的两类满足等可能性假设的概率模型：古典概率模型和几何概率模型；注意到在几何概率模型中，并不是所有的事件都能谈论概率，因此在古典概率模型和几何概率模型基础上升华出公理化的概率模型也就不难理解，尽管这一点在概率论学科的实际发展进程上整整耗费了100多年（从Laplace在1812年发表《分析概率论》开始，到Kolmogorov在1933年发表德文著作《概率论基础》，期间耗费了121年）。第3章是全书中一个极其特殊的章节，介绍了经典条件概率与事件独立性，是作者的“概率建模”观点在全书的第一次重要的落实[†]。我们基于古典概率模型，导出了古典概率模型中的条件概率的概念，并把这一概念外推至一般的公理化概率模型；同样基于古典概率模型讨论了如何为使用两套赌博道具的复杂赌博进行概率建模，由此导出了公理化概率论框架下概率空间的乘积空间的概念，并在此基础上定义了事件独立性的概念。这也很自然地回答了《测度论》中为何要讨论乘积可测空间与乘积测度空间、为何要探讨测度扩张问题，同时乘积空间也是初等概率论的后续课程中有关随机过程的轨道构造等理论探讨的必备手段；我们的讲义作为初学者的教材，在这些地方只能简单引用《测度论》等的相关结论，而把这些结论的略细致论述放在了下一章，即第4章[‡]。概率论的发展（特别是有关随机变量及其数学期望的公理化定义）受到其他多个数学分支（例如测度论、微积分，特别是Lebesgue积分理论）发展的影响与推动。为此，我们在全书中插入了另一个特殊的分析味道浓重的章节作为第4章，用于介绍二十世纪才发展起来的Lebesgue积分的理论；除了传统的欧氏空间中的Lebesgue积分理论，里面顺带介绍了一般测度的抽象Lebesgue积分的理论。其中，欧氏空间中的Lebesgue积分理论主要是为连续型随机变量的密度、分布律计算等做铺垫，一般测度的抽象Lebesgue积分理论主要是为数学期望、条件数学期望的定义做技术准备。后续我们把随机变量的介绍拆分成三个不连贯的章节进行：第5章先介绍随机变量及其分布律和随机变量的独立性的概念，之后介绍了随机变量中最容易理解的离散型随机变量的概念以及常见的离散型分布；第8章先介绍连续型随机变量及其密度函数的概念，之后介绍了概率微元法用于解决（光滑可逆或分片光滑可逆）变换下连续型随机变量的密度函数的计算，最后介绍了常见连续型分布及它们的密度函数；第9章介绍了一个奇异型随机变量的例子并讨论了随机变量的分布分类的问题，此外还介绍了一般分布函数的性质和随机变量的实现问题。在第5章，我们指出，随机变量的定义应当在可测函数的认同基础上做一个“补丁程序”，这与我们“概率建

^{*}Newton的经典著作是《自然哲学的数学原理》。

[†]但毫无疑问并不是整个概率论发展历史中的第一次；我个人觉得古典概率模型的建立可以认为是第一次虽然粗浅但也完美的“概率建模”，它第一次把模糊的自然语言中的概率在最容易理解的情况下予以了数学的精确化定义。

[‡]这类知识的教学置于《测度论》之类的课程更为适合，在本课程的教学由于教学时间等原因建议只做介绍而不展开讨论。

模”的精神是一致的。在第 6、7 章，我们分别介绍了概率论中两个重要的工具：数学期望与条件数学期望。在第 6 章的论述中，我们通过回顾历史上 Huygens 提出的期望收益的概念给出严谨的数学期望的定义，并通过引入保测映射的观点最终给出数学期望的计算公式。第 3 章定义的条件概率结合第 6 章刚刚定义的数学期望概念，在第 7 章中进一步发挥，导出了经典条件概率对应的条件数学期望，这个概念进一步拓广成更一般的条件数学期望的定义，由此引出更抽象的条件概率的概念；特别地，我们进一步可以探讨所谓的条件分布律，基于此我们最终给出了概率论中常用的一个随机变量关于另一个随机变量的条件数学期望的计算公式；这些内容请参见第 7 章的有关论述。以上多处概念的论述都遵循了物理学或数学中拓广概念的原则，并且介绍这些概念时都是尽力先从最容易理解的情形开始的。

本书最后三章是全书的高潮。在介绍了随机变量的几种收敛性概念后，第 10 章讨论了大数律（这是对本书第 1 章“频率稳定性”的呼应；另外，此处的 Chebyshev 不等式的一个推论也给出了前述随机变量的“补丁程序”的一种有效率的落实方法）。第 11 章通过介绍特征函数理论讨论了中心极限定理。为了呼应我们概率建模的观点，我们在最后一章，第 12 章（建议作为阅读材料，在实际教学中仅粗略介绍即可）中还给出了可以通过大样本数据用以检验模型准确性的几个定理：

- (1) 作为统计学基本定理的 Glivenko-Cantelli 定理。实务上我们可以基于这个理论，利用大样本数据做 P-P 图或 Q-Q 图，形象（或者说“定性”）地检验样本是否来自某个具体的分布；
- (2) Kolmogorov 定理；这个定理由于其证明需要更高级的技巧，在本书中就没有提供证明细节。实务上我们可以基于这个定理，构建 Kolmogorov 检验，“定量”地检验大样本是否来自于某个分布函数连续的分布；
- (3) Pearson 定理。实务上，基于 Pearson 定理，我们可以构建拟合优度检验，“定量”地检验大样本是否来自于某个简单的离散型分布；
- (4) Pearson-Fisher 定理。实务上，基于这个定理，我们可以构建关于独立性的列联表检验，基于大样本数据“定量”地判断二维数据之间是否具有独立性。

当然，以上四个定理只做到了对初等概率论中最重要的随机变量的分布律与独立性的检验。虽然在概率与统计等方向学者们几百年的研究积累下，现代概率论与现代统计学的理论已经非常丰富，发展得比较成熟，但贴近我们关心的一些更复杂的随机现象的概率建模理论以及对这样的概率模型的检验理论这两个方面仍然可能有待读者们去探索和发展。大数律和中心极限定理是概率论中的经典理论与研究范式。正如分析学中极限概念的引入与探讨是初等数学与高等数学之间的分水岭，我们特地以第 10、11 章节作为初等概率论课程的两个收尾章节。而第 12 章，最后一个收尾章节，则可视作我们概率建模观点下对数理统计理论的广告：除了该章节介绍的模型检验的理论，更多的对模型的检验问题留待读者在《统计学》的有关课程中去学习和研究；但另一方面，我们也不难看到，要论证上述 Kolmogorov 定理，需要随机过

程的语言以及泛函中心极限定理等更高端的概率理论，因此这也是对高等概率论课程的广告。

为了全书的完备性，我们还准备了三个附录章节：前两个附录章节用来补充探讨全书中因教学安排的原因未能详尽展开的一些证明细节；后一附录章节用来继续讨论在特殊场合能用来讨论随机变量的分布律的两种统计特征——矩和分布函数的 Laplace 变换。

作者希望这样的结构安排能够激发初学者学习概率论的热情与兴趣，提高学习的效率。当然，这样一来，全书虽有主线，但内容还是比较繁杂的*，对于学生的自学和教师的教学都提出了挑战。本书对学生的数学基础有较高的要求，相对而言更适合专业方向的学生学习。我的建议是，学生初次学习时要学会抓住主线，忽略一些对数学基础要求较多的证明细节，在学习概率论发展历史和概率理论体系的构建方法的过程中去学习和掌握有关理论，在数学基础提高后宜再次学习补上当初忽略的那些理论细节；对教师的具体授课而言，就需要教师根据实际需要调整教学内容。

很庆幸一路走来获得了一些经济援助、结识了一些尊敬可爱的师友，温暖了内心；因为内心有爱这才能在如今的教学岗位上做到持续用心与坚守底线。我大学阶段是一名来自农村的贫困生，幸得“晨兴助学金”资助，又得彼时舍友照顾介绍勤工俭学机会，稍解彼时困顿。彼时的老师们在生活上大多看淡物质享受与物质利益、却又能理解并在力所能及及时纾解学生们生活中的困苦；在科研上重视对未知的探索过程中的严谨与求真、求实；在教学上则一丝不苟、兢兢业业，讲到自己熟悉领域时神采飞扬，与学生的讨论中平等相处；所有这些风貌让我产生了以后要成为这样的教师的冲动。研究生阶段钱敏平教授开朗、乐观的个性感染了作为学生的我，并从她那里学到了“敝帚自珍”，珍视自己于黑暗中摸索出的各种收获；而刘培东教授在指导我论文的过程中表现出来的严谨与细致让我暗下决心，日后亦当如是。钱敏先生则被我自己视作精神教父，其为人、行事及在教育 and 科研上的表现与相关思想时常被我用来回忆观照自己；斯人已逝，唯记忆永存心底，怀念时不时涌现脑海。勇士屠恶龙的童话小说中，勇士成为英雄后由于种种原因成为了另一条恶龙；生活中的挫折、苟且多了，你也可能忘记你的诗与远方。我党提出要“不忘初心”；而我也时常心心念念提醒自己，至少不要成长为自己当初厌弃的模样。说实话，走上科研和教育的道路，对于在老家农村的老母及兄姐是有无数欠疚的，因为只能成为一个无法承担更多家庭责任的自了汉了。感恩家人的理解，感恩从事慈善事业的各界人士，感恩一路走来师友的教育、提携、理解与帮助。

讲义中的内容借鉴了众多师长的材料，如我的同学王岩华向我分享的何书元老师的《概率论》PPT 资料，我自己学生时代向钱敏平（《概率论》）、陈家鼎（《测度论》）、程士宏（《高等概率论》）等教授学习概率论相关课程的笔记。在复旦的教学过程中较长时期参考了我院同事应坚刚教授的《概率论》（与何萍教授合著），其他参考过的教材资料此处就不一一细说，见参考文献；而浙大的朋友赵敏智老师提供的分享以及与她的相关讨论尤为珍贵。特此向他们致谢。

对于这个讲义来说，如果没有教学的讲台，没有讲台下大部分学生们的

*但这是概率论学科本身特点的一部分，而不是作者故意如此编写以难为学生取乐。

宽容与理解，或许它就不会诞生、即使诞生也可能夭折。因此也感谢学院提供了这个教学舞台，感谢 2021 年秋季学期参加我课程的所有同学！

本讲义目前正文草稿基本完成，后续面临习题的添加、内容细节上补充与修订等细致工作。已有内容无疑会包含一些错误或疏漏不当之处，敬请各位同学、同行持续批评指正；也欢迎你们提出写作上的建议，分享你们觉得有趣的案例资料等。

谨以此作为讲义前言与导读。2021 年 11 月 13 日。

§ 1

引言：随机现象与概率论

现代人的衣食住行的方方面面充斥着“概率”、“机会”、“可能性”、“随机”等字眼。每天早晨起床，天气预报会告知你，今天是晴天还是阴天，降水概率是多少；学生们期末考试后会估计自己获得 A 类成绩的概率有多大；求职的时候，人们也会估计自己被录用的机会有多大；在农作物收获前，农民们有时会提前估计自己农田高产的可能性有多高。走在大街上，当你和朋友不知道去哪家饭店吃饭时，你可能会掏出一枚硬币，让它“随机”地为你挑选一家。我们自然语言中的“概率”、“机会”、“可能性”通常认为是同义词。那么到底什么是概率？在本章，我们无法完整回答这个问题。但我们希望学完本书，读者对此问题已有自己的答案与体悟。

几乎所有的概率论教科书都告诉我们，概率论是研究随机现象的数量规律的数学分支学科。那么什么是随机现象？我们通常所说的某个随机事件的概率又是什么含义呢？

1.1 随机现象与随机事件

在自然界和社会生活中，我们经常遇到两类现象：一类是确定性现象，它的结果是单一而确定的；一类是不确定性现象（也称为随机现象、偶然现象），它的结果是不唯一、并且每种结果通常都有可能发生的。

例 1.1. 以下一些现象或事件是确定的（即一定发生）：

- （在地球上观察）太阳东升西落；
- （地球上的）学校操场上向上抛出的帽子最终会掉落；
- （在地球上，常温常压条件下）人血管裂开了会流血。

以下一些现象或事件是不确定的（即不一定发生，也不一定不发生，具有偶然性或随机性）：

- 打靶，击中靶心；
- 抛掷一枚硬币，有头像的一面（正面）朝上；

- 买了彩票，中奖；
- 昨收盘买入一支股票，今日收盘涨停。

在自然语言中，我们有时会把“现象”和“事件”作为同义词而不加区分；但通常我们认为“现象”是可能重复发生/出现的，而很多时候有些事件（比如同一个人类个体的出生、初恋、第一次结婚、死亡）是不可能重复发生的，重复发生的其实仅仅是其中的某种特定现象（比如同一个人类个体的恋爱、结婚或不同个体的出生、恋爱、结婚、死亡）。因此，在本书中，为了精确化，我们在对随机事件、随机现象进行概率建模时，对“现象”与“事件”两个术语会略作区分：在同一个概率模型中，“现象”与“事件”很多时候不做区分；但在两个不同又互相关联的概率模型（其中一个概率模型认为是另一个的子概率模型）中，有时我们会把其中“小”模型中的 A 事件在另一个“大”的模型中说成是 A 现象。

例 1.2. 投掷一枚硬币一次的实验（称为“单次实验”），考虑获得硬币正面这个事件/现象 A ；在重复投掷这枚硬币 n 次的“复合实验”（称为“ n 次重复实验”）中就把 A 事件看成是 A 现象。这时可以讨论：

- A 现象在 n 次重复实验中总共发生了/观察到多少次？
- A 现象在 n 次重复实验中总共发生了/观察到 k 次的可能性有多大？
- 如果 A 现象发生了，最早发生/观察到该现象的那次实验是第几次？
- 在无穷次重复实验中，为了获得首次 A 现象的发生/观察，平均而言需要做多少次重复实验？

以后我们会看到：上述第一、第三个问题各定义了一个“随机变量”；第二个问题问的是第一个“随机变量”的具体取值概率；第四个问题问的是第二个“随机变量”的“数学期望”（自然语言中的“平均值”）。这些涉及的概念我们将逐步在数学上予以精确化的定义。

1.2 频率与概率

在中文词汇中，“机会”、“几率”、“概率”等都是“可能性”的同义词。而“可能性”高或低的判断，很多时候是人对同样条件下某种（可重复发生的）现象可能发生的频繁程度的主观感受，其中可能包含了他对此现象所收集的一些客观数据与经验。这个主观感受里面包含的客观数据与正确经验越多，判断所依据的推理理论越科学，则所下的判断就越准确；反之，主观感受中主观情绪等占比越多，错误经验越多，判断的准确性就越差。这也是**赌徒谬误**、**幸存者偏差**等诸多影响判断准确性的错误认知时有发生的原因。而本课程的学习将有助于在占有相同的已有数据的基础上*，通过科学的推理，形成更正确的直觉（经验），得到更准确的判断，从而减少不必要的错误认知，进而也减少不必要的损失或人际、社会冲突。

*此处，我们特意提一下，统计学课程的学习有助于更准确、有效地收集和分析数据。在概率论与数理统计学两个学科发展的早期，二者的发展是并轨、甚至混为一体的；早期的统计学专家一般也是概率学专家，反之亦然。目前通常认为概率论是数理统计学的重要基础。

从认知上来讲，关于某事件中可反复观察到的客观现象发生可能性高低的判断，人总是通过已经累积的历史事实（可能是个人的积累，也可能是他人的积累等），先感受、总结这类现象发生的频繁程度（精确地数量化后，称为**频率**，它是观察到的现象发生的次数与实际观察次数的比值），之后再据此判定具体某事中该现象发生的可能性高低；此处，这个可能性高低程度的数量化，就称为**概率**：有些事件是一定会发生的，我们称它们是必然事件，说它们发生的概率为 1；有些事件是不可能发生的，我们称它们是不可能事件，说它们发生的概率为 0。其余一些事件在量度它们发生的可能性大小（即概率）时，就让它们的概率取值介于 0 与 1 之间：事件发生的可能性越高，其概率值就越靠近 1；事件发生的可能性越低，其概率值就越靠近 0。

基于多次实验的历史数据，关于某现象的概率通常仍然是未知且难以精确计算的，在实践中就经常采用以频率代概率的办法，其中一个原因就是人们在实验次数足够多的重复实验中经常观察到/感受到的“**频率稳定性**”现象，这一现象后来被 Bernoulli 给予了严格的数学证明，见下例中论述：

例 1.3. *Jacob Bernoulli*（雅各布·伯努利，1654/12/27–1705/8/16；瑞士，原籍比利时）是第一个理论上证明了“频率稳定性”现象的数学家。借用投掷硬币的重复实验模型的语言，可以陈述如下：设单次投掷硬币时，正面朝上这一事件 A 发生（我们也认为此时实验成功）的概率为 $p_A = \mathbb{P}(A)$ ，记 n 次重复投掷硬币的实验中总共获得了 n_A 次正面，则此 n 次重复实验成功的频率为 $f_n(A) := \frac{n_A}{n}$ ，它满足：

$$\lim_{n \rightarrow \infty} f_n(A) = p_A.$$

上式中收敛的含义是依概率收敛，在本课程后半段将会介绍。

历史上有不少数学家进行过投掷硬币的实验，见表 1.1。

表 1.1: 历史上一些著名的投掷硬币的实验

| 实验者 | 投币次数 | 正面次数 | 反面次数 | 正面频率 (%) |
|-----------------------------------|-------|-------|-------|----------|
| G. Buffon (1701–1788; 法国) | 4040 | 2048 | 1992 | 50.69 |
| De Morgan (1806–1871; 印度) | 2048 | 1061 | 987 | 51.81 |
| De Morgan | 4092 | 2048 | 2044 | 50.05 |
| Carl Pearson (1857–1936; 英国) | 12000 | 6019 | 5981 | 50.16 |
| Carl Pearson | 24000 | 12012 | 11988 | 50.05 |
| Romanovsky (1879–1954; 苏联) | 80640 | 39699 | 40941 | 49.23 |
| W. Feller (1906–1970; 美籍克罗地亚裔) | 10000 | 4979 | 5021 | 49.79 |
| Vini | 30000 | 14994 | 15006 | 49.98 |

Bernoulli 证明的“频率的稳定性”被称为 Bernoulli（弱）大数律，它是概率论中第一个极限定理。在很长时间内，这被认为是概率的（一种）古典解释（称为“**频率解释**”），甚至在早期的一些概率论教材中作为概率的定义。在统计学中，这个大数律派生出了信奉大样本理论的频率学派。

1.3 概率论简史

从人类起源开始，人们无时无刻需要面对和处理随机现象、甚至利用随机现象做决策。远古时期的人类就有利用动物的跖骨、距骨或贝壳、龟壳等道具进行占卜来做决策的记载。在公元前 3000 年前的美索不达米亚就有所谓的“20 方块”的随机游戏（又称为“乌尔的国王游戏”）；公元前 2000 多年的埃及古墓中，已有正立方体的骰子。在 Aristotle（亚里士多德，公元前 384—前 322；古希腊）时代，人们已经认识到随机性/不确定性在客观世界中的普遍性。然而，一方面，由于技术上的障碍，这些古代的道具，也就是现代观点下的“随机数发生器”，材质上通常不是充分均匀和对称的，从而人们难以通过观察和累积使用这类道具的实验数据而得出科学的规律；更重要的是另一方面，由于观念上的制约，古人经常把偶尔的结果（随机结果）归结为“神的意志”，于是很自然不相信这些随机结果可能是完全的偶然、进而放弃寻找随机事件发生的稳定的频率等科学规律。

直到 15、16 世纪前后，由于科技的发展和观念的进步，人们才开始定量地研究随机性/不确定性，并尝试从中发现客观规律。历经几个世纪学者们的探索，特别是二十世纪在数学公理化的潮流下，概率论发展成为一门严谨的数学分支学科。时至今日，概率论的蓬勃发展和广泛应用，改变了人们自牛顿力学建立以来长期形成的确定性认知观念和思维方式，成为了人们探索未知世界（自然世界、人文社会、乃至精神世界）奥秘的有力工具。

具体而言，概率论的发展，根据有关资料，大致可以分为以下几个阶段*。

概率论的萌芽时期（15 世纪到 1654 年前）这一时期，环地中海地带商业往来频繁。概率论的萌芽在当时的经济强国意大利发端，代表性人物有 L. Pacioli（帕乔利，1447/?/?—1517/6/19；意大利）、G. Cardano（卡尔丹诺，1501/9/24—1576/9/21；意大利）、Galileo Galilei（伽利略，1564/2/15—1642/1/8；意大利）等。我们知道，在 14 世纪到 17 世纪的欧洲发生了一场反映新兴资产阶级要求的思想文化运动，史称“文艺复兴”；文艺复兴最先在意大利各城邦兴起，以后扩展到西欧各国，于 16 世纪达到顶峰，带来一段科学与艺术的革命时期，揭开了近代欧洲历史的序幕，被认为是中古时代和近代的分界。根据史料记载，1494 年 Pacioli 在威尼斯出版《算术书》[†]，其中提出所谓的“赌金分配问题”：两人决定赌博若干局，事先约定谁先赢得 6 局便算赢家，赢家获得全部赌金。如果在一个人赢 5 局，另一个人赢 2 局时因故终止赌博，应如何分配赌金才合理？他给出的答案是按照 5 : 2 分；这个答案引起诸多争议。半个多世纪后，Cardano 潜心研究赌博不输的方法，1564 年左右写了《赌博之书》[‡]，在书中他提出：投掷两枚骰子，以点数和的猜测准确与否作输赢，那么压几点的赢面最大？Cardano 认为 7 点最好；他对“赌金分配问题”也进行了讨论，给出了正确思路，但未能给出正确答案。书中论证了把几率定义为有利结局数目与不利结局数目之比的功效，为后来概率的古典定义做了铺垫。Galileo 则考虑了投掷三枚骰子的问题，计算出点数和 10 和 11 出现的机会相同，均有 27 种情形导致该点数和；而点数和 9 则有 25 种情形导致该结果。这一时期的概率论研究通常以数据统计为主要手

*此处分为 5 个阶段的观点纯属作者的个人观点，与大多数文献的分法不完全相同。

[†] 《Summa de arithmetica, geometria: Proportioni et proportionalita》。

[‡] 即《Liber de ludo aleae》，1663 年才正式出版。

段，主要研究保险、赌博、占卜等实际问题。



图 1.1: B. Pascal (1623-1662)



图 1.2: P. de Fermat (1601-1665)

古典概率论形成时期（1654–1713）一般认为，作为一门学科的概率论诞生于 17 世纪中叶，标志性的事件是 B. Pascal（帕斯卡，1623/6/19–1662/8/19；法国）和 P. de Fermat（费马，1601/8/17–1665/1/12；法国）在 1654 年左右关于赌徒分配赌金问题（亦即“赌金分配问题”，但此处原始问题是：两个赌徒一者胜 3 局，一者胜 1 局；约定谁先赢满 5 局谁获得全部赌金。问因不可抗力停止赌博后应如何公平地分配赌金？）的若干次通信讨论；而最初向 Pascal 提出这个问题的原名 Antoine Gombaud 的赌徒、业余数学家 Chevalier de Méré（1607/?/?–1684/12/29；法国）也因此而留名史册。在他们的讨论中，虽未明确给出概率的定义，但明确了赌徒获胜的机会是赢的情况数与所有可能情况数的比例，用多种方法对“赌金分配”问题给出了正确解答。

1655 年秋在巴黎游学的青年学者 C. Huygens（惠更斯，1629/4/14–1695/7/8；荷兰数学家、物理学家）听说他们的讨论后，也开始做与之相关的一些讨论；他通过朋友与 Fermat 在 1656 年建立通信联系，并在 1657 年发表《论赌博中的计算》*，书中第一次给出了加法定理、乘法定理以及数学期望的定义（由此暗含概率的古典定义）。该书一经出版，立即得到学术界的认可与重视，在欧洲作为概率论的标准教材长达 50 余年。

1713 年，J. Bernoulli（雅各布·伯努利，1655/1/6–1705/8/16；瑞士）的遗著《猜度术》†在他大侄子 Nicolaus I Bernoulli（尼古拉一世·伯努利，1687/10/20–1759/11/29；瑞士）的帮助下得以出版；其中 Bernoulli 建立了概率论中第一个极限结果—Bernoulli(弱)大数律；由此他认为，事件的概率应该定义为事件发生的极限频率。这本书的出版，是把概率论建立在稳固的数学基础上的首次尝试，标志着概率论成为一门独立的数学分支学科，因此 J. Bernoulli 也成为公认的概率论的奠基人。

这一时期的概率学家们主要研究离散型随机变量，研究手段以排列组合方法为主。

注 1.1. Blaise Pascal（帕斯卡，1623/6/19–1662/8/19）生于克拉蒙费朗、卒于巴黎，法国数学家、物理学家、哲学家、散文家。他有两个姐姐，大姐 Gilberte Pascal、二姐 Jacqueline Pascal。他父亲是一个小贵族，1631 年之前担任克拉蒙费朗的地方法官的职务、1639 年担任

*即《Tractatus de ratiociniis in aleae ludo》。

†即《Ars Conjectandi》。

鲁昂税务局长，是一位数学家（发现了 *Pascal* 螺线）和拉丁语学者。*Pascal* 从小体质虚弱，四岁丧母，五年后他的父亲辞去了法官职务（实际上是出售该职位，所得收入共 65,665 磅投资于国家债券，但到 1638 年政府违约，财富缩水为不超过 7,300 磅），全家搬到巴黎。由于发现自己的几个孩子（特别是 *B. Pascal*）都很聪明，他父亲亲自对他们进行教育，并经常带小 *Pascal* 参加巴黎的科学家集会（特别是“梅森学院”的沙龙活动）以开阔眼界。*Pascal* 也在此时表现出在数学上很高的天赋，11 岁时就写了一篇关于振动与声音的关系的文章，这使得他的父亲担心儿子会影响希腊和拉丁文的学习，于是禁止他在 15 岁前学习数学。一天，他父亲发现 *Pascal*（当时 12 岁）用一块煤在墙上独立证明三角形各角和等于两个直角。从那时起，*Pascal* 被允许学习 *Euclid* 几何。他 16 岁时发现著名的 *Pascal* 六边形定理：内接于一个二次曲线的六边形的三双对边的交点共线；17 岁时写成《圆锥曲线论》(1640)；1642 年为减轻父亲在税务计算上的劳动，他设计并制作了一台能自动进位的加减法计算装置，被认为是世界上第一台数字计算器，为以后的计算机设计提供了基本原理。1654 年他开始研究几个方面的数学问题：在无穷小分析上深入探讨了不可分原理，得出求不同曲线所围面积和重心的一般方法，并以积分学的原理解决了摆线问题，于 1658 年完成《论摆线》；他计算了三角函数和正切的积分，最早引入了椭圆积分。他的论文手稿对 *G. Leibniz*（莱布尼茨，1646/7/1–1716/11/14；德国）建立微积分学有很大启发。在研究二项式系数性质时，写成《算术三角形》向巴黎科学院提交，后收入他的全集，并于 1665 年发表。其中给出的二项式系数展开后人称为“帕斯卡三角形”（在我国称“杨辉三角形”或“贾宪三角”）。在与 *Fermat* 的通信中讨论赌金分配问题，对早期概率论的发展颇有影响。他还制作了水银气压计（1646），写了液体平衡、空气的重量和密度等方向的论文（1651–1654）。自 1655 年隐居修道院，写下《思想录》(1658) 等经典著作。

Pierre de Fermat（费马，1601/8/17–1665/1/12）生于博蒙德洛马涅（*Beaumont-de-Lomagne*），卒于喀斯特（*Castres*）。他是法国律师和业余数学家，被誉为“业余数学家之王”。他独立于 *R. Descartes*（笛卡尔，1596/3/31–1650/2/11；法国）发现了解析几何的基本原理，1630 年用拉丁文撰写了仅有八页的论文《平面与立体轨迹引论》，在 *Fermat* 去世 14 年以后才出版；*Descartes* 是从运动轨迹来寻找它的方程，而 *Fermat* 则是从方程出发来研究运动轨迹的，这正是解析几何基本原则的两个相对的方面。*Fermat* 建立了求切线、求极大值和极小值以及定积分方法，对微积分做出了重大贡献。*Fermat* 与 *Pascal* 通信讨论赌金分配问题，对概率论发展做出了贡献。他在数论中也有诸多贡献，最著名的是 *Fermat* 的最后定理；在我国一般称为 *Fermat* 大定理，最终由 *A. Wiles*（怀尔斯，1953/4/11–；英国）在 1995 年给出了证明。他还在光学中提出最小作用原理，也叫最短时间作用原理。



图 1.3: C. Huygens (1629-1695)



图 1.4: J. Bernoulli (1654-1705)

注 1.2. *Christiaan Huygens*（惠更斯，1629/4/14–1695/7/8）是荷兰物理学家、天文学家、数学家，生于海牙、卒于海牙。他是介于 *Galileo Galilei*（伽利略，1564/2/15–1642/1/8；意大利）与 *I. Newton*（牛顿，1643/1/4–1727/3/31；英国）之间一位重要的物理学先驱，是历史上最著名的物理学家之一，他对力学的发展和光学研究都有杰出的贡献，在数学和天文学方面也有卓越的成就，是近代自然科学的一位重要开拓者。他建立向心力定律，提出动量守恒原理，并改进了计时器。他在概率论和微积分方面也有成就。

Jacob Bernoulli（雅各布·伯努利，1654/12/27–1705/8/16；瑞士，原籍比利时）生于巴塞尔、卒于巴塞尔，是 *Bernoulli* 家族代表人物之一、该家族第一个数学家，是公认的概率论先

驱之一。他在数学上的贡献涉及微积分、微分方程、无穷级数求和、解析几何、概率论以及变分法等领域；他是最早使用极坐标系的数学家之一，积分一词也是 1690 年他首先使用，此外 $e = \lim_{n \rightarrow \infty} (1 + \frac{1}{n})^n$ 也是他发现的。*Bernoulli* 出身于商人世家。他毕业于巴塞尔大学，1671 年获艺术硕士学位，后来遵照父亲的意愿又取得神学硕士学位，但他却不顾父亲的反对，自学了数学和天文学。他在 1678 年、1681 年两次遍游欧洲，这使他接触到当时的许多一流数学家、科学家，其中包括 *G. W. Leibniz*（莱布尼兹，1646/7/1–1716/11/14；德国）；这丰富了他的知识、拓宽了他的兴趣。1682 年他重返巴塞尔，开始教授力学；1687 年成为巴塞尔大学数学教授，直至逝世。除进行数学研究工作外，他还广交学友，所写书信卷帙浩繁，是当时欧洲科学界一位颇有影响的人物。1699 年，*Bernoulli* 当选为巴黎科学院外籍院士；1701 年被柏林科学协会（后为柏林科学院）接纳为会员。许多数学成果与 *Bernoulli* 的名字相联系。例如悬链线问题（1690 年）、曲率半径公式（1694 年）、“*Bernoulli* 双纽线”（1694 年）、“*Bernoulli* 微分方程”（1695 年）、“等周问题”（1700 年）等。

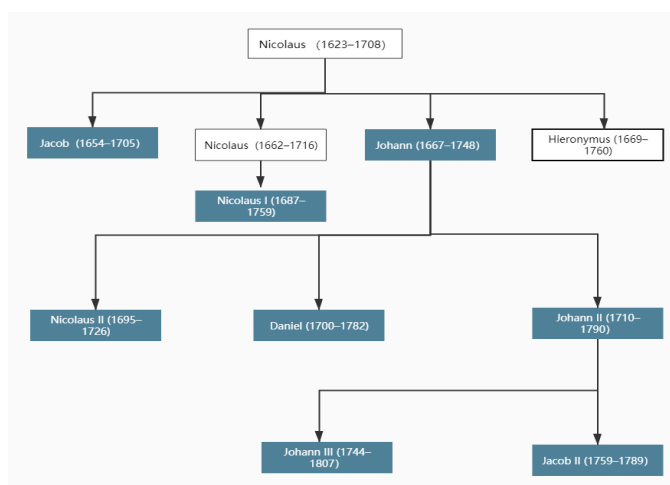


图 1.5: *Bernoulli* 家族（不完全）谱系图；蓝底色表示数学家

值得一提的是，*Bernoulli* 家族是一个数学家辈出的家族。除了 *Jacob Bernoulli* 外，三代 *Bernoulli* 家族共产生了 8 位数学家（据说在 17–18 世纪期间总共有 11 位）。其中比较著名的还有他的弟弟 *Johann Bernoulli*（约翰·伯努利，1667/8/6–1748/1/1；他的儿子 *Daniel Bernoulli* 和大数学家 *Leonhard Euler*（欧拉，1707/4/15–1783/9/18；瑞士）、*G. de l'Hôpital*（洛必达，1661/2/2–1704/2/2；法国）等都是他的学生）和侄子 *Daniel Bernoulli*（丹尼尔·伯努利，1700/2/8–1782/3/17）。这个家族中还有许多在其他非数学的学术/艺术领域的知名人物。

Jacob Bernoulli 培养的知名学生有：*Johann Bernoulli*（1667/8/6–1748/1/1；瑞士），*Jacob Hermann*（赫曼，1678/7/16–1733/7/11；瑞士），*Nicolaus I Bernoulli*（1687/10/20–1759/11/29；瑞士，*Jacob Bernoulli* 的侄子，圣彼得堡悖论的提出者）。

古典概率论发展时期（1713–1812）这一时期，首先是 *A. de Moivre*（棣莫弗，1667/5/26–1754/11/27；法国-英国）于 1718 年发表概率论专著《机遇论》（*The Doctrine of Chances*），首次定义了独立事件的“乘法原理”，基于二项分布 $B(n, \frac{1}{2})$ ，发现了新的概率极限定理—中心极限定理*以及正态分布[†]，开辟了概率论的新的研究方向；书中也明确提出了条件概率的概念，并为

*中心极限定理，Central Limit Theorem，这个英文名词是 Polyá 第一个提出的，用以强调 *De Moivre* 发现的这类概率极限定理在概率论中的核心地位。

[†]*C. F. Gauss*（高斯，1777/4/30–1855/2/23；德国数学家）在 1809 年也独立地提出了正态分布；但他在概率统计上的主要贡献是最小二乘法。

概率论发展了一套较为普遍的符号体系。这本书在概率论发展史上起着承前启后的作用，为后来 Laplace 提出分析概率论做了铺垫。



图 1.6: A. De Moivre (1667-1754)



图 1.7: T. Bayes (1701-1761)

之后，T. Bayes（贝叶斯，1701/??–1761/4/7；英国）在他的两篇遗作（分别于 1764、1765 年出版）中给出了著名的 Bayes 公式，提出了 Bayes 假设的概念；如今统计学中的 Bayes 方法和 Bayes 统计就发源于他的思想。G. Buffon（蒲丰，1707/9/7–1788/4/16；法国）则提出几何概率模型、Buffon 投针问题，发表在 1777 年的著作《偶然性的算术实验》中。由于通过他的投针试验法可以利用很多次随机投针试验算出 π 的近似值，所以特别引人瞩目；这也是最早的几何概率问题，同时这个计算 π 的近似方法也可视作如今的 Monte-Carlo 方法的鼻祖。另外 Daniel Bernoulli（丹尼尔·伯努利，



图 1.8: Buffon (1707-1788)



图 1.9: D. Bernoulli (1700-1782)

1700/2/8–1782/3/17；瑞士）在 1738 年首次将概率论用于人口统计，提出“正态分布误差理论”，并发表了第一张正态分布表；他还给出了圣彼得堡悖论的一种解决方案。L. Euler（欧拉，1707/4/15–1783/9/18，瑞士）对机遇游戏的概率计算和超几何级数进行了研究。

这一时期，由于 I. Newton（牛顿，1643/1/4–1727/3/31；英国）和 G. W. Leibniz（莱布尼兹，1646/7/1–1716/11/14；德国）发明的微积分在欧洲的逐渐深入传播和广泛使用，分析的手段逐渐融入到概率论的研究中，概率学家们也逐渐开始尝试研究连续型随机变量。

注 1.3. *Abraham de Moivre*（棣莫弗，1667/5/26–1754/11/27；法国-英国）生于法国香槟-阿登大区的维特里勒弗朗索瓦，卒于英国伦敦。1685年，棣莫弗与许多信仰新教的教友一道，参加了震惊欧洲的宗教骚乱；在这场骚乱中，他与许多人一起被监禁起来。正是在这一年，保护加尔文教徒的南兹敕令被撤销。随后，包括棣莫弗在内的许多有才华的学者由法国移居英国。据教会的材料记载，棣莫弗一直被监禁至1688年才获释，并于当年移居伦敦。但据20世纪60年代发现的一份当时的材料，1685年时棣莫弗已经到了英国。随后，棣莫弗一直生活在英国，他对数学的所有贡献全是在英国做出的。在那里，他成为 *Isaac Newton*、*Edmond Halley*、*James Stirling* 等的朋友。他以 *De Moivre* 公式和最早提出中心极限定理以及他的专著《机遇论》而知名。

Thomas Bayes（贝叶斯，1701/?/?–1761/4/7；英国）生于伦敦、卒于肯特郡的坦布里奇韦尔斯。他1719年入学爱丁堡大学，学习逻辑和神学；1722年回到伦敦，成为时任神父的父亲的助手；1734年搬到肯特郡的坦布里奇韦尔斯，担任 *Mount Sion* 教堂的神父，直至1752年；1742年他成为英国皇家学会会员。他是英国神学家、数学家、数理统计学家和哲学家。他在数学方面主要贡献是，首创将归纳推理法用于概率论基础理论，并提出了 *Bayes* 公式；他对统计推理的主要贡献是使用了“逆概率”这个概念，并把它作为一种普遍的推理方法提出来。他的这些工作大部分是他死后在1763年由 *Richard Price*（1723/2/23–1791/4/19；威尔士道德哲学家、新教传教士、数学家）整理发表在一本书中，书名是《*An Essay towards solving a Problem in the Doctrine of Chances*》。

Georges-Louis Leclerc Buffon（蒲丰，1707/9/7–1788/4/16）是法国数学家、自然科学家，生于蒙巴尔、卒于巴黎。他是几何概率的开创者，并以蒲丰投针问题闻名于世。

Daniel Bernoulli（丹尼尔·伯努利，1700/2/8–1782/3/17）是瑞士数学家、物理学家，*Bernoulli* 家族代表人物之一，被认为是该家族最著名的数学家之一。他生于荷兰格罗宁根、卒于瑞士的巴塞尔。他是 *Johann Bernoulli* 的儿子，*Jacob Bernoulli* 的侄子；他有一个哥哥 *Nicolaus II Bernoulli*（尼古拉二世·伯努利，1695/2/6–1726/7/31；瑞士）、一个弟弟 *Johann II Bernoulli*（约翰二世·伯努利，1710/5/18–1790/7/17；瑞士；他的儿子 *Johann III Bernoulli*（约翰三世·伯努利，1744/11/4–1807/7/13）、*Jakob II Bernoulli*（雅各布二世·伯努利，1759/10/17–1789/7/3）是第三代 *Bernoulli* 家族数学家）。他父亲最初要求他学习商科、后又要求他学习医科，*Daniel* 虽然退让了，但要求他父亲私下亲自教他数学（最后结果是他父亲和哥哥都教了他数学）。1715年他入学巴塞尔大学，后又到海德堡大学（1718）、斯特拉斯堡大学（1719）学习。1721年获得解剖学和植物学方面的博士学位。1725年他与他的哥哥在彼得大帝的邀请下到彼得堡科学院工作（但8个月后就染病发烧，1726年去世），被任命为生理学院士和数学院士；1727年，*L. Euler* 在他的导师 *Johann Bernoulli* 的请求下也来到彼得堡科学院工作，成为 *Daniel* 的助手。1733年 *Daniel* 回到了巴塞尔大学，先任解剖学和植物学教授。1738年出版著作《流体动力学》。1750年被选为英国皇家学会会员。*Daniel* 的研究工作几乎对当时的数学和物理学的前沿问题都有所涉及，特别突出的是他的数学在力学上的应用、尤其是流体力学和他概率和数理统计领域做的先驱工作。

近代分析概率论形成与发展时期（1812–1889）到1812年，*P. S. Laplace*（拉普拉斯，1749/3/23–1827/3/5；法国）出版了划时代的巨著《分析概率论》*，以强有力的分析工具处理概率论的基本内容，使以往零散的结果系统化，实现了从组合技巧向分析方法的过渡，开辟了概率论发展的新时期，标志着分析概率论正式形成。在该书中，*Laplace* 对一般的独立同分布 *Bernoulli* 随机变量列证明了 *De Moivre-Laplace* 中心极限定理。正是在这部书里，拉普拉斯明确给出了概率的古典定义：事件 *A* 的概率 $\mathbb{P}(A)$ 等于一次试验中有利事件 *A* 的可能的结果数与该试验中所有可能的结果数之比。

之后 *S. D. Poisson*（泊松，1781/6/21–1842/4/25；法国）在1837年发表著作《关于刑事案件和民事案件审判概率的研究》†。他提出了 *Poisson* 分布，论证了适当条件下二项分布收敛到 *Poisson* 分布；这个结果当时被称为小数字定律。另外，他还陈述了 *Poisson* 大数律。

*即《*Théorie analytique des probabilités*》。

†即《*Recherches sur la probabilité des jugements en matières criminelles et matière civile*》。

注 1.4. *Pierre Simon Laplace* (拉普拉斯, 1749/3/23–1827/3/5) 是法国数学家和物理学家。他生于法国西北部卡尔瓦多斯的博蒙昂诺日 (*Beaumont-en-Auge*)、卒于巴黎。1816 年 *Laplace* 被选为法兰西学院院士, 1817 年任该院院长。1812 年发表了重要的《分析概率论》一书, 在该书中总结了当时整个概率论的研究, 论述了概率在选举审判调查、气象等方面的应用, 导入“*Laplace* 变换”等。1799–1825 年出版 5 卷 16 册巨著《天体力学》; 在这部著作中第一次提出天体力学这一名词, 是经典天体力学的代表作。因此他被誉为法国的牛顿和天体力学之父。*Laplace* 有一句堪称经典的格言: “生活中最重要的问题, 绝大部分其实只是概率问题”。

Laplace 培养的知名学生有: *S. D. Poisson* (1781/6/21–1842/4/25; 法国) 和 *Napoleon Bonaparte* (拿破仑·波拿巴, 1769/8/15–1821/5/5; 法国); 前者是他的教子, 后者在一个时期是他的君主。

Siméon Denis Poisson (泊松, 1781/6/21–1842/4/25) 是法国数学家和物理学家。他生于皮蒂维耶 (*Pithiviers*)、卒于索镇 (*Sceaux*)。1798 年 *Poisson* 以第一名成绩考入巴黎综合理工学院深造, 很快受到 *Laplace* 和 *J.-L. Lagrange* (拉格朗日, 1736/1/25–1813/4/10; 法国) 的赏识。入学不到两年, 他已经发表了两本备忘录, 一本关于 *E. Bézout* (裴蜀, 1730/3/31–1783/9/27; 法国) 的消去法, 另外一本关于有限差分方程的积分的个数; 后一结果经 *S.-F. Lacroix* (拉克鲁瓦, 1765/4/28–1843/5/24; 法国) 和 *A.-M. Legendre* (勒让德, 1752/9/18–1833/1/9; 法国) 审查, 建议将它发表于《*Recueil des savants étrangers*》。*Poisson* 在 1800 年毕业后留校任教, 1802 年任副教授, 1806 年任教授。1808 年任法国经度局天文学家。1809 年巴黎理学院成立, 任该校数学教授。1812 年当选为巴黎科学院院士。*Poisson* 对积分理论、行星运动理论、热物理、弹性理论、电磁理论、位势理论和概率论都有重要贡献。有句名言通常归于他名下: “人生只有两样美好的事情—发现数学和教数学”。

Poisson 培养的知名学生有: *Michel Floréal Chasles* (查斯勒, 1793/11/15–1880/12/18) 和 *Joseph Liouville* (刘维尔, 1809/3/24–1882/9/8)。

从 *J. Bernoulli* 开始, 早期的概率学家们都信奉所谓的不充分理由原理: 如果因为无知, 使得我们没有办法判断哪种结果会比另外一种结果更容易出现时, 那么应该给予它们相同的概率。比如硬币, 由于不清楚硬币的哪一面更容易出现, 那么应该给予正面、反面相同的概率, 即都是 $1/2$; 比如骰子, 我们不清楚骰子的哪一面更容易出现时, 那么应该给予每一面相同的概率, 即都是 $1/6$ 。尽管 *Laplace* 在他的书中已经尝试对概率论中的一些基本概念 (特别是事件的概率这一重要概念) 给出定义, 但自从 *Buffon* 提出几何概率模型后, 人们发现了一些概率悖论。其中最著名的几何概率模型悖论是法国学者 *J. Bertrand* (贝特朗, 1822/3/11–1900/4/5) 于 1889 年提出的, 称为 **Bertrand 悖论**: 在半径为 r 的圆内随机选择弦, 计算弦长超过圆内接正三角形边长的概率; 根据“随机选择”的不同意义, 可以得到几种不同答案。

这类悖论的矛头直指概率的概念本身; 特别地, *Laplace* 的古典概率定义开始受到猛烈批评。此时, 无论是概率论的实际应用还是其自身发展, 都要求对概率论的逻辑基础作出更严格的考察。1900 年 *D. Hilbert* (希尔伯特, 1862/1/23–1943/2/14; 德国) 在巴黎的国际数学家大会提出了一系列问题 (*Hilbert* 的 23 个问题), 其中的第六个问题是物理学的公理化, 首当其冲的是概率和力学的公理化。在这个公理化的潮流下, 人们开始了概率的严谨化、特别是公理化的尝试。

现代概率论形成与发展时期 (1889–) 俄国数学家 *S. N. Bernstein* (伯恩斯坦, 1880/3/5–1968/10/26; 他的导师是 *C. E. Picard* 和 *D. Hilbert*)、奥地利犹太裔数学家 *R. E. von Mises* (冯·米西斯, 1883/4/19–1953/7/14; 样本空间概念的提出者) 最早尝试对概率论进行严格化, 他们都提出一些公理来作为概率论的前提, 但他们的公理理论都是不完善的。

作为测度论的奠基人之一, 法国数学家 *E. Borel* (波莱尔, 1871/1/7–1956/2/3) 在 1905 年首先将测度论方法引入概率论重要问题的研究, 并在 1909 年建立了 **Borel 强大数律**; 他的工作激发了数学家们沿这一崭新方向做

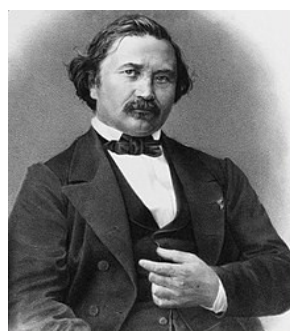


图 1.10: P.-S. Laplace (1749-1827)

图 1.11: J. Bertrand (1822-1900)

出一系列探索。其中，俄罗斯（或前苏联）的数学家有突出贡献。俄罗斯的圣彼得堡学派的缔造者和代表人物是 P. L. Chebyshev（切比雪夫，1821/5/26–1894/12/8），该学派重要成员有 A. A. Markov（1856/6/14–1922/7/20）和 A. Lyapunov（1857/6/6–1918/11/3）等；19 世纪末 20 世纪初俄罗斯的另一学派是以函数论研究为主（后期转向拓扑学）的莫斯科学派^{*}，其中 A. Y. Khintchine（1894/7/19–1959/11/18）是该学派的重要成员。这些数学家在概率方向发展出了一系列弱大数律、强大数律、重对数律等理论结果，提出特征函数方法、截断手术等理论方法，拓展出马氏链等新的研究领域。

莫斯科学派的超级巨星、前苏联数学家 A. Kolmogorov（柯尔莫哥洛夫，1903/4/25–1987/10/20）集前人研究之大成，最终完成概率论的公理化，其结果是 1933 年以德文出版的经典著作《概率论基础》。他在这部著作中建立起集合测度与事件概率的类比、积分与数学期望的类比、函数正交性与随机变量独立性的类比等等，这种广泛的类比终于赋予了概率论以演绎数学的特征。他提出了 6 条公理（三条关于 σ -代数，三条关于概率测度），整个概率论大厦可以从这 6 条公理出发建筑起来。这些公理本身在纯数学的测度论领域已经建立，Kolmogorov 的贡献在于发现无需添加更多的公理进入概率论中；因此 Kolmogorov 的公理体系很快获得了数学家们的普遍承认。由此概率论成为了一门严格的演绎科学，取得了与其他数学分支同等的地位，并通过集合论与其他数学分支密切地联系着。

概率论公理化完成以来，在诸多学者的贡献下，概率论在蓬勃发展的过程中逐步完善，关于随机变量独立和的强/弱大数律、中心极限定理、重对数律等经典理论到 1960 年前后发展成熟；除了这些经典研究范式，概率学家们又发展出不少新的研究领域和新的理论成果。局限于概率论在 20 世纪三、四十年代以来发展出来的新的重大理论工具，囿于编者学识，无法提供周全且准确的描述，只能略举若干重大事例如下：

（1）由于概率论的公理化，随机过程的研究获得了严格的理论基础和新的起点，成为现代概率论研究的一个重要主题。在 Markov 的工作基础上，随机过程理论得到蓬勃发展：1931 年 Kolmogorov 用分析方法奠定了一般马氏过

^{*}莫斯科学派早期的代表性人物有 Egorov（叶戈罗夫，1869–1931）和 Lusin（鲁金，1883–1950）等。

程的理论基础、1934 年 A. Khintchine（辛钦，1894/7/19–1959/11/18；前苏联）提出了平稳过程的有关理论、1948 年 P. Lévy（莱维，1886/9/15–1971/12/15；法国）提出并发展独立增量过程（即 Lévy 过程）；其中随机游动的有关研究尤其受到概率界的密切关注和长时间的推动。在此过程中，鞅论在 J. L. Doob 等的奠基性工作下也得到蓬勃发展和广泛应用。

（2）在 1827 年英国生物学家 R. Brown（布朗；1773/12/21–1858/6/10）观察到水面上的花粉粒子的 Brown 运动、1905 年 A. Einstein（爱因斯坦，1879/3/14–1955/4/18；德国，犹太裔物理学家）建立 Brown 运动的物理机制模型、N. Wiener（维纳，1894/11/26–1964/3/18；美国）等人完善 Brown 运动的数学定义等工作的基础上，1942 年日本数学家 K. Itô（伊藤清，1915/9/7–2008/11/10）引进随机积分与随机微分方程，创立了随机分析理论。

（3）在 Beurling 和 Deny 1958 年与 1959 年工作以及前人经典位势理论工作的基础上，1971 年日本的 M. Oshima、Y. Fukushima 首次利用有穷维空间的正则狄氏型（Dirichlet Form）构造了与之相联系的 Hunt 过程，使得狄氏型这个纯分析工具与随机分析建立了联系，并获得不少应用。之后诸多学者合作发展狄氏型理论。日本的 M. Oshima 和 Y. Fukushima、美国的 R. Gettoor、德国的 M. Röckner 和 S. Albeverio 以及我国的马志明院士（1948/1–）等在这方面有突出贡献，是该领域世界知名的专家。

（4）在 H. Cramér（克莱默，1893/9/25–1985/10/5；瑞典）、F. O. Lundberg（伦德伯格，1876/6/2–1965/12/31；瑞典）关于保险数学的研究工作基础上，1966 年 S. R. S. Varadhan（瓦拉当，1940/1/2–；印度-美国）创立了大偏差理论。

（5）P. Malliavin（马利亚万，1925/9/10–2010/6/3；法国）把确定性函数的变分拓展到随机过程的变分，创立了 Malliavin 分析理论。这是一个新兴的研究领域，联系着调和分析、泛函分析、概率、微分几何等多个数学分支。

（6）我国的彭实戈院士（1947/12/8–）与 E. Pardoux（巴赫杜，1947–；法国）在 1990 年合作，创立了倒向随机微分方程理论，在控制理论、金融数学等多个方向得到应用和发展。

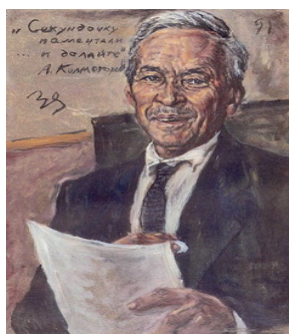


图 1.12: A. N. Kolmogorov (1903-1987)

注 1.5. A. Kolmogorov（柯尔莫哥洛夫，1903/4/25–1987/10/20）是前苏联的数学家，生于坦博夫。他 1925 年毕业于莫斯科大学，1930 年开始任莫斯科大学教授，1935 年获物理数学博士学位，导师是 Nikolai Luzin（鲁金，1883/12/9–1950/1/28；苏联/俄罗斯数学家，以 Luzin 定理、Luzin 空间、Luzin 集等闻名）。1939 年他被选为苏联科学院院士，1966 年当选苏联教

育科学院院士。他是 20 世纪最有影响的苏联数学家之一，是现代概率论的开拓者。他 1925 年以后与 Khintchine（辛钦，1894/7/19–1959/11/18）共同把实变函数论的方法应用于概率论，建立了测度论基础上的概率论公理化体系，不仅解决了概率论中的一系列难题，而且奠定了近代概率论的基础；1930 年以后，他着重研究了应用于具有连续时间变量的 Markov 过程的解析方法，发展了 Markov 过程的理论。他在数理逻辑（结构逻辑方面）、拓扑学（下同调理论）、力学（涡动性的静态理论）、微分方程、泛函分析、信息论等多方面做出了重要贡献。他关心数学教育，积极参与高等和中等学校数学教材的编写工作。由于他在概率论、调和分析及动力系统方面的出色工作，他荣获 1980 年的数学终身成就奖—Wolf 奖。在他的带领和影响下，上个世纪前苏联出现了一大批优秀的数学家。

Kolmogorov 直接培养的学生有（按年龄排列）：S. Nikolsky（1905/4/30–2012/11/9；俄罗斯数学家，在泛函分析、函数逼近等多个方向有奠基性贡献），B. Gnedenko（1912/1/1–1995/12/27；苏联数学家，以概率论、特别是极值理论中的研究成果著称，例如 Fisher–Tippett–Gnedenko 定理，1958 年爱丁堡国际数学家大会特邀报告人），I. Gelfand（1913/9/2–2009/10/5；苏联数学家，以群论、表示论、泛函分析方面的重要贡献闻名于世），A. Obukhov（1918/5/5–1989/12/3；俄罗斯物理学家、应用数学家，以湍流和大气物理学方面的统计理论研究著称，以他姓氏命名的结果有：Monin–Obukhov 相似性理论和 Monin–Obukhov 长度等），A. Yaglom（1921/3/6–2007/12/13，苏联-俄罗斯物理学家、数学家、统计学家、气象学家，以湍流的统计理论和随机过程理论方面的研究著称），A. Monin（1921/7/2–2007/9/22；俄罗斯物理学家、应用数学家、海洋学家，以湍流和大气物理学方面的统计理论研究著称），E. Dynkin（1924/5/11–2014/11/14；苏联-美国数学家），G. Barenblatt（1927/7/10–2018/6/22；俄罗斯数学家），R. Dobrushin（1929/7/20–1995/11/12；苏联-俄罗斯数学家，在概率论、数学物理、信息论等方面有重要贡献），Y. Prokhorov（1929/12/15–2013/7/16；俄罗斯数学家，以度量空间上测度族的胎紧性和列紧性关系的研究成果著称），V. Uspensky（1930/11/27–2018/6/27，俄罗斯数学家、语言学家、作家，他首创了俄罗斯的语言学教育改革），R. Minlos（1931/2/28–2018/1/9；苏联-俄罗斯数学家，在概率论和数学物理方面有重要贡献），A. Vitushkin（1931/6/25–2004/5/9；苏联数学家，盲人，以解析容度和其他数学分析方面的研究工作闻名），V. Alekseev（1932/6/17–1980/12/1；俄罗斯数学家，研究方向为天体力学和动力系统；1970 年 Nice 国际数学家大会邀请报告人），A. Shiryayev（1934/10/12–；苏联-俄罗斯数学家，以概率论、统计、金融数学方向的研究成果著称，俄罗斯科学院院士），Y. Sinai（1935/9/21–；俄罗斯数学家，以动力系统方向的研究成果闻名于世，1997 年获 Wolf 奖），V. Arnold（1937/6/12–2010/6/3；苏联-俄罗斯数学家，最知名结果是 Kolmogorov–Arnold–Moser 定理，1957 年就解决了 Hilbert 第 13 问题，2001 年获 Wolf 奖），E. Khazen，P. Martin-Löf（1942/5/8–；瑞典逻辑学家、哲学家、数理统计学家，斯德哥尔摩大学教授，以概率、统计、数理逻辑、计算机科学方面的基础研究而知名），L. Levin（1948/11/2–；苏联-美国数学家、计算机科学家，波士顿大学教授，以计算、算法复杂性中的随机性、平均复杂性方面的研究著称，与 Stephen Cook 独立地发现了 NP 完全问题的存在性），S. N. Artemov（1951/12/25–；俄罗斯-美国数学家，专长为逻辑及其应用，现为纽约城市大学教授）等。

习 题 1

习题 1.1. Pacioli 在《算术书》中提出所谓的“分赌金问题”：两人决定赌博若干局，事先约定谁先赢得 6 局便算赢家，从而获取全部赌金。如果在一个人赢 5 局，另一个人赢 2 局时因故终止赌博，应如何分配赌金才合理？他给出的答案是按照 5 : 2 分，也就是说按照赌徒实际已经赢下的局数比例来分配赌金。为什么这样的分配原则不合理？试举极端情况的例子来说明。

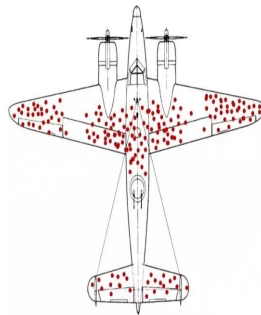
习题 1.2. (“三枚银币”骗局) 某人在街头设一赌局。他向观众出示了放在帽子里的三枚银币（记为甲、乙、丙），银币甲的两面涂了黑色，银币丙的两面涂了红色，银币乙一面涂了黑色，另一面涂了红色。游戏规则是：他让一个观众从帽子里任意取出一枚银币放到桌面上（这里不用“投掷银币”是为

了避免暴露银币两面的颜色)，然后由设局人猜银币另一面的颜色，如果猜中了，该参与者付给他 1 元钱，如果猜错了，他付给该参与者 1 元钱。试问：这一赌局是公平的吗？从直觉上看，无论取出的银币所展示的一面是黑色或红色，另一面是红色或黑色的概率都是 $1/2$ ，这一赌局似乎是公平的，但实际上不公平。为什么？

【提示：设局者只要每次“猜”背面和正面是同一颜色，他的胜算概率是 $2/3$ ，因为从这三张牌随机选取一枚银币，其两面涂相同颜色的概率就是 $2/3$ 。如果有许多人参与赌局，大概有 $1/3$ 的人会赢钱， $2/3$ 的人会输钱。】

习题 1.3. (Monty Hall 问题) 本题出自美国的一个电视游戏节目，问题的名字来自该节目的主持人蒙提·霍尔，20 世纪 90 年代曾在美国引起广泛和热烈的讨论。假定在台上有三扇关闭的门，其中一扇门后面有一辆汽车，另外两扇门后面各有一只山羊。主持人是知道哪扇门后面有汽车的。当竞猜者选定了一扇门但尚未开启它的时候，节目主持人去开启剩下两扇门中的一扇，露出的是山羊。主持人会问参赛者要不要改猜另一扇未开启的门。问题是：改猜另一扇未开启的门是否比不改猜赢得汽车的概率要大？

习题 1.4. (幸存者偏差) 本题内容源自知乎社区一位作者的文章，目的是让读者了解幸存者偏差 (Survivorship Bias) 这一逻辑推理上的误区。1940 年左右，在英国和德国进行的空战中，双方都损失了不少轰炸机和飞行员。因此当时英国军部研究的一大课题就是：在轰炸机的哪个部位装上更厚的装甲，可以提高本方飞机的防御能力，减少损失。由于装甲很厚，会极大的增加飞机的重量，不可能将飞机从头到尾全都用装甲包起来，因此研究人员需要做出选择，在飞机最易受到攻击的地方加上装甲。



当时的英国军方研究了那些从欧洲大陆空战中飞回来的轰炸机。如上图所示，飞机上被打到的弹孔主要集中在机身中央，两侧的机翼和尾翼部分。因此研究人员提议，在弹孔最密集的部分加上装甲，以提高飞机的防御能力。这一建议被美国军队统计研究部的统计学家 Abraham Wald 否决。Wald 连续写了 8 篇研究报告，指出这些百孔千疮的轰炸机是从战场上成功飞回来的“幸存者”，因此它们机身上的弹孔对于飞机来说算不上致命。要想救那些轰炸机飞行员的性命，更正确的方法应该是去研究那些被打中并坠毁的轰炸机。只有研究那些没有成功返航的“倒霉蛋”，才能有的放矢，找到这些飞机最脆弱的地方并用装甲加强。Wald 的建议后来被英国军方采纳，挽救了成千上万的飞行员性命。

你还知道哪些有关幸存者偏差的故事？

习题 1.5. (赌徒谬误) 本题的目的是介绍赌徒谬误 (Gambler's Fallacy)。据说，有些参加俄罗斯轮盘赌的赌徒有这样的一种策略：记录下每次轮盘转下

来的结果（红色或者黑色）。如果遇到连续多次是一个颜色（比如连续五次都是红色），那么赌徒就会果断出手，在下一把押上重注赌另一个颜色（比如黑色）。赌徒们认为，他们这样做的道理是 *Bernoulli* 已经论证过的“频率稳定性”（*Bernoulli* 弱大数律），也就是说频率具有某种回归性。你觉得有道理吗？

习题 1.6.（热手效应）与“赌徒佯谬”相对应的，另一个很多人容易犯的行为学错误叫做“热手效应”（*Hot Hand Fallacy*）。“热手效应”源于篮球运动。在篮球比赛中，有时候会发生这样的情况：某一位球员连续投中几个三分球。这时候其队友和教练都会认为这位球员的状态来了，即他的手开始“热”了。于是大家都会主动把球传给这位球员，好趁他手“热”的时候抓紧时间多投几个，为球队涨点分。“热手效应”在赌场里也很常见。比如在 21 点桌上，如果有一个玩家连续赢了庄家几把，那么在边上围观“飞苍蝇”的群众可能会产生这位玩家“手气非常好”的错觉，在接下来的时间里把更多的筹码赌在这位玩家上。问题在于，这种所谓的“热手效应”更多的只是大家的感觉而已，并没有可靠的证据支持（赌场中的）“热手效应”的存在。

习题 1.7.（圣彼得堡悖论）本题的目的是介绍圣彼得堡悖论；它是决策论中的一个悖论。数学家丹尼尔·伯努利（*Daniel Bernoulli*）的堂兄尼古拉·伯努利（*Nicolaus I Bernoulli*）在 1738 年提出了一个概率期望值悖论。设有如下一个需要付费 c 元参加的游戏：设定掷出正面或者反面为成功，游戏者如果第一次投掷成功，得奖金 2 元，游戏结束；第一次若不成功，继续投掷，第二次成功得奖金 4 元，游戏结束；这样，游戏者如果投掷不成功就反复继续投掷，直到成功，游戏结束。如果第 n 次投掷成功，得奖金 2^n 元，游戏结束。按照概率期望值的计算方法，将每一个可能结果的得奖值乘以该结果发生的概率即可得到该结果奖值的期望值。通常认为，合理的收费 c 就是游戏的期望值，它是所有可能结果的期望值之和。随着 n 的增大，以后的结果虽然概率很小，但是其奖值越来越大，每一个结果的期望值均为 1，所有可能结果的得奖期望值之和，即游戏的期望值，将为“无穷大”。按照概率的理论，多次试验的结果将会接近于其数学期望，也就是说合理的收费 $c = \infty$ ！但是正如 *Hacking(1980)* 所说：“没有人愿意花 25 元去参加一次这样的游戏。”这就出现了计算的期望值与实际情况的“矛盾”，问题在哪里？实际在游戏过程中，游戏的收费应该是多少？决策理论的期望值准则在这里还成立吗？这是不是给“期望值准则”提出了严峻的挑战？正确认识和解决这一矛盾对于人们认识随机现象、发展决策理论和指导实际决策无疑具有重大意义。你对此如何理解？历史上，*Daniel Bernoulli* 在提出这个问题的时候就给出一种解决办法，感兴趣的读者请自行检索有关信息。

习题 1.8.（辛普森悖论）本题的目的是介绍辛普森悖论，资料来源于百度百科；当人们尝试探究两种变量（比如新生录取率与性别）是否具有相关性的时候，会分别对之进行分组研究。然而，在分组比较中都占优势的一方，在总评中有时反而是失势的一方。该现象于 20 世纪初就有人讨论，但一直到 1951 年，*E.H.Simpson* 在他发表的论文中阐述此一现象后，该现象才算正式被描述解释。后来就以他的名字命名此悖论，即辛普森悖论。我们以下面的实例进行介绍。

“校长，不好了，有很多男生在校门口抗议，他们说今年研究发现女生录取率 42% 是男生 21% 的两倍，我们学校遴选学生有性别歧视”，校长满脸疑惑的问秘书：“我不是特别交代，今年要尽量提升男生录取率以免落人口实吗？”

秘书赶紧回答说：“确实有交代下去，我刚刚也查过，的确是有注意到，今年商学院录取率是男性 75%，女性只有 49%；而法学院录取率是男性 10%，女性为 5%。两个学院都是男生录取率比较高，校长这是我作的调查报告。”

| 学院 | 女生 申请 | 女生 录取 | 女生 录取率 | 男生 申请 | 男生 录取 | 男生 录取率 | 合计 申请 | 合计 录取 | 合计 录取率 |
|-----|----------|----------|-----------|----------|----------|-----------|----------|----------|-----------|
| 商学院 | 100 | 49 | 49% | 20 | 15 | 75% | 120 | 64 | 53.3% |
| 法学院 | 20 | 1 | 5% | 100 | 10 | 10% | 120 | 11 | 9.2% |
| 总计 | 120 | 50 | 42% | 120 | 25 | 21% | 240 | 75 | 31.3% |

“秘书，你知道为什么个别录取率男皆大于女，但是总体录取率男却远小于女吗？”

此例就是统计上著名的辛普森悖论 (*Simpson's Paradox*)。作为读者的你能理清这里悖论出现的原因吗？

§ 2

从古典概率模型、几何概率模型到概率论的公理

从本章起，我们将基于集合论来对关心的随机现象进行概率建模。如第1章所述，历史上人们最早进行研究的是在等可能性假设下的两类模型：**古典概率模型**和**几何概率模型**。关于它们的研究取得了成功，解决了许多现实问题；之后人们在几何概率模型中发现了 **Bertrand 悖论**，由此引发了人们对概率论的严谨化、特别是公理化的探索。

法国数学家 A. Weil (韦伊, 1906/05/06–1998/08/06) 说, “在课堂上干巴巴地讲述知识远不如阐述隐藏在这些知识背后的主要思想来得重要” [46]。而相关的学科史无疑浓缩了该学科方向的主要思想与发展脉络; “一个将历史知识和数学研究结合的最了不起的例子” [46] 是德国数学家 C. L. Siegel (西格尔, 1896/12/31–1981/04/04) 在 1930 年左右通过考证研究 G. F. B. Riemann (黎曼, 1826/09/17–1866/07/20; 德国) 有关解析数论的著作 (包括他去世那年未尽的论文散页) 而发现了如今称为 **Riemann-Siegel 公式** 的两个公式。秉承此理念, 在本章我们将依次介绍古典概率模型、几何概率模型、概率论的公理化及其相关的应用问题, 力图在一定程度上复现概率论学科的螺旋上升式的发展历程, 着力于阐述先辈隐藏在概率论的概念与理论背后的思想。

2.1 事件与集合

对随机现象的概率建模离不开 Cantor (康托尔, 1845/3/3–1918/1/6; 德国) 发展的集合论。在讨论某个随机事件或随机现象 A 时, 我们总是认为 (或假想) 相关的实验已经完成, 一个 (实验开始时 无法预知的) 实验结果 (“基本结果”) 已经出现了。在概率建模上, 我们总是把事件 A 对应于某个特定空间 Ω (称为**样本空间**) 的子集 (仍记作 A , 其中 $A \subset \Omega$)。此处样本空间 Ω 一般建模成所有可能的 “基本结果” 的全体。其中, 我们总假想有某个 (未知) 变量 ω (一个 “基本结果”, 也称为**样本点**, 代表此已完成实验的实验结果 $\omega \in \Omega$) 能刻画所关心的随机事件的一切有必要关心的细节, 这是有效的概率建模的基本要求; 而所谓的事件 A 或现象 A 只不过是具有特定的性质 A 的样本点全体 (此处即满足关系 $\omega \in A$ 的所有 ω)。也就是说, 我们有

下面的自然语言与集合论方式的概率建模的基本对应：

$$A \text{ 事件发生} \Leftrightarrow \omega \in A. \quad (2.1)$$

此处的 ω 理解为对随机事件进行概率建模时就已经出现的（某个“实验开始”时未知的）“基本结果”（从而不再发生变化）。当“实验完成”，“实验结果”已经出现时，就是我们通常所说的“靴子落地”、“尘埃落定”，对于已经知道发生了的实验结果（“基本结果”）本身，再谈它发生的可能性大小已经没有意义。因此，只有这个实验“基本结果”还未知的时候，我们才去谈论它导致某种现象 A 发生的可能性大小。我们所有的“实验开始”时的不确定性感受就来源于这个已经出现或将要出现的“基本结果”的无法预知性。随机事件 A 发生的可能性大小（不确定性感受）通过一定方法量化赋值就被认为是 A 事件发生的概率。因此，所谓概率实际上可以认为是在“实验开始”时（至少是在实验“基本结果”还未测定出来时）的一种对随机事件发生可能性大小的量化形式的预判。我们的概率论课程就是想介绍人类历史上逐步总结探索出来的对随机事件进行数学建模、合理推断随机事件发生的概率等统计规律的一种已经被大众广泛接受、经受了考验的科学理论与方法。但要注意，概率论在研究随机事件发生的概率等统计规律时，不关心“基本结果”的出现过程，从而进行对应的数学建模时通常忽略这方面的细节。

这里自然而然有两个特殊事件：必然事件 Ω 与不可能事件 \emptyset 。整个样本空间 Ω 是一个必然发生的事件：此时 $\omega \in \Omega$ 总成立，因此 Ω 也称为**必然事件**，建模时总赋予必然事件发生的概率值为 1；空集 \emptyset 是一个必然不发生的事件：此时 $\omega \in \emptyset$ 总是不可能成立，因此 \emptyset 也称为**不可能事件**，建模时总赋予不可能事件发生的概率值为 0。其他事件发生的可能性如果能讨论，在建模时就认为它能够谈论概率，并赋予它发生的概率值介于 0 与 1 之间：越靠近 1 的概率值，表示对应事件发生的可能性越高；越靠近 0 的概率值，表示对应事件发生的可能性越低。虽然对随机事件的概率赋值从现在来看具有很大的自由度，但实际上真正逻辑自洽的赋值最起码应该满足我们本章最后介绍的概率的公理；至于哪种逻辑自洽的概率赋值更合理，则需要结合实际问题进行讨论或通过实践来进行检验。本课程的重心在探讨对随机现象的概率空间建模、特别是在概率空间建模完成后如何来解答关心的随机事件的概率等问题，只在少数场合展开了对于概率建模的合理性的检验方面的讨论；最后一类问题是《统计学》的一个核心问题。

在本书中，关于集合之间的复合运算，我们将使用记号 $\cup, \cap, ^c$ ，分别表示集合的“并”、“交”、“补”运算；另外，我们也定义 $A \setminus B := A \cap B^c$ ，有时也记 $A - B := A \setminus B$ 。在对应关系(2.1)下，以下自然语言中的复合事件与建模中的相应集合的复合运算的对应关系也就很自然了：

- 必然事件 = Ω ， 不可能事件 = \emptyset ；
- A 的对立事件 = A^c ；
- A, B 两事件同时发生的复合事件 = $A \cap B$ ；
- A, B 两事件至少发生其一的复合事件 = $A \cup B$ ；
- A, B 两事件只发生其一的复合事件 = $A \Delta B := (A \setminus B) \cup (B \setminus A)$ 。

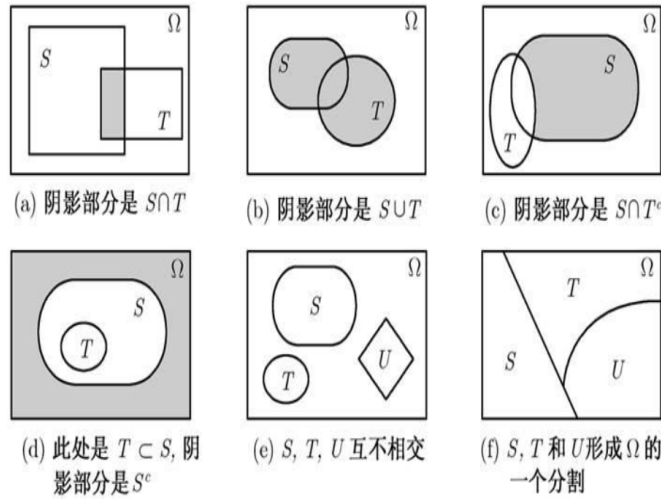


图 2.1: 事件的运算或关系的文氏图

其中 $A \cap B$ 有时也简单写作 AB , 即 $AB := A \cap B$ 。如果 $A \cap B = \emptyset$, 则称 A, B 两事件互斥或不交, 此时也改记 $A \cup B$ 为 $A \uplus B$, 以强调并运算中集合的两两不交性。如果一族集合 $\{A_\alpha \subset \Omega\}_{\alpha \in I}$ 满足:

$$A_\alpha \cap A_\beta = \emptyset, \forall \alpha, \beta \in I, \alpha \neq \beta,$$

就称 $\{A_\alpha\}_{\alpha \in I}$ 是两两不交的或互不相交的; 此时, 如果 $I \subset \mathbb{N}$, 我们就简称 $\{A_n\}_{n \in I}$ 为互不相交集列或不交集列。在概率论中, 互不相交的事件列也经常被称为互斥的事件列, 其含义就是它们不会两两同时发生。我们希望读者在学习本课程的过程中有意识地锻炼自己在自然语言和数学语言之间自由切换的能力; 这个能力有助于我们利用生活中形成的直观来学习本课程, 也有助于我们利用本课程学习的概率理论来理解或解决生活中的有关实际问题。

这里, 我们提醒读者: 在集合论中, 集合的多个并运算之间满足交换律和结合律; 集合的多个交运算之间也满足交换律和结合律; 交、并两种运算混合满足分配律; 补运算与并运算、补运算与交运算之间满足对偶律, 也称为 **De Morgan 律**。有关准确论述, 见下面的命题。

命题 2.1.1. 集合的交、并、补运算具有如下一些性质:

- (1) **交换律:** $A \cup B = B \cup A, A \cap B = B \cap A$;
- (2) **结合律:** $(A \cup B) \cup C = A \cup (B \cup C), (A \cap B) \cap C = A \cap (B \cap C)$;
- (3) **分配律:** $(A \cup B) \cap C = (A \cap C) \cup (B \cap C), (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$;
- (4) **De Morgan 律:** 对任意指标集 I ,

$$\left(\bigcup_{i \in I} A_i\right)^c = \bigcap_{i \in I} A_i^c, \quad \left(\bigcap_{i \in I} A_i\right)^c = \bigcup_{i \in I} A_i^c. \quad (2.2)$$

此外, 映射与集合运算之间有以下的一些结论。

命题 2.1.2. 设 $f: \Omega_1 \rightarrow \Omega_2$ 是一个映射, 那么

(1) 对任意 $A, B \subset \Omega_1$

$$f(A \cup B) = f(A) \cup f(B), \quad f(A \cap B) \subset f(A) \cap f(B);$$

(2) 对任意 $A, B \subset \Omega_2$

$$f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B),$$

$$f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B),$$

$$f^{-1}(A^c) = f^{-1}(A)^c.$$

另外, 我们给出集合列的极限概念如下。

定义 2.1.1. 设 $\{A_n\}_{n=1}^{\infty}$ 是空间 Ω 中子集列, $A \subset \Omega$ 。

(i) 设 $\{A_n\}_{n=1}^{\infty}$ 为单调上升的集合列, 即 $A_n \subset A_{n+1}, \forall n \geq 1$ 。称 $\{A_n\}_{n=1}^{\infty}$

的 (单调上升) 极限为 A , 记作 $A_n \nearrow A$, 如果 $A = \bigcup_{n=1}^{\infty} A_n$;

(ii) 设 $\{A_n\}_{n=1}^{\infty}$ 为单调下降的集合列, 即 $A_{n+1} \subset A_n, \forall n \geq 1$ 。称 $\{A_n\}_{n=1}^{\infty}$

的 (单调下降) 极限为 A , 记作 $A_n \searrow A$, 如果 $A = \bigcap_{n=1}^{\infty} A_n$;

(iii) 称 $\{A_n\}_{n=1}^{\infty}$ 的上极限集 (简称为上限集) 为 A , 记作 $A = \overline{\lim}_{n \rightarrow \infty} A_n$, 如

果 $A = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n$; 注意到此时恰好有

$$\overline{\lim}_{n \rightarrow \infty} A_n = \{\omega : \sum_{n=1}^{\infty} 1_{A_n}(\omega) = \infty\},$$

故在概率论中也经常简写为 $\{A_n \text{ i.o.}\} := \overline{\lim}_{n \rightarrow \infty} A_n$, 此处 *i.o.* 是英文 *infinitely occur* (或 *infinite often*) 的缩写。*这里, 对任意 $A \subset \Omega$, 1_A 称为集合/事件 A 的示性函数, 它的定义为

$$1_A(\omega) := \begin{cases} 1, & \text{当 } \omega \in A \text{ 时} \\ 0, & \text{当 } \omega \notin A \text{ 时} \end{cases}; \quad (2.3)$$

(iv) 称 $\{A_n\}_{n=1}^{\infty}$ 的下极限集 (简称为下限集) 为 A , 记作 $A = \underline{\lim}_{n \rightarrow \infty} A_n$, 如

果 $A = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n$;

(v) 称 $\{A_n\}_{n=1}^{\infty}$ 的极限为 A , 记作 $A = \lim_{n \rightarrow \infty} A_n$, 如果

$$A = \overline{\lim}_{n \rightarrow \infty} A_n = \underline{\lim}_{n \rightarrow \infty} A_n.$$

*概率论中还有 $\{A_n \text{ f.o.}\} := \{\omega : \sum_{n=1}^{\infty} 1_{A_n}(\omega) < \infty\}$ 的写法, *f.o.* 是英文 *finitely occur* (或 *finite often*) 的缩写。显然, 此时 $\underline{\lim}_{n \rightarrow \infty} A_n = \{A_n^c \text{ f.o.}\}$ 。用后面概率的语言来说, 即: 上限集是事件列中无穷多个发生的复合事件; 下限集是事件列中有限多个不发生的复合事件。

2.2 古典概率模型

2.2.1 古典概率模型简介

在古典概率模型中，我们要求

- 随机现象的所有可能的“基本结果”的数量是有限多个的；
- 认为这些“基本结果”是具有同等可能性发生的。

此时，我们把所有可能的“基本结果”的全体建模成一个有限集合 Ω （称为**样本空间**），每个“基本结果”对应于 Ω 中的一个点 ω （称为**样本点**）。显然，样本空间 Ω 的元素个数有限：

$$|\Omega| < \infty,$$

而任一事件 A 在建模下是样本空间的子集 $A \subset \Omega$ ，它的概率定义为 A 与 Ω 的元素个数之比：

$$\mathbb{P}(A) := \frac{|A|}{|\Omega|}. \quad (2.4)$$

在上述模型中，样本空间 Ω 的所有子事件 $A \subset \Omega$ 都可以谈论概率，即可以谈论概率的事件的全体是

$$\mathcal{F} = 2^\Omega := \{A : A \subset \Omega\}.$$

容易知道，在古典概率模型下，概率 $\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}$ 具有如下的简单性质：

- (i)（非负性）： $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$ ；
- (ii)（平凡性）： $\mathbb{P}(\emptyset) = 0$ ；
- (iii)（归一性）： $\mathbb{P}(\Omega) = 1$ ；
- (iv)（有限可加性）： 设 $\{A_n\}_{n=1}^N \subset \mathcal{F}$ 互不相交，则

$$\mathbb{P}\left(\biguplus_{n=1}^N A_n\right) = \sum_{n=1}^N \mathbb{P}(A_n);$$

- (v) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A), \forall A \in \mathcal{F}$ 。

例 2.1. 投掷一枚（质地均匀的）骰子，请问获得偶数点的概率有多大？

参考解答： 我们可以取骰子投掷出的点数作为基本结果，于是取

$$\Omega := \{1, 2, 3, 4, 5, 6\}$$

作为样本空间，并认为此时等可能性成立；此时“投掷一枚骰子获得偶数点”这一事件为

$$A := \{2, 4, 6\}.$$

因此所求概率为

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}.$$


□

2.2.2 计数方法简介

在中学阶段，我们已经学习过关于计数的两个原理——**加法原理与乘法原理**，自然语言的表述如下：

- **加法原理**：完成某件事有 r 类方法*，记第 i 类中有 n_i 种方法 ($i = 1, \dots, r$)，则完成该事总共有 $N = n_1 + \dots + n_r$ 种方法；
- **乘法原理**：完成某件事有 r 个步骤，设第 i 个步骤有 n_i 种方法 ($i = 1, \dots, r$) 实现，则完成该事总共有 $N = n_1 \cdot n_2 \cdot \dots \cdot n_r$ 种方法。

在学习了映射的概念之后，我们也有通过映射来计数的办法。

 **定理 2.2.1.** 设 $f: A \rightarrow B$ 是一个映射，其中 A, B 都是有限集合。那么

- (1) 当 f 是单射时， $|A| \leq |B|$ ；
- (2) 当 f 是满射时， $|A| \geq |B|$ ；
- (3) 当 f 是双射时， $|A| = |B|$ 。

在上述映射的计数方法观点下，之前的加法原理与乘法原理也可以借助集合论的语言简练而精确地表述如下：

 **定理 2.2.2.**（加法原理与乘法原理）

- (1) **加法原理**：当 $A \cap B = \emptyset$ 时， $|A \uplus B| = |A| + |B|$ ；
- (2) **乘法原理**： $|A \times B| = |A| \cdot |B|$ 。

例 2.2. 我们有以下一些排列组合计数的简单实例：

- (a) **（直线）排列**：从 $1, 2, \dots, n$ 中选出 $m \leq n$ 个数，排成一行的方法数是 $A_n^m = \frac{n!}{(n-m)!}$ ；
- (b) **组合**：从 $1, 2, \dots, n$ 中选出 $m \leq n$ 个数形成一个（无序）集合的方法数是 $C_n^m = \binom{n}{m} = \frac{n!}{m!(n-m)!}$ ；
- (c) **圆排列**： $1, 2, \dots, n$ 这 n 个数排成一圈的方法数是 $(n-1)!$ ；

例 2.3. 不定方程（其中 $r \geq 2, n, r \in \mathbb{N}$ ）

$$x_1 + \dots + x_r = n \quad (*)$$

的正整数解满足：

$1 \leq x_1 =: t_1 < x_1 + x_2 =: t_2 < \dots < x_1 + \dots + x_{r-1} =: t_{r-1} \leq n-1$
且 (x_1, \dots, x_r) 与 (t_1, \dots, t_{r-1}) 之间一一对应，因此

- (i) 不定方程 $(*)$ 的正整数解个数为 C_{n-1}^{r-1} ；
- (ii) 不定方程 $(*)$ 的非负整数解个数为 C_{n+r-1}^{r-1} 。

*这里要求不同类之间没有交叉，不会出现同时是 A 类又是 B 类的情况。

2.2.3 古典概率模型应用实例：(1) Polyá 坛子模型

例 2.4. Polyá 坛子模型（有放回） 坛内有 b 个黑球， r 个红球，每次随机地取一球，并放回。记 B_j 为“第 j 次取球时取到黑球”， $R_j = B_j^c$ 为“第 j 次取球时取到红球”，则：

$$(1) \mathbb{P}(R_j) = \frac{r}{b+r} =: p, \mathbb{P}(B_j) = \frac{b}{b+r} =: q = 1 - p, \forall j;$$

$$(2) \mathbb{P}(\text{连续}n\text{次取球，恰好总共取得}k\text{个红球}) = C_n^k p^k q^{n-k}.$$

参考解答：如果要使用古典概率模型来计算上述有关事件的概率，我们无法一劳永逸地通过建立一个样本空间就解决上述所有事件的概率计算。

(1) 为了计算 $\mathbb{P}(R_j)$ 和 $\mathbb{P}(B_j)$ ，我们可以如下方式建模：假想我们给黑球和红球编号为 $1, \dots, b+r$ ；记 $B = \{1, \dots, b\}$, $R := \{b+1, \dots, b+r\}$ 。那么编号 x 的球是黑球，当且仅当 $x \in B$ ；编号 x 的球是红球，当且仅当 $x \in R$ 。设第 k 次取球所得球的编号是 x_k 。于是下面的样本空间 Ω_j 设置能穷尽从第一次直到第 j 次取球的所有有必要关心的细节：令 $\Sigma := \{1, \dots, b+r\}$ ，并设置

$$\Omega_j := \Sigma^j = \{x = (x_1, \dots, x_j) : x_k \in \Sigma, k = 1, \dots, j\}.$$

这样设置的样本空间也符合我们的有放回规则下的“随机取球”的等可能性假定。于是在上述建模下

$$B_j = \Sigma^{j-1} \times B, \quad R_j = \Sigma^{j-1} \times R.$$

因此

$$\mathbb{P}(B_j) = \frac{|B_j|}{|\Omega_j|} = \frac{b}{b+r} =: p, \quad \mathbb{P}(R_j) = \frac{|R_j|}{|\Omega_j|} = \frac{r}{b+r} =: q = 1 - p;$$

(2) 为了计算 $\mathbb{P}(\text{连续}n\text{次取球，恰好总共取得}k\text{个黑球})$ ，我们可以使用样本空间 Ω_n ，于是

$$\begin{aligned} & \mathbb{P}(\text{连续}n\text{次取球，恰好总共取得}k\text{个黑球}) \\ &= \sum_{1 \leq j_1 < \dots < j_k \leq n} \mathbb{P}(\text{第}j_1, \dots, j_k\text{次均取得红球，其余}n-k\text{次均取得黑球}) \\ &= \sum_{1 \leq j_1 < \dots < j_k \leq n} \frac{r^k b^{n-k}}{(b+r)^n} = C_n^k p^k q^{n-k}. \end{aligned}$$

于是这里自然而然地出现了以后将介绍的二项分布的概率分布列。 \square

例 2.5. Polyá 坛子模型（无放回） 坛内有 b 个黑球， r 个红球，每次随机地取走一球，并不再放回。记 B_j 为第 j 次取到黑球， $R_j = B_j^c$ 为第 j 次取到红球，则：

$$(1) \mathbb{P}(B_j) = \frac{b}{b+r}, \mathbb{P}(R_j) = \frac{r}{b+r}, \forall 1 \leq j \leq b+r;$$

$$(2) \mathbb{P}(\text{连续}n\text{次取球，恰好总共取得}k\text{个黑球}) = \frac{C_b^k C_r^{n-k}}{C_{b+r}^n}, \text{ 其中, } 1 \leq n \leq b+r.$$

参考解答： (1) 在无放回的规则下，我们的概率空间设置略有差别。此处， Σ, B, R 的定义同上。但修改设定

$$\Omega_j := \{x = (x_1, \dots, x_j) \in \Sigma^j : \text{其中 } x_1, \dots, x_j \text{ 两两不同}\}.$$

此时在上述建模下

$$B_j = \{x \in \Omega_j : x_j \in B\}, \quad R_j = \{x \in \Omega_j : x_j \in R\}.$$

于是

$$|\Omega_j| = C_{b+r}^j \cdot j!, \quad |R_j| = C_r^1 C_{b+r-1}^{j-1} \cdot (j-1)!,$$

从而 $\mathbb{P}(R_j) = \frac{|R_j|}{|\Omega_j|} = \frac{r}{b+r}$ ，进而 $\mathbb{P}(B_j) = 1 - \mathbb{P}(R_j) = \frac{b}{b+r}$ 。

(2) 的解答留给读者。 \square

上例中结论 (1) 也说明了我们日常生活中用抓阄的方法来实现分配不充分的资源的“程序公平性”。

例 2.6. (抓阄的公平性) 假设有 r 项奖品，共有 $n \geq r$ 个人来共同分配（每人不允许获得两项及以上的奖品），因此采取抓阄的方式来实现：用 r 个红球代表奖品，用 $b = n - r$ 个黑球代表无奖品，然后采用无放回的方式取球 n 次。则这种抓阄方式在程序上是公平的：抓阄的次序不会影响分配的公平性。

以后我们会用条件概率的方法再次给出上述抓阄的公平性的证明。

2.2.4 古典概率模型应用实例：(2) 同生日问题

例 2.7. (同生日问题) 为方便计算，我们假定一年有 $N = 365$ 天；说两个人同一天生日，如果他们出生的日子是同一个月的同一天。现在假定某班级共有 $n = 50$ 位同学，请问该班级中有同学同一天生日的概率有多大？

参考解答： 我们认为每个人的生日等可能性地出现在 $N = 365$ 天中的任意一天。用 A_n 表示班级的 n 人中至少两人同生日。那么容易知道

$$p_n := \mathbb{P}(A_n) = 1 - \mathbb{P}(A_n^c) = 1 - \frac{n!C_N^n}{N^n} = 1 - \frac{N!}{(N-n)!N^n}.$$

在此基础上我们通过编程计算，得到下面的一张 p_n 的简单表格。

表 2.1: 同生日问题中的同生日概率值数列（部分）

| | | | | | | | | |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|
| n | 20 | 21 | 22 | 23 | 25 | 30 | 35 | 40 |
| p_n | 0.4114 | 0.4437 | 0.4757 | 0.5073 | 0.5687 | 0.7063 | 0.8144 | 0.8912 |
| n | 45 | 50 | 55 | 60 | 65 | 70 | 75 | 80 |
| p_n | 0.9410 | 0.9704 | 0.9863 | 0.9941 | 0.9977 | 0.9992 | 0.9997 | 0.9999 |

上面的表格告诉我们：如果班级人数超过 23 人，就有超过一半的概率有人同生日；到了 50 人的大班级，有两人同生日的概率已经高达 97%；到了 80 人的超大大班级，这个概率已经非常接近 1，可以认为有两个人同生日是近乎必然的事情了。 \square

注记 2.1. 结合上面的计算公式与表格，不难感受到，在一个人数只有 30 人的中型班级里，有两人生日至多相差一天的概率也会非常高。建议有兴趣的读者给出这个概率的计算公式，用计算机编程得出实际计算结果，并在自己所在的小团体中进行这方面的社会调查。这是一个不难但有意思的问题。

2.2.5 古典概率模型应用实例：(3) 唱票问题与反射原理

例 2.8. (唱票问题) 有且仅有甲乙两个候选人，已知甲获得了 m 票，乙获得了 n 票，其中 $m > n$ 。假定这 $m+n$ 张选票均匀混合，那么唱票过程中，甲的票数始终领先的概率有多大？

参考解答：取 $\Sigma := \{+1, -1\}$ 。对任意 $x = (x_1, \dots, x_{m+n}) \in \Sigma^{m+n}$ ，以 $x_k := +1$ 表示第 k 张票是支持甲的，否则 $x_k := -1$ 。于是

$$S_k = S_k(x) := x_1 + \dots + x_k$$

代表了到第 k 张选票唱完后甲领先乙的票数。我们可以建立如下的古典概率的样本空间：

$$\Omega := \{x \in \Sigma^{m+n} : S_{m+n}(x) = m - n\}.$$

所关心事件是

$$A := \{x \in \Omega : S_k(x) > 0, k = 1, 2, \dots, m+n\}.$$

容易知道，

$$|\Omega| = C_{m+n}^m = \frac{(m+n)!}{m!n!}.$$

问题的关键在于如何计算出 $|A|$ 。

事件 A 过于复杂。简单点，定义

$$A_1 := \{x \in \Omega : S_1(x) > 0\} \supset A.$$

则

$$A_1 = \{x \in \Sigma^{m+n} : x_1 = 1, S_{m+n}(x) = m - n\},$$

容易知道 $|A_1| = \frac{(m+n-1)!}{(m-1)!n!}$ 。

现在来看 $B := A_1 \setminus A = \{x \in A_1 : \exists j \in (1, m+n), S_j(x) \leq 0\}$ ，于是

$$B = \{x \in A_1 : \exists j \in (1, m+n), S_j(x) = 0\}.$$

我们想计算 $|B|$ ，为此构造一个合适的一一映射：

$$f : B \rightarrow f(B),$$

就会有 $|B| = |f(B)|$ ，当然我们希望后者容易计算。

我们用反射的方法来实现。首先对任意 $x \in B$ ，定义

$$\tau = \tau(x) := \inf\{k > 0 : S_k(x) = 0\}.$$

容易知道 $1 < \tau < m+n$ ，并且此时 $S_\tau = 0, \forall x \in B$ 。之后，我们定义映射 $\{S_j\}_{j=1}^{m+n} \mapsto \{\tilde{S}_j\}_{j=1}^{m+n}$ 如下：

$$\tilde{S}_k = S_k, \forall k \leq \tau,$$

约定 $\tilde{S}_0 := 0$ ，而当 $\tau < k \leq m+n$ 时，定义

$$\tilde{S}_k = -S_k.$$

进而再由此定义 $\tilde{x}_k := \tilde{S}_k - \tilde{S}_{k-1}, k = 1, \dots, m+n$ 。由此我们就给出了映射

$$x \in B \xrightarrow{f} \tilde{x} \in \Sigma^{m+n}.$$

不难知道 f 是单射，并且

$$f(B) = \{x \in \Sigma^{m+n} : x_1 = 1, S_{m+n}(x) = n - m, \exists j \in (1, m+n), S_j(x) = 0\}.$$

注意到 $x \in f(B)$ 时， $S_1 = 1 > 0, S_{n+m} = n - m < 0$ ，因此实际上有

$$f(B) = \{x \in \Sigma^{m+n} : x_1 = 1, S_{m+n}(x) = n - m\}.$$

从而 $|B| = |f(B)| = \frac{(m+n-1)!}{(n-1)!m!}$ 。于是

$$\begin{aligned} |A| &= |A_1 \setminus B| = |A_1| - |B| \\ &= \frac{(m+n-1)!}{(m-1)!n!} - \frac{(m+n-1)!}{(n-1)!m!} \\ &= \frac{(m+n-1)!}{m!n!} \cdot (m-n). \end{aligned}$$

结合 $|\Omega| = \frac{(m+n)!}{m!n!}$ ，立即得到

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{m-n}{m+n}.$$

□

上述问题的解答中通过构造反射 f 来实现计数，该方法的思想被称为**反射原理**。在概率论的一些高级课程中，读者将学习研究随机游动和 Brown 运动相关的一些问题，在那里这个方法发挥了重要作用。

2.2.6 古典概率模型应用实例：(4) 图论中的染色问题

Paul Erdős 首创把概率论应用于图论问题，他与 Renyi 的工作开创了随机图这一概率论与图论的交叉研究方向，对当今的复杂网络等研究方向有重要影响。下面的完全图的边染色问题（又称 Ramsey 问题）就来源于 Paul Erdős 关于 Ramsey 数的下界估计的工作。

例 2.9. (Ramsey 问题) 给定两个自然数 p, n ，其中 $3 \leq p \leq n$ 。如果

$$C_n^p < 2^{C_p^2-1}, \quad (2.5)$$

那么存在一种把完全图 K_n 的所有边染成红、蓝二色的染色方案，使得染色后的完全图 K_n 中找不到单色的完全子图 K_p 。也就是说，二染色的 Ramsey 数 $r_2(p, p) \geq \inf\{n : C_n^p \geq 2^{C_p^2-1}\}^*$ 。

参考解答：把完全图 K_n 的边编号为 $1, 2, \dots, N := C_n^2$ ；定义 $\Omega := \{0, 1\}^N$ 。如果 i 号边染成红色，则记 $\omega_i = 1$ ，否则记 $\omega_i = 0$ 。于是任意

$$\omega := (\omega_1, \dots, \omega_N) \in \Omega$$

就是一种染色方案；我们将把一种染色方案 $\omega \in \Omega$ 视作一个映射

$$\omega : \{1, \dots, N\} \rightarrow \{0, 1\}, \quad i \mapsto \omega_i.$$

*此处 $r_2(p, q)$ 代表一个完全图 K_n 为了在任意的边二染色方案下都有红色完全子图 K_p 或蓝色完全子图 K_q 的最小顶点数目 n ；据说目前仅已知 9 个精确的 Ramsey 数： $r_2(3, 3) = 6, r_2(3, 4) = 9, r_2(3, 5) = 14, r_2(3, 6) = 18, r_2(3, 7) = 23, r_2(3, 8) = 28, r_2(3, 9) = 36, r_2(4, 4) = 18, r_2(4, 5) = 25$ 。其他的能知道粗略取值范围，例如 $43 \leq r_2(5, 5) \leq 49$ 。

考虑以 Ω 为样本空间的古典概率模型（即每边等可能性地独立于其他边染两种颜色之一的随机染色模型）。考虑事件 A 为染色后的完全图 K_n 中存在单色的完全子图 K_p 。对于任一给定的 K_p 子图 B ，记它的边的序号形成的集合为 \tilde{B} ， $B \rightarrow \tilde{B}$ 是一一对应的；记 K_p 子图的全体为 $\mathcal{G}_{n,p}$ ，显然 $|\mathcal{G}_{n,p}| = C_n^p$ 。对任意 $B \in \mathcal{G}_{n,p}$ ， B 是单色子图这一事件为

$$\hat{B} := \{\omega \in \Omega : \omega|_{\tilde{B}} \equiv 0 \text{ 或 } \omega|_{\tilde{B}} \equiv 1\},$$

它满足 $\mathbb{P}(\hat{B}) = 2 \cdot \left(\frac{1}{2}\right)^{C_p^2} = 1/2^{C_p^2-1}$ 。注意到

$$A = \bigcup_{B \in \mathcal{G}_{n,p}} \hat{B},$$

我们有

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{B \in \mathcal{G}_{n,p}} \hat{B}\right) \\ &\leq \sum_{B \in \mathcal{G}_{n,p}} \mathbb{P}(\hat{B}) \\ &\leq C_n^p / 2^{C_p^2-1} < 1. \end{aligned}$$

于是 A^c 非空，即存在一种染色方案，使得染色后的完全图 K_n 中找不到单色的完全子图 K_p 。□

2.3 几何概率模型

2.3.1 几何概率模型简介

在几何概率模型中，我们要求：

- 随机现象的所有可能的“基本结果”的数量是不可数的无穷多个；
- 认为这些“基本结果”是具有同等可能性发生的；
- 所有可能的“基本结果”的全体可以建模成某个欧氏空间 \mathbb{R}^d 中的一个“体积”有限区域 Ω （称为**样本空间**），每个“基本结果”对应于 Ω 中的一个点 ω （称为**样本点**）。

此时，

- (1) 样本空间 $\Omega \subset \mathbb{R}^d$ 是一个“体积”有限的区域： $0 < |\Omega| < \infty$ ，这也意味着 Ω 必须是 \mathbb{R}^d 中的可测集（通常要求是 **Borel** 可测集）；
- (2) 样本空间 Ω 中元素 ω 称为样本点，认为不同样本点具有同等大小的可能性发生，事件 $A \subset \Omega$ 并不是都可以谈论概率，能谈论概率的事件全体可取为 $\mathcal{F} := \Omega \cap \mathcal{B}^d$ （其中 \mathcal{B}^d 代表 \mathbb{R}^d 上的 **Borel** σ -代数，参见第4章 §4.1），它是 Ω 的 **Borel** 可测子集的全体；

(3) $A \in \mathcal{F}$ 时, A 事件发生的概率定义为“区域” A 的体积与全空间 Ω 的体积之比:

$$\mathbb{P}(A) := \frac{|A|}{|\Omega|}. \quad (2.6)$$

2.3.2 几何概率模型应用实例

例 2.10. (约会问题) 甲乙二人约定在 11:00-12:00 之间到某商场碰面, 先到者等后来者 20 分钟, 如果没有等到就自行离去。请问此二人约会成功的概率?

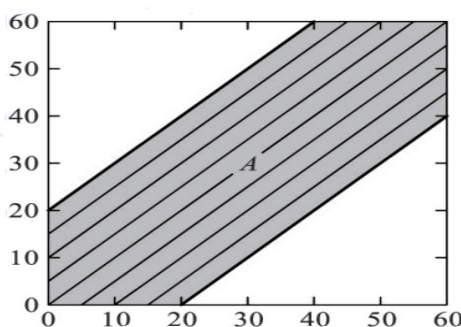


图 2.2: 约会问题图解

参考解答: 为方便, 以 $x, y \in [0, 60]$ 表示甲、乙的到场时间 (实际表示他们在 11 点过 x 分或 y 分的时候到达商场)。于是可设置约会问题的样本空间为

$$\Omega := [0, 60]^2 = \{(x, y) : x, y \in [0, 60]\}.$$

则根据题设, “约会成功” 这一事件 A 可以表示为

$$A := \{(x, y) \in [0, 60]^2 : |x - y| \leq 20\}.$$

容易看到, $|\Omega| = 60^2, |A^c| = 40^2$, 于是

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = 1 - \frac{|A^c|}{|\Omega|} = 1 - \frac{40^2}{60^2} = \frac{5}{9}. \quad \square$$

例 2.11. (Buffon 投针问题) 在地面上有无数条等间隔的平行线 (设间隔为 L), 某人拿着一个材质均匀的短针 (其长度记作 ℓ) 随机地投向地面, 请问短针与这组平行线相交 (只需要与平行线中任意一条相交, 我们就说针与这组平行线相交) 的概率? (为方便, 此处假设 $0 < \ell < L$)

参考解答: 落在地面时, 短针的中点与最近的平行线的距离 $d \in [0, L/2]$ 和短针与平行线的夹角 $\theta \in [0, \pi]$ 这两个参数能完整描述短针落在两条相邻平行线之间的相对位置。因此, 我们不妨设样本空间为

$$\Omega := \{(d, \theta) : d \in [0, L/2], \theta \in [0, \pi]\}.$$

短针与最近的平行线相交这一事件 A 可以表达为:

$$A := \{(d, \theta) \in \Omega : d / \sin \theta \leq \ell/2\}.$$

显然 $|\Omega| = \frac{L}{2} \cdot \pi = \frac{\pi L}{2}$, $|A| = \int_0^\pi \frac{\ell}{2} \sin \theta d\theta = \ell$. 于是

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{2\ell}{\pi L}. \quad \square$$

注记 2.2. Buffon 投针问题的意义与历史地位

(1) 历史上第一个几何概率问题;

(2) 记短针与平行线相交概率为 p , 则圆周率有表达式: $\pi = \frac{2\ell}{pL}$. 因此, 历史上自 Buffon 开始就有许多人通过做投针实验, 以短针与平行线相交的频率 \hat{p} 来代替概率 p 的方式, 测算圆周率 $\pi \approx \frac{2\ell}{\hat{p}L}$, 并且发现在大样本下该方法确实具有一定的精度, 参见下面的表格 (注意, 不同实验者的数据之间没有可比性; Buffon 的实验除外, 表格中最后一列参数 ℓ/L 是我们根据前人的表单反推算出来的);

表 2.2: 历史上用蒲丰投针实验估计圆周率的实验记录

| 实验者 | 时间 | 投掷次数 | 相交次数 | 圆周率估计值 | ℓ/L 的推算值 |
|-----------|--------|------|--------|-----------|---------------|
| Buffon | 1777 年 | 2212 | 704 | 3.142 | 0.500 |
| Wolf | 1850 年 | 5000 | 2532 | 3.1596 | 0.800 |
| Smith | 1855 年 | 3204 | 1218.5 | 3.1554 | 0.600 |
| De Morgan | 1860 年 | 600 | 382.5 | 3.137 | 1.000 |
| Fox | 1884 年 | 1030 | 489 | 3.1595 | 0.750 |
| Lazzarini | 1901 年 | 3408 | 1808 | 3.1415929 | 0.833 |
| Reina | 1925 年 | 2520 | 859 | 3.1795 | 0.542 |

(3) 上述众多实验事实印证了我们对该问题的前述几何概率建模的合理性, 是马克思主义理论中“实践是检验真理的 (唯一) 标准”的一个很好例子;

(4) 上述通过做随机实验的办法来计算确定性数学量或物理量等的方法, 就是如今著名的 Monte-Carlo 方法。因而, Buffon 投针实验可以认为是 Monte-Carlo 方法的鼻祖。

2.3.3 Bertrand 悖论

Bernoulli 在总结前人关于古典概率模型的研究时, 提出了不充分理由原理 (Insufficient Reason Principle): 如果因为无知, 使得我们没有办法判断哪种 (基本) 结果会比另一种 (基本) 结果更容易出现, 那么应该给予它们相同的概率。比如:

(1) 硬币: 由于不清楚硬币哪一面更容易出现, 应当给予硬币的正面、反面相同的概率, 即 $\frac{1}{2}$;

(2) 骰子: 由于不清楚骰子不同点数的面哪一面更容易出现, 应当给予骰子的每一面相同的概率, 即 $\frac{1}{6}$ 。

Laplace 在他的著作中进一步把不充分理由原理的使用范围扩充到几何概率模型中, 提出“未知的概率均为等概率”, 依此原则建立了他的分析概率论, 并且这一理论确实取得了辉煌的成绩。因此整个 19 世纪人们都对此原则确

信无疑，直至 Bertrand 于 1888 年在他的著作《Calcul des probabilités》中提出了以他姓氏冠名的悖论。

例 2.12. (Bertrand 悖论) 在单位圆上随机的取一条弦，求此弦长超过圆的内接正三角形边长的概率？（至少有三种对随机取弦的理解，从而有至少三个不同的答案。）

Bertrand 悖论中“随机的取一条弦”的三种理解

- (1) 随机地在单位圆周上取两点，从而确定一条弦；
- (2) 在单位圆内随机地选一点，把它作为弦的中点，从而确定一条弦；
- (3) 在单位圆上随机地选一条直径，之后在这条直径上随机地选一个点作为弦的中点，从而确定一条弦。

Bertrand 悖论的解答 (1): 我们在单位圆周上选定一点 P 后，可以在单位圆周上找到另外两点 P_2, P_3 ，使得 P, P_2, P_3 三点按逆时针顺序等间距的落在单位圆周上。此时再随机选另一点 Q ，弦 PQ 要大于单位圆的内接正三角形边长的充要条件是： Q 落在圆弧 $\widehat{P_2P_3}$ 的内部，这部分圆弧占单位圆周总弧长的 $1/3$ 。因而所求事件的概率为 $1/3$ 。□

Bertrand 悖论的解答 (2): 单位圆内弦长大于单位圆的内接正三角形边长的充要条件是：弦的中点落在与单位圆同心， $1/2$ 为半径的同心圆的内部。于是所求事件的概率为

$$\mathbb{P}(A) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}. \quad \square$$

Bertrand 悖论的解答 (3): 在先随机取直径，之后再随机在直径上取点作为弦中点的情况下，弦长大于单位圆的内接正三角形边长的充要条件是：弦的中点落在选定的直径上的与单位圆心距离不超过 $1/2$ 的线段内部。因此所求事件的概率为

$$\mathbb{P}(A) = \frac{1}{2}. \quad \square$$

后人还类似地构造出了 Bertrand 悖论的其他版本。下面例子取自知乎社区的“马同学”关于 Bertrand 悖论的回答。

例 2.13. 有一家锯木厂，它会把木头切成不同的木方，木方的截面都是正方形，边长会在 $1 \sim 3$ 尺之间浮动*；那么根据几何概率，该锯木厂生产出来的正方形边长在 $1 \sim 2$ 尺之间的概率为多少？

两种不同概率值的回答：这里，木方的长度和面积都是未知的。

如果对“长度”来运用“不充分理由原理”，也就是认同“长度”上的等可能性假设，则所求概率为

$$\frac{2-1}{3-1} = \frac{1}{2}.$$

*原回答的问题中，此处用词是“随机浮动”，特修改为“浮动”，去掉了“随机”字眼。

如果对“面积”来运用“不充分理由原则”，也就是认同“面积”上的等可能性假设，则所求概率为

$$\frac{2^2 - 1^2}{3^2 - 1^2} = \frac{3}{8}.$$

本问题中悖论出现的根源就在于对“等可能性”（通常也对应“随机”这一字眼）的不同解读。□

注记 2.3. (Bertrand 悖论的意义与历史地位) Bertrand 悖论的提出，促使人们开始认真思考概率概念本身，并严谨地检查概率论的理论基础。1900 年 Hilbert 提出他的 23 个数学问题，并倡导公理化运动，此后概率论方向的先贤也开始了概率的公理化的探索。

在概率论的公理化完成后，这个悖论也自然消失，不成其为悖论，而被认为是由于自然语言对随机理解的歧义的原因造成了某种意义上互相矛盾的多个答案。

2.4 概率的公理化及其性质

2.4.1 概率的公理

历史上，有多人参与了概率论的公理化进程，最终由 Kolmogorov 集大成，完成了概率论的公理化；他提出了总计 6 条公理，其中前三条本质上只是用来界定能谈论概率的事件的全体（即事件域或 σ -代数）。因此，现代概率论中，关于概率的公理化一般会如下表述。

定义 2.4.1. 设 Ω 是一个非空集。 \mathcal{F} 是 Ω 的一些子集形成的集族，它称为 Ω 上的事件域或 σ -代数，如果它满足：

(i) $\Omega \in \mathcal{F}$;

(ii) 补运算封闭：如果 $A \in \mathcal{F}$ ，则 $A^c \in \mathcal{F}$;

(iii) 可列并运算封闭：如果 $A_n \in \mathcal{F}, n \in \mathbb{N}$ ，则 $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}$ 。

上面的三条性质都有相应的概率解释： \emptyset 代表不可能事件、 Ω 代表必然事件，因此都是可以谈论概率的（即 $\emptyset, \Omega \in \mathcal{F}$ ）；如果一个事件 A 可以谈论概率（即 $A \in \mathcal{F}$ ），那么它的对立事件 A^c 也可以谈论概率（即 $A^c \in \mathcal{F}$ ）；如果事件列 $\{A_n\}_1^\infty$ 可以谈论概率（即 $A_n \in \mathcal{F}, \forall n \geq 1$ ），那么事件列 $\{A_n\}_1^\infty$ 中至少发生某一个的复合事件也可以谈论概率（即 $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$ ）。

当 \mathcal{F} 是 Ω 上的 σ -代数时，二元组 (Ω, \mathcal{F}) 也被称为一个可测空间或可测结构；有时也把 σ -代数 \mathcal{F} 称为 Ω 上的一个可测结构。

给定指标集 I ，如果对任意 $\alpha \in I$ ， \mathcal{F}_α 都是 Ω 上的 σ -代数，那么

$$\bigcap_{\alpha \in I} \mathcal{F}_\alpha := \{A : A \in \mathcal{F}_\alpha, \forall \alpha \in I\}$$

也是一个 σ -代数。因此，任给 Ω 上的子集族 \mathcal{E} ，存在唯一一个包含 \mathcal{E} 的最小 σ -代数，这个包含 \mathcal{E} 的最小 σ -代数就被称为是由 \mathcal{E} 生成的 σ -代数，记作 $\sigma(\mathcal{E})$ （为了强调生成 σ -代数时取的全空间为 Ω ，有时也记作 $\sigma_\Omega(\mathcal{E})$ ）。

概率公理：设 Ω 是一个非空集， \mathcal{F} 是其上的 σ -代数， $\mathbb{P}: \mathcal{F} \rightarrow \mathbb{R}$ 是一个集函数，我们称 Ω 为样本空间，称 $\omega \in \Omega$ 为样本点，称 \mathcal{F} 为事件域，称 $A \in \mathcal{F}$ 为事件，称 $\mathbb{P}(A)$ 为事件 A 的概率，相应地 \mathbb{P} 称为概率测度，称 $(\Omega, \mathcal{F}, \mathbb{P})$ 为概率空间（有时也说它是一个概率结构），如果三元组 $(\Omega, \mathcal{F}, \mathbb{P})$ 满足如下公理：

公理 1.（非负性） $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$;

公理 2.（归一性） $\mathbb{P}(\Omega) = 1$;

公理 3.（ σ -可加性/可数可加性） 设 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ 互不相交，则

$$\mathbb{P}\left(\biguplus_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

整个现代概率论的大厦就是在事件域的定义及其上述三条概率公理基础上通过逻辑演绎方式发展出来的用于理解和处理现实生活中各种随机现象的一套数学模型与理论方法。

注记 2.4. 上述概率测度的公理中，如果公理 2 去掉，仅保留公理 1（非负性）和公理 3（ σ -可加性），同时附加要求 \emptyset 的测度值为 0，对应的数学对象称为（非负）测度（只不过一般容许全空间的测度值取无穷值）；此时的三元组 $(\Omega, \mathcal{F}, \mathbb{P})$ 称为测度空间（有时也说它是一个测度结构），详见第 4 章的定义 4.2.1。另外，在概率论及测度论中，我们一般约定 $0 \cdot \infty := 0$ ，这个约定通常仅限于测度及积分的计算中。

2.4.2 概率的基本性质

概率的公理告诉我们，概率测度 \mathbb{P} 具有下面三条性质：

性质（1）（非负性） $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$;

性质（2）（归一性） $\mathbb{P}(\Omega) = 1$;

性质（3）（ σ -可加性） 设 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ 互不相交，则

$$\mathbb{P}\left(\biguplus_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

在上面性质（3）中，取 $A_1 = \Omega, A_n = \emptyset, n \geq 2$ ，则根据性质（1）、（2），我们有

性质（4）（平凡性） $\mathbb{P}(\emptyset) = 0$;

在上面性质（3）中，取 $A_{n+k} = \emptyset, k \geq 1$ ，则根据性质（4），并注意到 $\Omega = A \uplus A^c$ ，我们有

性质（5）（有限可加性） 对任意给定的 $n \geq 1$ ，若 $\{A_k\}_{k=1}^n \subset \mathcal{F}$ 且互不相交，则

$$\mathbb{P}\left(\biguplus_{k=1}^n A_k\right) = \sum_{k=1}^n \mathbb{P}(A_k).$$

特别的， $\mathbb{P}(A^c) = 1 - \mathbb{P}(A), \forall A \in \mathcal{F}$;

在上面性质（5）中，注意到 $A \subset B$ 时， $B = A \uplus (B \setminus A)$ ，我们有

性质（6）（单调性）若 $A, B \in \mathcal{F}$ 且 $A \subset B$ ，则 $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$ ，特别的， $\mathbb{P}(A) \leq \mathbb{P}(B)$ ；

进一步，我们不难证明下面几条性质：

性质（7）（次可列可加性）若诸 $A_n \in \mathcal{F}$ ，则

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n);$$

性质（8）（下连续性）若诸 $A_n \in \mathcal{F}$ 且 $A_n \uparrow A$ ，则 $A \in \mathcal{F}$ ，并且

$$\mathbb{P}(A) = \mathbb{P}\left(\lim_{n \rightarrow \infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n); \quad (2.7)$$

性质（9）（上连续性）若诸 $A_n \in \mathcal{F}$ 且 $A_n \downarrow A$ ，则 $A \in \mathcal{F}$ ，并且

$$\mathbb{P}(A) = \mathbb{P}\left(\lim_{n \rightarrow \infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n); \quad (2.8)$$

特别的， \mathbb{P} 在 \emptyset 处上连续：若诸 $A_n \in \mathcal{F}$ 且 $A_n \downarrow \emptyset$ ，则

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0.$$

以上这些概率的基本性质非常重要，读者应该认真学习掌握，并在之后的课程学习中通过例题、习题等方式锻炼自己，尽力达到熟练掌握、灵活应用的程度。

2.4.3 Jordan 公式

设 $A, B \in \mathcal{F}$ 。利用上一节导出的概率的基本性质，注意到

$$A = (A \cap B) \uplus (A \setminus B), B = (A \cap B) \uplus (B \setminus A)$$

以及

$$A \cup B = (A \setminus B) \uplus (A \cap B) \uplus (B \setminus A),$$

我们得到

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B), \\ \mathbb{P}(B) &= \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A), \\ \mathbb{P}(A \cup B) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A). \end{aligned}$$

于是如下公式成立

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \quad (2.9)$$

一般的，设 $\{A_k\}_{k=1}^n \subset \mathcal{F}$ ，我们有下面的 Jordan 公式（有些文献也称为 Poicaré 恒等式）

$$\mathbb{P}\left(\bigcup_{k=1}^n A_k\right) = \sum_{r=1}^n (-1)^{r-1} \sum_{1 \leq j_1 < \dots < j_r \leq n} \mathbb{P}\left(\bigcap_{p=1}^r A_{j_p}\right). \quad (2.10)$$

该公式可以通过数学归纳法给出证明，此处从略。

例 2.14. (伯努利-欧拉装错信封问题*) 某人写了 n 封信与 n 个信封, 黑暗中信与信封曾散落在地后重新收起; 他在黑暗中完成了把信逐一装入信封的工作。请问: 这些信全部装错的概率有多大?

参考解答: 记 A_k 为第 k 封信装对了信封, 则关心的事件为 $B_n := (\bigcup_{k=1}^n A_k)^c$ 。

容易知道, 给定 $1 \leq r \leq n$ 及 $1 \leq j_1 < \cdots < j_r \leq n$

$$\mathbb{P}(\bigcap_{p=1}^r A_{j_p}) = \frac{(n-r)!}{n!}.$$

于是根据 Jordan 公式

$$\begin{aligned} \mathbb{P}(B_n) &= 1 + \sum_{r=1}^n (-1)^r \sum_{1 \leq j_1 < \cdots < j_r \leq n} \mathbb{P}(\bigcap_{p=1}^r A_{j_p}) \\ &= 1 + \sum_{r=1}^n (-1)^r C_n^r \cdot \frac{(n-r)!}{n!} \\ &= \sum_{r=0}^n \frac{(-1)^r}{r!}. \end{aligned}$$

上述计算给出了这些信全部装错的概率。

很显然, 当 $n \rightarrow \infty$ 时 $\mathbb{P}(B_n) \rightarrow \frac{1}{e} \approx 0.367879$, 即当信的数量充分多时, 信全部装错有不小的概率。□

习 题 2

习题 2.1. 我们可以用函数极限的观点来看待集合的极限问题。求证: 在全集是 Ω 时,

$$(1) A_n \nearrow A \Leftrightarrow 1_{A_n} \nearrow 1_A, A_n \searrow A \Leftrightarrow 1_{A_n} \searrow 1_A;$$

$$(2) \overline{\lim}_{n \rightarrow \infty} 1_{A_n} = 1 \overline{\lim}_{n \rightarrow \infty} A_n, \text{ 且 } \omega \in \overline{\lim}_{n \rightarrow \infty} A_n \Leftrightarrow \sum_{n=1}^{\infty} 1_{A_n}(\omega) = \infty;$$

$$(3) \underline{\lim}_{n \rightarrow \infty} 1_{A_n} = 1 \underline{\lim}_{n \rightarrow \infty} A_n, \text{ 且 } \omega \in \underline{\lim}_{n \rightarrow \infty} A_n \Leftrightarrow \sum_{n=1}^{\infty} 1_{A_n^c}(\omega) < \infty;$$

$$(4) A_n \rightarrow A \Leftrightarrow 1_{A_n} \rightarrow 1_A;$$

$$(5) (\overline{\lim}_{n \rightarrow \infty} A_n) \setminus (\underline{\lim}_{n \rightarrow \infty} A_n) = \overline{\lim}_n (A_n \Delta A_{n+1});$$

$$(6) \overline{\lim}_{n \rightarrow \infty} \{f_n \geq C\} \subset \{\overline{\lim}_{n \rightarrow \infty} f_n \geq C\}, \overline{\lim}_{n \rightarrow \infty} \{f_n \leq C\} \subset \{\underline{\lim}_{n \rightarrow \infty} f_n \leq C\},$$

$$\underline{\lim}_{n \rightarrow \infty} \{f_n \geq C\} \subset \{\underline{\lim}_{n \rightarrow \infty} f_n \geq C\}, \underline{\lim}_{n \rightarrow \infty} \{f_n \leq C\} \subset \{\overline{\lim}_{n \rightarrow \infty} f_n \leq C\}.$$

*此处的伯努利是 Nicolaus I Bernoulli.

习题 2.2. 设 T_n 是集合 $\Omega_n := \{1, 2, \dots, n\}$ 的不同划分 (或分割; *Partition*) 的数目 (本质上它是 Ω_n 上所有不同的 σ -代数的数目, 因为对于离散的样本空间, 一个分割对应生成一个 σ -代数; 反之亦然), 易知 $T_1 = 1, T_2 = 2$ 。请证明: T_n 满足如下递推公式:

$$T_{n+1} = 1 + \sum_{k=1}^n C_n^k \cdot T_k.$$

【提示: Ω_{n+1} 的任一分拆可以按如下方式得到: 考虑 Ω_n 中选 k 个元素单独成一个集合并加入 $n+1$; 剩下的 $n-k$ 个元素继续进行分拆。】

习题 2.3. 令 $\Sigma := \{0, 1\}$, 对任意 $1 \leq k \leq n$, 令

$$A := \{x = (x_1, \dots, x_n) \in \Sigma^n : \sum_{i=1}^n x_i \geq k\}.$$

计算 A 中元素个数 $|A|$ 。

习题 2.4. 证明: 一个只有有限个元素的 σ -代数一定恰有 2^n 个元素, 其中 n 为某整数。此时该 σ -代数中可以找到一个全空间的分割, 它恰有 n 个元素, 并且这个分割生成这个 σ -代数。

习题 2.5. 设 Ω 为一个无穷集合 (可列或不可数), $\mathcal{E} := \{A \subset \Omega : A \text{ 可数或 } A^c \text{ 可数}\}$ 。则 \mathcal{E} 是 σ -代数。

习题 2.6. 设 Ω 为非空集合。令 $\mathcal{E} := \{\{\omega\} : \omega \in \Omega\}$ 。证明: 由 \mathcal{E} 生成的 σ -代数为 $\mathcal{F} := \{A \subset \Omega : A \text{ 或 } A^c \text{ 为可数集}\}$ 。

习题 2.7. 当 Ω 是不可数集时, 证明上一习题中的 σ -代数 \mathcal{F} 不是可数生成的, 即不存在可数个集合 $A_n \in \mathcal{F}, n \geq 1$, 使得 $\mathcal{F} = \sigma(\{A_n : n \geq 1\})$ 。

习题 2.8. 设 $\mathcal{B}_1, \mathcal{B}_2$ 是空间 Ω 上的两个 σ -代数, 问集系

$$\mathcal{B}_1 \cap \mathcal{B}_2 := \{A : A \in \mathcal{B}_1 \text{ 且 } A \in \mathcal{B}_2\},$$

$$\mathcal{B}_1 \cup \mathcal{B}_2 := \{A : A \in \mathcal{B}_1 \text{ 或 } A \in \mathcal{B}_2\}$$

是否是 σ -代数?

习题 2.9. 设 $\mathcal{B}_1, \mathcal{B}_2$ 分别是空间 Ω_1, Ω_2 上的 σ -代数 (假定其中 $\mathcal{B}_1, \mathcal{B}_2$ 都有超过 2 个元素, 即不是平凡的 σ -代数), 问集系

$$\mathcal{B}_1 \times \mathcal{B}_2 := \{A_1 \times A_2 : A_1 \in \mathcal{B}_1, A_2 \in \mathcal{B}_2\}$$

是否是 σ -代数?

习题 2.10. 设 $\{A_1, A_2, \dots, A_n\}$ 为 Ω 的一个覆盖, \mathcal{F} 是这个覆盖生成的 σ -代数, 试证明: $\#(\mathcal{F}) \leq 2^{2^n-1}$ 。

习题 2.11. 给定 \mathcal{E} 为 Ω 上集族及一个非空集 $\Omega_0 \subset \Omega$ 。记

$$\Omega_0 \cap \mathcal{E} := \{\Omega_0 \cap A : A \in \mathcal{E}\},$$

它是 Ω_0 上的集族, 它在 Ω_0 上生成的 σ -代数记作 $\sigma_{\Omega_0}(\Omega_0 \cap \mathcal{E})$ 。那么

$$\sigma_{\Omega_0}(\Omega_0 \cap \mathcal{E}) = \Omega_0 \cap \sigma_{\Omega}(\mathcal{E}).$$

请证明上面的结论。

习题 2.12. 设 \mathcal{E} 为 $\Omega_0 \subset \Omega$ 上的代数。你能给出 $\sigma_{\Omega}(\mathcal{E})$ 与 $\sigma_{\Omega_0}(\mathcal{E})$ 的联系吗?

习题 2.13. 设有 A_1, A_2, A_3 三个事件。请用集合表达下述复合事件：

- (i) 三个事件中，仅 A_1 发生；
- (ii) 三个事件中，仅 A_3 没有发生；
- (iii) 三个事件都发生；
- (iv) 三个事件中至少有一个发生；
- (v) 三个事件中至少有两个发生；
- (vi) 三个事件中仅有一个发生；
- (vii) 三个事件中仅有两个发生；
- (viii) 三个事件中不多于两个事件发生。

习题 2.14. 在概率空间中证明：

$$\lim_{n \rightarrow \infty} \overline{\mathbb{P}}(A_n) \leq \mathbb{P}(\overline{\lim_{n \rightarrow \infty}} A_n), \quad \lim_{n \rightarrow \infty} \mathbb{P}(A_n) \geq \mathbb{P}(\lim_{n \rightarrow \infty} A_n).$$

习题 2.15. 有 n 对夫妇列席参加一个会议，求下列两种不同情况下恰有 k 对夫妇的夫妻座位相邻的概率、至少有 k 对夫妇的夫妻座位相邻的概率：

- (1) 所有 $2n$ 个座位排成一排；
- (2) 所有 $2n$ 个座位排成一圈，形成圆桌会议。

习题 2.16. 一个集合 $A \subset \mathbb{N}$ 称为具有渐近密度 $d \in [0, 1]$ ，如果

$$\lim_{n \rightarrow \infty} \frac{\#(A \cap (0, n])}{n} = d.$$

\mathbb{N} 中具有渐近密度的集合全体记作 \mathcal{E} 。它是 σ -代数吗？

习题 2.17. 设 A, B 为两事件，证明：“对称差” $A \Delta B := (A \setminus B) \cup (B \setminus A)$ 表示了事件“ A, B 中有且仅有其中一者发生”，并且

$$\mathbb{P}(A \Delta B) = \mathbb{P}(A) + \mathbb{P}(B) - 2\mathbb{P}(A \cap B).$$

习题 2.18. 证明： $\mathbb{P}(A \cap B \cap C) \geq \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - 2$ ，并用归纳法证

$$\text{明：} \mathbb{P}\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{i=1}^n \mathbb{P}(A_i) - (n-1).$$

习题 2.19. 给定事件 A_1, \dots, A_n 。令 $S_0 := 1$,

$$S_r := \sum_{1 \leq k_1 < k_2 < \dots < k_r \leq n} \mathbb{P}(A_{k_1} \cap \dots \cap A_{k_r}), \quad 1 \leq r \leq n.$$

记 $N_A := \sum_{k=1}^n 1_{A_k}$ 。

- (1) $\{N_A = m\}$ 为事件“ A_1, \dots, A_n 中有且仅有其中 m 件发生”。于是

$$\mathbb{P}(N_A = m) = \sum_{r=m}^n (-1)^{r-m} C_r^m S_r.$$

特别的, $\{N_A = 0\} = \Omega \setminus (A_1 \cup \cdots \cup A_n)$,

$$\mathbb{P}(N_A = 0) = \sum_{r=0}^n (-1)^r S_r;$$

(2) $\{N_A \geq m\}$ 为事件 “ A_1, \cdots, A_n 中至少有其中 m 件发生”, 其概率为

$$\mathbb{P}(N_A \geq m) = \sum_{r=m}^n (-1)^{r-m} \cdot C_{r-1}^{m-1} \cdot S_r.$$

特别的, “ A_1, \cdots, A_n 中至少有其中一件发生” 这一事件是 $\{N_A \geq 1\} = A_1 \cup \cdots \cup A_n$, 其概率为

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{r=1}^n (-1)^{r-1} S_r; \quad (\text{Poincaré 恒等式})$$

(3) *Fréchet* 不等式: $\frac{S_{r+1}}{C_{n+1}^{r+1}} \leq \frac{S_r}{C_n^r}, \quad r = 0, 1, \cdots, n-1;$

(4) *Gumbel* 不等式: $\frac{C_{n-1}^{r+1} - S_{r+1}}{C_{n-1}^r} \leq \frac{C_n^r - S_r}{C_{n-1}^{r-1}}, \quad r = 1, \cdots, n-1.$

习题 2.20. 设 A, B, C 为概率空间中的事件, 试证明如下 “三角不等式”:

$$\mathbb{P}(A \Delta B) \leq \mathbb{P}(A \Delta C) + \mathbb{P}(C \Delta B).$$

习题 2.21. 在概率空间的事件域中定义 $d(A, B) := \frac{\mathbb{P}(A \Delta B)}{\mathbb{P}(A \cup B)} \cdot 1_{\{\mathbb{P}(A \cup B) > 0\}}$. 求证: $d(A, B) \leq d(A, C) + d(C, B)$.

习题 2.22. 设 A, B 为概率空间中的事件, 如果 $\mathbb{P}(A \cap B) = 0$ 则称 A, B 几乎互不相交; 事件列 $\{A_n : n \geq 1\}$ 几乎互不相交, 如果存在事件 Γ 使得 $\mathbb{P}(\Gamma) = 0$ 且事件列 $\{A_n \setminus \Gamma : n \geq 1\}$ 互不相交。试证明: 如果事件列 $\{A_n : n \geq 1\}$ 两两几乎互不相交, 那么它们几乎互不相交, 并且

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

习题 2.23. 设 $\{A_k\}_{k=1}^n$ 为 $n \geq 3$ 个事件, 具有如下性质: (1) 三个及以上事件不可能同时发生; (2) 单个事件不可能发生。试证明:

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \frac{1}{2} \cdot [\mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)].$$

习题 2.24. 设 $\{A_k\}_{k=1}^n$ 为 $n \geq 3$ 个事件, 具有如下性质: (1) 至少有一个事件发生; (2) 三个及以上事件不可能同时发生; (3) $\mathbb{P}(A_i) = p, \mathbb{P}(A_i \cap A_j) = q, i \neq j$. 试证明: $p \geq \frac{1}{n}, q \leq \frac{2}{n(n-1)}$.

习题 2.25. n 个不同编号的球随机地放入 m 个不同编号的盒子中 ($2 \leq m \leq n$), 求

- (1) 恰有一个空盒的概率;
- (2) 没有空盒的概率。

习题 2.26. 在一条线段上随机地取两点, 把该线段分成了三段。求这三段子线段能组成三角形的概率。

习题 2.27. 赌徒还曾向 *Pascal* 提出以下概率大小的比较问题，请你思考：

(1) 投掷 4 次骰子至少得到一个 6 的概率，与投掷两个骰子 24 次至少得到一次一双 6 的概率，哪个大？

(2) 投掷 $6n$ 个骰子得到至少 n 个 6 的概率，与投掷 $6(n+1)$ 个骰子得到至少 $n+1$ 个 6 的概率，哪个大？

习题 2.28. 有穿过长管道的六根长绳，绳子的头尾分别露在管道的两端。小明抓住它们的头，两两配对共打 3 个死结；小军抓住它们的尾，也两两配对打 3 个死结。请问，最终六根长绳形成一个环的概率？

习题 2.29. 运用例 2.9 中方法，证明：当 $C_n^p < 3^{C_p^2-1}$ 时，存在对完全图 K_n 的一种边三染色方案，使得染色后其中找不到单色子图 K_p 。

习题 2.30. (1) 对完全图 K_4 的边进行随机二染色，求其中能找到单色子图 K_3 的概率；

(2) 对完全图 K_5 的边进行随机二染色，求其中能找到单色子图 K_3 的概率。

习题 2.31. 把单位圆周染成红白二色，其中白色比重为 q 。试证明：如果 $q < \frac{1}{4}$ ，那么不论如何染色，总能在染色后的圆周上找到均为红色的四点，它们构成单位圆的内接正方形。

【提示：把单位圆周等同于 $[0, 1]$ ，考虑模 1 意义下的加法，设 $A \subset [0, 1]$ 满足 $\text{Leb}(A) > \frac{3}{4}$ ，论证 $\text{Leb}(A \cap (A + \frac{1}{4}) \cap (A + \frac{2}{4}) \cap (A + \frac{3}{4})) > 0$ 。】

§ 3

经典条件概率与事件独立性

自本章开始，我们总假设 $(\Omega, \mathcal{F}, \mathbb{P})$ 是一个概率空间，并且在无特殊必要的场合将不再申明这一点；至于为何可以如此地作假设，请参见本章注记 3.3。我们将通过逻辑演绎与类比，引入一些新的概念，逐步展开介绍概率的有关理论。

3.1 经典条件概率

3.1.1 条件概率空间与经典条件概率的定义

在现实生活中，一件事情发生后，总是或多或少能提供给我们一些信息，这些信息会导致我们判断其他事件是否发生的可能性大小发生变化：天空中飘来乌云（积雨云）后我们通常会认为天将下雨的可能性高于原来晴空万里、乌云未至的情况下的天将下雨的可能性；原来装有 2 个黑球、2 个白球的袋子中，摸走一个黑球后，接下来再次摸球摸中黑球的概率就相比最初的时候降低了。这些都是我们生活中的经验和直观就能感知的结论。事件 A 发生的条件下，另一事件 B 发生的概率，我们称为**条件概率**，符号上记作 $\mathbb{P}(B|A)$ ，它的取值通常应该与不知道 A 是否发生的时候给出的无条件概率 $\mathbb{P}(B)$ 有所区别。问题是，我们该如何合理地在数学上精确定义我们直观中的条件概率这一模糊概念呢？

为了借助直观来理解这一问题，我们有必要来看看最简单的古典概率模型 $(\Omega, \mathcal{F}, \mathbb{P})$ 。此时 Ω 是一个具有有限个样本点的样本空间，其中诸样本点具有等可能性作为一个基本结果出现，能谈论概率的事件全体为事件域 $\mathcal{F} = 2^\Omega$ ，等可能性假设意味着 $\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}, \forall \omega \in \Omega$ 。当事件 $A \subset \Omega$ 发生时，我们拟推断另一事件 $B \subset \Omega$ 发生的条件概率。显然，当我们谈论事件 $A \subset \Omega$ 发生时，一个（一般情况下未知的）基本结果 $\omega \in A$ 已经出现，如果我们还要求事件 $B \subset \Omega$ 发生（此时 $\omega \in B$ ），则本质上我们要求 $\omega \in A \cap B$ ，即事件 $A \cap B$ 发生，它是 A 事件的子事件。很明显，我们应该考虑以 A 作为样本空间的一个新的概率空间（此处假定 $\mathbb{P}(A) > 0$ ），某个非常自然的子概率空间 $(A, \mathcal{F}_A, \mathbb{P}_A)$ ，其中应有

$$\mathcal{F}_A = A \cap \mathcal{F} := \{A \cap B : B \in \mathcal{F}\}.$$

问题在于，我们该如何选择定义 \mathbb{P}_A 使得 $\mathbb{P}_A(A \cap B)$ 就可以认为是条件概率 $\mathbb{P}(B|A)$ ？由于我们认为概率空间 $(A, \mathcal{F}_A, \mathbb{P}_A)$ 是大的概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 的非常自然的子空间，大空间满足等可能性假设，理所当然地我们认为小的子概率空间 $(A, \mathcal{F}_A, \mathbb{P}_A)$ 也满足等可能性假设（这是一种非常合理的遗传性），于是 $(A, \mathcal{F}_A, \mathbb{P}_A)$ 也是古典概率模型，

$$\mathbb{P}(B|A) := \mathbb{P}_A(A \cap B) = \frac{|A \cap B|}{|A|} = \frac{|A \cap B|/|\Omega|}{|A|/|\Omega|} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

上述简单易懂理解情形的讨论立即诱使我们给出一般的概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中条件概率的定义：当 $A, B \in \mathcal{F}$ 满足 $\mathbb{P}(A) > 0$ 时，我们以 $\mathbb{P}(B|A)$ 表示事件 A 发生的条件下事件 B 发生的**条件概率**，它定义为如下数值

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \quad (3.1)$$

上述讨论的方式、方法是自然科学（特别是物理学）从容易理解的特例中的已知概念、具体性质出发，派生更一般的抽象概念的常见套路，读者应自行多多体悟。

以上讨论也给出了一般的概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 的**子概率空间** $(A, \mathcal{F}_A, \mathbb{P}_A)$ 的合理定义（其中 $A \in \mathcal{F}$ 满足 $\mathbb{P}(A) > 0$ ），即此时 $\mathcal{F}_A := A \cap \mathcal{F}$ ，且

$$\mathbb{P}_A(\cdot) := \mathbb{P}(\cdot|A).$$

由此， $(A, \mathcal{F}_A, \mathbb{P}_A)$ 也可称为**条件概率空间**。请读者验证如上定义的 $(A, \mathcal{F}_A, \mathbb{P}_A)$ 确实是一个概率空间。在第 7 章探讨条件数学期望、条件概率时，我们将会再次涉及此处的这个条件概率空间。

例 3.1. 原来装有 2 个黑球、2 个白球的袋子中，第一次摸走一个黑球后，再次摸一个球，请问第二次能摸中黑球的概率是多少？

参考解答：在原来装有 2 个黑球、2 个白球的袋子中，第一次摸走一个黑球后，袋子中剩下 1 个黑球、2 个白球，从而第二次摸一个球，能摸中黑球的概率很自然是 $\frac{1}{3}$ ；这是直观就能简单理解的。

因为袋子中总共有 4 个球，在无放回取球的规则下，我们现在故意用连续摸 4 次球的概率空间建模的观点来看这个问题。把黑球编号为 1, 2 并记 $B = \{1, 2\}$ ，白球编号为 3, 4 并记 $W := \{3, 4\}$ ，这时候可以建模为样本空间如下的古典概率模型

$$\Omega := \{(\omega_1, \dots, \omega_4) : \omega_1, \dots, \omega_4 \text{ 是 } 1, 2, 3, 4 \text{ 的一个排列}\}.$$

此时， $A_1 := \{\omega_1 \in B\}$ 代表第一次摸走了一个黑球， $A_2 := \{\omega_2 \in B\}$ 代表第二次摸走了一个黑球。容易算出 $|\Omega| = 4! = 24$ ， $|A_1| = 12$ ， $|A_1 \cap A_2| = 4$ 。于是 $\mathbb{P}(A_1) = \frac{12}{24} = \frac{1}{2}$ ， $\mathbb{P}(A_1 \cap A_2) = \frac{4}{24} = \frac{1}{6}$ ，

$$\mathbb{P}(A_2|A_1) = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} = \frac{1/6}{1/2} = \frac{1}{3}. \quad \square$$

对上述问题，读者也可尝试建立连续摸 2 次球的古典概率模型进行解答。

例 3.2. 已知某夫妇有两个孩子。某次外出活动中，该夫妇仅带出来其中一个孩子，是女孩。请问该夫妇的另一个孩子也是女孩的概率有多大？（假定小孩出生时，生儿、育女是等概率的。）

参考解答：由于假设生儿、育女是等概率的，我们可以取样本空间为

$$\Omega = \{(\text{男}, \text{男}), (\text{男}, \text{女}), (\text{女}, \text{男}), (\text{女}, \text{女})\},$$

用以表示两个孩子按照老大、老二排列形成的组合全体。于是已经发生的事件 A 是“观察到老大或老二是女孩”，即 $A = \Omega \setminus \{(\text{男}, \text{男})\}$ ；关心的事件 B 是“老大、老二都是女孩”，即 $B = \{(\text{女}, \text{女})\}$ ，关心的概率其实是条件概率 $\mathbb{P}(B|A)$ ，其中 $A \cap B = B$ 。因此

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{1/4}{3/4} = \frac{1}{3}. \quad \square$$

例 3.3. (*Monty Hall 问题*, 又称*三门问题*、*山羊汽车问题*) 以下是 *Monty Hall* 问题的一个著名的叙述 (*Steve Selvin* 于 1975 年 2 月寄给 *American Statistician* 杂志的叙述的改编版本)，来自 *Craig F. Whitaker* 于 1990 年寄给 *Parade Magazine* (*《展示杂志》*) *Marilyn vos Savant* (玛丽莲·沃斯·莎凡特) 专栏的信件。假设你正在参加一个游戏节目，你被要求在三扇门中选择一扇：其中一扇后面有一辆车；其余两扇后面则是山羊。你选择了一扇门，假设是一号门，然后知道门后面有什么的主持人蒙提霍尔 (*Monty Hall*)，开启了另一扇后面有山羊的门，假设是三号门。他然后问你：“你想选择二号门吗？”转换你的选择是否能够增加你获得汽车奖品的概率？

三门问题的错误概率解答：按两羊一车在三门中的排列是等可能性的假定

$$\begin{aligned} \mathbb{P}(\text{主持人打开的三号门} = \text{羊}) &= \frac{2}{3}, \\ \mathbb{P}(\text{一号门} = \text{车}, \text{主持人打开的三号门} = \text{羊}) &= \frac{1}{3}, \end{aligned}$$

因此

$$\begin{aligned} &\mathbb{P}(\text{一号门} = \text{车} | \text{主持人打开的三号门} = \text{羊}) \\ &= \frac{\mathbb{P}(\text{一号门} = \text{车}, \text{主持人打开的三号门} = \text{羊})}{\mathbb{P}(\text{主持人打开的三号门} = \text{羊})} = \frac{1}{2}. \end{aligned}$$

这表明换不换的（条件）概率都是 $\frac{1}{2}$ 。 \square

但上述乍看貌似道理满满的解答其实是对问题的误读。我们通常应该解读为，主持人总是打开剩下两门中对应山羊的其中某一扇门，这样事件“主持人打开的三号门 = 羊”在你选了一号门后，可以解读为确定性事件！也就是说在这种理解下，上述条件概率

$$\mathbb{P}(\text{一号门} = \text{车} | \text{主持人打开的三号门} = \text{羊}) = \frac{1}{3} < \frac{1}{2}.$$

因此，应该转换选择。

哪怕主持人并没有“总是打开剩下两门中对应山羊的其中某一扇门”的意图，而只是“刚巧打开的三号门对应是山羊”，我们也可以给出另外两种解释来说明应该转换选择。这两种解释，一者是借助直观，另一者则是穷尽几种情况分析，并呈现为图片形式。

三门问题的一种直观解答（改编自知乎社区的一位回答者）：实际上，你选了一扇门后，剩下的 2 扇门中总是至少有一扇对应于山羊；这个步骤可以理解为由你来指派把三个门分成 1 个门和 2 个门的两组。问题转化为：为了更

高的概率获得汽车奖品，你该选择 1 个门的组还是 2 个门的组？毫无疑问应该是后者，并且对应汽车奖品概率是 $2/3$ ，而原选择对应汽车奖品概率为 $1/3$ ；主持人打开剩下两扇门中对应山羊的 (某个或唯一一个) 门的行为只不过更方便你更换选择后直接去兑换汽车奖品（如果中奖的话）而已。 □

三门问题的另一种概率解答（图片来自维基百科）：

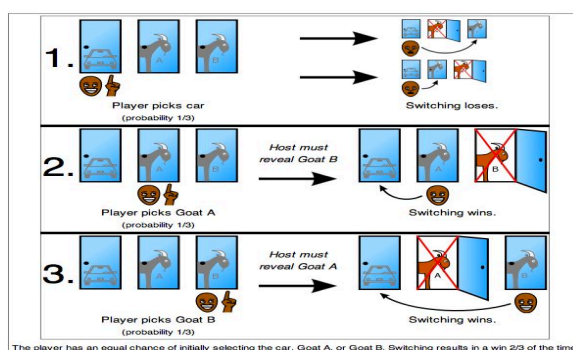


图 3.1: 三门问题的图示解答

3.1.2 乘法公式

根据(3.1)，我们有如下两个事件同时发生的概率的一个计算公式：

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A). \quad (3.2)$$

上述公式称为乘法公式。

下面有趣的例子来自 [34]；这提供了条件概率的乘法公式和概率测度的上连续性的一个很好的应用，同时也提供了我们对涉及“无穷”的问题的一种合理解释。

例 3.4. (Littlewood “无穷”悖论问题) 假设有一个无限大的空坛子以及无限个分别带有编号 1、2、3、... 的球。考虑以下实验 ①：在差 1 分到 12 点钟的时候（我们称之为第 1 步），将 1 至 10 号球放入坛中，并把 10 号球取出（总假设放球、取球不花费时间）；在差 $\frac{1}{21}$ 分到 12 点钟的时候（我们称之为第 2 步），将 11 至 20 号球放入坛中，并把 20 号球取出；在差 $\frac{1}{22}$ 分到 12 点钟的时候，将 21 至 30 号球放入坛中，并把 30 号球取出；一般的，在差 $\frac{1}{2^n}$ 分到 12 点钟的时候（我们称之为第 $n+1$ 步），将 $10 * n + 1$ 至 $10 * (n + 1)$ 号球放入坛中，并把 $10 * (n + 1)$ 号球取出；...

问题 1：在实验 ① 中，到 12 点钟整的时刻，坛子里有多少个球？

问题 1 的答案很明显，那时坛子中将会有无穷多个球，因为所有编号的球，除了 10 的倍数编号的球外，都在 12 点钟之前放入、并一直呆在坛子中而没有被再次取出。

现在对上述放球、取球方案进行调整，称为实验 ②：一般的，对任意 $n \geq 0$ ，总是在差 $\frac{1}{2^n}$ 分到 12 点钟的时候（第 $n+1$ 步），将 $10 * n + 1$ 至 $10 * (n + 1)$ 号球放入坛中，并把 $n + 1$ 号球取出。

问题 2：在实验 ② 中，到 12 点钟整的时刻，坛子里有多少个球？

令人惊讶的是，问题 2 的答案是：到 12 点钟整的时刻，坛子里一个球也没有！理由是：任何编号为 n 的球都在第 n 步中被取出、并且从此之后再也没有被重新放入坛子里。*读者需要设法理清这里反直觉的现象出现的原因。

最后我们设计另外一种放球之后再进行随机取球的方案，称为实验 ③：一般的，对任意 $n \geq 0$ ，总是在差 $\frac{1}{2^n}$ 分到 12 点钟的时候（第 $n+1$ 步），将 $10 \cdot n + 1$ 至 $10 \cdot (n+1)$ 号球放入坛中，并在把坛内的球（此时坛内有 $9 \cdot (n+1) + 1$ 球）混合均匀后随机的从中取出一球。

问题 3：在实验 ③ 中，到 12 点钟整的时刻，坛子里有多少个球？

问题 3 的解答：我们将证明：最后坛子里一个球也没有；数学上，我们是通过证明“到 12 点钟整的时刻，坛中有球”这一事件是几乎不可能事件来实现这个说理的。具体来说，我们将证明，对任意编号为 n 的球，它在坛子中的概率是 0。

先考虑 1 号球。记 E_n 表示“在第 n 步的放球、取球动作完成后，1 号球仍然在坛子中”这一事件。记 E_∞ 为“到 12 点钟时 1 号球仍然在坛子中”这一事件。容易知道（这里推荐使用条件概率诱导的乘法公式： $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A)$ ）

$$\mathbb{P}(E_n) = \frac{9}{10} \cdot \frac{18}{19} \cdots \frac{9n}{9n+1} = \left[\prod_{k=1}^n \left(1 + \frac{1}{9k}\right) \right]^{-1},$$

并且 $E_\infty = \bigcap_{n=1}^{\infty} E_n, E_n \searrow E_\infty$ 。由概率的上连续性

$$\begin{aligned} \mathbb{P}(E_\infty) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(E_n) = \lim_{n \rightarrow \infty} \left[\prod_{k=1}^n \left(1 + \frac{1}{9k}\right) \right]^{-1} \\ &\leq \lim_{n \rightarrow \infty} \left[1 + \sum_{k=1}^n \frac{1}{9k} \right]^{-1} = 0. \end{aligned}$$

现在对任意 $i \geq 1$ ，记 F_i 为“到 12 点钟时 i 号球仍然在坛子中”这一事件；到 12 点钟时坛子不空即事件 $F := \bigcup_{i=1}^{\infty} F_i$ 。仿照上面方法可以证明总有 $\mathbb{P}(F_i) = 0$ 。于是

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} F_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(F_i) = 0.$$

因此，到 12 点钟时，事件 F 的概率为 0，亦即坛子空的概率为 1。 \square

3.1.3 全概率公式

现在假设有两两互斥的事件组 $\{A_k\}_{k=1}^N \subset \mathcal{F}$ （其中 $N \leq \infty$ ）满足

$$\Omega = \bigsqcup_{k=1}^N A_k.$$

*这被 J. Littlewood (1885/6/9–1977/9/6; 英国) 称为 an infinity paradox, 见 [24]; 因此我们称之为 Littlewood “无穷” 悖论。

我们称此事件组 $\{A_k\}_{k=1}^N$ 为**完备事件组**。此时，对任意事件 $B \in \mathcal{F}$ ，我们有

$$B = B \cap \left(\bigcup_{k=1}^N A_k \right) = \bigcup_{k=1}^N (A_k \cap B).$$

于是

$$\begin{aligned} \mathbb{P}(B) &= \sum_{k=1}^N \mathbb{P}(A_k \cap B) \\ &= \sum_{k=1}^N \mathbb{P}(A_k) \cdot \mathbb{P}(B|A_k). \end{aligned}$$

因此我们有下面的定理。

定理 3.1.1.（全概率公式） 如果 $\{A_k\}_{k=1}^N$ 是完备事件组，那么

$$\mathbb{P}(B) = \sum_{k=1}^N \mathbb{P}(A_k) \cdot \mathbb{P}(B|A_k). \quad (3.3)$$

注意，对于两两互斥的事件组 $\{A_k\}_{k=1}^N \subset \mathcal{F}$ ，如果

$$\mathbb{P}(B \setminus \bigcup_{k=1}^N A_k) = 0,$$

则上述全概率公式(3.3)仍然成立。

例 3.5.（选举得票率问题） 本问题来源于 [44]。假设一个学生团体即将进行选举，有特雷茜和保罗两个候选人。民调显示，在左撇子和右撇子这两类选民（并假定只有这两类选民）中，对两位候选人的支持率存在重大差异。可能是特雷茜大力提倡教室里的左手桌的政策缘故，左撇子选民中，75% 支持特雷茜，只有 25% 支持保罗。然而右撇子选民并不信服：在这些选民中，60% 支持保罗，只有 40% 支持特雷茜。假设左撇子占投票人数的 20%，那么谁会赢得选举？用不同（但却等价的）方式来表述该问题：如果你随机地问一个投票的人，他/她投票给特雷茜的概率有多大？

参考解答：我们用 $L =$ 左撇子, $R =$ 右撇子, $T =$ 投票特雷茜, $B =$ 投票保罗，由全概率公式

$$\begin{aligned} \mathbb{P}(T) &= \mathbb{P}(L)\mathbb{P}(T|L) + \mathbb{P}(R)\mathbb{P}(T|R) \\ &= 20\% \times 75\% + (1 - 20\%) \times 40\% = 47\%. \end{aligned}$$

从而 $\mathbb{P}(B) = 1 - \mathbb{P}(T) = 53\%$ ，即特雷茜得票率为 47%，比保罗的得票率低 6%，民调结果表明保罗将赢得选举。

民调中，特雷茜输掉选举的原因是，左撇子选民在整体选民中的比例小，她提倡的政策吸引的选民受众面太窄；为了赢得选举，她还应当设法在占大多数的右撇子选民中扩大影响、提升支持率。□

对于上面的例子，假设左撇子选民占比由 20% 变动至 40%，其他数据不变，请读者通过计算判断谁会赢得选举。

例 3.6.（敏感性调查问题） 设某校想要对研究生论文抄袭现象进行社会调查。如果直接就此问题进行问卷调查，就是说，要被调查者直说是否发生过

抄袭，即使这样的调查是无记名的，也会使被调查者感到尴尬，从而得不到如实的回答。现在设计如下方案可使被调查者更乐意作出真实的回答：在一个箱子里放进 1 个红球和 1 个白球。被调查者在摸到球后记住颜色并立刻将球放回，然后根据球的颜色是红和白分别回答第一个问题或第二个问题：(1) 你的生日是否在 7 月 1 日以前？(2) 你做论文时是否有过抄袭行为？回答时只要在一张预备好的白纸上打 \checkmark 或打 \times ，分别表示是或否。假定被调查者有 150 人，统计出有 60 个 \checkmark 。问题：有抄袭行为的比率大概是多少？

参考解答：注意到生日在 7 月 1 日以前的人群比例大致是 50%，我们使用简略易懂的记号可以把题设中的已知条件表述为：

$$\begin{cases} \mathbb{P}(\text{红}) = \mathbb{P}(\text{白}) = 0.5, \\ \mathbb{P}(\checkmark|\text{红}) = \mathbb{P}(\times|\text{红}) = 0.5, \\ \mathbb{P}(\checkmark) = \frac{60}{150} = 0.4. \end{cases}$$

根据全概率公式

$$\mathbb{P}(\checkmark) = \mathbb{P}(\text{白})\mathbb{P}(\checkmark|\text{白}) + \mathbb{P}(\text{红})\mathbb{P}(\checkmark|\text{红}),$$

于是所求为

$$\begin{aligned} \mathbb{P}(\checkmark|\text{白}) &= \frac{\mathbb{P}(\checkmark) - \mathbb{P}(\text{红}) \cdot \mathbb{P}(\checkmark|\text{红})}{\mathbb{P}(\text{白})} \\ &= \frac{0.4 - 0.5 \cdot 0.5}{0.5} = 0.3. \end{aligned}$$

因此，此社会调查估计的抄袭行为比例为 30%。

□

注记 3.1. 对于涉及隐私等私密性、敏感性问题的社会调查的调查问卷的设计，上述敏感性调查问题给出了一个极其聪明的调查设计方案，是统计调查领域的经典案例。

下面另一个类似的经典案例则来自 Ross [34, Chapter 3, Example 3h]。

例 3.7. (同卵双生问题) 双胞胎可能是同卵双生或异卵双生的。同卵双生也叫单卵双生，是由一个受精卵分裂成为两个完全一样的部分发育而来，因而同卵双胞胎含有相同的基因。异卵双生又叫二卵双生，是由两个受精卵植入子宫发育而来。异卵双胞胎就像不同时间出生的兄弟姐妹一样，基因多少是有些一样的。为了知道双胞胎中同卵双胞胎的比例，加利福尼亚州洛杉矶市指派了一位统计学家来研究该问题。这位统计学家最初要求该市每家医院对双胞胎做记录，同时对是否同卵双生做标记。然而，医院告诉他，要判断一个新生儿是否是同卵双生并不是一件简单的事，这关系到父母是否愿意自费给孩子做这项复杂而又昂贵的 DNA 检验。经过一番思考之后，统计学家只让医院提供标记着双胞胎是否是相同性别的所有双胞胎数据列表。当数据表明，约有 64% 的双胞胎是性别相同时，统计学家就宣称：约有 28% 的双胞胎是同卵双生的。他是如何得到这个结论的？

参考解答：由于生物学上的结论，同卵双生的双胞胎总是同性别的；而异卵双生的双胞胎就相当于普通的兄弟姐妹们，因此他们性别相同的概率为 $1/2$ 。我们使用简略易懂的记号可以把题设中的已知条件表述为：

$$\begin{cases} \mathbb{P}(\text{同性}|\text{同卵}) = 1, \\ \mathbb{P}(\text{同性}|\text{异卵}) = 0.5, \\ \mathbb{P}(\text{双生同性}) = 0.64, \end{cases}$$

根据全概率公式

$$\begin{aligned}\mathbb{P}(\text{双生同性}) &= \mathbb{P}(\text{同卵})\mathbb{P}(\text{同性}|\text{同卵}) + \mathbb{P}(\text{异卵})\mathbb{P}(\text{同性}|\text{异卵}) \\ &= \mathbb{P}(\text{同卵}) + [1 - \mathbb{P}(\text{同卵})] \cdot \frac{1}{2} \\ &= 0.5 + 0.5 * \mathbb{P}(\text{同卵}).\end{aligned}$$

于是所求为

$$\mathbb{P}(\text{同卵}) = \frac{\mathbb{P}(\text{双生同性}) - 0.5}{0.5} = \frac{0.64 - 0.5}{0.5} = 28\%.$$

因此，同卵双生在双胞胎中的比例为 28%。 \square

例 3.8.（抓阄的公平性） 假设有 r 项奖品，共有 $n \geq r$ 个人来共同分配（每人不允许获得两项及以上的奖品），因此采取抓阄的方式来实现：用 r 个红球代表奖品，用 $b = n - r$ 个黑球代表无奖品，然后采用无放回的方式取球 n 次。则这种抓阄方式是公平的，抓阄的次序不会影响公平性。

参考解答： 记初始 b 个黑球， r 个红球时考虑的无放回取球问题对应概率测度为 $\mathbb{P}_{b,r}$ ，以 R_j 表示第 j 次取到红球这一事件。于是容易知道 $\mathbb{P}_{b,r}(R_1) = \frac{r}{b+r}$ ， $\mathbb{P}_{b,r}(R_1^c) = \frac{b}{b+r}$ 对所有 $b, r \geq 1$ 成立。现在用归纳法证明 $\mathbb{P}_{b,r}(R_j) = \frac{r}{b+r}$ ， $\mathbb{P}_{b,r}(R_j^c) = \frac{b}{b+r}$ 对所有 $b, r \geq 1, 1 \leq j \leq b+r$ 成立：事实上，当 $j+1 \leq b+r$ 时

$$\begin{aligned}\mathbb{P}_{b,r}(R_{j+1}) &= \mathbb{P}_{b,r}(R_1) \cdot \mathbb{P}_{b,r}(R_{j+1}|R_1) + \mathbb{P}_{b,r}(R_1^c) \cdot \mathbb{P}_{b,r}(R_{j+1}|R_1^c) \\ &= \frac{r}{b+r} \mathbb{P}_{b,r-1}(R_j) + \frac{b}{b+r} \mathbb{P}_{b-1,r}(R_j) \quad (\text{模型假定}) \\ &= \frac{r}{b+r} \cdot \frac{r-1}{b+r-1} + \frac{b}{b+r} \cdot \frac{r}{b+r-1} \quad (\text{归纳假设}) \\ &= \frac{r}{b+r}.\end{aligned}$$

进而 $\mathbb{P}_{b,r}(R_{j+1}^c) = 1 - \mathbb{P}_{b,r}(R_{j+1}) = \frac{b}{b+r}$ 。由此最终结论成立。 \square

上例中，通过考虑第一步/第一次（取球）的所有可能，建立对应的全概率公式，进而建立递推公式的思想非常自然，也非常有用。历史上，Pascal 和 Fermat 对于赌金分配问题的讨论中给出了若干种解答，其中就有这种建立递推公式的解答思路，我们介绍如下。

例 3.9.（赌金分配问题） 设甲乙赌徒进行公平的赌博（即二人单局获胜概率相同；假定没有平局）。事先约定谁先赢得 6 局便算赢家，赢家获得全部赌金。如果在甲赢 5 局，乙赢 2 局时因故终止赌博，应如何分配赌金才合理？

参考解答： 令 $f_6(x, y) := \mathbb{P}_{x,y}$ (甲先赢得 6 局) 表示甲、乙分别赢了 x 局和 y 局时继续赌下去甲最终先赢得 6 局的条件概率。通过考虑后续的第一局的胜负情况，建立 $f_6(x, y)$ 的递推公式如下：

$$f_6(x, y) = \frac{1}{2}[f_6(x+1, y) + f_6(x, y+1)], \quad 0 \leq x, y \leq 5.$$

注意到

$$f_6(x, 6) = 0, f_6(6, x) = 1, \quad \forall 0 \leq x \leq 5,$$

并由对称性得到

$$f_6(x, x) = \frac{1}{2}, \quad \forall 0 \leq x \leq 5,$$

我们可以计算 $f_6(5, 2)$ 如下

$$\begin{aligned} f_6(5, 2) &= \frac{1}{2}[f_6(6, 2) + f_6(5, 3)] \\ &= \frac{1}{2} + \frac{1}{4}[f_6(6, 3) + f_6(5, 4)] \\ &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8}[f_6(6, 4) + f_6(5, 5)] \\ &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16}. \end{aligned}$$

进而甲、乙两赌徒赌金分配比例为 $f_6(5, 2) : [1 - f_6(5, 2)] = 15 : 1$ 。

如果令 $g_6(x, y) := \mathbb{P}_{x,y}(\text{乙先赢得 6 局})$ 表示甲、乙分别赢了 x 局和 y 局时继续赌下去乙最终先赢得 6 局的条件概率，也可以类似计算 $g_6(5, 2)$ ，结果是一致的，计算效率略高些。请读者自行尝试。 \square

作为一个对应的练习，请读者参见本章习题 3.4 探讨的赌金分配问题的变种问题——赌徒破产模型。

例 3.10. (Polyá 坛子模型：推广模型 A) 坛内有 b 个黑球， r 个红球，每次随机地取走一球，观察取出的球的颜色并放回坛内与之同色的球 $a+1 \geq 0$ 个，则 $a = -1$ 时相当于无放回模型， $a = 0$ 时相当于有放回模型， $a \geq 1$ 时就是每次增加同色球 a 个的模型。根据首次取球的颜色，利用全概率公式，归纳法容易证明：

- $\mathbb{P}(B_j) = \frac{b}{b+r}, \mathbb{P}(R_j) = \frac{r}{b+r}, \forall j$ 满足 $b+r+ja \geq 0$;
- $\mathbb{P}(\text{连续 } n \text{ 次取球，恰取得 } k \text{ 个黑球}) = \frac{[b+(k-1)a]!_{a,k} [r+(n-k-1)a]!_{a,n-k}}{[b+r+(n-1)a]!_{a,n}}$ ，其中， $n!_{a,k} := n(n-a) \cdots [n-(k-1)a]$ ；此处对 b, r, a, n, k 有一些要求（乘积中分母不出现 0 和负数，分子不出现负数）。

上例中结论的证明留给读者去完成。

例 3.11. (Polyá 坛子模型：推广模型 B) 坛内有 b 个黑球， r 个红球，每次随机地取走一球，记录下取出的球的颜色（黑色记录为 $I = 0$ ，红色记录为 $I = 1$ ）并放回坛内黑、红球各 $b_I + (1-I), r_I + I$ 个 ($I = 0, 1$)。当 $b_0 = r_1 = a, b_1 = r_0 = 0$ 时，此模型就变成上例中的模型。本推广模型中相关事件的概率计算可以通过条件概率的方法实现，只不过计算及结果的表达更为复杂，此处从略。


3.1.4 Bayes 公式

现在假定有完备事件组 $\{A_k\}_{k=1}^N \subset \mathcal{F}$ 以及满足 $\mathbb{P}(B) > 0$ 的事件 $B \in \mathcal{F}$ ，我们关心 $\mathbb{P}(A_i|B)$ 的概率大小，此时根据条件概率的定义、乘法公式以及全

概率公式

$$\begin{aligned}\mathbb{P}(A_i|B) &= \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(B|A_i)}{\sum_{k=1}^N \mathbb{P}(A_k) \cdot \mathbb{P}(B|A_k)}.\end{aligned}$$

于是我们有下面的 Bayes 公式：

 **定理 3.1.2.** (*Bayes 公式*) 如果 $\{A_k\}_{k=1}^N$ 是完备事件组，并且 $\mathbb{P}(B) > 0$ ，那么

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(B|A_i)}{\sum_{k=1}^N \mathbb{P}(A_k) \cdot \mathbb{P}(B|A_k)}. \quad (3.4)$$

上述 Bayes 公式(3.4)中，通常把 $\mathbb{P}(A_i)$ 称为先验概率， $\mathbb{P}(A_i|B)$ 称为后验概率；把 $\{\mathbb{P}(A_k)\}_{k=1}^N$ 称为先验分布（列）， $\{\mathbb{P}(A_k|B)\}_{k=1}^N$ 称为后验分布（列）。

例 3.12. (*选举得票率问题后续*) 本例是例 3.5 的后续。考虑两个后续问题：

问题 A: 假设你在一个自助厅，观察到坐在对面的人（他/她是选民）用左手拿勺子，并据此推断出他/她是左撇子。那么他/她支持特雷茜的概率有多大？

问题 B: 假设你在一个自助厅，观察到坐在对面的人（他/她是选民）带有一个印着“我 ♥ 特雷茜”的徽章，并据此推断出他/她支持特雷茜。那么他/她是左撇子的概率有多大？

参考解答： 问题 A 的答案根据例 3.5 中数据应该是 $\mathbb{P}(T|L) = 75\%$ 。我们知道问题 B 实际上在问 $\mathbb{P}(L|T) = ?$ ；在给出答案之前，我们要指出这两个问题并不是同一个问题，即一般而言

$$\mathbb{P}(T|L) \neq \mathbb{P}(L|T).$$

认为这两个概率值相同是人们在估计概率时非常容易犯的错误。

现在我们根据 Bayes 公式来计算

$$\begin{aligned}\mathbb{P}(L|T) &= \frac{\mathbb{P}(L)\mathbb{P}(T|L)}{\mathbb{P}(T)} \\ &= \frac{20\% \times 75\%}{47\%} \approx 0.32.\end{aligned}$$

这比 $\mathbb{P}(T|L) = 75\%$ 小得多。 □

例 3.13. (*常规赛表现与季后赛问题*) [44] 的三位作者在 2012 年某周的报纸上阅读到一位在线专栏的作家在哀叹（美国职业橄榄球比赛）该赛季达拉斯牛仔队常规赛 1:4 的糟糕开局的评论：“只有 4% 的季后赛球队在常规赛中以 1:4 开局，这意味着牛仔队只有 1/25 的概率进入季后赛”；作者察觉出了在线专栏作家犯的错误，你们察觉出来了吗？这正是上一例题中我们指出的那个人们在估计概率时非常容易犯的错误：用 A 代表“球队常规赛中前 5 局只赢 1 局”，用 M 代表“球队进入季后赛”，一般而言 $\mathbb{P}(M|A) \neq \mathbb{P}(A|M)$ 。[44] 的作者们告诉我们，那时总共有 32 支 NFL 球队，其中有 12 支进入季后赛，但没有再去查阅更多有关比赛的历史资料，从而按照等可能性假设得到一般情况下一支球队进入季后赛的概率： $\mathbb{P}(M) = \frac{12}{32} = \frac{3}{8}$ 。

为了估算牛仔队进季后赛的概率 $\mathbb{P}(M|A)$ ，三位作者首先按下面的方式估算了 $\mathbb{P}(A)$ ：简单假定每场比赛按照投掷硬币来决定输赢，那么一支球队在常规赛前五场比赛中只赢一场的概率为

$$\mathbb{P}(A) = C_5^1 \left(\frac{1}{2}\right)^5 = \frac{5}{32}.$$

我们已知 $\mathbb{P}(A|M) = 4\% = \frac{1}{25}$ ，因此牛仔队进季后赛的概率可以估计为

$$\mathbb{P}(M|A) = \frac{\mathbb{P}(M) \cdot \mathbb{P}(A|M)}{\mathbb{P}(A)} = \frac{\frac{3}{8} \cdot \frac{1}{25}}{5/32} = 9.6\%$$

这是三位作者对于专栏作家犯的的错误的一种纠正方案，作者们在书中吐槽道，“牛仔队并没有（专栏作家估计得）这么糟”，尽管“事实上，他们的确很糟糕...”。

例 3.14.（医学诊断问题） 本题来源于网络上的一篇博文。张某为了解自己患上了 X 疾病的可能性，去医院作常规血液检查。其结果居然为阳性，他赶忙到网上查询。根据网上的资料，血液检查实验是有误差的，这种实验有“1% 的假阳性率和 1% 的假阴性率”（真的患病者得到阴性结果称为假阴性，未患病的人得到阳性结果称为假阳性）。即在得病的人中做实验，有 1% 的概率是假阴性，99% 是真阳性。而在未得病的人中做实验，有 1% 的概率是假阳性，99% 是真阴性。于是张某根据这种解释，估计他自己得了 X 疾病的概率为 99%。张某的推理是，既然只有 1% 的假阳性率，那么，99% 都是真阳性，那我已被感染 X 病的概率便应该是 99%。张某咨询了医生，医生说：“99%？哪有那么大的感染几率啊。99% 是测试的准确性，不是你得病的概率。你忘了一件事：这种 X 疾病的正常比例是不大的，1000 个人中只有一个人有 X 病。” 张某不放心，又做了一个尿液检查，进一步检查他患上了 X 疾病的可能性，其结果仍然为阳性，尿液检查的实验有“5% 的假阳性率和 5% 的假阴性率”。

- 张某初始计算感染 X 病的概率是 99%，问题出在哪？
- 张某在血液检查之后感染 X 病的概率是多少？
- 张某在血液和尿液检查之后得 X 病的概率是多少？
- 假设根据张某的家族遗传信息，他得 X 病的概率是 1%，请问结合血液和尿液检查结果，张某得 X 病的概率是多少？

参考答案： a) 在这个例子中，张某由于没有认识到 X 疾病在人群中的患病率对自己患病率的影响，从而得出了错误的结论。换言之，虽然题中“真阳率 + 假阳率 = 100%”（这是题中数据的巧合，不是必然如此），不代表所有人都是患 X 病的，也不代表所有人去检测都会得到阳性的诊断。张某的错误在于没理解假阳性和假阴性的真实含义，产生了两点误会：其一是误以为总有“真阳率 + 假阳率 = 100%”（再次向读者强调，这仅仅是题中数据的巧合，不是必然如此）；其二是了解

真阳率 := $\mathbb{P}(\text{阳}|\text{患})$, 假阳率 := $\mathbb{P}(\text{阳}|\text{不患})$,

误以为“真阳率 = $\mathbb{P}(\text{患}|\text{阳})$, 假阳率 = $\mathbb{P}(\text{不患}|\text{阳})$ ”。更数学化地说，一般而言， $\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$ 。

b) 此处，使用 Bayes 公式计算，得到

$$\begin{aligned}\mathbb{P}_{\text{血}}(\text{患}) &:= \mathbb{P}(\text{患}|\text{血阳}) = \frac{\mathbb{P}(\text{患})\mathbb{P}(\text{血阳}|\text{患})}{\mathbb{P}(\text{患})\mathbb{P}(\text{血阳}|\text{患}) + \mathbb{P}(\text{不患})\mathbb{P}(\text{血阳}|\text{不患})} \\ &= \frac{\frac{1}{1000} * 99\%}{\frac{1}{1000} * 99\% + \frac{999}{1000} * 1\%} \approx 9\%.\end{aligned}$$

c) 在上一步血检阳性的条件下，张某

患病：不患病

的先验分布已经由原来的 $\frac{1}{1000} : \frac{999}{1000}$ 变成了后验分布 $9\% : 91\%$ 。在进一步做尿检为阳性的条件下，前述后验分布成为此处的先验分布；特别注意，由于血检和尿检是两种不同的检测手段，应有

$\mathbb{P}_{\text{血}}(\text{尿阳}|\text{患}) = \mathbb{P}(\text{尿阳}|\text{患}) = 95\%$, $\mathbb{P}_{\text{血}}(\text{尿阳}|\text{不患}) = \mathbb{P}(\text{尿阳}|\text{不患}) = 5\%$, 因而

$$\begin{aligned}\mathbb{P}_{\text{血}}(\text{患}|\text{尿阳}) &= \frac{\mathbb{P}_{\text{血}}(\text{患})\mathbb{P}_{\text{血}}(\text{尿阳}|\text{患})}{\mathbb{P}_{\text{血}}(\text{患})\mathbb{P}_{\text{血}}(\text{尿阳}|\text{患}) + \mathbb{P}_{\text{血}}(\text{不患})\mathbb{P}_{\text{血}}(\text{尿阳}|\text{不患})} \\ &= \frac{9\% * 95\%}{9\% * 95\% + 91\% * 5\%} \approx 65\%.\end{aligned}$$

b)、c) 两步的过程可以用图 3.2–3.3 表示。

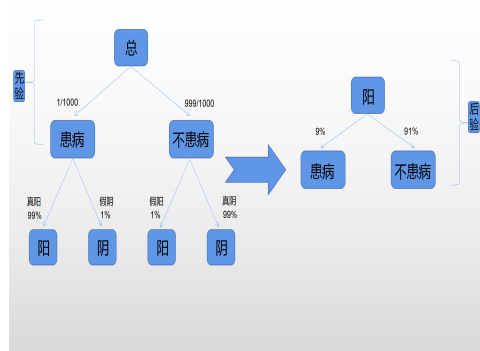


图 3.2: 血检

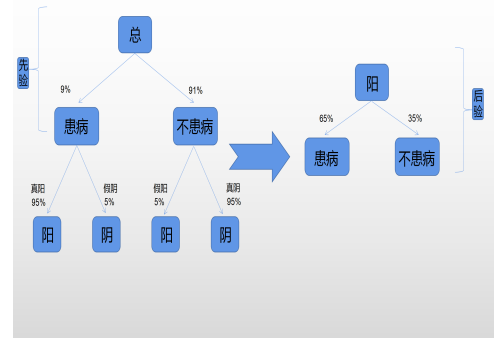


图 3.3: 基于血检后的尿检

d) 在此步骤中，由于检查了张某的家族遗传信息，对他的疾病诊断不应该再使用广大人群患病的先验分布 $\frac{1}{1000} : \frac{999}{1000}$ ，而应该使用 $\frac{1}{100} : \frac{99}{100}$ 作为先验分布，重复 b)、c) 两步的计算如下：

$$\begin{aligned}\mathbb{P}_{\text{血}}(\text{患}) &:= \mathbb{P}(\text{患}|\text{血阳}) = \frac{\mathbb{P}(\text{患})\mathbb{P}(\text{血阳}|\text{患})}{\mathbb{P}(\text{患})\mathbb{P}(\text{血阳}|\text{患}) + \mathbb{P}(\text{不患})\mathbb{P}(\text{血阳}|\text{不患})} \\ &= \frac{1\% * 99\%}{1\% * 99\% + 99\% * 1\%} = 50\%,\end{aligned}$$

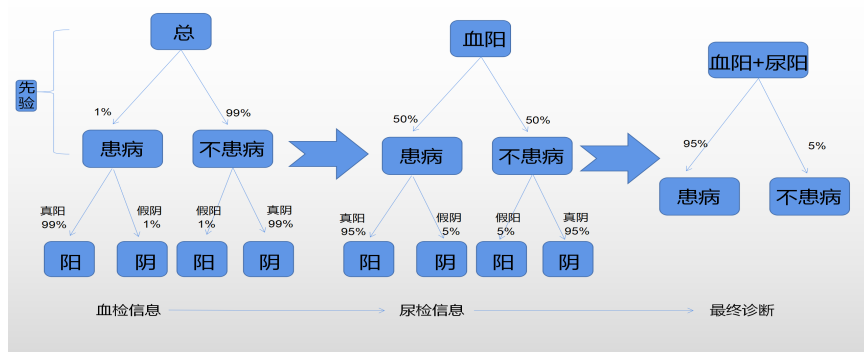


图 3.4: 基于遗传信息后的血检 + 尿检

$$\begin{aligned}
 \mathbb{P}_{\text{血}}(\text{患}|\text{尿阳}) &= \frac{\mathbb{P}_{\text{血}}(\text{患})\mathbb{P}_{\text{血}}(\text{尿阳}|\text{患})}{\mathbb{P}_{\text{血}}(\text{患})\mathbb{P}_{\text{血}}(\text{尿阳}|\text{患}) + \mathbb{P}_{\text{血}}(\text{不患})\mathbb{P}_{\text{血}}(\text{尿阳}|\text{不患})} \\
 &= \frac{50\% * 95\%}{50\% * 95\% + 50\% * 5\%} = 95\%.
 \end{aligned}$$

上述计算过程可以用图 3.4 表示。

□

注记 3.2. 我们再次分析一下上述医学诊断问题。容易知道

$$\begin{aligned}
 \mathbb{P}(\text{患}|\text{阳}) &= 1 / \left[1 + \frac{1 - \mathbb{P}(\text{患})}{\mathbb{P}(\text{患})} \cdot \frac{1 - \mathbb{P}(\text{阴}|\text{不患})}{\mathbb{P}(\text{阳}|\text{患})} \right] \\
 &= 1 / \left[1 + \frac{1 - \text{人群发病率}}{\text{人群发病率}} \cdot \frac{1 - \text{真阴率}}{\text{真阳率}} \right].
 \end{aligned} \tag{3.5}$$

也就是说，如果某疾病的人群发病率较低，那么必须检测手段的灵敏度、精确度极高才能确保最终诊断时阳性报告对应于高概率患病的诊断结论。

在公共卫生领域一般建议医生不要对普通人群进行罕见疾病的检查，因为此时即使检测结果呈阳性，极有可能是假阳性。例如，除非有很强的家族遗传史或其他原因，医生不建议 20 多岁或 30 多岁的女性进行乳房 X 光检查以检测乳腺癌。随着人们开始普遍认识到假阳性的问题，2004 年日本停止了对婴儿进行神经母细胞癌的全民检查*。

下面的例子来源于 Ross [34, Chapter 3, Example 3g]。

例 3.15.（是否作弊问题） 1965 年 5 月，在阿根廷的首都布宜诺斯艾利斯举行的世界桥牌锦标赛上，英国一对著名的桥牌手 Terrence Reese 和 Boris Schapiro 被指控作弊，说是他们用手指作暗号暗示他们红桃的张数；两位选手都否认这一指控。事后，英国桥牌协会举行了一个听证会，听证会按法律的程序进行，既包含控方，也包含辩方，双方都有目击证人。在接下来的调查过程中，控方检查了 Reese 和 Schapiro 打的几乎牌，并声称他们的打法与

*本段落引自 [44]，略有改动。

通过作弊已知了红桃张数的打法是吻合的。针对这一观点，辩方律师指出，他们的打法也同样与标准打法一致。然而，控方指出，既然他们的打法与其作弊的假设是一致的，那么就应该是支持这种假设。你如何理解控方的理由？

参考解答：这是一个新的证据（在本例中是“选手对特定情况下桥牌的打法”，我们记作 E ）如何影响某个特定假设（在本例中是“选手作弊”，我们记作 H ）成立的概率的问题。与上述注记类似， H 相当于那里的“患病”假设， E 相当于那里的“阳性报告”证据，因此

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(H)}{\mathbb{P}(H) + [1 - \mathbb{P}(H)] \cdot \frac{\mathbb{P}(E|H^c)}{\mathbb{P}(E|H)}}. \quad (3.6)$$

我们关心有了新证据后，何时可以推断 $\mathbb{P}(H|E) > \mathbb{P}(H)$ ，因为这表示新证据倾向于支持假设成立。根据上面方程，假定 $\mathbb{P}(H) \in (0, 1)$ ，很容易知道：

$$\mathbb{P}(H|E) > \mathbb{P}(H) \Leftrightarrow \mathbb{P}(E|H) > \mathbb{P}(E|H^c).$$

因此，在本例中，除非在“选手作弊”假设下这种打法的可能性大于“选手没作弊”假设下这种打法的可能性，“选手的打法支持他们是作弊的假设”就不应当认为成立。但显然，从控方与辩方关于证据的陈述中，可以知道， $\mathbb{P}(E|H) > \mathbb{P}(E|H^c)$ 不成立，因而控方关于新证据支持作弊的假设这一断言是无效的。□

我们定义事件 A 的**优势比**为 $\frac{\mathbb{P}(A)}{\mathbb{P}(A^c)}$ 。容易知道，在有新的证据 E 后，假设 H 的优势比也会发生如下的变化：

$$\frac{\mathbb{P}(H|E)}{\mathbb{P}(H^c|E)} = \frac{\mathbb{P}(H)}{\mathbb{P}(H^c)} \cdot \frac{\mathbb{P}(E|H^c)}{\mathbb{P}(E|H)}. \quad (3.7)$$

这意味着，如果证据支持假设，那么有了证据后的优势比就会大于原来的优势比。因此比值 $\frac{\mathbb{P}(E|H^c)}{\mathbb{P}(E|H)}$ 可以视作证据 E 对假设 H 的支持强度：比值越大，就认为证据支持假设的强度越大；当这个比值不超过 1 时，就应认为证据不支持假设。

3.2 乘积概率空间与事件独立性

古典概率论起源于古人关于赌博的数学建模探讨。一个公理化的概率模型（比如古典概率模型）可以想象为使用某种赌博道具一次的赌博的概率建模。但现实生活中，很多时候人们赌博未必只使用一种赌博道具，比如人们可以同时使用骰子和布克牌或麻将等道具进行赌博；即使使用同一种赌博道具，赌博中也未必只使用一次，比如我们有时会抛掷一枚硬币多次来进行一些复杂的随机决策。这意味着我们有必要探讨：如何对同时（或先后）使用两种赌博道具（通常这两种赌博道具之间没有关联性）进行的“复杂赌博”做概率建模？探讨的结果表明，我们应当使用概率空间的“乘积”的手段，也就是两个概率空间的**乘积概率空间**这一工具来进行概率建模。

3.2.1 乘积概率空间

我们假定 $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k), k = 1, 2$ 分别是对使用道具 1 一次的赌博和使用道具 2 一次的赌博的概率建模；我们假定这两种赌博道具的各自使用之间没有关联性：使用道具 1 得到的结果不会影响使用道具 2 得到的结果，反之亦然。我们的问题是：如何利用前述概率建模，实现对先后（或同时）使用道具 1、道具 2 各一次的“复杂赌博”进行概率建模？

简单起见，我们不妨假定两个概率模型 $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k), k = 1, 2$ 都是古典概率模型，都满足等可能性假设。那么为了实现对先后使用道具 1、道具 2 各一次的复杂结果的详尽刻画，很自然我们应当考虑乘积空间 $\Omega_1 \times \Omega_2$ ，其中的点 $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$ 就能描述清楚：使用道具 1 得到的“基本结果”是 ω_1 ，使用道具 2 得到的“基本结果”是 ω_2 ；并且我们认为 ω_1, ω_2 是两个自由变量或独立变量，这一点也切合“两种赌博道具之间没有关联性”的假定。由此可以取 $\Omega := \Omega_1 \times \Omega_2$ 作为先后使用道具 1、道具 2 各一次的“复杂赌博”的样本空间。这诱导我们去发展概率空间的乘积这一概念。

形式地，我们可以把先后使用道具 1、道具 2 各一次的“复杂赌博”建模为“乘积”概率空间

$$(\Omega, \mathcal{F}, \mathbb{P}) := (\Omega_1, \mathcal{F}_1, \mathbb{P}_1) \times (\Omega_2, \mathcal{F}_2, \mathbb{P}_2).$$

问题是，我们该如何在新的“样本空间” $\Omega = \Omega_1 \times \Omega_2$ 上合理地定义能够谈论概率的事件的全体 \mathcal{F} 以及对应的概率测度 $\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}$ ？

我们先来看乘积空间 $\Omega = \Omega_1 \times \Omega_2$ 中能够谈论概率的事件。对 $k = 1, 2$ ，设 $A_k \in \mathcal{F}_k$ 是概率空间 $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$ 中能够谈论概率的事件。很自然，“在使用道具 1、道具 2 各一次后 A_1 事件和 A_2 事件都发生了”这样的复杂事件应该是能够谈论概率的，而这一复杂事件在上述数学建模中对应于 $\Omega_1 \times \Omega_2$ 中事件 $A_1 \times A_2$ ，这样形态的事件在概率论或测度论中被称为可测矩形。可测矩形的全体是下面的集族：

$$\mathcal{F}_1 \times \mathcal{F}_2 := \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}. \quad (3.8)$$

容易知道，上述集族一般而言不是 σ -代数。因此我们可以取上述集族生成的 σ -代数

$$\mathcal{F}_1 \otimes \mathcal{F}_2 := \sigma(\mathcal{F}_1 \times \mathcal{F}_2) \quad (3.9)$$

作为 $\Omega = \Omega_1 \times \Omega_2$ 上的事件域 \mathcal{F} 。上述 $\mathcal{F}_1 \otimes \mathcal{F}_2$ 称为乘积 σ -代数，亦即我们有下面的乘积可测空间或乘积可测结构

$$(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2) = (\Omega_1, \mathcal{F}_1) \times (\Omega_2, \mathcal{F}_2). \quad (3.10)$$

现在我们来定义所谓的乘积概率测度 $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$ 。由于我们已经假定 $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k), k = 1, 2$ 都是古典概率模型，满足等可能性假设，很自然应该认为新的概率空间 $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$ 仍然满足等可能性，也是古典概率模型，也就是说，对任意 $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$

$$\begin{aligned} \mathbb{P}_1 \times \mathbb{P}_2(A_1 \times A_2) &= \frac{|A_1 \times A_2|}{|\Omega_1 \times \Omega_2|} = \frac{|A_1| \cdot |A_2|}{|\Omega_1| \cdot |\Omega_2|} \\ &= \frac{|A_1|}{|\Omega_1|} \cdot \frac{|A_2|}{|\Omega_2|} = \mathbb{P}_1(A_1) \cdot \mathbb{P}_2(A_2). \end{aligned}$$

因此，乘积概率测度 $\mathbb{P}_1 \times \mathbb{P}_2$ 在可测矩形 $A_1 \times A_2$ 上的合理赋值是

$$\mathbb{P}_1 \times \mathbb{P}_2(A_1 \times A_2) := \mathbb{P}_1(A_1) \cdot \mathbb{P}_2(A_2), \quad A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2. \quad (3.11)$$

很幸运的是，上述在全体可测矩形上的赋值方式就唯一确定了这个乘积概率测度 $\mathbb{P}_1 \times \mathbb{P}_2$ ；此处我们不探讨这一点，对此感兴趣的读者请移步第4章去学习和应用 Carathéodory 定理（见定理 4.2.2）来论证这一点；在附录 A 中我们给出了证明此处这一结果的证明。读者也可以去参考其他《测度论》教材的对应内容。

通过以上论述，我们在 $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k), k = 1, 2$ 都是古典概率模型的假定下给出了乘积概率空间的合理定义。由此外推至一般的概率空间 $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k), k = 1, 2$ ，我们定义乘积概率空间（有时也称为独立乘积概率空间、乘积概率结构）

$$(\Omega, \mathcal{F}, \mathbb{P}) := (\Omega_1, \mathcal{F}_1, \mathbb{P}_1) \times (\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$$

如下：样本空间取 $\Omega := \Omega_1 \times \Omega_2$ ；事件域取 $\mathcal{F} := \mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ ，其中 $\mathcal{F}_1 \times \mathcal{F}_2$ 的定义见(3.8)；概率测度取 $\mathbb{P} := \mathbb{P}_1 \times \mathbb{P}_2$ ，其中 $\mathbb{P}_1 \times \mathbb{P}_2$ 由(3.11)决定，有时也称 $\mathbb{P}_1 \times \mathbb{P}_2$ 为 \mathbb{P}_1 与 \mathbb{P}_2 的独立乘积。

以上探讨了两个概率空间的乘积空间的定义（期间也本质上给出了两个可测结构的乘积可测结构的定义）；原则上可以类似定义任意多个概率空间（或可测结构、测度结构）的乘积概率空间（或乘积可测结构、乘积测度结构），此处不再赘述；同样请对此感兴趣的同学移步学习《测度论》的有关理论。

注记 3.3. 在我们发展出了乘积概率空间的技巧后，原则上我们就可以有足够丰富的概率空间用以容纳讨论所关心的各种随机现象。因此几乎所有的概率论教材在介绍完公理化的概率空间后，除非确有必要指出样本空间、事件域、概率测度等的具体构造的场合，一般情形下我们总是默认概率空间已经给定、却又不指明，然后在此基础上直接进行有关随机事件、随机现象（包括后续将介绍的随机变量）的陈述，并通过逻辑推理与计算给出关心的概率问题的解答。

3.2.2 事件独立性的定义

在上一小节的讨论中，我们给出了乘积概率空间

$$(\Omega, \mathcal{F}, \mathbb{P}) := (\Omega_1, \mathcal{F}_1, \mathbb{P}_1) \times (\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$$

的合理定义。在那里，容易看到：概率空间 $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ 中的 A_1 事件（ $A_1 \in \mathcal{F}_1$ ）应该对应于概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中的 \tilde{A}_1 事件，其中 $\tilde{A}_1 := A_1 \times \Omega_2$ ；概率空间 $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ 中的 A_2 事件（ $A_2 \in \mathcal{F}_2$ ）应该对应于概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中的 \tilde{A}_2 事件，其中 $\tilde{A}_2 := \Omega_1 \times A_2$ 。此时

$$A_1 \times A_2 = \tilde{A}_1 \cap \tilde{A}_2,$$

即它代表 \tilde{A}_1 、 \tilde{A}_2 两事件同时发生这一复合事件。此时

$$\begin{aligned} \mathbb{P}(\tilde{A}_1 \cap \tilde{A}_2) &= \mathbb{P}_1 \times \mathbb{P}_2(A_1 \times A_2) \\ &= \mathbb{P}_1(A_1) \cdot \mathbb{P}_2(A_2) \\ &= \mathbb{P}_1 \times \mathbb{P}_2(A_1 \times \Omega_2) \cdot \mathbb{P}_1 \times \mathbb{P}_2(\Omega_1 \times A_2) \\ &= \mathbb{P}(\tilde{A}_1) \cdot \mathbb{P}(\tilde{A}_2). \end{aligned}$$

我们把两事件 \tilde{A}_1, \tilde{A}_2 之间的上述关系提炼为概率论中特有的事件独立性这一概念。

定义 3.2.1. 给定概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中两事件 $A, B \in \mathcal{F}$ 。称它们相互独立（或 A 与 B 独立），如果

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \quad (3.12)$$

更一般的，给定事件列 $\{A_k\}_{k=1}^N \subset \mathcal{F}$ （其中 $2 \leq N \leq \infty$ ），称 $\{A_k\}_{k=1}^N$ 相互独立，如果对任意 $n \in \mathbb{N}, 2 \leq n \leq N$ 及 $1 \leq i_1 < i_2 < \cdots < i_n \leq N$ ，总有

$$\mathbb{P}\left(\bigcap_{r=1}^n A_{i_r}\right) = \prod_{r=1}^n \mathbb{P}(A_{i_r}). \quad (3.13)$$

容易看到，对于 $A, B \in \mathcal{F}$ ，如果 $\mathbb{P}(A) > 0$ ，那么 A 与 B 独立当且仅当

$$\mathbb{P}(B|A) = \mathbb{P}(B). \quad (3.14)$$

亦即，两事件的相互独立性也等价于：“条件概率”就是“无条件概率”。

更一般的，我们可以定义子 σ -代数之间相互独立的概念。

定义 3.2.2. 设 $\mathcal{G}_1, \mathcal{G}_2$ 都是 \mathcal{F} 的子 σ -代数。称 \mathcal{G}_1 与 \mathcal{G}_2 相互独立，如果

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2), \forall A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2.$$

同样， $n \geq 2$ 个子 σ -代数 $\mathcal{G}_i \subset \mathcal{F}, i = 1, \dots, n$ 称为相互独立，如果

$$\mathbb{P}\left(\prod_{i=1}^n A_i\right) = \prod_{i=1}^n \mathbb{P}(A_i), \forall A_i \in \mathcal{G}_i, i = 1, \dots, n. \quad (3.15)$$

同样容易看到， A 与 B 相互独立，当且仅当 $\sigma(A)$ 与 $\sigma(B)$ 相互独立；此处 $\sigma(A)$ 代表由 A 生成的 σ -代数，即 $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$ 。

3.2.3 重复实验与条件实验

我们假定单个（随机）实验对应的概率模型是 $(\Omega, \mathcal{F}, \mathbb{P})$ 。此处我们关心两个事件 $A, B \in \mathcal{F}$ ；在重复实验或条件实验场景，我们将分别称它们为 A, B 现象。

所谓重复实验就是在同样的实验条件下，与过往实验独立/不关联地去反复做实验，并记录下各次实验的“基本结果”。在数学建模上，它对应于概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 的无穷多次独立乘积空间 $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}}) = (\Omega, \mathcal{F}, \mathbb{P})^\infty$ 。此时 $\tilde{\omega} = (\omega_1, \omega_2, \dots) \in \tilde{\Omega} = \Omega^\infty$ 是各次实验的实验记录结果的序列。

所谓 A -条件实验，则是重复实验中，实验结果首次出现 A -现象的那次实验；此处我们假定 $p := \mathbb{P}(A) > 0$ 。 A -条件实验的实际实验序号为

$$\tau_A = \tau_A(\tilde{\omega}) := \inf\{n \geq 1 : \omega_n \in A\},$$

满足 $\tilde{\mathbb{P}}(\tau_A = n) = pq^{n-1}$ ，其中 $q := 1 - p$ ，进而 $\tilde{\mathbb{P}}(\tau_A = \infty) = 0$ 。

例 3.16.（条件概率的另一种解释*）我们可以计算

$$I := \tilde{\mathbb{P}}(A\text{-条件实验中观察到 } B \text{ 现象})$$

*本例来源于浙江大学的赵敏智老师分享的资料，此处有所改写。特作此说明并致谢。

如下：

$$\begin{aligned}
 I &= \tilde{\mathbb{P}}(\omega_{\tau_A} \in B) = \sum_{n=1}^{\infty} \tilde{\mathbb{P}}(\tau_A = n, \omega_{\tau_A} \in B) \\
 &= \sum_{n=1}^{\infty} \tilde{\mathbb{P}}(\omega_j \notin A, j = 1, \dots, n-1, \omega_n \in A \cap B) \\
 &= \sum_{n=1}^{\infty} [\prod_{j=1}^{n-1} \tilde{\mathbb{P}}(\omega_j \notin A)] \cdot \tilde{\mathbb{P}}(\omega_n \in A \cap B) \\
 &= \sum_{n=1}^{\infty} q^{n-1} \cdot \mathbb{P}(A \cap B) = \frac{\mathbb{P}(A \cap B)}{1 - q} \\
 &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \mathbb{P}(B|A).
 \end{aligned}$$

这告诉我们， A 条件实验中观察到 B 现象的概率恰好是条件概率 $\mathbb{P}(B|A)$ 。

在第一章中，我们回顾概率论的历史的时候曾经说，材质均匀的“随机数发生器”的制造难度是直到 15 世纪前后才出现概率论的萌芽的部分原因。现在我们说，在概率论得到较成熟的发展后，基于上述条件概率的解释，即使是使用材质不均匀的“随机数发生器”，我们也有办法实现严格精确的可能性，见下面的例题。

例 3.17. 假设甲乙二人需要等概率地分配一张电影票的所有权，手边只有一枚不知道是否均匀的硬币。如何设计分配方案以做到公平合理？

参考解答：假设投掷硬币的记录用正/反来表达得到的是正面或反面；连续两次投掷的结果，如果第一次正面，第二次反面，就记作 (正, 反)；其他记号类似理解。显然有

$$\mathbb{P}((\text{正}, \text{反})) = \mathbb{P}((\text{反}, \text{正})),$$

进而

$$\mathbb{P}((\text{正}, \text{反}) | (\text{正}, \text{反}) \text{ 或 } (\text{反}, \text{正})) = \frac{1}{2}.$$

根据前述条件概率的另一种解释，我们两次、两次地投掷这枚硬币，直到出现 (正, 反) 或 (反, 正) 的结果为止。我们约定出现 (正, 反) 记录则把电影票分配给甲，出现 (反, 正) 记录则把电影票分配给乙。这就是一个公平的分配方案。□

在 Littlewood “无穷” 悖论问题（例 3.4）的解答中，我们认为零概率事件在实际生活中是不会发生的；如果某事件发生了，而按照我们的概率建模该事件是零概率事件，毫无疑问我们应该质疑我们的概率建模的正确性。人们在长期的实践中进一步总结出如下的小概率事件原理：概率很小的事件在一次随机实验中实际上几乎不会发生。反过来，如果在一个随机实验中有多种（互斥）现象 A_1, \dots, A_n 都可能发生，而在单次随机实验中发生了现象 A_i ，则认为该现象 A_i 应该是所有这些现象 $\{A_i\}_{i=1}^n$ 中发生概率最大者，这被称为极大似然原理：某事件 A 真实发生了，而我们的概率建模如果带有未知的参数，则该参数在实际应用中应当选择为使得事件 A 发生的概率达

到最大的那个参数值，这种估计方法被称为极大似然估计。在本讲义中，极大似然估计的具体案例参见第 5 章的例 5.10、例 5.11 及第 8 章例 8.16。

例 3.18. * 某接待站在某一周曾接待过 12 次来访；已知所有这 12 次接待都是在周二和周四进行的。问：是否能推断接待时间是有规定的？

参考解答：假设接待站的接待时间没有规定，且来访之间是相互独立的，从而各来访是等可能性的发生在一周的任何一天。于是，12 次接待来访都发生在周二、周四的概率为

$$\left(\frac{2}{7}\right)^{12} \approx 2.959 \times 10^{-7}.$$

这是一个非常小的概率，但这一小概率的事件在一次随机实验中竟然发生了，根据小概率事件原理有理由怀疑假设的正确性，从而推断接待站不是每天都接待来访者，即认为其接待时间是有规定的。□

下面的例子表明，发生的概率极小的事件/现象在大量的重复实验中会变成近乎必然事件。这对人口密集城市提出了安全出行、安全生产等方面的高水平管理需求。

例 3.19. † 春节燃放烟花爆竹是我国延续了一千多年的传统风俗，早已成为我国历史文化的一部分。但燃放烟花爆竹也常常引发意外，造成惨剧。假设每次燃放烟花爆竹引发火灾的概率是十万分之一。如果春节期间上海有 100 万人次燃放烟火爆竹，请问全市没有因此引发火警的概率是多少？

参考解答：取 $n = 10^6$ ， A_j 表示第 j 次燃放烟花爆竹没有引发火警，其中 $1 \leq j \leq n$ 。于是春节期间全市没有因 n 人次燃放烟火爆竹而引起火灾这一事件为

$$B := \bigcap_{j=1}^n A_j.$$

此处应认为事件列 $\{A_j\}_{j=1}^n$ 相互独立。按题目中约定， $\mathbb{P}(A_j) = 1 - 10^{-5}$ 。于是

$$\mathbb{P}(B) = (1 - 10^{-5})^n = (1 - 10^{-5})^{10^6} \approx 4.54 \times 10^{-5}.$$

也就是说，几乎不太可能不引发火警。□

注记 3.4. 有些文献区分主观概率和客观概率。通常认为，客观概率反映了事物的客观属性，它不因决策者的因素不同而不同，即与决策者的个别解释无关；客观概率的一个重要特点是通过大量重复试验，或在有限集合中事件中的个体数量与总体比例来定义概率。但是，并不是所有事件都能做重复实验（或者做重复实验的代价巨大）；在实际管理活动中，又往往要求人们对这些事件的可能性进行估计，并根据估计的可能性进行决策。这种要求就导致了主观概率概念的形成与应用。*L. J. Savage* 认为，主观概率 (*subjective probability*) 或个人概率 (*personal probability*) 是一种见解，是合理的信念的测度；这是某人对特定事件发生的可能性的信念（或意见、看法）的度量，即他认为的事件发生的可能性。主观概率反映事物的主观属性，它的确定允许与决策者的主观个别解释有关，它是由决策者的经验知识以及对客观情况

*本例来源于 [48, 例 1.3.8]。

†本例来源于何书元老师的《概率论》PPT 资料（或他编著的《概率论基础》），略有改动。

的了解，利用相关信息进行分析、推理、综合判断而设定的，不是主观臆测。主观概率和客观概率的定义反映了哲学上的两种不同的观点：即主观概率论主义者（贝叶斯主义者）和客观概率论主义者（非贝叶斯主义者）。客观概率主义者认为，概率是系统的固有的客观性质，是在相同条件下重复试验时频率的极限。

在编者看来，这种区分的意义不大，无非是有些情况下，大家有共同的认知（比如投掷均匀硬币，正面朝上的概率与反面朝上的概率是均等的），而另一些情况下，大家的认知差别巨大，从而本质上我们每一个认知主体所说的概率都是条件概率。在随机的世界里，我们对事物的看法、观点通常都是基于已经发生的事件、已经获得的信息。编者认为，这一定程度上可以解释为已介绍的经典的条件概率 $\mathbb{P}(\cdot|A)$ 以及后续将介绍的更一般的条件概率 $\mathbb{P}(\cdot|\mathcal{G})$ 。也就是说，每个逻辑自洽的理性人的概率空间实际上都是非常个性化的，是某个共有的概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 关于其个性化的子-代数 \mathcal{G} （代表他物理上已经获得、且心理上接受的信息）的条件概率空间；而很多时候理性人所做的判断或决策通常是基于其个性化的条件概率测度结合极大似然原理或 Bayes 思想等公认的原则。由于信息上的不对称，现实生活中在很多问题上人们意见经常相左、甚至发生冲突。这也是为什么加强信息交流、传播以获得更多共识在现代社会中越来越重要。

习 题 3

习题 3.1. 借用习题 2.22 中定义的几乎互不相交概念，进一步定义几乎完备事件组概念如下：如果 $\{A_n\}_{n=1}^N \subset \mathcal{F}$ 是几乎互不相交的事件列，并且 $\bigcap_{n=1}^N A_n$ 是几乎必然事件，则称 $\{A_n\}_{n=1}^N$ 是几乎完备事件组。请证明：对于几乎完备事件组，全概率公式和 Bayes 公式都成立。

习题 3.2. 证明：(1) 对于正实数列 $\{0 \leq x_n \leq 1\}_{n=1}^\infty$ ，我们有

$$\sum_{n=1}^{\infty} x_n = \infty \Rightarrow \prod_{n=1}^{\infty} (1 - x_n) = 0.$$

如果进一步有 $\inf\{1 - x_n : n \geq 1\} > 0$ ，那么

$$\sum_{n=1}^{\infty} x_n = \infty \Leftrightarrow \prod_{n=1}^{\infty} (1 - x_n) = 0.$$

(2) 利用上面的结果证明

$$\sum_{n=2}^{\infty} \mathbb{P}(A_n | \bigcap_{k=1}^{n-1} A_k^c) = \infty \Rightarrow \mathbb{P}(\bigcap_{k=1}^{\infty} A_k^c) = 0.$$

习题 3.3. 假设甲乙丙三人需要等概率地分配一张电影票的所有权，手边有一枚均匀的硬币。如何设计分配方案以做到公平合理。

习题 3.4. (赌徒破产模型) 设甲乙赌徒进行公平的赌博（即二人单局获胜概率相同；假定没有平局）。开局初甲乙两人筹码数量分别为 a, b ，每局的输家向赢家支付一个筹码，某方输光筹码则赌局结束。试求甲在这场赌博中的获胜概率，并说明当 $a/b \rightarrow 0$ 时甲赌徒输光的概率趋于 1。

【提示：两人总筹码数为 $K := a + b$ ，令 $f(x) := \mathbb{P}^x(\text{甲最终赢得所有筹码})$ 表

示甲初始筹码为 x 时在这场赌博中的获胜概率，通过考虑第一局的胜负情况，建立 $f(x)$ 的递推公式，并注意到 $f(0) = 0, f(K) = 1$ 而进行求解，得到 $f(a) = \frac{a}{K} = \frac{a}{a+b}$ 。】

习题 3.5. 设 A_1, A_2 是某实验 A 中不相容的两个现象。仿照例 3.16 说明，在 A -重复实验中现象 A_1 先于现象 A_2 发生的概率是

$$\frac{\mathbb{P}(A_1)}{\mathbb{P}(A_1) + \mathbb{P}(A_2)} = \mathbb{P}(A_1 | A_1 \uplus A_2).$$

习题 3.6. 设 E, H, G 均为正概率事件。证明：

$$\frac{\mathbb{P}(H|E)}{\mathbb{P}(G|E)} = \frac{\mathbb{P}(H)}{\mathbb{P}(G)} \cdot \frac{\mathbb{P}(E|H)}{\mathbb{P}(E|G)}.$$

假如在得到“新的证据” E 前，假设 H 成立的可能性是假设 G 成立的可能性的 3 倍；而当假设 G 成立时“新的证据” E 出现的可能性是假设 H 成立时的 2 倍。请问：当“新证据” E 出现时，哪种假设更可能成立？

习题 3.7. 本题来源于 [20]，试图用概率方法证明 Riemann 的 Zeta 函数基于欧拉 (Euler) 的素数分解定理的公式。Zeta 函数定义为

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

记 $\mathcal{P} := \{p : p \text{ 为素数}\}$ 及 $\mathcal{P}_n := \{p \leq n : p \text{ 为素数}\}$ 。给定 $s > 1$ ，定义 $\Omega := \mathbb{N}$ 上的概率

$$\mathbb{P}(\{n\}) := \frac{1}{\zeta(s) \cdot n^s}.$$

在 [34] 中，这个概率分布被称为 ζ -分布，它曾被意大利经济学家 V. Pareto 用于描述某个给定国家的家庭收入的分布；有时也称为 Zipf 分布，因为 G. K. Zipf 把这一分布运用到不同领域的更广泛问题中，从而推广了它的运用。请证明：

- (1) $\mathbb{P}(m\mathbb{N}) = \frac{1}{m^s}, \forall m \in \mathbb{N}$;
- (2) 事件列 $\{p\mathbb{N}\}_{p \in \mathcal{P}}$ 是相互独立的，从而事件列 $\{(p\mathbb{N})^c\}_{p \in \mathcal{P}}$ 也是相互独立的；
- (3) 利用容斥原理（或 Poincaré 恒等式）计算

$$\frac{1}{\zeta(s)} = \mathbb{P}(\{1\}) = \mathbb{P}\left(\bigcap_{p \in \mathcal{P}} (p\mathbb{N})^c\right) = \lim_{n \rightarrow \infty} \mathbb{P}\left(\bigcap_{p \in \mathcal{P}_n} (p\mathbb{N})^c\right),$$

由此完成证明

$$\zeta(s) = \prod_{p \in \mathcal{P}} \left[1 - p^{-s}\right]^{-1}, \quad \forall s > 1.$$

§ 4

Lebesgue 积分理论简介

为了后续内容的严谨性与逻辑自洽性，我们在本章简略介绍 Lebesgue 积分理论。欧氏空间上的 Lebesgue 积分理论是数学分析中的 Riemann 积分理论的推广，在现代数学理论中占有重要地位，它的应用也极其广泛；本章在介绍这一积分理论的同时，也顺带介绍一般测度空间的 Lebesgue 积分理论，因为它们本质上具有完全类似的形式逻辑。

本章的内容一部分来自《实变函数论》，一部分来自《测度论》，较大部分的内容没有提供证明。对概率论初学者，本章的有关理论论述只了解即可，初次学习时不必深究其具体论证。但 Carathéodory 扩张定理、 σ -有限预测度的扩张唯一性定理和 Lebesgue 积分的定义流程建议了解，Levi 单调收敛定理、Fatou 引理、Lebesgue 控制收敛定理和 Fubini 定理建议熟练掌握，Radon-Nikodym 定理建议了解并会使用。

4.1 可测集与可测函数

4.1.1 Borel 可测集与一般可测集

在一维欧氏空间 \mathbb{R} 中，我们总是使用所有开区间生成的拓扑，其中拓扑中的集合都称为开集，而所有开集生成的 σ -代数就称为（一维）Borel σ -代数，记作 \mathcal{B} ； \mathcal{B} 中的任一元素 $A \in \mathcal{B}$ 都称为是（一维）Borel 可测集。以下集族

$$\begin{aligned}\mathcal{E}_1 &:= \{(a, b) : a, b \in \mathbb{R}, a < b\}, \mathcal{E}_2 := \{(a, b] : a, b \in \mathbb{R}, a < b\}, \\ \mathcal{E}_3 &:= \{[a, b] : a, b \in \mathbb{R}, a < b\}, \mathcal{E}_4 := \{(a, b) : a, b \in \mathbb{Q}, a < b\}, \\ \mathcal{E}_5 &:= \{(a, b] : a, b \in \mathbb{Q}, a < b\}, \mathcal{E}_6 := \{[a, b] : a, b \in \mathbb{Q}, a < b\}\end{aligned}$$

生成的 σ -代数都是 \mathcal{B} 。这一结论在把上述集族 $\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6$ 的表示中的 \mathbb{Q} 替换为其他在 \mathbb{R} 中稠密的子集 $D \subset \mathbb{R}$ 时仍然成立。

在 $n \geq 2$ 维欧氏空间 \mathbb{R}^n 中，我们可以使用 \mathbb{R} 的拓扑的 n 重乘积拓扑以及乘积 σ -代数，即可以定义乘积可测结构 $(\mathbb{R}^n, \mathcal{B}^n) := (\mathbb{R}, \mathcal{B})^n$ （参见第 3 章中乘积 σ -代数的定义）*，此处的 n 重乘积 σ -代数 \mathcal{B}^n 就称为 n 维 Borel σ -代

*这里略有符号的滥用，即此处 $\mathcal{B}^n := \sigma(\mathcal{B} \times \cdots \times \mathcal{B})$ ，而不作 $\mathcal{B} \times \cdots \times \mathcal{B}$ 理解。

数, \mathcal{B}^n 中的任一元素 $A \in \mathcal{B}^n$ 都称为是 n 维 Borel 可测集。

一般的, 如果 (E, τ_E) 是一个拓扑空间*, τ_E 是 E 中所有开集的全体 (即所谓的拓扑结构), 那么 τ_E 生成的 σ -代数 $\mathcal{B}_E := \sigma(\tau_E)$ 也可以视作一种 Borel σ -代数, 其中的任意子集 $A \in \mathcal{B}_E$ 也称为 E 的 Borel 可测集。例如, $D \subset \mathbb{R}^n$ 是一个 Borel 可测集, 那么 D 上诱导的自然拓扑为

$$\tau_D := \{D \cap V : V \subset \mathbb{R}^n \text{ 是开集}\}.$$

此时容易知道, D 上的 Borel σ -代数为

$$\mathcal{B}_D = \sigma(\tau_D) = D \cap \mathcal{B}^n = \{D \cap A : A \in \mathcal{B}^n\}.$$

在欧氏空间或拓扑空间中, 如未指明可测结构, 则我们通常采纳 Borel 可测结构作为默认的可测结构。

更一般的, 设 (Ω, \mathcal{F}) 是一个给定的可测结构 (亦即 \mathcal{F} 是 Ω 上的 σ -代数), 则任意子集 $A \subset \Omega$ 称为可测集 (或 \mathcal{F} -可测集), 如果 $A \in \mathcal{F}$ 。

4.1.2 Borel 可测函数与一般的可测映射

我们首先给出欧氏空间上的 Borel 可测函数的概念。

定义 4.1.1. 任意给定一个非空的 Borel 可测集 $D \subset \mathbb{R}$ 。通常, 我们把 $f : D \rightarrow \mathbb{R}$ 称为 (一元) Borel 可测函数 (简称可测函数), 如果对任意 $x \in \mathbb{R}$

$$\{t \in D : f(t) \leq x\} \in \mathcal{B}_D.$$

这也等价于 $f^{-1}(A) \in \mathcal{B}_D, \forall A \in \mathcal{B}$; 此性质也可简写为 $f^{-1}(\mathcal{B}) \subset \mathcal{B}_D$ 。

对于整数 $n \geq 2$ 及任意给定的非空 Borel 可测集 $D \subset \mathbb{R}^n$, 我们把 $f : D \rightarrow \mathbb{R}$ 称为 (n 元) Borel 可测函数 (简称可测函数), 如果对任意 $x \in \mathbb{R}$

$$\{t \in D : f(t) \leq x\} \in \mathcal{B}_D.$$

这也等价于 $f^{-1}(A) \in \mathcal{B}_D, \forall A \in \mathcal{B}$; 此性质同样简写为 $f^{-1}(\mathcal{B}) \subset \mathcal{B}_D$ 。但要注意, 此处 f 已经是 n 元函数了, $f^{-1}(A) \subset \mathbb{R}^n$ 。

一般的, 我们还可以给出欧氏空间之间的 Borel 可测映射的概念, 但此时我们一般称为 Borel 可测 (向量值) 函数, 即仍然简称可测函数。

定义 4.1.2. 对于自然数 $m, n \geq 1$ 及任意给定一个非空的 Borel 可测集 $D \subset \mathbb{R}^n$, 我们把 $f : D \rightarrow \mathbb{R}^m$ 称为 (n 元 m 维向量值) Borel 可测函数 (简称可测函数), 如果

$$f^{-1}(A) \in \mathcal{B}_D, \forall A \in \mathcal{B}^m$$

上述性质也可简写为 $f^{-1}(\mathcal{B}^m) \subset \mathcal{B}_D$ 。

我们定义更一般的可测映射和可测函数如下。

定义 4.1.3. 设有两个给定的可测结构 $(\Omega_k, \mathcal{F}_k), k = 1, 2$ 以及它们之间的一个映射

$$T : \Omega_1 \rightarrow \Omega_2,$$

*空间 E 上拓扑 τ_E 是 E 上的集族, 满足: (i) $\emptyset, E \in \tau_E$; (ii) 任意并运算封闭: $A_\alpha \in \tau_E, \alpha \in I \Rightarrow \bigcup_{\alpha \in I} A_\alpha \in \tau_E$; (iii) 有限交运算封闭: $A, B \in \tau_E \Rightarrow A \cap B \in \tau_E$ 。 $A \in \tau_E$ 时, 称 A 为开集, 称 A^c 为闭集。

称 T 是可测映射（更确切的， $\mathcal{F}_1/\mathcal{F}_2$ -可测映射），如果

$$T^{-1}(A) \in \mathcal{F}_1, \forall A \in \mathcal{F}_2.$$

上述性质也可以简单记为 $T^{-1}(\mathcal{F}_2) \subset \mathcal{F}_1$ ，其中不难知道

$$T^{-1}(\mathcal{F}_2) := \{T^{-1}(A) : A \in \mathcal{F}_2\}$$

是一个 σ -代数。

在上述定义中，如果 $(\Omega_2, \mathcal{F}_2) = (\mathbb{R}, \mathcal{B})$ ，则称上述可测映射 $T : \Omega_1 \rightarrow \Omega_2$ 为（实）可测函数；如果 $(\Omega_2, \mathcal{F}_2) = (\mathbb{R}^m, \mathcal{B}^m)$ ，则称上述可测映射 $T : \Omega_1 \rightarrow \Omega_2$ 为（实）向量值可测函数。在不引起歧义的情况下，此二者都简称可测函数；为了指明定义域 Ω_1 中采纳的可测结构，有时更确切地称它们为 \mathcal{F}_1 -可测函数。

容易知道， \mathbb{R} 上的连续函数、单调函数都是 Borel 可测函数。一般的，设 (Ω, \mathcal{F}) 是一个可测空间，对任意 $A \subset \Omega$ ，示性函数 $1_A : \Omega \rightarrow \mathbb{R}$ 是一个可测函数，当且仅当 $A \in \mathcal{F}$ ；参见示性函数的定义(2.3)。

定义 4.1.4. 设 (Ω, \mathcal{F}) 是一个给定的可测结构。 Ω 上的可测函数 f 称为简单函数，如果存在 $\{a_i\}_{i=1}^N \subset \mathbb{R}$ 及 $\{A_i\}_{i=1}^N \subset \mathcal{F}$ ，使得

$$f = \sum_{i=1}^N a_i 1_{A_i}.$$

简单函数的全体记作 \mathcal{S} ；非负简单函数的全体记作 \mathcal{S}_+ 。

4.2 Lebesgue 测度与一般可测空间中的非负测度

4.2.1 测度、预测度、外测度的定义

欧氏空间中的 Lebesgue 测度是我们直观很容易理解的测度；在第 2 章中，可测集 $A \subset \mathbb{R}^n$ 的“体积”（ $n=1$ 时叫“长度”， $n=2$ 时叫“面积”）被记作 $|A|$ ，它就是我们通常说的可测集 A 的 Lebesgue 测度值。后来在公理化框架下我们定义了概率测度，在那里我们顺带提及了测度的概念。这里我们重新给出测度的更详细、准确的定义如下。

定义 4.2.1. 给定一个可测空间 (Ω, \mathcal{F}) ， $\mu : \mathcal{F} \rightarrow \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ 称为这个可测结构上的（非负）测度，如果它满足

- (1)（非负性 + 平凡性） $\mu(A) \geq 0, \forall A \in \mathcal{F}$ ，且 $\mu(\emptyset) = 0$ ；
- (2)（ σ 可加性/可列可加性）设 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ 互不相交，则

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

现在设 μ 为 (Ω, \mathcal{F}) 上测度，此时称 $(\Omega, \mathcal{F}, \mu)$ 是一个测度空间 或一个测度结构。如果 $\mu(\Omega) < \infty$ ，则称 μ 为 (Ω, \mathcal{F}) 上有限测度。如果存在 $\Omega_n \nearrow \Omega$ 使得 $\mu(\Omega_n) < \infty, \forall n$ ，则称 μ 为 (Ω, \mathcal{F}) 上 σ -有限测度。如果 Ω 是一个拓扑空间， \mathcal{B}_Ω 是它的 Borel σ -代数， μ 为 $(\Omega, \mathcal{B}_\Omega)$ 上测度，则称 μ 为 Borel 测度；进一步，如果对 Ω 中任意紧集 $K \subset \Omega$ ，总有 $\mu(K) < \infty$ ，则称 μ 为 Ω 上 Radon 测度。

例 4.1.（计数测度） 给定一个非空集合 Ω ，对它的任意子集 $A \subset \Omega$ ，可以考虑 A 的元素个数 $\#(A)$ ，则

$$\#(\cdot) : 2^\Omega \rightarrow \bar{\mathbb{R}}$$

是一个测度，称为计数测度。在古典概率模型中我们把 $\#(A)$ 记作了 $|A|$ 。

例 4.2.（体积测度） 在欧氏空间 \mathbb{R}^d 中，我们有“体积测度”的直观概念（ $d=1$ 时叫“长度”， $d=2$ 时叫“面积”， $d=3$ 时叫“体积”），对任意 Borel 子集 $A \subset \mathbb{R}^d$ ，我们后续将严格定义它在 \mathbb{R}^d 中的“体积”，记作 $\text{Vol}(A)$ 或 $\text{Vol}^{(d)}(A)$ ，有时也记作 $\text{Leb}(A)$ 、 $\text{Leb}^{(d)}(A)$ ；它满足：

$$\text{Vol}^{(d)}((a_1, b_1] \times \cdots \times (a_d, b_d]) = \prod_{i=1}^d (b_i - a_i), \text{ 其中 } a_i < b_i, i = 1, 2, \dots, d.$$

在几何概率模型中我们把 $\text{Vol}(A)$ 仍然记作了 $|A|$ 。

下面例子中奇怪的测度不是我们感兴趣的测度：

例 4.3.（奇异测度） 设 Ω 是非空集合， \mathcal{F} 是其上 σ -代数，定义 $\mu(\emptyset) := 0$ 及

$$\mu(A) := \infty, \forall A \in \mathcal{F}, A \neq \emptyset,$$

则 $\mu : \mathcal{F} \rightarrow \bar{\mathbb{R}}$ 是一个测度，称为奇异测度。

类似于概率测度的性质，不难知道 (Ω, \mathcal{F}) 上测度 μ 具有以下性质：

性质（1）（非负性） $\mu(A) \geq 0, \forall A \in \mathcal{F}$ ；

性质（2）（平凡性） $\mu(\emptyset) = 0$ ；

性质（3）（ σ -可加性） 设 $\{A_n\}_{n=1}^\infty \subset \mathcal{F}$ 互不相交，则

$$\mu\left(\biguplus_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty \mu(A_n).$$

性质（4）（有限可加性） 对任意给定的 $n \geq 1$ ，若 $\{A_k\}_{k=1}^n \subset \mathcal{F}$ 且互不相交，则

$$\mu\left(\biguplus_{k=1}^n A_k\right) = \sum_{k=1}^n \mu(A_k).$$

性质（5）（单调性） 若 $A, B \in \mathcal{F}$ 且 $A \subset B$ ，则 $\mu(A) \leq \mu(B)$ ；

性质（6）（次可列可加性） 若诸 $A_n \in \mathcal{F}$ ，则

$$\mu\left(\bigcup_{n=1}^\infty A_n\right) \leq \sum_{n=1}^\infty \mu(A_n);$$

性质（7）（下连续性） 若诸 $A_n \in \mathcal{F}$ 且 $A_n \uparrow A$ ，则 $A \in \mathcal{F}$ ，并且

$$\mu(A) = \mu\left(\lim_{n \rightarrow \infty} A_n\right) = \lim_{n \rightarrow \infty} \mu(A_n); \quad (4.1)$$

性质（8）（上连续性） 若诸 $A_n \in \mathcal{F}$ 且 $A_n \downarrow A$ ，并且存在 $n_0 < \infty$ 使得 $\mu(A_{n_0}) < \infty$ ，则 $A \in \mathcal{F}$ ，并且

$$\mu(A) = \mu\left(\lim_{n \rightarrow \infty} A_n\right) = \lim_{n \rightarrow \infty} \mu(A_n); \quad (4.2)$$

特别的, μ 在 \emptyset 处上连续: 若诸 $A_n \in \mathcal{F}$ 且 $A_n \downarrow \emptyset$, 并且存在 $n_0 < \infty$ 使得 $\mu(A_0) < \infty$, 则

$$\lim_{n \rightarrow \infty} \mu(A_n) = 0.$$

这里, 我们指出: 测度这个概念是物理学中一类特殊的物理量—**广延量** (比如质量、体积、内能等物理量) 的数学抽象, 其中测度的 σ -可加性就是这类特殊物理量的广延性的数学抽象描述。

如果函数 $\mu: \mathcal{F} \rightarrow \mathbb{R}$ 仅具有性质 (2) 与性质 (3), 则称 μ 是 (Ω, \mathcal{F}) 上的**符号测度**。称此符号测度 μ 是有限的, 如果对于任意 $A \in \mathcal{F}$, $|\mu(A)| < \infty$; 类似可以定义 σ -有限的符号测度。可测集 $A \in \mathcal{F}$ 称为 **μ -正集 (负集)**, 如果其任意可测子集的 μ 测度均为非负 (非正)。

设 μ 为 \mathcal{F} 上符号测度。易知, 可加性蕴含: 或者 $\mu(A) < \infty, \forall A \in \mathcal{F}$; 或者 $\mu(A) > -\infty, \forall A \in \mathcal{F}$ 。对任意 $A \in \mathcal{F}$, 定义

$$\mu^+(A) := \sup\{\mu(B) : B \subset A, B \in \mathcal{F}\}, \quad \mu^- := (-\mu)^+.$$

从而 μ^+, μ^- 均为非负的集函数 (因为 $\mu(\emptyset) = 0$)。下面的定理说明, μ^+, μ^- 实际上是非负测度, 即符号测度总是两个非负测度的差。

定理 4.2.1. (Hahn 分解与 Jordan 分解) 设 (Ω, \mathcal{F}) 为可测空间, μ 为 \mathcal{F} 上一个符号测度。那么

- (1) 正集的可列并是正集;
- (2) $A \in \mathcal{F}$ 是正集 $\Leftrightarrow \mu^-(A) = 0$; $A \in \mathcal{F}$ 是负集 $\Leftrightarrow \mu^+(A) = 0$;
- (3) μ^+, μ^- 具有单调性;
- (4) 存在 $H \in \mathcal{F}$ 使得: H 为正集, 而 H^c 为负集; 集合 H 称为 **Hahn 集**。空间分解 $\Omega = H \cup H^c$ 就称为 **Hahn 分解**;
- (5) $\mu^+(A) = \mu(A \cap H), \mu^-(A) = -\mu(A \cap H^c), \forall A \in \mathcal{F}$ 。因而 μ^+, μ^- 都是测度 (分别称为符合测度 μ 的**正部**和**负部**), 且其中至少有一个是有限测度;
- (6) $\mu = \mu^+ - \mu^-$, 这称为 **Jordan 分解**。如果存在测度 μ_1, μ_2 使得 $\mu = \mu_1 - \mu_2$, 那么 $\mu^+ \leq \mu_1, \mu^- \leq \mu_2$; 记 $|\mu| := \mu^+ + \mu^-$, 它称为是 μ 的**全变差测度**, $|\mu|(\Omega)$ 称为 μ (在 Ω 上) 的**全变差**;
- (7) **Hahn 集**通常不唯一, 但在只相差一个 $|\mu|$ -零测集的意义下唯一。

概率论的初学者对上述定理的陈述略作了解即可; 下面提供的证明只是出于全书封闭性的考量。

定理 4.2.1 的证明: (1)–(3) 的证明留作习题。以下我们证明 (4)–(6)。

不妨设 $\mu(A) < \infty, \forall A \in \mathcal{F}$, 且 μ 不是零测度。令 $b := \sup\{\mu(A) : A \in \mathcal{F}\}$; 则 $0 \leq b \leq \infty$ 。我们先证明下面的断言 1:

断言 1: 对任意 $\beta \in [0, b)$, 存在正集 $A \in \mathcal{F}$ 使得 $\mu(A) \geq \beta$ 。

$b = 0$ 是平凡情况; 以下设 $b > 0$ 。由 b 的定义, 存在 $B \in \mathcal{F}$, $\mu(B) > \beta$ 。若 B 是正集, 则取 $A = B$ 。否则 $0 < \mu^-(B) \leq \infty$; 进而, 对任意 $a_1 > 0$ 满足 $\frac{\mu^-(B)}{2} \wedge 1 < a_1 < \mu^-(B)$, 存在 $E_1 \subset B$, $\mu(E_1) < -a_1$, 从而对 $B_1 := B \setminus E_1$,

$\mu(B_1) > \beta_1 := \beta + a_1$ 。若 B_1 为正集，则记 $A = B_1$ 。否则，对任意 $a_2 > 0$ 满足 $\frac{\mu^-(B_1)}{2} \wedge 1 < a_2 < \mu^-(B_1)$ ，存在 $E_2 \subset B_1$ ， $\mu(E_2) < -a_2$ ，从而对 $B_2 := B_1 \setminus E_2 = B \setminus (E_1 \dot{\cup} E_2)$ ， $\mu(B_2) > \beta_2 := \beta + a_1 + a_2$ 。如此进行下去，若有限步内总是不能得到所需的正集，则可得到 B 中一系列互不相交的可测集 $\{E_n\}_{n=1}^\infty$ 和一系列正实数 $\{a_n\}_{n=1}^\infty$ 以及 $B_n := B \setminus (\bigcup_{k=1}^n E_k)$ ，满足：
 $\mu^-(B_n) > 0, \mu(E_n) < -a_n, \forall n \geq 1$ ，其中

$$0 < \min\left(\frac{\mu^-(B_n)}{2}, 1\right) < a_n < \mu^-(B_n).$$

令 $E := \bigcup_{n=1}^\infty E_n, A := B \setminus E$ 。若 $\mu^-(A) > 0$ ，注意到 $A \subset B_n, \forall n \geq 1$ ，有 $\mu^-(B_n) \geq \mu^-(A)$ ，从而 $a_n \geq a := \min\left(\frac{\mu^-(A)}{2}, 1\right) > 0$ ，进而

$$\mu(E) = \sum_{n=1}^\infty \mu(E_n) \leq \sum_{n=1}^\infty (-a_n) = -\infty,$$

即 $0 \leq \beta < \mu(B) = \mu(A) + \mu(E) = -\infty$ ，矛盾。因此断言 1 成立。

我们进一步证明下面的断言 2：

断言 2： $b < \infty$ 且 $\exists H \in \mathcal{F}, \mu(H) = b$ ，且 H 可取为 Hahn 集。

由断言 1，我们可取可测正集列 $\{A_n\}_{n=1}^\infty$ ，使得 $\lim_{n \rightarrow \infty} \mu(A_n) = b$ 。现在令

$H := \bigcup_{n=1}^\infty A_n$ 。 H 显然为正集，且 $\mu(H) \geq b$ ，故 $b = \mu(H) < \infty$ 。另外，对 H^c 的任何可测子集 A ，有 $b \geq \mu(H \cup A) = b + \mu(A)$ ，即 $\mu(A) \leq 0$ ，从而 H^c 是负集。于是 H 是 Hahn 集，且满足断言 2。

由 Hahn 集 H 的存在性，我们知道

$$\begin{aligned} \mu^+(A) &:= \sup\{\mu(B) : B \in \mathcal{F}, B \subset A\} \\ &= \sup\{\mu(B \cap H) + \mu(B \cap H^c) : B \in \mathcal{F}, B \subset A\} \\ &\leq \sup\{\mu(B \cap H) : B \in \mathcal{F}, B \subset A\} = \mu(A \cap H). \end{aligned}$$

但显然应有

$$\mu^+(A) := \sup\{\mu(B) : B \in \mathcal{F}, B \subset A\} \geq \mu(A \cap H).$$

因此， $\mu^+(A) = \mu(A \cap H)$ 。同理 $\mu^-(A) = -\mu(A \cap H^c)$ 。

现在定理中的结论 (5)–(6) 很容易验证了，留作习题。

我们现在回过头来看 Hahn 集的唯一性问题。设 H_1, H_2 均为 Hahn 集，由我们的假定，有 $0 \leq \mu(H_1) = \mu(H_2) < \infty$ 。另外，

$$\mu(H_1 \setminus H_2) = \mu(H_1 \cap H_2^c) = \mu^+(H_2^c) = -\mu^-(H_1),$$

从而 $\mu(H_1 \setminus H_2) = 0$ ，同理 $\mu(H_2 \setminus H_1) = 0$ 。注意到

$$\begin{aligned} |\mu|(H_1 \Delta H_2) &= \mu^+(H_1 \Delta H_2) + \mu^-(H_1 \Delta H_2) \\ &= \mu(H_1 \cap (H_1 \Delta H_2)) - \mu(H_1^c \cap (H_1 \Delta H_2)) \\ &= \mu(H_1 \setminus H_2) - \mu(H_2 \setminus H_1) = 0, \end{aligned}$$

可见 Hahn 集虽然一般不唯一，但在相差一个 $|\mu|$ -零测集的意义下唯一。 \square



图 4.1: H. Hahn(1879-1934)



图 4.2: C. Jordan(1838-1922)

注 4.1. *Hans Hahn* (哈恩, 1879/9/27–1934/7/24; 奥地利/匈牙利) 生于维也纳, 卒于维也纳。他的父亲是当时的电话局高级官员。他 1898 年成为维也纳大学的学生, 起初学习法律, 1899 年转学数学, 之后游学于斯特拉斯堡大学、慕尼黑大学和哥廷根大学; 1902 年在维也纳获得博士学位。上面提及的 *Hahn* 集就是以他的名字命名的。他的著名工作有泛函分析中的 *Hahn-Banach* 定理、一致有界原理 (共鸣定理; 也被称为 *Banach-Steinhaus* 定理), 抽象代数中的 *Hahn* 嵌入定理, 测度论中的 *Hahn-Kolmogorov* 定理 (有时也称为 *Carathéodory-Fréchet* 扩张定理、*Carathéodory-Hopf* 扩张定理、*Hopf* 扩张定理、*Hahn-Kolmogorov* 扩张定理)、*Vitali-Hahn-Saks* 定理, 拓扑方面的 *Hahn-Mazurkiewicz* 定理等。

Hahn 培养的知名学生有: *K. Menger* (1902/1/13–1985/10/5; 奥地利-美国数学家, 以 *Menger* 定理知名, *Sierpinski* 方块/海绵实际是他发现的, 应称为 *Menger* 方块/海绵), *W. Hurewicz* (1904/6/29–1956/9/6; 波兰数学家, 以高阶同伦群、纤维的长正则同伦序列和 *Hurewicz* 定理知名), *K. Gödel* (哥德尔, 1906/4/28–1978/1/14; 奥地利/匈牙利-美国数学家、逻辑学家、分析哲学家, 以 *Gödel* 不完备定理、*Gödel* 完备定理等成果闻名于世) 等。

Camille Jordan (约当, 1838/1/5–1922/1/22; 法国) 生于里昂, 卒于巴黎; 大学毕业于巴黎综合理工学院。他以 *Jordan* 曲线定理、*Jordan* 标准型、*Jordan* 矩阵等知名, 他在把 *Galois* 理论推进到数学的主流上做了不少贡献; 他写的教材《*Cours d'analyse*》影响很大。上面提及的 *Jordan* 分解也以他的姓氏冠名; 1920 年的斯特拉斯堡国际数学家大会上他是特邀报告人。他的姓名容易与其他两位也姓 *Jordan* 的德国学者混淆: *Wihelm Jordan* (1842/3/1–1899/4/17; 德国几何学家, 以 *Gauss-Jordan* 消去算法知名)、*Pascual Jordan* (1902/10/18–1980/7/31; 德国物理学家, 在量子力学、量子场论、矩阵力学等方面有贡献, *Jordan* 代数、*Jordan-Brans-Dicke* 理论、*Jordan and Einstein* 标架、*Jordan* 映射、*Jordan-Wigner* 变换等是以他的姓氏命名的)。

我们还需要给出预测度的概念。

定义 4.2.2. 给定非空集 Ω 及其上一个非空的子集族 \mathcal{E} (未必是 σ -代数), 其中 $\emptyset \in \mathcal{E}$ 。 $\mu: \mathcal{E} \rightarrow \mathbb{R}$ 称为 \mathcal{E} 上的 (非负) 预测度, 如果它满足

(1) (非负性 + 平凡性) $\mu(A) \geq 0, \forall A \in \mathcal{E}$, 且 $\mu(\emptyset) = 0$;

(2) (σ 可加性/可列可加性) 设 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{E}$ 互不相交, 并且 $\biguplus_{n=1}^{\infty} A_n \in \mathcal{E}$,

则

$$\mu\left(\biguplus_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

我们进一步给出外测度的概念。

定义 4.2.3. 给定非空集 Ω 。 $\mu : 2^\Omega \rightarrow \bar{\mathbb{R}}$ 称为 Ω 上的外测度，如果它满足

(1) (非负性 + 平凡性) $\mu(A) \geq 0, \forall A \subset \Omega$ ，并且 $\mu(\emptyset) = 0$;

(2) (次可列可加性) 设 $\{A_n\}_{n=1}^\infty$ 为 Ω 中集列，那么

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n).$$

(3) (单调性) 若 $A \subset B \subset \Omega$ ，则 $\mu(A) \leq \mu(B)$.

4.2.2 Carathéodory 扩张与 Lebesgue 测度

现在我们可以来定义 \mathbb{R} 中的 Lebesgue 测度了。取

$$\mathcal{E} := \{(a, b] : -\infty \leq a \leq b \leq +\infty\}.$$

定义 $\mu : \mathcal{E} \rightarrow \bar{\mathbb{R}}$ 如下：

$$\mu((a, b]) = b - a, \text{ 其中 } -\infty \leq a \leq b \leq +\infty. \quad (4.3)$$

显然 μ 是 \mathcal{E} 上的预测度。我们可以如下诱导外测度 $\mu^* : 2^\mathbb{R} \rightarrow \bar{\mathbb{R}}$ ：

$$\mu^*(A) := \inf\left\{\sum_{n=1}^{\infty} \mu(A_n) : A \subset \bigcup_{n=1}^{\infty} A_n, A_n \in \mathcal{E}, n = 1, 2, \dots\right\}.$$

接着，定义集族

$$\mathcal{L} := \{A \subset \mathbb{R} : \mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c), \forall E \subset \mathbb{R}\}. \quad (4.4)$$

在实变函数论中，已经证明了上述集族 \mathcal{L} 是一个 σ -代数，并且 $\mathcal{E} \subset \mathcal{L}$ 。 \mathcal{L} 中元素称为 Lebesgue 可测集。由 $\mathcal{E} \subset \mathcal{L}$ 立即知道 $\mathcal{B} = \sigma(\mathcal{E}) \subset \mathcal{L}$ 。在此基础上，实变函数论中也证明了

$$\mu^* : \mathcal{L} \rightarrow \bar{\mathbb{R}}$$

是一个测度，称为 \mathbb{R} 上 Lebesgue 测度。进而 $\text{Leb} := \mu^*|_{\mathcal{B}}$ 也是一个测度，仍然称为 \mathbb{R} 上 Lebesgue 测度，它可以认为是预测度 $\mu : \mathcal{E} \rightarrow \bar{\mathbb{R}}$ 的自然延拓。这种自然延拓的方法被称为 Carathéodory 扩张。

\mathbb{R}^n 上的 Lebesgue 测度被定义为 \mathbb{R} 上 Lebesgue 测度 Leb 的 n 重乘积测度 $\text{Leb}^{(n)} := \text{Leb} \times \dots \times \text{Leb}$ ；为方便，在不产生歧义时， \mathbb{R}^n 上的 Lebesgue 测度仍然记作 Leb 。在第 2 章的几何概率模型中，所谓可测集 A 的体积测度，就是它的 Lebesgue 测度，即

$$|A| := \text{Leb}(A), A \in \mathcal{B}^n.$$

给定一个单调递增、右连续函数 $F : \mathbb{R} \rightarrow \mathbb{R}$ ，把(4.3)替换为

$$\mu_F((a, b]) = F(b) - F(a), \text{ 其中 } -\infty \leq a \leq b \leq +\infty, \quad (4.5)$$

并同样使用 Carathéodory 扩张的程序，可以得到一个新的测度 $\mu_F : \mathcal{B} \rightarrow \bar{\mathbb{R}}$ ，它称为 F 诱导的分布测度。

上述 Carathéodory 扩张的流程在抽象空间中仍然可以使用。为此我们先介绍以下几个集合系/集合类的概念。

定义 4.2.4. \mathcal{E} 称为 π -系 (π -system), 如果 $A, B \in \mathcal{E}$ 蕴含了 $A \cap B \in \mathcal{E}$ 。在此基础上, π -系 \mathcal{E} 进一步称为半环 (Semi-ring)*, 如果 $A, B \in \mathcal{E}$ 蕴含了: 存在 \mathcal{E} 中两两不交的有限集合列 $\{A_k \in \mathcal{E}\}_{k=1}^n$, 使得 $A \setminus B = \biguplus_{k=1}^n A_k$ 。

注意: 对半环 \mathcal{E} , $A, B \in \mathcal{E}$ 时, 未必有 $A \setminus B \in \mathcal{E}$! 即, 对于半环, 集合的差运算未必封闭。

如果半环定义中的后一条修改成“集合的差运算封闭”, 再补充上“集合的并运算封闭”就得到了环的概念, 陈述如下:

定义 4.2.5. \mathcal{E} 称为环 (Ring), 如果它满足:

- (i) $A, B \in \mathcal{E}$, 则 $A \cap B \in \mathcal{E}$;
- (ii) $A, B \in \mathcal{E}$, 则 $A \cup B \in \mathcal{E}$;
- (iii) $A, B \in \mathcal{E}$, 则 $A \setminus B \in \mathcal{E}$ 。

定义 4.2.6. \mathcal{E} 称为代数 (Algebra) 或域 (Field), 如果它满足:

- (i) $\Omega \in \mathcal{E}$;
- (ii) 如果 $A \in \mathcal{E}$, 则 $A^c \in \mathcal{E}$;
- (iii) 如果 $A, B \in \mathcal{E}$, 则 $A \cup B \in \mathcal{E}$ 。

显然代数对集合并、交、补、差运算有限封闭。此外上述 (i)–(iii) 也可换为

- (i') $\Omega \in \mathcal{E}$;
- (ii') 如果 $A, B \in \mathcal{E}$, 则 $A \setminus B \in \mathcal{E}$ 。

事实上, $A^c = \Omega \setminus A$, $A \cup B = (A^c \cap B^c)^c = \Omega \setminus [(\Omega \setminus A) \setminus B]$ 。

容易知道, 代数就是包含了全空间 Ω 的环; 因此有些文献也称环为预代数 (pre-algebra)。

例 4.4. $\Omega = \mathbb{R}$, $\mathcal{E}_1 := \{(a, b] : -\infty < a \leq b < \infty\}$,

$\mathcal{E}_2 := \{E = \bigcup_{k=1}^n A_k : n \in \mathbb{N}, \{A_k\}_{k=1}^n \subset \mathcal{E}_1 \text{ 是两两不交的集合列}\}$ 。

则 \mathcal{E}_1 是 π -系、半环, 但不是环; \mathcal{E}_2 是环, 但不是代数。 \mathcal{E}_2 恰好是 \mathcal{E}_1 生成的环: $\mathcal{E}_2 = R(\mathcal{E}_1)$ 。

在很多场合, 我们所关心的测度只是在某个特定的半环上有良好的定义, 比如我们之前讲的 \mathbb{R} 上的 Lebesgue 测度就是如此。但由半环生成的环结构更好 (对集合的有限次的并、交、补运算都封闭), 并且其上的测度取值理论上也可借助半环上的测度取值简单计算出来。因此为了表述的简洁, 在研究测度时, 以下我们总是考虑环, 而不再提半环。

*在 [13] 中有比半环定义略苛刻的 Semialgebra 的定义, 其中后一条件替换成了: 集合系中元素的补集可以表达成集合系内部有限个元素的不交并; 这蕴含了集合系中存在全空间的有限覆盖的要求。因此, 本书例 4.4 中集合类不满足 [13] 中的 Semialgebra 的要求。

定义 4.2.7. 设 \mathcal{E} 为 Ω 上的环。设

$$\mu: \mathcal{E} \rightarrow \bar{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{\infty\} = [0, \infty].$$

是 \mathcal{E} 上的预测度。如果 $\bar{\mu}$ 是 $\mathcal{F} := \sigma(\mathcal{E})$ 上测度，且 $\bar{\mu}|_{\mathcal{E}} = \mu$ ，则称 $\bar{\mu}$ 为 μ 的一个扩张或延拓。

类似于 σ -有限测度的定义，环 \mathcal{E} 上预测度 μ 称为 σ -有限的，如果存在 $A_n \in \mathcal{E}, n \geq 1$ 使得 $\Omega = \bigcup_{n \geq 1} A_n$ 且 $\mu(A_n) < \infty$ ；有限预测度类似定义。

很自然要问：在环上定义的预测度，能否延拓成环生成的 σ -代数上的一个测度？进一步，如果能延拓，得到的测度是否是唯一的？以下的 Carathéodory 定理先肯定的回答了延拓存在性的问题。后续的定理则告诉我们，当预测度具有 σ -有限性时，延拓也是唯一的。对这两个定理的证明感兴趣的读者请参考有关测度论的专著。

给定空间 Ω 上一个环 \mathcal{E} 及其上一个预测度 μ ，定义 μ^* 如下： $\forall E \in 2^\Omega$

$$\mu^*(E) := \inf \left\{ \sum_n \mu(A_n) : A_n \in \mathcal{E}, n \geq 1, E \subset \bigcup_n A_n \right\}. \quad (4.6)$$

（约定 $\inf \emptyset = \infty$ 。）显然 μ^* 具有性质（5）中陈述的单调性，并且容易验证它具有性质（4）中陈述的次可列可加性，因此 μ^* 是 Ω 上一个外测度。此时显然有 $\mu^*|_{\mathcal{E}} = \mu$ 。对任意 $A \subset \Omega$ ，称 A 为 μ^* -可测，如果对于任意 $E \subset \Omega$ ，总有

$$\mu^*(E) = \mu^*(A \cap E) + \mu^*(A^c \cap E).$$

记 \mathcal{F}_* 为 Ω 上 μ^* -可测集的全体。

定理 4.2.2.（Carathéodory 扩张定理）给定空间 Ω 上一个环 \mathcal{E} 及其上一个预测度 μ ，外测度 μ^* 及集族 \mathcal{F}_* 如上定义。则

- (1) $(\Omega, \mathcal{F}_*, \mu^*)$ 是一个测度空间；
- (2) $\mathcal{F}_* \supset \mathcal{F} = \sigma(\mathcal{E}) \supset \mathcal{E}$ ， $\mu^*|_{\mathcal{E}} = \mu$ ，从而 $\mu^*|_{\mathcal{F}}$ 是 μ 的一个扩张，称之为 Carathéodory 扩张；
- (3) \mathcal{F}_* 包含了所有 μ^* -零测集，因而 $(\Omega, \mathcal{F}_*, \mu^*)$ 是一个完备的测度空间。



图 4.3: C. Carathéodory (1873-1950)

注 4.2. 希腊数学家 *Constantin Carathéodory*（卡拉西奥多里，1873/9/13–1950/2/2）生于柏林，卒于慕尼黑；他原籍希腊，祖先数代前移居土耳其埃迪尔内（*Edirne*），他的父亲是土耳其驻圣彼得堡、柏林等地的外交官。1891–1895 年 *Carathéodory* 在比利时的军事学院学习，毕业后受雇于英国政府到埃及参加艾斯尤特（*Asyut*）水坝的建设。1900 年返回柏林研究数学。1902 年到哥廷根，在 *H. Minkowski*（闵可夫斯基，1864/6/22–1909/1/12）指导下于 1904 年取得博士学位。后在哥廷根、波恩、汉诺威、布雷斯劳、柏林等地任教。1920 年被希腊政府召回士麦那筹建大学。1922 年士麦那被土耳其人焚毁，他领导学校将图书馆移至雅典，并在雅典大学任教。1924 年应邀到慕尼黑大学接替 *C. L. F. von Lindemann*（1852–1939）任教授。他曾是《数学年刊》杂志的编辑。

Carathéodory 在数学上有多方面的贡献。他发展了变分法，把光滑曲线的理论推广到有角曲线上，特别提出解曲线场的概念。他重新研究变分法与一阶偏微分方程的关系，并应用于解拉格朗日问题。在函数论方面，研究函数值分布论，简化了在单位圆上单连通域的保形变换的主要定理，给出了边界对应的理论。在测度论方面，进行了公理化研究，所提出的测度扩张方法被大学教科书普遍采用。此外，对热力学公理化和狭义相对论也有贡献。

Carathéodory 培养的知名学生有：*H. Rademacher*（1892/4/3–1969/2/7，德国出生的美国数学家，研究方向为数学分析和数论），*P. Finsler*（1894/4/11–1970/4/29；德国-瑞士数学家，以 *Finsler* 空间、*Hadwiger-Finsler* 不等式等知名），*H. Boerner*（1906/7/11–1982/6/3；德国数学家，研究方向为变分、复分析、群表示理论），*E. Peschl*（1906/9/1–1986/6/9；德国数学家，研究方向为几何复分析、偏微分方程、多变量复分析等），*G. Aumann*（1906/11/11–1980/8/4；德国数学家，以 *General topology*、*Contact relations* 方向的研究而知名），*W. Seidel*（1907/12/21–1981/1/12；俄罗斯出生的德国-美国数学家，以 *Seidel* 类闻名），*N. Terzioğlu*（1912/?/?–1976/9/20；土耳其数学家）等。

定理 4.2.2 的证明. 上面的定理本质上源于 *Lebesgue* 测度的构造，其证明除了需要把 *Lebesgue* 测度替换为一般测度以外，所有细节也毫无二致。但为了读者的便利，我们还是给出它的证明。

(1) 我们需要证明 \mathcal{F}_* 是一个 σ -代数，且 μ^* 是其上的测度。显然， $\emptyset, \Omega \in \mathcal{F}_*$ ，且 \mathcal{F}_* 对补集运算封闭。故只需进一步验证 \mathcal{F}_* 对可列并运算封闭。

事实上， \mathcal{F}_* 对有限并运算封闭，因为若 $A, B \in \mathcal{F}_*$ ，则对任意 $E \subset \Omega$

$$\begin{aligned} \mu^*(E) &= \mu^*(A \cap E) + \mu^*(A^c \cap E) \\ &= \mu^*(A \cap E) + [\mu^*(B \cap (A^c \cap E)) + \mu^*(B^c \cap (A^c \cap E))] \\ &= [\mu^*(A \cap (A \cup B) \cap E) + \mu^*(A^c \cap (A \cup B) \cap E)] + \mu^*((A \cup B)^c \cap E) \\ &= \mu^*((A \cup B) \cap E) + \mu^*((A \cup B)^c \cap E), \end{aligned}$$

由此 $A \cup B \in \mathcal{F}_*$ 。

下面只需再验证 \mathcal{F}_* 对不相交集列的可列并运算封闭。设 $\{A_n\}_{n=1}^\infty$ 是 \mathcal{F}_* 中互不相交的集列。令

$$B_n := \bigcup_{k=1}^n A_k, \quad A := \bigcup_{k=1}^\infty A_k,$$

由外测度的单调性，

$$\begin{aligned} \mu^*(E) &= \mu^*(E \cap B_n) + \mu^*(E \cap B_n^c) \\ &\geq \mu^*(E \cap B_n) + \mu^*(E \cap A^c) \\ &= \mu^*(E \cap B_n \cap B_{n-1}^c) + \mu^*(E \cap B_n \cap B_{n-1}^c) + \mu^*(E \cap A^c) \\ &= \mu^*(E \cap B_{n-1}) + \mu^*(E \cap A_n) + \mu^*(E \cap A^c) \\ &= \cdots = \sum_{k=1}^n \mu^*(E \cap A_k) + \mu^*(E \cap A^c). \end{aligned}$$

由 n 的任意性以及 μ^* 的次可列可加性

$$\mu^*(E) \geq \sum_{k=1}^{\infty} \mu^*(E \cap A_k) + \mu^*(E \cap A^c) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c).$$

因此 $A \in \mathcal{F}_*$ 。同时，从上述证明过程也看到， μ^* 在 \mathcal{F}_* 上具有有限可加性。结合 μ^* 的次可列可加性，立即导出 μ^* 在 \mathcal{F}_* 上具有可列可加性。由此， μ^* 是 (Ω, \mathcal{F}_*) 上测度。

(2) 设 $A \in \mathcal{E}$ 。 $\forall E \subset \Omega$ ，往证

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c),$$

即 $A \in \mathcal{F}_*$ 。

如果 $\mu^*(E) = \infty$ ，由次可加性，

$$\mu^*(E) \leq \mu^*(E \cap A) + \mu^*(E \cap A^c),$$

立即得到 $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$ 。

因此，以下我们设 $\mu^*(E) < \infty$ 。于是对 $\forall \varepsilon > 0$ ，存在 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{E}$ 满足

$$E \subset \bigcup_{n=1}^{\infty} A_n, \quad \sum_{n=1}^{\infty} \mu^*(A_n) \leq \mu^*(E) + \varepsilon.$$

因而

$$\begin{aligned} \mu^*(E) &\leq \mu^*(E \cap A) + \mu^*(E \cap A^c) \\ &\leq \mu^*\left(\left(\bigcup_{n \geq 1} A_n\right) \cap A\right) + \mu^*\left(\left(\bigcup_{n \geq 1} A_n\right) \cap A^c\right) \\ &= \mu^*\left(\bigcup_{n \geq 1} (A_n \cap A)\right) + \mu^*\left(\bigcup_{n \geq 1} (A_n \cap A^c)\right) \\ &\leq \sum_{n \geq 1} [\mu^*(A_n \cap A) + \mu^*(A_n \cap A^c)] \\ &\leq \sum_{n \geq 1} [\mu(A_n \cap A) + \mu(A_n \cap A^c)] \\ &= \sum_{n \geq 1} \mu(A_n) \leq \mu^*(E) + \varepsilon. \end{aligned}$$

由 $\varepsilon > 0$ 的任意性， $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$ ，即 $A \in \mathcal{F}_*$ 。因此 $\mathcal{E} \subset \mathcal{F}_*$ ，进而由于 \mathcal{F}_* 是 σ -代数，由单调类定理（见定理 A.2.2）， $\mathcal{E} \subset \sigma(\mathcal{E}) = \mathcal{F} \subset \mathcal{F}_*$ 。

(3) 设 $B \in \mathcal{F}_*$ 为零测集，即 $\mu^*(B) = 0$ 。设 $A \subset B$ ，则

$$0 \leq \mu^*(A) \leq \mu^*(B) = 0,$$

即 $\mu^*(A) = 0$ 。 $\forall E \subset \Omega$ ，有 $E \cap A \subset A \subset B$ ，从而 $\mu^*(E \cap A) = 0$ 。进而

$$\mu^*(E \cap A^c) \leq \mu^*(E) \leq \mu^*(E \cap A) + \mu^*(E \cap A^c) = \mu^*(E \cap A^c),$$

因此 $\mu^*(E) = \mu^*(E \cap A^c) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$ 。进而 $A \in \mathcal{F}_*$ 。由此可见 $(\Omega, \mathcal{F}_*, \mu^*)$ 是一个完备的测度空间。 \square

需要指出的是，Carathéodory 扩张在一般情形下并不总是唯一的扩张。

例 4.5. 取 \mathbb{R} 上环

$$\mathcal{E} := \left\{ \bigcup_{i=1}^n (a_i, b_i] : n \geq 1, a_i \leq b_i, 1 \leq i \leq n \right\}.$$

则 $\mathcal{B} = \sigma(\mathcal{E})$ 。对任意 $A \in \mathcal{E}$ ，定义

$$\mu(A) := \begin{cases} 0, & A = \emptyset, \\ \infty, & A \neq \emptyset. \end{cases}$$

则 μ 的 *Carathéodory* 扩张是 $(\mathbb{R}, \mathcal{B})$ 上奇异测度，对任意 $A \in \mathcal{F}$ ， $\mu(A)$ 取值情况同上。但是计数测度 $\#(\cdot)$ 无疑也是 μ 的一个扩张，它不同于 *Carathéodory* 扩张。本例中， μ 在环 \mathcal{E} 上不是 σ -有限的。

定理 4.2.3. (测度扩张的唯一性) 给定 Ω 上一个环 \mathcal{E} 及其上一个预测度 μ 。如果这个预测度是 σ -有限的，那么它在 $\mathcal{F} = \sigma(\mathcal{E})$ 上的扩张是唯一的。

证明. 设 μ 有两个扩张 μ_1, μ_2 。任取 $\Omega_0 \in \mathcal{E}$ 使得 $\mu(\Omega_0) < \infty$ ，注意到习题 2.11，我们断言：

$$\mu_1(A) = \mu_2(A), \forall A \in \Omega_0 \cap \mathcal{F}.$$

事实上，记 $\Omega_0 \cap \mathcal{F} = \{\Omega_0 \cap A : A \in \mathcal{F}\}$ ，它可以视为 Ω_0 上的 σ -代数；令

$$\mathcal{A}_0 := \{A \in \Omega_0 \cap \mathcal{F} : \mu_1(A) = \mu_2(A)\}.$$

容易验证：(i) $\Omega_0 \cap \mathcal{E} \subset \mathcal{A}_0 \subset \Omega_0 \cap \mathcal{F}$ ；(ii) \mathcal{A}_0 是 Ω_0 上 Dynkin 系。由于 $\Omega_0 \cap \mathcal{E}$ 是 Ω_0 上环，由单调类定理，

$$\mathcal{A}_0 \supset \lambda_{\Omega_0}(\Omega_0 \cap \mathcal{E}) = \sigma_{\Omega_0}(\Omega_0 \cap \mathcal{E}) = \Omega_0 \cap \sigma_{\Omega}(\mathcal{E}) = \Omega_0 \cap \mathcal{F}.$$

因此 $\mathcal{A}_0 = \Omega_0 \cap \mathcal{F}$ ，即断言成立。

由预测度 μ 的 σ -有限性，存在单调上升集合列 $\{\Omega_n : n \geq 1\} \subset \mathcal{E}$ 满足 $\Omega_n \nearrow \Omega$ 且 $\mu(\Omega_n) < \infty$ 。于是对任意 $A \in \mathcal{F}$ ， $\Omega_n \cap A \in \Omega_n \cap \mathcal{F}$ ，

$$\mu_1(A) = \lim_{n \rightarrow \infty} \mu_1(\Omega_n \cap A) = \lim_{n \rightarrow \infty} \mu_2(\Omega_n \cap A) = \mu_2(A).$$

从而 $\mu_1 = \mu_2$ 在 \mathcal{F} 上成立。 \square

4.3 欧氏空间中的 Lebesgue 积分

4.3.1 Lebesgue 积分的定义

现在假定 $E \subset \mathbb{R}^d$ 是 Borel 可测集。 $f : E \rightarrow \mathbb{R}$ 是 Borel 可测函数。当 f 满足“适当条件”（后续将明确何为“适当条件”）时，我们将给出积分 $\int_E f d\text{Leb}$ 的定义；传统上，这个积分将被更简单地记作

$$\int_E f dx := \int_E f d\text{Leb},$$

并把它理解为

$$\int_E f dx = \int 1_E \cdot f dx.$$

Lebesgue 本人对他的这一积分思想做过一个生动有趣的描述：“我必须偿还一笔钱。如果我从口袋中随意地摸出来各种不同面值的钞票，逐一地还给债

主直到全部还清，这就是 **Riemann** 积分；不过我还有另外一种做法，就是把钱全部拿出来并把相同面值的钞票放一起，然后再一起付出应还的数目，这就是我的积分” [54]。

\mathbb{R}^d 上 **Borel** 可测函数关于其上的 **Lebesgue** 测度 **Leb** 的 **Lebesgue** 积分定义的出发点是，把 **Borel** 可测集 A 的测度值（体积） $\text{Leb}(A) = |A|$ 视作可测函数 1_A 关于测度 **Leb** 的积分：

$$\int 1_A d\text{Leb} = \int 1_A dx := \text{Leb}(A) = |A|. \quad (4.7)$$

在此基础上，我们意图保持积分的线性性质；聪明的读者应该看到了此处与泛函中的 **Hahn-Banach** 延拓方法上的相似之处。

对任意非负简单函数 $f \in \mathcal{S}_+$ ，存在有限集 $\{a_k\}_{k=1}^n \subset \bar{\mathbb{R}}_+$ 及互不相交的可测集 $A_1, \dots, A_n \in \mathcal{F}$ ，使得 $\mathbb{R}^d = \bigcup_{k=1}^n A_k$ 且

$$f = \sum_{k=1}^n a_k 1_{A_k}.$$

此时我们定义

$$\int f dx := \sum_{k=1}^n a_k \cdot |A_k|,$$

称之为 f 的 **Lebesgue** 积分。此处约定 $0 \cdot \infty = 0$ 。不难验证，映射

$$\mathcal{S}_+ \ni f \mapsto \int f dx$$

是单调且线性的。

对于一般的非负可测函数 $f: \mathbb{R}^d \rightarrow \bar{\mathbb{R}}_+$ ，定义

$$\int f dx := \sup \left\{ \int g dx : 0 \leq g \leq f, g \in \mathcal{S}_+ \right\},$$

称之为 f 的 **Lebesgue** 积分。另一种等价、但相对容易理解的处理是使用函数逼近的思想：对任意非负可测函数 $f: \mathbb{R}^d \rightarrow \bar{\mathbb{R}}_+$ ，可以找到非负简单函数 $f_n \uparrow f$ (比如可以取 $f_n := \frac{\lfloor 2^n \cdot f \rfloor}{2^n} \wedge n$)，定义

$$\int f dx := \lim_{n \rightarrow \infty} \int f_n dx;$$

但在此过程中需要论证极限与序列 f_n 的选取无关，以保证定义是良性的。容易知道，在非负可测函数类上， $f \mapsto \int f dx$ 仍然具有单调性和线性性质。

对于一般的可测函数 $f: \mathbb{R}^d \rightarrow \bar{\mathbb{R}}$ ，注意到 $f = f^+ - f^-$ ，其中 f^+, f^- 分别为 f 的正部和负部，当 $\int f^+ dx, \int f^- dx$ 两者中至少有一个取有限值时，称 f 的 **Lebesgue** 积分有意义（或有定义），记作

$$\int f dx := \int f^+ dx - \int f^- dx.$$

当 $\int f^+ dx, \int f^- dx$ 两者均取有限值时，称 f 是 **Lebesgue** 可积的，记作

$f \in L^1(\mathbb{R}^d)$ 或更简单的 $f \in L^1$ ；这也等价于 $\int |f| dx < \infty$ 。

另外， f 在 $E \in \mathcal{F}$ 上的 Lebesgue 积分记为

$$\int_E f dx := \int 1_E \cdot f dx.$$

4.3.2 Lebesgue 积分与极限的交换

本小节我们探讨欧氏空间上的 Lebesgue 积分与极限的交换问题，即探讨

$$\lim_{n \rightarrow \infty} \int f_n dx = \int \lim_{n \rightarrow \infty} f_n dx$$

成立的（充分）条件，最终得到 Levi 单调收敛定理、Fatou 引理、Lebesgue 控制收敛定理三个重要结论。它们配合本节中最后一小节的 Fubini 定理基本上就能解决大多数实际问题中提出的积分与极限的交换问题（含两重积分的交换问题）。

从积分的定义出发不难验证积分的单调性：如果 $f \leq g$ 且两个函数的积分都存在，则

$$\int f dx \leq \int g dx.$$

定理 4.3.1.（单调收敛定理/Levi 定理）设 $\{f_n\}_{n=1}^\infty$ 为一列非负可测函数，且存在可测函数 f 使得 $f_n \nearrow f$ 。那么

$$\lim_{n \rightarrow \infty} \int f_n dx = \int f dx.$$

证明. 由单调性，显然数列 $\{\int f_n dx\}_{n=1}^\infty$ 有（广义）非负极限，且

$$\int f dx \geq \lim_{n \rightarrow \infty} \int f_n dx.$$

不妨设 $\int f dx < \infty$ 。现在任取非负简单函数 $g \leq f$ 以及正实数 $\lambda \in (0, 1)$ ，令 $A_n := \{x \in \mathbb{R}^d : f_n(x) \geq \lambda \cdot g(x)\}$ 。注意到 $g \leq f, f_n \leq f, f_n \nearrow f$ ，显然有 $A_n \nearrow \mathbb{R}^d$ 。于是

$$\int f_n dx \geq \int f_n \cdot 1_{A_n} dx \geq \lambda \int g \cdot 1_{A_n} dx.$$

由于 g 是非负简单函数，

$$\lim_{n \rightarrow \infty} \int f_n dx \geq \lim_{n \rightarrow \infty} \lambda \int g \cdot 1_{A_n} dx = \lambda \int g dx;$$

最后的等号利用了积分在 \mathcal{S}_+ 上的线性性质及测度 Leb 的下连续性。由 λ 的任意性，

$$\lim_{n \rightarrow \infty} \int f_n dx \geq \int g dx.$$

再由 $\int f dx$ 的定义立得

$$\lim_{n \rightarrow \infty} \int f_n dx \geq \int f dx.$$

$\int f dx = \infty$ 情形的证明可以类似实现，细节留给读者。 □

利用上面的单调收敛定理，容易证明积分具有线性性：如果 f, g 均可积，则对任意实数 α, β ,

$$\int [\alpha f + \beta g] dx = \alpha \int f dx + \beta \int g dx.$$



图 4.4: Levi(1875–1961)

注 4.3. 上面定理中的 *Levi* 是意大利数学家 *Beppo Levi*（列维，1875/5/14–1961/8/28），而不是法国数学家 *Paul Lévy*（莱维，1886/9/15–1971/12/15），也不是意大利数学家 *E. E. Levi*（莱维，1883/10/8–1917/10/28）或另一个意大利数学家、物理学家 *Tullio Levi-Civita*（列维-奇维塔，1873/3/29–1941/12/20）。*Beppo Levi* 以其单调收敛定理闻名，他出生在 *Turin*，21 岁获得博士学位，几年后在 *Plaisance* 大学任教授，后加入 *Cagliari* 大学。由于他的犹太人身份，1938 年被墨索里尼政府废除公职。之后他移民到阿根廷，在当地建立数学系，创办数学杂志。

上面的单调收敛定理说明了：在被积函数列非负且单调上升这一特殊情况下，积分与极限的交换。如果被积函数列仅仅有非负性，而没有单调性质时，下面的 *Fatou* 引理就非常有用。

定理 4.3.2. (*Fatou* 引理) 设 $\{f_n\}_{n=1}^{\infty}$ 为一列非负可测函数，那么

$$\liminf_{n \rightarrow \infty} \int f_n dx \geq \int \liminf_{n \rightarrow \infty} f_n dx.$$

证明. 令 $g_n := \inf\{f_k : k \geq n\}$ 。于是 $\{g_n\}_{n=1}^{\infty}$ 是一个单调递增的非负可测函数列，且 $g_n \leq f_n$ 。于是由单调收敛定理

$$\int (\liminf_{n \rightarrow \infty} f_n) dx = \int (\lim_{n \rightarrow \infty} g_n) dx = \lim_{n \rightarrow \infty} \int g_n dx \leq \liminf_{n \rightarrow \infty} \int f_n dx.$$

□

下面的 *Lebesgue* 控制收敛定理是用来处理函数极限与积分交换关系的重要工具，需要重点掌握，学会应用。

定理 4.3.3. (*Lebesgue* 控制收敛定理) 设 $\{f_n\}_{n=1}^{\infty}$ 为一列几乎处处（或依测度）收敛于 f 的可测函数，且存在可测函数 $g \in L^1$ 使得 $|f_n| \leq g$ 。那么

$$\lim_{n \rightarrow \infty} \int f_n dx = \int \lim_{n \rightarrow \infty} f_n dx.$$

证明. 记 $f = \lim_{n \rightarrow \infty} f_n$ 。因 $\{g + f_n\}_{n=1}^{\infty}$ 与 $\{g - f_n\}_{n=1}^{\infty}$ 都是非负可测函数列，对前者利用 Fatou 引理及 g 的可积性

$$\int (g+f)dx = \int \left[\lim_{n \rightarrow \infty} (g+f_n) \right] dx \leq \lim_{n \rightarrow \infty} \int (g+f_n)dx = \int gdx + \lim_{n \rightarrow \infty} \int f_n dx.$$

$$\text{由此 } \int f dx \leq \lim_{n \rightarrow \infty} \int f_n dx.$$

同理对 $\{g - f_n\}_{n=1}^{\infty}$ 利用 Fatou 引理可得 $\lim_{n \rightarrow \infty} \int f_n dx \leq \int f dx$ 。由此

$$\lim_{n \rightarrow \infty} \int f_n dx = \int f dx. \quad \square$$

注 4.4. 上面定理中出现的 *Fatou* 和 *Lebesgue* 都是法国数学家。

微积分从它诞生时起，主要是处理光滑曲线和可导函数的。但 1872 年 7 月 18 日德国大数学家 *Weierstrass* 在柏林科学院的一次讲演中给出了一族处处连续、处处不可导的函数

$$f(x) = \sum_{n=0}^{\infty} b^n \cos(a^n \pi x).$$

Direchlet 同样举出了在无理点取值 0、有理点取值 1 的极端病态函数；众多的病态函数破坏了 *Riemann* 意义下的可积性。

Henri Léon Lebesgue (勒贝格, 1875/6/28–1941/7/26) 生于博韦 (*Beauvais*), 卒于巴黎。他在 1894 年考入巴黎高等师范学校, 1897 年大学毕业后在该校图书馆工作了两年。他是法国数学家 *E. Borel* (波莱尔, 1871/1/7–1956/2/3) 的学生。*Lebesgue* 从 1899 年到 1902 年在 *Nancy* 的一所中学任教, 虽然工作繁忙, 但仍孜孜不倦地研究有关积分的理论, 并于 1902 年发表了博士论文“积分、长度、面积” (*Intégrale, longueur, aire*), 同年在 *Sorbonne* 巴黎大学理学院通过博士学位。在这篇文章中, *Lebesgue* 创立了后来以他的名字命名的积分理论, 它能很好的讨论病态函数的积分问题。*Lebesgue* 积分一出现就用来研究三角级数, 由此人们可以得到函数可展为三角级数的更弱的充分条件, 接着导数的概念也得到推广, 一门微积分的延续学科—实变函数论在 *Lebesgue* 手中诞生了。

Lebesgue 的工作一开始并未得到人们的一致赞成, 反对病态函数的人到处都是; 当时很多人认为病态函数是“一种变态的不健康的函数”、“一种学究式的数学游戏”。法国著名数学家 *Poincaré* (庞加莱, 1854/4/29–1912/7/17) 认为, “逻辑有时产生怪物, 过去人们为了一个实际的目的而创造一个新的函数; 今天人们为了说明先辈在推理方面的不足而故意造出这些函数来, 而从这些函数所能推出来的也就是仅此而已。”当时, 批评病态函数最严厉的是法国数学家 *C. Hermite* (厄尔米特, 1822/12/24–1901/1/14), 他在一封信中说“我怀着惊恐的心情对不可导函数的令人痛惜的祸害感到厌恶” [60]。直到 1910 年, *Lebesgue* 才被同意进入巴黎大学理学院工作, 1921 年起才进入法兰西学院任教授, 次年进入巴黎科学院。这时 *Lebesgue* 已经 47 岁, 距离他 27 岁发表博士学位论文已经足足 20 年。

到上世纪 30 年代, *Lebesgue* 积分已经成熟, 并在概率论、谱理论、泛函空间等方面获得广泛应用。时至今日, 连工程师也不得不接触病态函数、谈论抽象积分。历史是一面镜子, 科学体系中内部矛盾所提出的理论问题, 有时会成为这门学科的生长点。但理论问题也不是越抽象越好, *Lebesgue* 就曾经这样告诫自己: “搞出过于一般的理论, 数学将会变成没有内容的漂亮的形式, 而这将会很快死亡”。

Lebesgue 培养的知名学生有: *P. Montel* (1876/4/29–1975/1/22; 法国数学家, 专长为复分析中的全纯函数, 以 *Montel* 定理、*Montel* 空间等闻名), *Z. Janiszewski* (1888/6/12–1920/1/3; 波兰数学家, 以 *Janiszewski* 定理和 *Brouwer–Janiszewski–Knaster* 连续统闻名), *G. de Rham* (1903/9/10–1990/10/9; 瑞士数学家, 因对微分拓扑方向的贡献而闻名于世) 等。

Pierre Joseph Louis Fatou (法图, 1878/2/28–1929/8/9) 生于洛里昂 (*Lorient*), 卒于波尔尼谢 (*Pornichet*)。他在 1898 年就学于巴黎高等师范学校, 1907 年获博士学位, 指导老师是 *P. Painlevé* (1863/12/5–1933/10/29; 法国数学家、政治家, 以微分方程理论中的 *Painlevé* 性质等闻名); 之后长期在巴黎天文台供职。对 *Taylor* 级数、亚纯函数、*Lebesgue* 积分等方面的问题有论述, 得到 *Lebesgue* 积分的一些基本结果; 在复动力系统中, *Fatou* 集 (*Julia* 集的余集) 就是以他的名字命名。在天文学上用概率方法对恒星与行星的位置测定、双星测量、天文仪器常数等进行过研究。



图 4.5: Lebesgue(1875–1941)



图 4.6: Fatou(1878–1929)

4.3.3 Lebesgue 积分与 Riemann 积分

在分析学中我们曾经学习过欧氏空间上的 Riemann 积分的定义。现在，欧氏空间上可测函数关于 Lebesgue 测度（记为 Leb ）的积分也已给出定义，称之为 Lebesgue 积分。很自然要研究它们之间的关系。

本小节我们将证明：有界区间上 Riemann 可积蕴涵 Lebesgue 可积，且此时两种积分值相同。

定理 4.3.4. 设 $I = [a, b]$ 为 \mathbb{R} 中有界区间， $f : I \rightarrow \mathbb{R}$ 为 Riemann 可积函数。那么 f 也是 Lebesgue 可积的，并且

$$\int_I f d\text{Leb} = \int_a^b f(x) dx.$$

证明. 回忆 Riemann 积分的定义，在 Darboux 大和、Darboux 小和的极限值存在、有限且相等的前提下，Riemann 积分就定义为这一共同极限值，并称对应函数 Riemann 可积。由于 f 是 Riemann 可积的，因此它在区间 $I = [a, b]$ 上有界。于是不妨假设 $f \geq 0$ （否则考虑 $f + \|f\|_\infty$ ）。设有一列逐渐加细的分割 $\{\Delta_n : n \geq 1\}$ ，其中 Δ_n 设为

$$a = t_0^{(n)} < t_1^{(n)} < \cdots < t_n^{(n)} = b.$$

定义 $\|\Delta_n\| := \max\{t_k^{(n)} - t_{k-1}^{(n)} : k = 1, \dots, n\}$ 。下面定义

$$\begin{aligned} \bar{f}_n(t) &:= \sum_{k=1}^n \sup f([t_{k-1}^{(n)}, t_k^{(n)})) \cdot 1_{\{t_{k-1}^{(n)} \leq t < t_k^{(n)}\}} + f(b) \cdot 1_{\{t=b\}}, \\ \underline{f}_n(t) &:= \sum_{k=1}^n \inf f([t_{k-1}^{(n)}, t_k^{(n)})) \cdot 1_{\{t_{k-1}^{(n)} \leq t < t_k^{(n)}\}} + f(b) \cdot 1_{\{t=b\}}. \end{aligned}$$

显然，Darboux 大和为

$$U_n := \sum_{k=1}^n \sup f([t_{k-1}^{(n)}, t_k^{(n)})) \cdot [t_k^{(n)} - t_{k-1}^{(n)}] = \int_I \bar{f}_n d\text{Leb},$$

Darboux 小和为

$$L_n := \sum_{k=1}^n \inf f([t_{k-1}^{(n)}, t_k^{(n)})) \cdot [t_k^{(n)} - t_{k-1}^{(n)}] = \int_I \underline{f}_n d\text{Leb}.$$

由于 Δ_{n+1} 是 Δ_n 的加细, 显然有 $\underline{f}_n \leq \underline{f}_{n+1} \leq f \leq \bar{f}_{n+1} \leq \bar{f}_n$. 从而存在可测函数 \underline{f}, \bar{f} 使得 $\underline{f}_n \uparrow \underline{f}$ 以及 $\bar{f}_n \downarrow \bar{f}$. 由此 $\underline{f}_n \leq \underline{f} \leq f \leq \bar{f} \leq \bar{f}_n$. 继而由 f 的 Riemann 可积性, $\lim_{n \rightarrow \infty} U_n = \lim_{n \rightarrow \infty} L_n =: \int_a^b f(x) dx$, 我们有

$$\int_I (\bar{f} - \underline{f}) d\text{Leb} \leq \int_I (\bar{f}_n - \underline{f}_n) d\text{Leb} = U_n - L_n \rightarrow 0.$$

这表明 $\bar{f} = \underline{f}$ 在 I 上 Leb-a.e. 成立, 进而 $\bar{f} = \underline{f} = f$ 在 I 上 Leb-a.e. 成立. \bar{f}, \underline{f} 的 Lebesgue 可积性说明了 f 的 Lebesgue 可积性, 由此立即知道定理中方程成立. \square

上面的定理说明, 在有界闭区间上, 一个 Riemann 可积函数 (比如连续函数) 一定是 Lebesgue 可积的, 并且这两种积分值相等; 在此意义下, 人们通常认为 Lebesgue 积分是 Riemann 积分的推广*; Lebesgue 实际上证明了, 在有界闭区间上, 一个函数是 Riemann 可积的, 当且仅当它的不连续点全体是 Lebesgue 零测集, 从而彻底解决了 Riemann 可积函数的刻画. 因而, 人们在书写积分表达式时也经常把微元 dx 视作 $d\text{Leb}$ 而不加区分是 Lebesgue 积分还是 Riemann 积分. 但实际上严格意义下来说, 这个书写习惯略有瑕疵.

下面就是一个 Riemann 可积但 Lebesgue 积分不存在的例子, 原因就在于它不是有界闭区间上的积分.

例 4.6. 考虑函数 $f(x) := \frac{\sin x}{x}, x \in \mathbb{R}$ (在 $x = 0$ 处, $f(0) := 1$). 设 Leb 是 \mathbb{R} 上 Lebesgue 测度, 则易验证 $\text{Leb}(f^+) = \text{Leb}(f^-) = \infty$, 从而 f 的 Lebesgue 积分不存在. 而由数学分析中的结论, f 是 Riemann 可积的, 且

$$\int_{-\infty}^{\infty} \frac{\sin x}{x} dx = \pi. \quad (4.8)$$

上面的积分结果就是著名的 *Riemann-Lebesgue* 引理. 我们在特征函数的唯一性定理的证明部分需要用到这一结果.

注 4.5. Isaac Newton (牛顿, 1643/1/4–1727/3/31; 英国)、Gottfried Wilhelm Leibniz (莱布尼兹, 1646/7/1–1716/11/14; 德国) 在 17 世纪末就各自独立地给出了定积分的定义. (根据 Newton 周围的人所述, Newton 要比 Leibniz 早几年得出他的方法, 但在 1693 年以前 Newton 几乎没有发表任何相关内容, 并直至 1704 年他才给出了其完整的叙述. 其间, Leibniz 已在 1684 年发表了他的方法的完整叙述. 此外, Leibniz 的符号和“微分法”被欧洲大陆全面地采用, 在大约 1820 年以后, 英国也采用了该方法.) 但是定积分的现代数学定义却是用 Riemann 和的极限给出.

G. F. B. Riemann (黎曼, 1826/9/17–1866/7/20) 是德国著名的数学家, 他在数学分析和微分几何方面作出过重要贡献, 开创了黎曼几何, 并且给后来爱因斯坦的广义相对论提供了数学基础; 数论中著名的黎曼猜想也是他提出的. Riemann 出生于汉诺威王国 (今德国) 的小镇布列斯伦茨 (Breselenz), 父亲是当地的路德会牧师. 1846 年, Riemann 进入哥廷根大学学习哲学和神学. 在此期间他去听了一些数学讲座, 包括 C. F. Gauss (高斯, 1777/4/30–1855/2/23; 德国) 关于最小二乘法的讲座. 在得到父亲的允许后, 他改学数学, 成为 Gauss 晚年的学生. 1847 年春, Riemann 转到柏林大学, 投入 C. G. J. Jacobi (雅可比, 1804/12/10–1851/2/18; 德国)、P. G. L. Dirichlet (狄利克雷, 1805/2/13–1859/5/5; 德国) 和 J. Steiner (施泰纳, 1796/3/18–1863/4/1; 瑞士) 门下; 两年后他回到哥廷根. 1851 年, 他在柏林大学获博士学位; 1854 年, 成为哥廷根大学的讲师, 1857 年, 升为哥廷根大学的编外教授. 1859 年, 他接替 Dirichlet 成为教授. 1866 年 7 月 20 日, 他在第三次去意大利休养的途中因肺结核在塞拉斯卡 (Selasca) 去世.

Riemann 培养的知名学生有: E. Selling (1834/11/5–1920/1/31; 德国数学家, 计算机发明者之一)、G. A. Roch (1839/12/9–1866/11/21; 德国数学家, 以 Riemann-Roch 定理闻名).

*借用 [20, pp. 97] 中观点: Riemann 积分需要考虑逐渐加细的垂直条带, 反映定义域中的几何特性; 而 Lebesgue 积分需要考虑逐渐加细的水平条带, 反映值域中的几何特性. 而对于被积函数而言, 值域的几何性质通常好于定义域的. 这是“推广”能成功的底层逻辑.

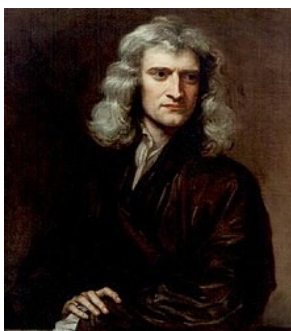


图 4.7: Newton(1643-1727)



图 4.8: Leibniz(1646-1716)



图 4.9: Riemann(1826-1866)



图 4.10: Darboux(1842-1917)

J. G. Darboux (达布, 1842/8/14–1917/2/23) 是法国数学家, 在数学分析 (积分、偏微分方程) 和微分几何 (曲线和曲面的研究) 等方向做出了重要贡献。他生于尼姆 (Nîmes), 卒于巴黎。*Darboux* 幼年丧父, 家境清贫, 但他勤奋好学, 上完中学后, 1861 年以名列榜首的骄人成绩考入巴黎高等师范学校学习, 大学四年级时就脱颖而出, 发表了一篇关于正交曲面的论文, 1864 年大学毕业后留校任教并攻读博士学位, 1866 年 7 月获博士学位, 导师是 *Michel Floréal Chasles* (查斯勒, 1793/11/15–1880/12/18; 法国数学家, *Poisson* 的学生, 以射影几何中的多个 *Chasles* 定理知名)。之后他相继在两所中学及法兰西学院、巴黎大学、索邦大学、巴黎师范学校任教, 1880 年起成为巴黎师范学校的教授。1889–1903 年任巴黎大学理学院院长, 后任名誉院长。1884 年当选为法国科学院院士。1895 年被选为彼得堡科学院通讯院士。同时还被聘为英国皇家学会会员和其他国家科学会的会员, 并荣获国内外许多大学的名誉学位。

Darboux 培养的知名学生有: *E. Picard* (皮卡, 1856/7/24–1941/12/11; 法国数学家, 以他的姓氏命名的数学结果有: *Picard* 函子、*Picard* 群、*Picard* 定理、*Picard-Lefschetz* 公式、*Picard-Lindelöf* 定理等), *T. Stieltjes* (斯蒂尔切斯, 1856/12/29–1894/12/31; 荷兰数学家, 矩问题的先驱, 对连分数研究也有贡献, 以 *Riemann-Stieltjes* 积分闻名), *E. Goursat* (古尔萨, 1858/5/21–1936/11/25; 法国数学家, 二十世纪初出版了数学分析、复分析方面的经典教材《*Cours d'analyse mathématique*》), *S. Zarembka* (1863/10/3–1942/11/23; 波兰数学家, 工程师), *E. Cartan* (嘉当, 1869/4/9–1951/5/6; 法国数学家, 在 *Lie* 群、微分系统和微分几何方面有奠基性的工作), *E. Borel* (波莱尔, 1871/1/7–1956/2/3; 法国数学家、政治家), *G. Țițeica* (1873/10/4–1939/2/5; 罗马尼亚数学家, 在几何学方面有重要贡献, 被认为是微分几何罗马尼亚学派的创始人) 等。

4.3.4 微积分基本定理的探讨

在定义了 Lebesgue 积分后，对于 \mathbb{R} 上任意可积的可测函数 ρ 以及 $a, b \in \mathbb{R}, a < b$ ，我们可以定义

$$F(x) := \int_a^x \rho dx, x \in [a, b].$$

实变函数告诉我们，此时 F 的导函数 F' 在 Lebesgue 几乎处处意义下存在，并且 $F' = \rho$ 几乎处处成立。于是此时有

$$\int_a^b F' dx = F(b) - F(a),$$

即微积分基本定理成立；只不过我们把左边的积分按照 Lebesgue 积分而不是 Riemann 积分的意义来理解。也就是说，微积分基本定理在更广的意义下成立。

现在我们的问题是，对什么样的 F ，下述形态的微积分基本定理成立：

$$\int_a^x F' dx = F(x) - F(a), \quad \forall x \in [a, b]. \quad (4.9)$$

实变函数论最终给出的回答是：当 F 是 $[a, b]$ 上的有界变差函数时，导函数 F' Lebesgue 几乎处处存在；进一步，微积分基本定理(4.9)成立，当且仅当 F 是 $[a, b]$ 上的绝对连续函数；此处有关概念将在下文中给出。

定义 4.3.1. 给定实值函数 $f : [a, b] \rightarrow \mathbb{R}$ ，做划分 $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ ，并定义

$$V_\Delta(f) := \sum_{i=1}^n |f(x_i) - f(x_{i-1})|.$$

令

$$V_a^b(f) := \sup\{V_\Delta(f) : \Delta \text{ 为 } [a, b] \text{ 的划分}\}.$$

我们称 $V_a^b(f)$ 为 f 在 $[a, b]$ 上的全变差，当这个全变差有限时，我们称 f 为 $[a, b]$ 上的有界变差函数。 $[a, b]$ 上的有界变差函数全体记作 $BV([a, b])$ 。

实变函数的理论告诉我们

定理 4.3.5. 给定 \mathbb{R} 中实数 $a < b$ 。以下结论成立：

- (1) 如果 $f : [a, b] \rightarrow \mathbb{R}$ 是单调函数，那么 $f \in BV([a, b])$ ；
- (2) $BV([a, b])$ 是一个线性空间；
- (3) 如果 $f : [a, b] \rightarrow \mathbb{R}$ 是 Lipschitz 函数，那么 $f \in BV([a, b])$ ，并且

$$V_a^b(f) \leq \text{Lip}(f) \cdot (b - a);$$

- (4) 如果 $a < c < b$ ， $f \in BV([a, b])$ ，则

$$V_a^b(f) = V_a^c(f) + V_c^b(f);$$

- (5) (Jordan 分解定理) $f \in BV([a, b])$ ，当且仅当存在单调递增函数 g, h 使得 $f = g - h$ 。事实上，当 $f \in BV([a, b])$ 时，可以取

$$g(x) := \frac{1}{2}[V_a^x(f) + f(x)], \quad h(x) := \frac{1}{2}[V_a^x(f) - f(x)].$$

(6) 如果 $f \in \text{BV}([a, b])$, 那么 f' 几乎处处存在, 并且 f' 可积。当 f 为单调上升函数时,

$$\int_a^x f' dx \leq f(x) - f(a), \quad \forall x \in [a, b].$$

设 μ 是 $[a, b]$ 上的一个只取有限值的符号测度, 定义 $f(x) := \mu([a, x]), x \in [a, b]$, 则: (1) 当 μ 是非负测度时, f 是单调递增、右连续函数; (2) 对于有限的符号测度 μ , 总有 $f \in \text{BV}([a, b])$; (3) 通过 μ 的 Hahn 分解可以建立 f 的 Jordan 分解: 我们可以取上述定理 (5) 中的函数 g, h 如下:

$$g(x) = \mu^+([a, x]), \quad h(x) = \mu^-([a, x]), \quad \forall x \in [a, b].$$

定义 4.3.2. 给定实值函数 $f: [a, b] \rightarrow \mathbb{R}$. 如果对任给 $\varepsilon > 0$, 存在 $\delta > 0$, 使得: 对任意有限个不交开区间的并

$$I = \bigcup_{i=1}^n (a_i, b_i) \subset [a, b],$$

当 $|I| < \delta$ 时就有

$$\sum_{i=1}^n |f(b_i) - f(a_i)| < \varepsilon,$$

那么我们称 f 是 $[a, b]$ 上的绝对连续函数。

不难知道, (i) 绝对连续函数是连续函数; (ii) 绝对连续函数也是有界变差函数。

对于单调递增、右连续函数 F , 我们可以定义测度 μ_F :

$$\mu_F((\alpha, \beta]) = F(\beta) - F(\alpha), \quad a \leq \alpha < \beta \leq b.$$

Carathéodory 扩张的理论告诉我们上述定义给出了 $[a, b]$ 上的一个有限测度。此时, F 是绝对连续的, 也等价于测度 μ_F 具有如下的一种“连续性”: 对任意开集列 $I_n \subset [a, b]$, $|I_n| \rightarrow 0$ 时也必定有 $\mu_F(I_n) \rightarrow 0$ 。回忆物理学中物质的密度概念: 单位体积内的质量称为密度; 这一点态密度的概念将意味着, 具有密度函数的物质应当具有如下的“连续性”: 当以任一固定点为中心的区域体积足够小时, 该区域内的物质的质量也必定充分小。这一物质的“连续性”是该物质的密度函数存在的必要条件。实变函数论及测度论告诉我们, 该条件也是充分的, 并且这一条件有更简单的等价形式 (的静态) 刻画: 测度 μ_F 关于 Lebesgue 测度绝对连续, 记作 $\mu_F \ll \text{Leb}$ 。

定义 4.3.3. \mathbb{R}^d 上 (σ -有限的) 测度 μ 关于 Lebesgue 测度绝对连续, 记作 $\mu \ll \text{Leb}$, 如果 Lebesgue 零测集都是 μ -零测集, i.e., $\forall A \in \mathcal{B}^n, \text{Leb}(A) = 0$ 蕴含了 $\mu(A) = 0$ 。

\mathbb{R}^d 上 (σ -有限的) 符号测度 μ 关于 Lebesgue 测度绝对连续, 记作 $\mu \ll \text{Leb}$, 如果全变差测度满足 $|\mu| \ll \text{Leb}$ 。

在下面定理中我们仅表述了一维的结果; 据说这个定理是 G. Vitali (维塔利, 1875/8/26–1932/2/29; 意大利) 的贡献。

定理 4.3.6. 给定 \mathbb{R} 中实数 $a < b$ 。设 $F : [a, b] \rightarrow \mathbb{R}$ 是实函数。那么存在 $\rho \in L^1([a, b])$ 使得

$$\int_a^x \rho(x) dx = F(x) - F(a), \quad \forall x \in [a, b]$$

的充分必要条件是： F 是绝对连续函数。此时，Lebesgue 几乎处处成立 $\rho = F'$ ，进而微积分基本定理成立：

$$\int_a^x F' dx = F(x) - F(a), \quad \forall x \in [a, b].$$

4.3.5 重积分与累次积分—Fubini 定理

在 Riemann 积分框架下，数学分析中研究了何时重积分可以化为累次积分，由此也探讨了两重积分能够交换积分次序的条件；在那里的讨论是比较繁杂的。在 Lebesgue 积分框架下，也有类似的重积分化为累次积分以及两重积分何时能交换积分次序的问题，但此时的讨论就非常清爽，最终的结果统一成如下的 Fubini 定理。

定理 4.3.7. (Fubini 定理) 设 $f : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$ 是非负可测函数（或 $f \in L^1(\mathbb{R}^{m+n})$ ），那么

$$\begin{aligned} \int_{\mathbb{R}^m \times \mathbb{R}^n} f(x, y) dx dy &= \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}^m} f(x, y) dx \right] dy \\ &= \int_{\mathbb{R}^m} \left[\int_{\mathbb{R}^n} f(x, y) dy \right] dx. \end{aligned}$$

证明思想： 为了方便，记 \mathbb{R}^n 中的 Lebesgue 测度为 $\text{Leb}^{(n)}$ 。这样，定理的结论就可以表述为：

$$\int f(x, y) d\text{Leb}^{(m)} \times \text{Leb}^{(n)} = \int \left[\int f(x, y) d\text{Leb}^{(m)} \right] d\text{Leb}^{(n)} \quad (4.10)$$

$$= \int \left[\int f(x, y) d\text{Leb}^{(n)} \right] d\text{Leb}^{(m)}. \quad (4.11)$$

Step 1. 对于 $A \in \mathcal{B}^m, B \in \mathcal{B}^n$ ，改写下方方程

$$\text{Leb}^{(m)} \times \text{Leb}^{(n)}(A \times B) = \text{Leb}^{(m)}(A) \cdot \text{Leb}^{(n)}(B)$$

为积分

$$\begin{aligned} \int 1_{A \times B}(x, y) d\text{Leb}^{(m)} \times \text{Leb}^{(n)} &= \int \left[\int 1_{A \times B}(x, y) d\text{Leb}^{(m)} \right] d\text{Leb}^{(n)} \\ &= \int \left[\int 1_{A \times B}(x, y) d\text{Leb}^{(n)} \right] d\text{Leb}^{(m)}. \end{aligned}$$

即方程(4.10)、(4.11)对于 $f = 1_{A \times B}$ 成立。单调类方法将能证明方程(4.10)、(4.11)对于 $f = 1_E$ 成立，只要 $E \in \mathcal{B}^{m+n}$ 。

Step 2. 利用 Lebesgue 积分的定义与性质，立即知道方程(4.10)、(4.11)对于非负可测函数 f 成立；

Step 3. 进一步，利用 $f = f^+ - f^-$ ，立即知道方程(4.10)、(4.11)对于可测函数 $f \in L^1(\mathbb{R}^{m+n})$ 成立。 \square



图 4.11: Fubini(1879-1943)

注 4.6. *Guido Fubini*（富比尼，1879/1/19–1943/6/6；意大利）生于威尼斯，卒于美国纽约。他中学毕业后到比萨继续深造。1901年起先后在热那亚（*Genoa*）大学和都灵（*Turin*）大学讲授数学分析；1938年因法西斯政府的民族歧视法令而被迫离职，次年移居美国，任职于普林斯顿高等研究院。

4.4 抽象测度的 Lebesgue 积分

4.4.1 抽象测度的 Lebesgue 积分的定义

上一节欧氏空间的 Lebesgue 积分的理论可以几乎完全平行地发展为一般的测度空间上的 Lebesgue 积分理论。当然，在微积分基本定理的处理上就略微不太一样了；测度论中在此处的对应是发展出了所谓的（一个测度关于另一个测度的）Radon-Nikodym 导数来代替通常的（函数的）导数，从而也有类似于微积分基本定理的探讨；有关讨论见下一小节。

给定测度空间 $(\Omega, \mathcal{F}, \mu)$ 。我们按照以下流程定义可测函数 $f: \Omega \rightarrow \mathbb{R}$ 关于测度 μ 的积分（有时也记作 $\mu(f)$ ）

$$\int f d\mu = \mu(f)$$

如下：

首先定义

$$\int 1_A d\mu := \mu(A), \quad \forall A \in \mathcal{F}.$$

其次，对任意非负简单函数 $f \in \mathcal{S}_+$ ，存在 $\{a_n\}_{n=1}^N \subset \mathbb{R}_+$ 及不交集列 $\{A_n\}_{n=1}^N \subset \mathcal{F}$ ，使得

$$f = \sum_{n=1}^N a_n 1_{A_n},$$

此时定义

$$\int f d\mu = \sum_{n=1}^N a_n \mu(A_n).$$

之后，对任意非负可测函数 f ，定义

$$\int f d\mu := \sup \left\{ \int g d\mu : 0 \leq g \leq f, g \in \mathcal{S}_+ \right\}.$$

最后，对任意可测函数 f ，注意到

$$f = f^+ - f^-,$$

当

$$\min \left\{ \int f^+ d\mu, \int f^- d\mu \right\} < \infty$$

时，就定义

$$\int f d\mu := \int f^+ d\mu - \int f^- d\mu;$$

否则称 f 关于 μ 的积分不存在。当

$$\max \left\{ \int f^+ d\mu, \int f^- d\mu \right\} < \infty$$

时，就称 f 关于 μ 可积，简记为 $f \in L^1(\mu)$ 。

4.4.2 抽象测度的 Lebesgue 积分的极限交换问题

本小节我们探讨抽象测度的 Lebesgue 积分下的两类极限交换问题：其一是

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int \lim_{n \rightarrow \infty} f_n d\mu$$

成立的（充分）条件的探讨；其二是抽象测度的重积分何时可以化成累次积分的问题，也就是两重积分的交换问题。

测度论的理论告诉我们，积分 $\int f d\mu$ 关于 f 具有线性性质，并且上一节中的 Levi 单调收敛定理、Fatou 引理、Lebesgue 控制收敛定理和 Fubini 定理仍然成立；我们把它们罗列如下：

定理 4.4.1. (单调收敛定理/Levi 定理) 给定测度空间 $(\Omega, \mathcal{F}, \mu)$ 。设 $\{f_n\}_{n=1}^\infty$ 为一列非负可测函数，且存在可测函数 f 使得 $f_n \nearrow f$ 。那么

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu.$$

定理 4.4.2. (Fatou 引理) 给定测度空间 $(\Omega, \mathcal{F}, \mu)$ 。设 $\{f_n\}_{n=1}^\infty$ 为一列非负可测函数，那么

$$\liminf_{n \rightarrow \infty} \int f_n d\mu \geq \int \liminf_{n \rightarrow \infty} f_n d\mu.$$

定义 4.4.1. 给定测度空间 $(\Omega, \mathcal{F}, \mu)$ ， $\{f_n\}_{n=1}^\infty$ 为其上一列可测函数。称它们几乎处处收敛于 f （记作 $f_n \xrightarrow{a.e.} f$ 或 $f_n \xrightarrow{\mu-a.e.} f$ 或 $f_n \rightarrow f, \mu - a.e.$ ），如果

$$\mu(\{\omega : \lim_{n \rightarrow \infty} f_n(\omega) \neq f(\omega)\}) = 0.$$

当 μ 是一个概率测度时，有时也称几乎处处收敛为以概率 1 收敛或几乎必然收敛 (*almost sure convergence*)，记作 $f_n \xrightarrow{a.s.} f$ 。

如果

$$\lim_{n \rightarrow \infty} \mu(\{\omega : |f_n(\omega) - f(\omega)| \geq \varepsilon\}) = 0, \quad \forall \varepsilon > 0,$$

则称 $\{f_n\}_{n=1}^{\infty}$ 依测度收敛到 f , 记作 $f_n \xrightarrow{\mu} f$; 当 μ 是一个概率测度时, 上述收敛也称依概率收敛。

定理 4.4.3. (Lebesgue 控制收敛定理) 给定测度空间 $(\Omega, \mathcal{F}, \mu)$ 。设 $\{f_n\}_{n=1}^{\infty}$ 为一列几乎处处收敛 (或依测度收敛) 于 f 的可测函数, 且存在可测函数 $g \in L^1(\mu)$ 使得 $|f_n| \leq g$ 。那么

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int \lim_{n \rightarrow \infty} f_n d\mu.$$

定理 4.4.4. (Fubini 定理) 给定两个 σ -有限的测度空间 $(\Omega_k, \mathcal{F}_k, \mu_k), k = 1, 2$ 。设 $f: \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ 是非负可测函数 (或 $f \in L^1(\mu_1 \times \mu_2)$), 那么

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} f d\mu_1 \times \mu_2 &= \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1) \right] d\mu_2(\omega_2) \\ &= \int_{\Omega_1} \left[\int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2) \right] d\mu_1(\omega_1). \end{aligned}$$

4.4.3 微积分基本定理的推广: Radon-Nikodym 定理

为了后续讨论的需要, 我们给出一个测度关于另一个测度绝对连续和两个测度相互奇异的观念。

定义 4.4.2. 给定 (Ω, \mathcal{F}) 上两个 (σ -有限的) 测度 μ, ν 。称 μ 关于 ν 绝对连续, 记作 $\mu \ll \nu$, 如果任意 ν -零测集都是 μ -零测集, 即

$$(\nu(A) = 0 \Rightarrow \mu(A) = 0), \forall A \in \mathcal{F}.$$

如果 $\mu \ll \nu$ 且 $\nu \ll \mu$, 则称 μ, ν 相互等价, 记作 $\mu \sim \nu$ 。

称 μ, ν 为相互奇异的, 记作 $\mu \perp \nu$, 如果存在 $H \in \mathcal{F}$, 使得 $\mu(H) = 0$ 且 $\nu(H^c) = 0$ 。

上述概念对于符号测度也可以定义, 只不过其实质是对应的全变差测度之间的关系, 我们对此不再赘述。

下面的结果是测度论中非常深刻的一个结果, 本质上它是抽象测度版本的微积分基本定理。

定理 4.4.5. (Radon-Nikodym 定理) 设 μ, ν 分别是可测空间 (Ω, \mathcal{F}) 上的测度与符号测度, 二者均 σ -有限。设 $\mu \ll \nu$ 。则存在 ν -几乎处处有限的可测函数 f 使得 $\mu = f \cdot \nu$, 即

$$\mu(A) = \int_A f d\nu, \forall A \in \mathcal{F}.$$

函数 f 称为 μ 关于 ν 的 Radon-Nikodym 导数, 记作 $\frac{d\mu}{d\nu} := f$ 。在 ν -几乎处处相等的意义下, Radon-Nikodym 导数是唯一的。

证明. 为简单起见, 不妨设 ν 是测度而不是符号测度 (否则 $\nu = \nu^+ - \nu^-$, 且此时 $\nu^+ \ll \mu, \nu^- \ll \mu$, 分别考虑 ν^+, ν^- 关于 μ 的 R-N 导数即可); 进一步不妨设 μ, ν 都是有限测度。(读者请思考: 为什么可以这样假设?)

为简化记号，对任意可测函数 h ，令 $h \cdot \mu$ 代表一个测度，它定义为

$$(h \cdot \mu)(A) := \int_A h d\mu.$$

现在定义

$$\mathcal{H} := \{h \geq 0 : h \text{ 关于 } \mathcal{F} \text{ 可测, 且 } h \cdot \mu \leq \nu\}.$$

令 $b := \sup\{\mu(h) : h \in \mathcal{H}\}$ 。

我们首先证明 \mathcal{H} 关于函数运算 \vee 封闭。事实上，对任意 $h_1, h_2 \in \mathcal{H}$ ，记 $h := h_1 \vee h_2$ ，有 $h = h_2 \cdot 1_{\{h_1 \leq h_2\}} + h_1 \cdot 1_{\{h_1 > h_2\}}$ ，进而

$$\begin{aligned} h \cdot \mu &= (h_2 \cdot 1_{\{h_1 \leq h_2\}}) \cdot \mu + (h_1 \cdot 1_{\{h_1 > h_2\}}) \cdot \mu \\ &\leq 1_{\{h_1 \leq h_2\}} \cdot \nu + 1_{\{h_1 > h_2\}} \cdot \nu = \nu, \end{aligned}$$

因此 $h \in \mathcal{H}$ 。

于是存在递增列 $\{h_n\}_{n=1}^\infty \subset \mathcal{H}$ ，使得 $b = \lim_{n \rightarrow \infty} \mu(h_n)$ 。令 $g = \lim_{n \rightarrow \infty} h_n$ ，由单调收敛定理， $g \in \mathcal{H}$ 且 $b = \mu(g) = (g \cdot \mu)(\Omega) \leq \nu(\Omega) < \infty$ 。

下面证明 $\nu = g \cdot \mu$ 。令 $\gamma = \nu - g \cdot \mu$ ，由于 $g \in \mathcal{H}$ ， γ 是一个有限测度。任给 $n \geq 1$ ，设 D_n 是符号测度 $\gamma - \frac{1}{n}\mu$ 的 Hahn 集。于是对任意 $A \in \mathcal{F}$

$$\nu(A \cap D_n) \geq g \cdot \mu(A \cap D_n) + \frac{1}{n}\mu(A \cap D_n), \quad (4.12)$$

$$\gamma(A \cap D_n^c) \leq \frac{1}{n}\mu(A \cap D_n^c). \quad (4.13)$$

(4.12) 亦即 $1_{D_n} \cdot (g + \frac{1}{n}) \cdot \mu \leq 1_{D_n} \cdot \nu$ 。从而对 $g_n := g + \frac{1}{n} \cdot 1_{D_n}$ ，有

$$\begin{aligned} g_n \cdot \mu &= g \cdot \mu + \frac{1}{n} \cdot 1_{D_n} \cdot \mu \\ &= 1_{D_n^c} \cdot g \cdot \mu + 1_{D_n} \cdot g \cdot \mu + \frac{1}{n} \cdot 1_{D_n} \cdot \mu \\ &= 1_{D_n^c} \cdot g \cdot \mu + 1_{D_n} \cdot (g + \frac{1}{n}) \cdot \mu \\ &\leq 1_{D_n^c} \cdot \nu + 1_{D_n} \cdot \nu = \nu. \end{aligned}$$

由此 $g_n \in \mathcal{H}$ 。而由 g 的极大性，应有 $\mu(D_n) = 0$ 。令 $D = \bigcup_{n=1}^\infty D_n$ ，有 $\mu(D) = 0$ ，进而 $\gamma(D) = 0$ 。另一方面，由(4.13)，

$$\gamma(D^c) \leq \gamma(D_n^c) \leq \frac{1}{n}\mu(D_n^c) \leq \frac{1}{n}\mu(\Omega) \rightarrow 0,$$

于是 $\gamma(D^c) = 0$ 。因此 $\gamma = 0$ ，即 $\nu = g \cdot \mu$ 。

显然 Radon-Nikodym 导数 $f := g$ 是几乎处处有限的，并且容易看出在 μ -几乎处处相等的意义下，Radon-Nikodym 导数是唯一的。□

基于上述结果，我们指出如下结果成立；这是我们前述的质量与密度两个物理概念的内在数学上的抽象表达。

定理 4.4.6. 给定 $(\mathbb{R}^d, \mathcal{B}^d)$ 上两个 σ -有限的测度 μ, ν 。则 μ 关于 ν 绝对连续的充分必要条件是如下版本的“ μ 关于 ν 的连续性”成立： $\forall V_n \in \mathcal{B}^d$ ， $V_n \downarrow$ ，且 $\nu(V_n) \downarrow 0$ 时，就有 $\mu(V_n) \rightarrow 0$ 。

当上述“ μ 关于 ν 的连续性”成立时，对 ν -a.e. $x \in \mathbb{R}^d$

$$\rho(x) := \frac{d\mu}{d\nu}(x) = \lim_{r \downarrow 0} \frac{\mu(B(x, r))}{\nu(B(x, r))}.$$

此时

$$\mu(A) = \int_A \rho(x) d\nu(x), \quad \forall A \in \mathcal{B}^d.$$

上述定理中 $d = 1$ 时的结果（同时取 $\nu = \text{Leb}$ ），就对应了上一节中对微积分基本定理的探讨。



图 4.12: Vitali(1875-1932)



J. I. Radon

图 4.13: Radon(1887-1956)



图 4.14: Nikodym(1887-1974) 与 Banach(1892-1945) 的纪念雕像（波兰，Kraków）

注 4.7. *Giuseppe Vitali*（维塔利，1875/8/26–1932/2/29）是意大利数学家、力学家，生于拉韦纳（*Ravenna*），卒于博洛尼亚（*Bologna*）；他最早研究可测函数的性质，建立了 *Vitali* 覆盖定理、解析函数列的极限函数列的解析性定理、积分与极限符号可交换的定理。他在 1905 年引入了绝对连续函数的概念，并且证明了

$$\int_a^x f(t) dt = F(x) - F(a), \quad \forall x \in [a, b]$$

成立的充分必要条件是 $F'(x) = f(x)$ a.e. 于 $[a, b]$ ，且 $F(x)$ 在 $[a, b]$ 上绝对连续的。这样，*Lebesgue* 积分扩大了使微积分基本定理成立的函数类。

Johann Radon（拉东，1887/12/16–1956/5/25）是奥地利数学家，生于捷克斯洛伐克的波西米亚（*Bohemia*，现名为 *Decin*），1947 年他受聘为维也纳大学教授，并选为奥地利科学院院士，后卒于奥地利的维也纳。在分析学上他提出了 *Radon* 测度和 *Radon* 积分。他 1913 年把 *Vitali* 的上述工作推广到定义在 n 维欧氏空间中的 *Borel* 测度的情形。

Otto Marcin Nikodym (尼科迪姆, 1887/8/13–1974/5/4) 是美籍波兰数学家, 生于波兰 Zablotów 郊区, 1948 年移民到美国任教于 Kenyon 学院, 1966 年退休后移居纽约直至逝世。他于 1929 年把上述 Radon 的工作进一步推广到一般测度空间中的积分情形。

由 Radon-Nikodym 定理, 我们有下面的简单推论。

◆ 推论 4.4.1. 设 μ, ν 分别是可测空间 (Ω, \mathcal{F}) 上的 σ -有限测度与符号测度, 其中 $\nu = f \cdot \mu$, f 是一个广义实值可测函数。那么

- (i) $|\nu(\Omega)| < \infty$ 等价于 $f \in L^1(\mu)$;
- (ii) ν 是 σ -有限的等价于 $\mu(\{|f| = \infty\}) = 0$ 。

下面的 Lebesgue 分解定理则给出了一般的（符号）测度在给定参考测度下的结构。

☞ 定理 4.4.7. (**Lebesgue 分解**) 设 μ, ν 分别是可测空间 (Ω, \mathcal{F}) 上的测度与符号测度, 二者均 σ -有限。则可将 ν 唯一的表示为 $\nu = \nu_1 + \nu_2$, 使得 ν_1, ν_2 均为 σ -有限的符号测度, 且 $\nu_1 \ll \mu, \nu_2 \perp \mu$ 。

证明. 类似于 Radon-Nikodym 定理的证明, 不妨设 μ, ν 均为有限测度。令

$$\mathcal{L} := \{f \in L^1(\Omega, \mathcal{F}, \mu) : f \cdot \mu \leq \nu\}.$$

则 \mathcal{L} 有下面一些性质:

- (i) $0 \in \mathcal{L}$, 从而 \mathcal{L} 非空;
- (ii) 如果 $f_1, f_2 \in \mathcal{L}$, 则 $f_1 \vee f_2 \in \mathcal{L}$ 。这是因为

$$\begin{aligned} (f_1 \vee f_2) \cdot \mu &= 1_{\{f_1 \leq f_2\}} \cdot f_2 \cdot \mu + 1_{\{f_1 > f_2\}} \cdot f_1 \cdot \mu \\ &\leq 1_{\{f_1 \leq f_2\}} \cdot \nu + 1_{\{f_1 > f_2\}} \cdot \nu = \nu; \end{aligned}$$

- (iii) 如果 $f_n \in \mathcal{L}, f_n \nearrow f$, 则 $f \in \mathcal{L}$ 。事实上, 容易知道

$$g_n := f_n + f_1^- \geq f_1 + f_1^- = f_1^+ \geq 0$$

是一个非负单调函数列, 由此 $g_n \nearrow f + f_1^- \geq 0$, 根据单调收敛定理, 易知 $f \in L^1(\Omega, \mathcal{F}, \mu)$ 且 $f \cdot \mu \leq \nu$, 进而 $f \in \mathcal{L}$ 。

根据以上性质, \mathcal{L} 中有极大元 $f \geq 0$, 且 $\nu_1 := f \cdot \mu \leq \nu$ 。记 $\nu_2 := \nu - f \cdot \mu$, 于是 $\nu = \nu_1 + \nu_2$, 且 $\nu_1 \ll \mu$ 。以下证明 $\nu_2 \perp \mu$ 。事实上, 令 D_n 为 $\nu_2 - \frac{1}{n}\mu$ 的 Hahn 集。于是对任意 $A \in \mathcal{F}$

$$\nu_2(A \cap D_n) \geq \frac{1}{n}\mu(A \cap D_n), \quad \nu_2(A \cap D_n^c) \leq \frac{1}{n}\mu(A \cap D_n^c).$$

由此

$$(f + \frac{1}{n} \cdot 1_{D_n}) \cdot \mu \leq f \cdot \mu + 1_{D_n} \cdot \nu_2 \leq \nu_1 + \nu_2 = \nu,$$

即 $f + \frac{1}{n} \cdot 1_{D_n} \in \mathcal{L}$ 。由 f 的极大性, $\mu(D_n) = 0$ 。记 $D := \bigcup_{n=1}^{\infty} D_n$, 则 $\mu(D) = 0$ 。

另一方面,

$$\nu_2(D^c) \leq \nu_2(D_n^c) \leq \frac{1}{n}\mu(D_n^c) \leq \frac{1}{n}\mu(\Omega).$$

令 $n \rightarrow \infty$ 即知 $\nu_2(D^c) = 0$ 。由此 $\nu_2 \perp \mu$ 。 \square

习 题 4

习题 4.1. 设 μ 是 (Ω, \mathcal{F}) 上的符号测度。求证：它是有限符号测度的充要条件是 $|\mu(\Omega)| < \infty$ 。

习题 4.2. 设 μ 是 (Ω, \mathcal{F}) 上测度。对任意 $0 < p < \infty$, 以及可测函数 $f: \Omega \rightarrow \mathbb{R}$, 定义

$$\|f\|_p := \left[\int |f|^p d\mu \right]^{1/p},$$

并定义

$$L^p(\Omega, \mathcal{F}, \mu) := \{f: f \text{ 是可测函数, 且 } \|f\|_p < \infty\}.$$

$L^p(\Omega, \mathcal{F}, \mu)$ 也简记为 $L^p(\mu)$ 或 L^p 。求证：对 $1 \leq p < \infty$, 以下成立：

(1) 对任意 $a \in \mathbb{R}, f \in L^p$, $\|af\|_p = |a| \cdot \|f\|_p$;

(2) (Hölder 不等式) 对任意 $f \in L^p, g \in L^q$, 其中 $\frac{1}{p} + \frac{1}{q} = 1$ (此时称 p, q 共轭)

$$\|f \cdot g\|_1 \leq \|f\|_p \cdot \|g\|_q;$$

(3) (Minkowski 不等式/三角不等式) 对任意 $f, g \in L^p$,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p;$$

(4) 对任意 $f, g \in L^p$, 我们认为 $f = g$ 如果 $\mu(\{f \neq g\}) = 0$ 。在这种认同下 $(L^p, \|\cdot\|_p)$ 是一个赋范空间：它满足上面的 (1)、(3) 及

$$\|f\|_p = 0 \Leftrightarrow f = 0 \quad \mu\text{-a.e.};$$

(5) 称 $\{f_n\}_{n=1}^\infty \subset L^p$ 为 L^p 中的 Cauchy 列, 如果

$$\lim_{n, m \rightarrow \infty} \|f_n - f_m\|_p = 0.$$

对于上述 Cauchy 列, 存在 $f \in L^p$, 使得

$$\lim_{n \rightarrow \infty} \|f_n - f\|_p = 0.$$

这说明了 $(L^p, \|\cdot\|_p)$ 的完备性。于是 $(L^p, \|\cdot\|_p)$ 是一个 Banach 空间 (即完备的赋范空间)。

习题 4.3. 设 μ 是 (Ω, \mathcal{F}) 上测度。对 $p = \infty$, 以及可测函数 $f: \Omega \rightarrow \mathbb{R}$, 定义

$$\|f\|_\infty := \inf\{M \geq 0 : \mu(|f| > M) = 0\},$$

称 $\|f\|_\infty$ 为 f 的本质范数或无穷范数, 并定义

$$L^\infty(\Omega, \mathcal{F}, \mu) := \{f: f \text{ 是可测函数, 且 } \|f\|_\infty < \infty\}.$$

有时也把 $L^\infty(\Omega, \mathcal{F}, \mu)$ 简记为 $L^\infty(\mu)$ 或 L^∞ 。求证：上面习题中结论对 $p = \infty$ 仍然成立。

习题 4.4. 设 μ 是 (Ω, \mathcal{F}) 上有限测度。对 $1 \leq p < \infty$, 以及可测函数 $f: \Omega \rightarrow \mathbb{R}$, 定义

$$L_*^p(\Omega, \mathcal{F}, \mu) := \{f: f \text{ 是可测函数, 且 } \sup_{t>0} t^{p-1} \mu(|f| > t) < \infty\}.$$

有时也把 $L_*^p(\Omega, \mathcal{F}, \mu)$ 简记为 $L_*^p(\mu)$ 或 L_*^p 。求证： $L^p \subset L_*^p \subset L^{p-\varepsilon}, \forall \varepsilon > 0$ 。

§ 5

随机变量 (I)

接下来我们将介绍概率论中最重要且基础的概念：随机变量及其分布律。这是初等概率论的核心概念；概率论中几乎所有的演绎与推理本质上都是立足于分布律的。为了教学安排的方便，我们拆分成了三个章节，分别是本章（第 5 章）、第 8 章和第 9 章，中间插入了第 6、7 章用以介绍数学期望和条件数学期望及条件分布律。

有了随机变量这个概念，我们就可以定义随机向量，借助它们我们可以很方便地探讨随机事件，进而发展出更多的概念与理论。不同于现行的大部分初等概率论教材，在本章我们通过令人信服的例子指出，我们应当在“随机变量是可测函数”这一定义的基础上施行一个合理的“补丁程序”，以确保逻辑的严谨性以及概率论整体理论体系的自治性：这个“补丁程序”本质上也是对两个随机变量在几乎处处相等意义下的认同，这一点在概率论的高级课程中是常规设定。

在随机变量/向量概念的基础上，我们介绍了它们的分布律的概念，目前来说就是（联合）分布函数或相应的分布测度。在此基础上，我们介绍多个随机变量/向量之间的相互独立性概念，并给出独立性概念的性质与等价刻画。

最后，在本章我们介绍离散型随机变量及常见离散型分布，并给出两个利用离散型分布进行极大似然估计的经典应用案例。

5.1 随机变量、随机向量及其分布律

5.1.1 随机变量与随机向量的定义

定义 5.1.1. 设 $(\Omega, \mathcal{F}, \mathbb{P})$ 为概率空间。此时可测函数 $\xi : \Omega \rightarrow \mathbb{R}$ 也称为随机变量，这等价于要求

$$\{\xi \leq x\} \in \mathcal{F}, \forall x \in \mathbb{R}. \quad (5.1)$$

用自然语言来说，当函数 $\xi : \Omega \rightarrow \mathbb{R}$ 满足：对所有 $x \in \mathbb{R}$ ，事件 $\{\xi \leq x\}$ 都能谈论概率，就称 ξ 是随机变量。这也等价于要求 $\xi^{-1}(\mathcal{B}) \subset \mathcal{F}$ ，此处 \mathcal{B} 是 \mathbb{R} 上的 Borel σ -代数（见第 4 章 §4.1），

$$\xi^{-1}(\mathcal{B}) := \{\xi^{-1}(B) : B \in \mathcal{B}\}$$

有时也称为 ξ 生成的 σ -代数, 记作 $\sigma(\xi) = \xi^{-1}(\mathcal{B})$; 它本质上就是 “能用 ξ 描写的事件” 的全体。

此时函数 $F: \mathbb{R} \rightarrow \mathbb{R}$

$$F(x) := \mathbb{P}(\xi \leq x), \quad \forall x \in \mathbb{R}$$

称为 ξ 的分布函数 (简称分布); 此时记 $\xi \sim F$, 读作: ξ 服从分布 F , 或 ξ 的分布函数是 F 。

显然, 随机变量引入的一个重要目的就是为了更方便且精确地描述各种各样的与随机现象关联的实际问题中的 “能谈论概率的那些 (随机) 事件”。例如, 在投掷一个均匀的骰子一次的概率空间的建模中, 借助引入 ξ 为骰子投掷出的点数, 我们可以用 $\{\xi = k\}$ 代表事件 “骰子投掷出点数 k ”, $k = 1, \dots, 6$; 在一次性投掷三枚均匀的骰子的概率空间的建模中, 借助引入 ξ_k 为第 k 枚骰子投掷出的点数, $k = 1, 2, 3$, 我们可以很方便、精确地描写更复杂的事件 “三枚骰子投掷出的点数和为 n ” 这一事件为 $\{\xi_1 + \xi_2 + \xi_3 = n\}$ 。这种借助引进 “随机变量” 来描述事件的方法, 在概率论发展的初期人们早已无意识地普遍使用了; 特别地, 很多现实问题中这种 “随机变量” 也是很自然出现, 且被人们关心, 第 1 章的例 1.2 中有关问题就是一个明证; 只不过在现代概率论中, 我们对 “随机变量” 的数学内涵进行明确, 并对其使用加以规范。也就是说, 概率论对随机变量概念的使用经历了一个从无意识到有意识、从不自觉到自觉、从不规范到规范的过程; 这种对成功经验的分析、理解、吸收与内化过程在科学的发展、竞技体育的训练等过程中是非常普遍的。我们对数学的学习过程也应尽力遵循这种规律, 方能事半功倍。

定义 5.1.2. 给定 $n \in \mathbb{N}$, 当 ξ_1, \dots, ξ_n 都是随机变量时, 我们称

$$\xi = (\xi_1, \dots, \xi_n)^T: \Omega \rightarrow \mathbb{R}^n$$

为 n 维随机向量。此时, 函数 $F: \mathbb{R}^n \rightarrow \mathbb{R}$

$$F(x) := \mathbb{P}(\xi_1 \leq x_1, \dots, \xi_n \leq x_n), \quad \forall x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$$

称为 ξ 的联合分布函数 (也简称分布); 对应地记 $\xi \sim F$, 读作: ξ 服从分布 F , 或 ξ 的联合分布函数是 F 。

设 $\xi_i \sim F_i, i = 1, \dots, n$, 如果 $\xi \sim F$, 则显然通过令 $x_j \rightarrow +\infty, j \in \{1, \dots, n\} \setminus \{i\}$, 可以从联合分布函数 F 取极限得到 ξ_i 的分布函数 F_i , 这可以简单地记为

$$F_i(x_i) = F(\underbrace{\infty, \dots, \infty}_{i-1 \text{ 个}}, x_i, \underbrace{\infty, \dots, \infty}_{n-i \text{ 个}}).$$

这里, 每个 F_i 都称为联合分布函数 F 的 (一维) 边缘分布函数; 准确地说, F_i 是 F 的第 i 维边缘分布函数。更一般的, 对于 $1 \leq d < n$ 以及 $1 \leq j_1 < j_2 < \dots < j_d \leq n$, $(\xi_{j_1}, \dots, \xi_{j_d})$ 的联合分布函数 F_{j_1, \dots, j_d} 称为 F 的 d 维边缘分布函数 (简称边缘分布函数;)。

这里同样有 n 维随机向量 ξ 生成的 σ -代数的概念, 它实际上是

$$\sigma(\xi) = \xi^{-1}(\mathcal{B}^n) := \{\xi^{-1}(B) : B \in \mathcal{B}^n\}.$$

此时有 $\sigma(\xi) = \sigma(\bigcup_{i=1}^n \sigma(\xi_i))$ 。

对于随机变量/向量 ξ (设它的值域为 \mathbb{R}^n), 我们可以定义它的 (联合) 分

布测度 μ_ξ 如下：

$$\mu_\xi(A) := \mathbb{P}(\xi \in A), \text{ 其中 } A \text{ 为 } \mathbb{R}^n \text{ 中可测集.}$$

此时 μ_ξ 是可测空间 $(\mathbb{R}^n, \mathcal{B}^n) := (\mathbb{R}, \mathcal{B})^n$ 上的 (Borel) 概率测度；我们也记 $\xi \sim \mu_\xi$ ，读作： ξ 服从分布 μ_ξ ，或 ξ 的分布测度是 μ_ξ 。对于随机向量，此处也有边缘（分布）测度的说法：例如 ξ_i 的分布测度 μ_{ξ_i} 就称为 μ_ξ （或 ξ ）的第 i 维边缘分布测度。

在上述定义中，有几处偷懒的记号需要指出：事件 $\{\xi \leq x\}$ 的准确描述是 $\{\omega \in \Omega : \xi(\omega) \leq x\}$ ，事件 $\{\xi \leq x\}$ 的概率被简单地记作 $\mathbb{P}(\xi \leq x)$ 。这些记号上的偷懒都是概率论长期发展过程中形成的约定俗成的书写习惯，需要读者铭记于心并灵活使用。另外，我们一般使用 26 个英文字母中排列较前的大写字母（如 A, B, C, D 等）代表事件，用 26 个英文字母中排列较后的大写字母（如 U, V, W, X, Y, Z 等）及一些小写希腊字母（例如 ξ, η, ζ 等）表示随机变量，用小写英文字母表示具体值。

上面通过可测函数的概念给出了随机变量的严格定义，它要求随机变量不能取无穷值；但这个定义其实过于苛刻，从而在实际应用中不方便使用。下面的例子就说明了这一点。

例 5.1. 设概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 是对投掷一枚均匀硬币无穷多次的概率模型（比如取 $(\Omega, \mathcal{F}, \mathbb{P}) = (\Sigma, 2^\Sigma, \mu)^\infty$ ，其中 $\Sigma = \{0, 1\}, \mu(\{0\}) = \mu(\{1\}) := \frac{1}{2}$ ）；第 n 次投掷该硬币，获得正面时，我们记 $X_n = 1$ ，否则记为 $X_n = 0$ 。一般的，我们会关心为了获得正面，所需要的最少的投掷次数 τ （参见例 1.2 的第三个问题或例 3.16 中的 τ_A ）。它在数学上可以表述为

$$\tau := \inf\{n \geq 1 : X_n = 1\}.$$

不难知道，对任意自然数 $n \geq 1$

$$\mathbb{P}(\tau = n) = \mathbb{P}(X_1 = \cdots = X_{n-1} = 0, X_n = 1) = \frac{1}{2^n},$$

从而 $\mathbb{P}(\tau = \infty) = 0$ 。在初等概率论中一般认为 τ 是随机变量，服从的分布是参数为 $1/2$ 的几何分布。但是在逻辑上我们无法排除 $\{\tau = \infty\}$ 非空，即

$$\{X_n = 0, \forall n \geq 1\} \neq \emptyset$$

的可能。也就是说，在严格意义上，此处 τ 的取值空间不是 \mathbb{N} （或 \mathbb{R} ），而应把 τ 视作 $\tau : \Omega \rightarrow \bar{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$ （或更一般的 $\tau : \Omega \rightarrow \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ ）。

由此我们给出随机变量定义的如下“补丁程序”：

定义 5.1.3. 设 $(\Omega, \mathcal{F}, \mathbb{P})$ 为概率空间。以 $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ 表示广义实数空间。概率空间上的广义实值函数

$$\xi : \Omega \rightarrow \bar{\mathbb{R}}$$

称为是可测的，如果对任意 $x \in \mathbb{R}$ ， $\{\xi \leq x\} \in \mathcal{F}$ 。此时广义实值可测函数 ξ 也称为广义实值随机变量；当它额外满足

$$\mathbb{P}(|\xi| = \infty) = 0$$

时，我们才称 ξ 是随机变量。

随机向量的补丁程序完全类似。

在上述补丁程序的精神下，两个随机变量 ξ, η 认为是同一个随机变量，或者说， $\xi = \eta$ 几乎处处成立，如果 $\mathbb{P}(\xi \neq \eta) = 0$ ，亦即 $\mathbb{P}(\xi = \eta) = 1$ 。于是容易看到，概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上的随机变量的全体构成一个代数空间（线性空间且乘法运算封闭）；并且当随机变量 ξ, η 满足 $\mathbb{P}(\eta = 0) = 0$ 时， ξ/η 也是随机变量。在第 3 章例 3.4 对问题 3 的回答“到 12 点钟整的时刻，坛子里一个球也没有”的理解也遵循了这个补丁程序的精神。

本书在前言中就已明确，（公理化的）概率空间这一套语言是我们对随机现象的数学建模方法。“反复投掷一枚均匀硬币，硬币首次出现正面时的投掷次数”（例 5.1 中的 τ ）是一个自然语言定义的非常明确的“随机变量”，两个人对此进行的概率空间建模完全可以不同，但应确保这一“随机变量”在不同的概率建模中本质是相同的，从而对这一“随机变量”描写的相关随机事件的概率解答应该是一致的，这就是上述补丁程序的实质精神。这种精神在数学期望的计算公式的推导过程（见第 6 章的 §6.3）中有集中体现。这就像我们对空间建立坐标系一样，一个明确的空间点在不同坐标系下可以有不同的坐标表示，但不同坐标系下描写同一个运动的方程之间应该有一种协调的性质。

不难知道，一维随机变量的分布函数具有如下性质。

命题 5.1.1. 设 F 是随机变量 ξ 的分布函数，即 $F(x) := \mathbb{P}(\xi \leq x)$ 。那么

- (i) $F: \mathbb{R} \rightarrow \mathbb{R}$ 是单调递增函数；
- (ii) F 是右连续函数；
- (iii) $F(-\infty) := \lim_{x \rightarrow -\infty} F(x) = 0, F(+\infty) := \lim_{x \rightarrow +\infty} F(x) = 1$ 。

随机变量（或随机向量）的分布函数（或联合分布函数）是随机变量（或随机向量）的统计特征，被认为是对应随机变量（或随机向量）的分布律的一种表现形式；分布测度也是分布律的一种表现形式；本章将介绍的离散型随机变量的概率分布列、第 8 章连续型随机变量/向量的密度函数/联合密度函数，以及第 11 章介绍的特征函数都是分布律的表现形式。这些分布律的表现形式都能决定对应随机变量的统计特征（本质上就是随机变量的分布测度）。在附录 C 中，进一步介绍了矩问题和 Laplace 变换，其目的也在于介绍特殊情况下用来刻画分布律的其他一些非常规的手段。

注记 5.1. 随机变量 ξ 的分布函数 F 与它的分布测度 μ_ξ 之间是一一对应的。事实上，给定分布函数 F ，可以定义

$$\mu_F((a, b]) := F(b) - F(a), \text{ 其中 } a < b, a, b \in \mathbb{R}. \quad (5.2)$$

上式进一步通过 Carathéodory 扩张的方式唯一确定了 $(\mathbb{R}, \mathcal{B})$ 上的一个概率测度 μ_F ，参见第 4 章。反过来，也容易看到

$$F(x) = \mu_F((-\infty, x]), \text{ 其中 } x \in \mathbb{R}.$$

当 $\xi \sim F$ 时， $\mu_\xi = \mu_F$ 。

为了方便，本书约定（随机）向量都指（随机）列向量，除非特殊申明。欧氏空间中两向量 x, y 的标准内积我们也简单的以 $x \cdot y$ 表达；有时也用 $\langle x, y \rangle$ 或 $x^T y = y^T x$ （其中 x^T 表示 x 的转置）表达。

当 (E, \mathcal{B}_E) 为一般的可测空间时，可测映射 $\xi: (\Omega, \mathcal{F}) \rightarrow (E, \mathcal{B}_E)$ 有时也称为**随机元**，它等价于要求： $\xi^{-1}(A) := \{\xi \in A\} \in \mathcal{F}, \forall A \in \mathcal{B}_E$ （这也可以概括为 $\xi^{-1}(\mathcal{B}_E) \subset \mathcal{F}$ ；同样， $\sigma(\xi) := \xi^{-1}(\mathcal{B}_E)$ 称为 ξ 生成的 σ -代数）。

注记 5.2. 概率论中使用随机变量这一术语代替可测函数这一术语的原因，编者揣测有两个：其一，可测函数的概念容易让人联想到定义域和映射法则，而概率论中的随机变量通常不知道明确的定义域（因为概率论中经常不明确说出样本空间的构造），通常也更不清楚映射法则；其二，可测函数容易让人联想到确定性，而概率论中的随机变量通常仅仅可以知道统计规律，通常无法确定、也无需关心其映射法则，具有不确定性。

5.1.2 随机变量/向量的相互独立性

在第 3 章中我们发展了事件的相互独立性概念；现在我们引进随机变量后，就可以完全类似地发展随机变量/向量的相互独立性概念。

定义 5.1.4. 设 X, Y 是两个随机变量/向量。称 X, Y **相互独立**，如果

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y), \forall x, y \quad (5.3)$$

成立。上式中，在 X 是随机向量的场合， x 也理解为对应维度的向量，同时事件 $\{X \leq x\}$ 被理解为“ X 的各维分量对应小于等于 x 的各维分量同时发生”这样一个复合事件。

类似于事件列的相互独立性，随机变量列/向量列 $\{X_n\}_{n=1}^N$ 称为**相互独立**的，如果对任意 $r \in \mathbb{N}, 2 \leq r \leq N$ ，以及任意的 $1 \leq i_1 < \cdots < i_r \leq N$

$$\mathbb{P}(X_{i_1} \leq x_1, \cdots, X_{i_r} \leq x_r) = \mathbb{P}(X_{i_1} \leq x_1) \cdots \mathbb{P}(X_{i_r} \leq x_r), \forall x_1, \cdots, x_r$$

成立。但此处容易知道，当 $N < \infty$ 时，随机变量列/向量列 $\{X_n\}_{n=1}^N$ **相互独立**，当且仅当

$$\mathbb{P}(X_1 \leq x_1, \cdots, X_N \leq x_N) = \prod_{k=1}^N \mathbb{P}(X_k \leq x_k), \forall x_1, \cdots, x_N. \quad (5.4)$$

如果两个随机变量或随机向量 X, Y 具有相同的分布，我们就称它们**同分布**，记作 $X \stackrel{d}{=} Y$ 。

设随机变量列/向量列 $\{X_n\}_{n=1}^N$ 相互独立，如果它们具有相同的分布函数（或联合分布函数） F ，即 $X_n \sim F, \forall n$ ，则称随机变量列/向量列 $\{X_n\}_{n=1}^N$ **独立同分布于 F** ，记作 $\{X_n\}_{n=1}^N \stackrel{\text{i.i.d.}}{\sim} F$ 。在统计学中此时把 $\{X_n\}_{n=1}^N$ 称为来自分布 F 的（ N 个）**简单样本**，同时把 N 称为**样本量**。

我们有下面近乎显然、但非常实用的推论。

◆ 推论 5.1.1. 设 $X \stackrel{d}{=} Y$ ，那么 $\varphi(X) \stackrel{d}{=} \varphi(Y)$ ，其中 φ 为使 $\varphi(X), \varphi(Y)$ 有意义的可测函数。

我们指出，随机变量列/向量列 $\{X_n\}_{n=1}^N$ 相互独立，等价于子 σ -代数列 $\{\sigma(X_n)\}_{n=1}^N$ 相互独立，其中 $\sigma(X)$ 表示 X 生成的 σ -代数。请参见第 3 章中定义 3.2.2。

下面的定理给出了随机变量/向量独立性的刻画与重要性质。

☞ 定理 5.1.1. 设 X, Y 是两个随机变量/向量，其中 $X \sim \mu_X, Y \sim \mu_Y$ 。则

(1) X, Y 相互独立, 当且仅当

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B), \forall A, B \text{ 可测}. \quad (5.5)$$

这也等价于

$$(X, Y) \sim \mu_X \times \mu_Y; \quad (5.6)$$

(2) X, Y 相互独立, 且 φ, ψ 是可测函数, 则 $\varphi(X), \psi(Y)$ 也相互独立。

证明: 为证明书写的方便, 我们假定 X, Y 是一维随机变量; 随机向量情形的证明完全类似。

先证明 (1) 的必要性部分。当 X, Y 相互独立时, (5.3) 成立, 由此基于附录中的单调类方法可以得到

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B), \forall A, B \in \mathcal{B}.$$

此即 (5.5)。这也意味着 (X, Y) 的联合分布测度 $\mu_{X,Y}$ 满足:

$$\mu_{X,Y}(A \times B) = \mu_X(A) \cdot \mu_Y(B) = \mu_X \times \mu_Y(A \times B), \forall A, B \in \mathcal{B}.$$

因此 $\mu_{X,Y} = \mu_X \times \mu_Y$ 。此即 (5.6)。

再证明 (1) 的充分性部分。假设 (5.5) 成立。在 (5.5) 中特殊取

$$A = (-\infty, x], B = (-\infty, y],$$

由 x, y 任意性, 立即得到 X, Y 相互独立。

现在假设 $(X, Y) \sim \mu_X \times \mu_Y$, 那么对任意 $x, y \in \mathbb{R}$

$$\begin{aligned} \mathbb{P}(X \leq x, Y \leq y) &= \mathbb{P}((X, Y) \in (-\infty, x] \times (-\infty, y]) \\ &= \mu_X \times \mu_Y((-\infty, x] \times (-\infty, y]) \\ &= \mu_X((-\infty, x]) \cdot \mu_Y((-\infty, y]) \\ &= \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y). \end{aligned}$$

因此 X, Y 相互独立。

现在来证明 (2)。当 X, Y 相互独立时我们知道, 对任意 $A, B \in \mathcal{B}$, 利用 (1) 中的刻画, 对任意 $x, y \in \mathbb{R}$

$$\begin{aligned} \mathbb{P}(\varphi(X) \leq x, \psi(Y) \leq y) &= \mathbb{P}(X \in \varphi^{-1}((-\infty, x]), Y \in \psi^{-1}((-\infty, y])) \\ &= \mathbb{P}(X \in \varphi^{-1}((-\infty, x])) \cdot \mathbb{P}(Y \in \psi^{-1}((-\infty, y])) \\ &= \mathbb{P}(\varphi(X) \leq x) \cdot \mathbb{P}(\psi(Y) \leq y). \end{aligned}$$

因此 X, Y 相互独立。 □

5.2 离散型分布

5.2.1 离散型随机变量的定义

定义 5.2.1. 设 ξ 是一个随机变量/向量; 如果 a 使得 $\mathbb{P}(\xi = a) > 0$, 就称 a 是 ξ 的 (分布的) 原子。如果存在可数集 $D = \{a_i\}_{i=1}^N$ (此处通常认为此集合为有序集, 或者说是一个数列; 此处允许 $N = \infty$), 使得 D 是 ξ 的所有原子所成集合, 并且

$$\mathbb{P}(\xi \in D) = 1,$$

就称 ξ 是离散型随机变量/向量。此时，记

$$p_i := \mathbb{P}(\xi = a_i) > 0,$$

称数列 $\{p_i\}_{i=1}^N$ （或点列 $\{(a_i, p_i)\}_{i=1}^N$ ）为 ξ 的概率分布列，并记

$$\xi \sim \begin{pmatrix} a_1 & \cdots & a_N \\ p_1 & \cdots & p_N \end{pmatrix}.$$

显然上面的概率分布列 $\{p_i\}_{i=1}^N$ 应满足*

$$\sum_{i=1}^N p_i = 1. \quad (5.7)$$

有时我们也用下面的图表来表达上述分布律：

$$\begin{array}{c|ccc} \xi & a_1 & \cdots & a_N \\ \hline \mathbb{P} & p_1 & \cdots & p_N \end{array}$$

对于两个离散型随机变量/向量 ξ, η ，设

$$p_{i,j} := \mathbb{P}(\xi = a_i, \eta = b_j), i = 1, \cdots, M, j = 1, \cdots, N,$$

其中 $\{a_i\}_{i=1}^M$ 互不相同， $\{b_j\}_{j=1}^N$ 也互不相同，并且

$$\sum_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N}} p_{i,j} = 1,$$

则称 $\{(a_i, b_j; p_{i,j}) : 1 \leq i \leq M, 1 \leq j \leq N\}$ 为随机向量 (ξ, η) 的联合概率分布列，此处一般要求

$$p_{i,\cdot} := \mathbb{P}(\xi = a_i) = \sum_{j=1}^N p_{i,j} > 0, \quad p_{\cdot,j} := \mathbb{P}(\eta = b_j) = \sum_{i=1}^M p_{i,j} > 0.$$

有时也用下面的图表来表达上述 (ξ, η) 的分布律：有时进一步在上述表格的

$$\begin{array}{c|ccc} \xi \backslash \eta & b_1 & \cdots & b_N \\ \hline a_1 & p_{1,1} & \cdots & p_{1,N} \\ a_2 & p_{2,1} & \cdots & p_{2,N} \\ \vdots & \vdots & \cdots & \vdots \\ a_M & p_{M,1} & \cdots & p_{M,N} \end{array}$$

每行、每列的末端分别配置上对应的 $p_{i,\cdot}, p_{\cdot,j}$ 值，这种图表就称为列联表，如下所示：

当 ξ 是离散型随机变量或离散型随机向量时，我们也称它的（联合）分布函数和（联合）分布测度是离散型的。

在上述定义中，如果 D 是单点集 $\{a\}$ ，即 $\mathbb{P}(\xi = a) = 1$ ，则称 ξ 服从单点分布或退化分布，对应分布测度记作 δ_a ，称为 a 点处的 Dirac 测度；此时可记 $\xi \sim \delta_a$ ，这样的随机变量也被称为退化随机变量。当 D 是两点集 $\{a, b\}$ ，

*在不追求严谨性的情况下，说数列 $\{p_i\}_{i=1}^N$ 是概率分布列，如果这个数列是非负的，且满足(5.7)，即容许某些 $p_i = 0$ 。联合概率分布列情形也做类似理解。

| $\xi \backslash \eta$ | b_1 | \cdots | b_N | |
|-----------------------|---------------|----------|---------------|---------------|
| a_1 | $p_{1,1}$ | \cdots | $p_{1,N}$ | $p_{1,\cdot}$ |
| a_2 | $p_{2,1}$ | \cdots | $p_{2,N}$ | $p_{2,\cdot}$ |
| \vdots | \vdots | \cdots | \vdots | \vdots |
| a_M | $p_{M,1}$ | \cdots | $p_{M,N}$ | $p_{M,\cdot}$ |
| | $p_{\cdot,1}$ | \cdots | $p_{\cdot,N}$ | |

即 $\mathbb{P}(\xi \in \{a, b\}) = 1$ 且 $\mathbb{P}(\xi = a) > 0, \mathbb{P}(\xi = b) > 0$ 时，我们称 ξ 服从**两点分布**。

对于离散型随机变量/向量之间的独立性，我们有下面更简单的利用概率分布列的判据，这使得在离散型随机变量的相互独立性的判断中列联表非常有用；有关证明留给读者。

◆ **推论 5.2.1.** 设 X, Y 都是离散型随机变量/向量，那么它们相互独立当且仅当

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y), \quad \forall x, y.$$

5.2.2 常见离散型分布

以下我们介绍一些常见的离散型分布。

例 5.2. (离散型均匀分布) 给定 $N \in \mathbb{N}$ 且 $N \geq 2$ 。设 $X \sim \begin{pmatrix} a_1 & \cdots & a_N \\ \frac{1}{N} & \cdots & \frac{1}{N} \end{pmatrix}$ 。我们称 X 服从 $\{a_1, \dots, a_N\}$ 上的**离散均匀分布**，记作 $X \sim U(\{a_1, \dots, a_N\})$ 。

例 5.3. (0-1 两点分布) 设 $A \in \mathcal{F}$ ，且 $p := \mathbb{P}(A) \in [0, 1]$ ；记 $q := 1 - p$ 。则 $\xi := 1_A$ 是一个随机变量，并且

$$\mathbb{P}(\xi = 1) = p, \mathbb{P}(\xi = 0) = q.$$

(1) 当 $p = 0$ 或 $p = 1$ 时， ξ 的分布是退化的（单点分布）， $\xi \sim \delta_p$ ；

(2) 当 $p \in (0, 1)$ 时， ξ 的分布也称为**0-1 两点分布**，记作 $\xi \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$ ，有时也记作 $\xi \sim B(1, p)$ ，称为 *Bernoulli* 两点分布。

以下给定 $p \in (0, 1)$ ，此时总记 $q := 1 - p$ 。

例 5.4. (Bernoulli 二项分布) 给定 $n \in \mathbb{N}$ ，设 $\{X_k\}_{k=1}^n \stackrel{\text{i.i.d.}}{\sim} B(1, p)$ ，那么 $S_n := X_1 + \cdots + X_n$

满足

$$\mathbb{P}(S_n = k) = C_n^k p^k q^{n-k}, \quad k = 0, 1, \dots, n. \quad (5.8)$$

因为

$$1 = (p + q)^n = \sum_{k=0}^n C_n^k p^k q^{n-k},$$

所以 S_n 的分布律称为（Bernoulli）**二项分布**，记作 $S_n \sim B(n, p)$ 。设某项实验成功的概率为 p ， S_n 可以解释为该实验重复做 n 次后，实验成功的总次数。

例 5.5. (几何分布) 现在设 $\{X_k\}_{k=1}^{\infty} \stackrel{\text{i.i.d.}}{\sim} B(1, p)$, 定义

$$\tau := \inf\{n \geq 1 : X_n = 1\}.$$

那么 τ 满足

$$\mathbb{P}(\tau = k) = pq^{k-1}, k = 1, 2, \dots. \quad (5.9)$$

τ 的分布律称为几何分布, 记作 $\tau \sim \text{Geo}(p)$ 。设某项实验成功的概率为 p , τ 可以解释为: 为获得一次成功的实验所需重复的实验次数。

例 5.6. (Pascal 分布与负二项分布) 给定 $r \in \mathbb{N}$ 。现在设

$$\{\tau_k\}_{k=1}^{\infty} \stackrel{\text{i.i.d.}}{\sim} \text{Geo}(p),$$

定义 $N_r := \tau_1 + \dots + \tau_r$ 。那么 N_r 满足

$$\mathbb{P}(N_r = \ell) = C_{\ell-1}^{r-1} p^r q^{\ell-r}, \ell = r, r+1, \dots. \quad (5.10)$$

N_r 的分布律称为 Pascal 分布, 记作 $N_r \sim \text{Pascal}(r, p)$ 。设某项实验成功的概率为 p , N_r 可以解释为: 为获得累计 r 次成功的实验所需重复的实验次数。

现在令 $\tilde{N}_r = N_r - r$, 那么 \tilde{N}_r 满足

$$\mathbb{P}(\tilde{N}_r = \ell) = C_{r+\ell-1}^{\ell} p^r q^{\ell}, \ell = 0, 1, \dots. \quad (5.11)$$

对任意实数 α 与非负整数 ℓ , 引进记号 $C_{\alpha}^{\ell} := \frac{\alpha(\alpha-1)\dots(\alpha-\ell+1)}{\ell!}$, 由于

$$C_{r+\ell-1}^{\ell} p^r q^{\ell} = C_{-r}^{\ell} p^r (-q)^{\ell},$$

\tilde{N}_r 的分布律称为负二项分布, 记作 $\tilde{N}_r \sim \text{NB}(r, p)$ 。设某项实验成功的概率为 p , \tilde{N}_r 可以解释为: 为获得累计 r 次成功的实验所需忍受的重复实验失败次数。

例 5.7. (超几何分布) 我们用 Polyá 的坛子模型来解释本例将介绍的超几何分布: 从有 b 个黑球、 r 个红球的坛子里无放回地取 $n \leq b+r$ 个球所获得的黑球数量记作 X_n , 则 X_n 的概率分布列为

$$\mathbb{P}(X_n = k) = \frac{C_b^k C_r^{n-k}}{C_{b+r}^n}, \ell_1 := (n-r)^+ \leq k \leq \ell_2 := b \wedge n. \quad (5.12)$$

此时, 我们记 $X_n \sim H(b, r; n)$, 称之为超几何分布。

注意到例 2.4, 当 $p = \frac{b}{b+r}$ 时, 从有 b 个黑球、 r 个红球的坛子里有放回地取 $n \leq b+r$ 个球所获得的黑球数量记作 X , 则 $X \sim B(n, p)$ 。对上例中的 X_n , 如果 $p \in (0, 1)$, $b \rightarrow \infty, r \rightarrow \infty$ 且 $\frac{b}{b+r} \rightarrow p$, 则对固定的 $0 \leq k \leq n$

$$\mathbb{P}(X_n = k) = \frac{C_b^k C_r^{n-k}}{C_{b+r}^n} \rightarrow C_n^k p^k (1-p)^{n-k}.$$

因此上例中的分布被称为超几何分布纯粹是历史原因 (源于 Euler 对超几何级数的研究), 它本质上是与 Bernoulli 二项分布对应的 Polyá 坛子模型无放回抽样版本。

下例中的分布是与 Pascal 分布对应的 Polyá 坛子模型无放回抽样版本。

例 5.8. (Pascal 分布的 Polyá 的坛子模型无放回抽样版本) 在有 b 个黑球、 r 个红球的坛子里无放回地取球。为了取到 $s \leq b$ 个黑球所需的最少取球次数记作 Y_s , 则 Y_s 的概率分布列为

$$\mathbb{P}(Y_s = n) = \frac{s}{n} \cdot \frac{C_b^s C_r^{n-s}}{C_{b+r}^n}, s \leq n \leq r+s. \quad (5.13)$$

例 5.9. (Poisson 小数定理与 Poisson 分布) 设随机变量列 $\{X_n\}_{n=1}^{\infty}$ 满足 $X_n \sim B(n, p_n)$, 其中 p_n 满足 $\lim_{n \rightarrow +\infty} np_n = \lambda \in (0, \infty)$, 那么

$$\lim_{n \rightarrow +\infty} \mathbb{P}(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots. \quad (5.14)$$

上述结论称为 *Poisson 小数定理*。

$\{\frac{\lambda^k}{k!} e^{-\lambda}\}_{k=0}^{\infty}$ 是一个概率分布列, 对应分布称为参数为 $\lambda > 0$ 的 *Poisson 分布*, 记作 $\text{Poisson}(\lambda)$ 。于是随机变量 $Y \sim \text{Poisson}(\lambda)$ 时

$$\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots. \quad (5.15)$$

注记 5.3. 现实世界中许多量的分布规律近似服从 *Poisson 分布*。

例如, 1910 年著名科学家 *Rutherford* (卢瑟福) 和 *Geiger* (盖革) 观察了放射性物质钋 (*Polonium*) 放射 α 粒子的情况; 他们进行了 $N = 2608$ 此观测, 每次观察 7.5 秒, 一共观测到 10094 个 α 粒子, 下面的表格就是他们的实验数据的拟合情况 (表格来自于 [40, 第 77 页])

表 5.1: *Rutherford* 与 *Geiger* 放射性粒子数观察频率与拟合概率

| 放射粒子数 X | 观察到次数 n_k | 频率 $\nu_k = \frac{n_k}{N}$ | 拟合概率 (参数 $\lambda=3.87$) $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ |
|--------------|----------------|-------------------------------|--|
| 0 | 57 | 0.022 | 0.021 |
| 1 | 203 | 0.078 | 0.081 |
| 2 | 383 | 0.147 | 0.156 |
| 3 | 525 | 0.201 | 0.201 |
| 4 | 532 | 0.204 | 0.195 |
| 5 | 408 | 0.156 | 0.151 |
| 6 | 273 | 0.105 | 0.097 |
| 7 | 139 | 0.053 | 0.054 |
| 8 | 45 | 0.017 | 0.026 |
| 9 | 27 | 0.010 | 0.011 |
| 10 | 10 | 0.004 | 0.004 |
| ≥ 11 | 6 | 0.002 | 0.003 |
| 总计 | $N=2608$ | 0.999 | 1.000 |

统计数据表明, 某段高速公路上一年的交通事故数、某市场一天中到达的顾客次数、某办公室一天中收到的电话数等都近似服从 *Poisson 分布*。

又如, 有人统计了 1500-1931 年共 $N = 432$ 年间的战争, 一年中爆发战争的频率与 *Poisson 分布* 的概率接近, 下面的表格就是这些数据的拟合情况 (参见 [45, 第 63 页])

5.2.3 两个极大似然估计的例子

下面我们举两个对离散型分布的分布参数做极大似然估计的例子。对于其他一些离散型分布族, 读者掌握了极大似然估计的思想后, 也可以思考应当如何进行它们的分布参数的极大似然估计。

表 5.2: 1500–1931 年间发生的重要战争年度频次与概率拟合

| 年度战争次数 X | 观察到的年数 n_k | 频率 $\nu_k = \frac{n_k}{N}$ | 拟合概率（参数 $\lambda=0.69$ ） $\mathbb{P}(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$ |
|---------------|-----------------|-------------------------------|---|
| 0 | 223 | 0.516 | 0.502 |
| 1 | 142 | 0.329 | 0.346 |
| 2 | 48 | 0.111 | 0.119 |
| 3 | 15 | 0.035 | 0.028 |
| ≥ 4 | 4 | 0.009 | 0.005 |
| 总计 | N=432 | 1.000 | 1.000 |

例 5.10.（频率代替概率的合理性） 设某项实验成功的概率为 p , S_n 为该实验重复做 n 次后，实验成功的总次数，于是 n 次实验成功的频率为

$$f_n := \frac{S_n}{n}.$$

我们知道, $S_n \sim B(n, p)$, 亦即

$$\mathbb{P}(S_n = k) = C_n^k p^k (1-p)^{n-k}.$$

于是 $\mathbb{P}(S_n = k)$ 作为 p 的函数的最大值点 \hat{p} 为

$$\hat{p} = \arg \max_{p \in [0,1]} C_n^k p^k (1-p)^{n-k} = \frac{k}{n}.$$

也就是说, 当 $S_n = k$ 时, 亦即知道该实验重复做 n 次总共成功 k 次时, 我们对实验成功概率 p 的一种合理估计为 $\hat{p} = \frac{k}{n} = f_n$. 上述估计方法就是所谓的极大似然估计方法。这个例子说明, 在实际生活中, 基于极大似然估计的思想, 使用实验成功的频率作为实验成功概率的估计是有道理的; 后文中给出的 *Bernoulli* 弱大数律 (以及 *Kolmogorov* 强大数律) 进一步说明, 在大样本下这样的估计值能逼近真实的参数值。至于这样做估计的误差对应的近似分布/极限分布, 可由后文的中心极限定理给出。

例 5.11.（估算池塘里的鱼） 现有一个池塘, 内有若干成鱼; 假设共 N 条, 其中 N 未知待估计。常用的估计方法如下: 首先从池塘中捕捞出 M 条成鱼, 都做好标记, 并重新放回池塘。过几天后 (假定期间成鱼数量不改变, 且之前作的标记不会脱落), 再次进行捕捞, 捞出了 n 条成鱼, 其中有 $X = m$ 条是之前做过标记的成鱼。那么统计学家认为, 如下定义的 \hat{N} 就是一个合理的估计量:

$$\hat{N} := \frac{nM}{X}.$$

其逻辑是: 此时 $X \sim H(M, N-M; n)$, $X = m$ 的概率值

$$\mathbb{P}(X = m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}$$

作为 N 的函数, 最大值点 \hat{N} 就在点 $\frac{nM}{m}$ 附近, 亦即

$$\hat{N} = \arg \max_N \mathbb{P}(X = m) = \frac{nM}{m}.$$

请读者自行论证这一点。由此 $\hat{N} := \frac{nM}{X}$ 就是此模型下的极大似然估计。

注记 5.4. 历史上, 上述巧妙的估计方法最早是 *Laplace* 在 1786 年为了估计法国人口时提出的。上述统计方法经后人发展成为所谓的“捕获再捕获抽样”理论, 广泛应用于生态学等领域。上述方法, 本质上是二重抽样; 只不过第一重抽样的目的是给子样本“贴标签”再放回总样本中。由于有可能二次抽样时没有获得贴有标签的子样本, 即 $X = 0$, 从而统计方法失效, 人们提出所谓的“逆抽样”的方法: 在第二次抽样时, 并不对 n 的数量作出规定, 而是一直抽到曾做过记号的单元数达到 s (它是预先给定的值) 个为止。假设第二次抽样的样本量为 n (它是随机的, 分布律参见例 5.8), 此时可以同样用估计量 $\hat{N} := \frac{nM}{s}$ 作为总体样本量的估计, 该估计仍然是极大似然估计。

习 题 5

习题 5.1. 试证明: 一个连续的分布函数 $F: \mathbb{R} \rightarrow \mathbb{R}$ 是一致连续的。

习题 5.2. 设 $\xi \sim F$, 其中 $F: \mathbb{R} \rightarrow \mathbb{R}$ 是分布函数。那么 F 在 $x = a$ 处连续的充分必要条件是 $\mathbb{P}(\xi = a) = 0$ 。

习题 5.3. 设 $X_k \sim F_k, k = 1, 2$, 又设 $(X_1, X_2) \sim F$ 是联合分布函数。那么 F 是二元连续函数的充分必要条件是边缘分布函数 F_1, F_2 都是连续函数。

习题 5.4. 设 $D \subset \mathbb{R}$ 是一个稠密子集, F 是 \mathbb{R} 上分布函数, 求证: F 由 $F_D := F|_D$ 唯一决定。

习题 5.5. 设有随机向量 (X, Y) , 若存在 \mathbb{R} 上函数 F 使得对任何 x, y , 有 $\mathbb{P}(X \leq x, Y \leq y) = F(x \wedge y)$, 求证: $X = Y$ a.s.。

习题 5.6. 设 $X \sim B(1, p), p \in (0, 1)$ 。求 X 的分布函数 F , 并计算 $Y := F(X)$ 的分布。

习题 5.7. 设有两个分布函数 F_0, F_1 以及一个常数 $p \in (0, 1)$ 。再设有随机变量 $X_0 \sim F_0, X_1 \sim F_1, I \sim B(1, p)$, 并且此三个随机变量相互独立。试求证: $X_I \sim (1-p)F_0 + pF_1$ 。

习题 5.8. 证明几何分布有下面的无记忆性: 设 $X \sim \text{Geo}(p)$, 则

$$\mathbb{P}(X > n + m | X > n) = \mathbb{P}(X > m), \forall n, m \in \mathbb{Z}_+.$$

§ 6

数学期望

数学期望是概率论中基于随机变量概念发展的一套重要工具。它本质上与概率空间中的概率测度是等价的，但性质更为优良：数学期望作为算子是线性的；而概率测度最重要的性质是 σ -可加性，这是非线性的。因此数学期望这一工具引入到概率论中后，更多的数学中的分析技巧得以借此途径施展，从而大大促进了概率论的发展。

在本章，我们通过回顾历史上 Huygens 的贡献，借助期望收益的直观，首先对离散型随机变量定义了数学期望，之后通过借助实变函数论中 Lebesgue 积分的定义流程框架，实现对一般的随机变量的数学期望的定义。最后我们进一步借助保测映射的观点，证明了通常大部分初等概率论教材中给出的随机变量的数学期望的计算公式。

这样一套定义与说理是略显冗长的，但在目前的学科发展形势下来看，对于概率论专业的学生是有了了解的必要的；对于非概率论专业的学生，也可以在学习时跳过此定义流程，（或对前述定义流程略作了解后）直接采纳最终导出的计算公式。

6.1 数学期望的定义

在第 4 章，我们已经定义了一般测度空间上的可测函数的抽象 Lebesgue 积分；由于随机变量 ξ 是概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中的可测函数，很自然也就定义了积分

$$\int \xi d\mathbb{P},$$

在概率论中更习惯把它称作 ξ 的**数学期望**，同时把它书写为 $\mathbb{E}\xi$ ，亦即

$$\mathbb{E}\xi = \int \xi d\mathbb{P}. \quad (6.1)$$

但为了读者的方便，我们在本章还是通过回顾概率论发展的历史重新给出随机变量的数学期望的定义，并通过逻辑演绎基于随机变量的分布函数给出随机变量的数学期望计算公式。

历史上，Pascal 和 Fermat 在 1654 年左右进行了若干次（据说是七次）通信，讨论了著名的赌金分配问题；在那个问题的讨论中，Pascal 和 Fermat 确

立了分配赌金的原则：赌徒应当按照最终获胜的概率的比值来分配赌金。尽管 Pascal 和 Fermat 并没有给出概率的定义，但在他们的讨论过程中实质上暗含了等可能性假设下的古典概率模型的概率定义。在 1655 年，正在巴黎游学的青年时期的 Huygens 听说了他们关于赌金分配问题的讨论，也开始研究这个问题，最终于 1657 年发表了《论赌博中的计算》一书；在此书中，Huygens 进一步把 Pascal 和 Fermat 的分配赌金的原则推广为：赌徒应当按照各自期望收益的多少来进行赌金的分配。这个原则相比 Pascal 和 Fermat 给出的原则，适用范围大大拓广，可以不再局限于原始的赌金分配问题所限定的赌博类型。我们今天要讲的数学期望，本质上就来源于 Huygens 给出的期望收益的定义。

仍然局限于原始的赌金分配问题，其中涉及的两个赌徒为方便分别称为甲、乙赌徒。用今天的术语来讲，设随机变量 ξ 代表甲赌徒在继续赌下去直至决出最后的赢家时的真实收益，设 A 代表甲赌徒是最终赢家这一事件，用 K 代表最终约定分配的总赌金，由原始的赌金分配问题中的规则，显然应有

$$\xi = K \cdot 1_A \sim \begin{pmatrix} 0 & K \\ 1 - \mathbb{P}(A) & \mathbb{P}(A) \end{pmatrix}.$$

朴素的直觉告诉我们（这也同时是 Huygens 选择的定义），此时甲赌徒的期望收益（有时也称为平均收益）为

$$\mathbb{E}\xi := K\mathbb{P}(A).$$

也就是说，Huygens 的期望收益分配原则与 Pascal、Fermat 的最终获胜概率分配原则在原始的赌金分配问题中本质上是等价的。但这里蕴含了我们定义数学期望的根基与出发点：对任意可测集 $A \in \mathcal{F}$ ， 1_A 是一个随机变量，它的数学期望定义为

$$\mathbb{E}1_A := \mathbb{P}(A). \quad (6.2)$$

我们将在承认算子 \mathbb{E} 的线性性质的基础上，结合上面的方程给出一般的随机变量的数学期望的定义。

尽管赌博是不好的行为，我们还是选择借用这个概念的外衣来进行解释。在选择把随机变量 ξ 解释为随机世界里你参与某项赌博结束时的真实收益，把 $\mathbb{E}\xi$ 解释为你参与这项赌博的期望收益后，数学期望算子 \mathbb{E} 的线性性质就是我们的直观就能理解的了：第一，原则上你可以参加多项赌博，整体期望收益应该是各项赌博期望收益的加和：

$$\mathbb{E}[\xi_1 + \xi_2] = \mathbb{E}\xi_1 + \mathbb{E}\xi_2;$$

第二，原则上你可以使用杠杆放大原始真实收益的倍数，相应的期望收益也应当会放大相应的倍数：

$$\mathbb{E}[k\xi] = k\mathbb{E}\xi, \text{ 其中 } k \text{ 为常数.}$$

在上述理解的基础上，我们按以下流程来实现随机变量的数学期望 $\mathbb{E}\xi$ 的定义：

- **Step 1.** 对任何非负简单随机变量 ξ （即非负简单函数，亦即只取有限个非负实值的离散型随机变量；这种随机变量的全体将记作 \mathcal{S}_+ ），存

在两两不同的非负实数列 $\{a_i\}_{i=1}^N$, 使得

$$\xi = \sum_{i=1}^N a_i 1_{\{\xi=a_i\}},$$

此时, 基于(6.2), 定义

$$\mathbb{E}\xi := \sum_{i=1}^N a_i \mathbb{P}(\xi = a_i).$$

不难验证, 在此步骤下, \mathbb{E} 具有线性性质: $\forall \xi, \eta \in \mathcal{S}_+, c \in \mathbb{R}_+$

$$\mathbb{E}[c\xi] = c\mathbb{E}\xi, \quad \mathbb{E}[\xi + \eta] = \mathbb{E}\xi + \mathbb{E}\eta;$$

- **Step 2.** 对于一般的非负随机变量 ξ , 定义

$$\mathbb{E}\xi := \sup\{\mathbb{E}\eta : 0 \leq \eta \leq \xi, \eta \in \mathcal{S}_+\}.$$

这里需要注意, 由于是使用上确界给出的定义, 有可能出现 $\mathbb{E}\xi = \infty$ 。

- **Step 3.** 对于一般的随机变量 ξ , 注意到

$$\xi = \xi^+ - \xi^-,$$

当 $\mathbb{E}[\xi^+], \mathbb{E}[\xi^-]$ 至少有一者是有限值时, 定义

$$\mathbb{E}\xi := \mathbb{E}[\xi^+] - \mathbb{E}[\xi^-],$$

称之为随机变量 ξ 的数学期望。此时称 ξ 的数学期望有意义。当 $\mathbb{E}[\xi^+], \mathbb{E}[\xi^-]$ 都是有限值时, 此时 $\mathbb{E}\xi$ 也是一个有限值, 称 ξ 的数学期望存在, 有时也称 ξ 可积, 记作 $\xi \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ 或简单的 $\xi \in L^1(\mathbb{P})$, 或更简单的 $\xi \in L^1$ 。当 $\mathbb{E}[\xi^+], \mathbb{E}[\xi^-]$ 都是无穷值时, 称 ξ 的数学期望无意义。

在约定 $0 \cdot \infty = 0$ 后, 上述定义流程对于广义实值随机变量也可以操作。显然, 离散型随机变量的数学期望有下面的计算公式:

定理 6.1.1. 设 $\xi \sim \begin{pmatrix} a_1 & \cdots & a_N \\ p_1 & \cdots & p_N \end{pmatrix}$, 那么当 $\mathbb{E}\xi$ 有意义时,

$$\mathbb{E}\xi = \sum_{i=1}^N a_i p_i. \quad (6.3)$$

特别的, 对任意可测函数 φ , 如果 $\mathbb{E}[\varphi(\xi)]$ 有意义, 那么

$$\mathbb{E}[\varphi(\xi)] = \sum_{i=1}^N \varphi(a_i) p_i. \quad (6.4)$$

6.2 数学期望的性质

数学期望具有很多性质, 以下我们仅罗列它的一些基本性质而不给出证明 (以下 ξ, η 为随机变量), 对有关证明细节感兴趣的读者请参见附录 B:

性质 (0) $\mathbb{P}(A) = \mathbb{E}[1_A], \forall A \in \mathcal{F};$

性质（1）（线性性质）设 $a, b \in \mathbb{R}$ 且 $\mathbb{E}\xi, \mathbb{E}\eta, a\mathbb{E}\xi + b\mathbb{E}\eta$ 均有意义，则

$$\mathbb{E}[a\xi + b\eta] = a\mathbb{E}\xi + b\mathbb{E}\eta;$$

性质（2）（正算子/单调性）如果 $\xi \geq 0$ a.s., 那么 $\mathbb{E}\xi \geq 0$ 。一般的，设 $\mathbb{E}\xi, \mathbb{E}\eta$ 有意义，且 $\xi \leq \eta$ a.s., 那么 $\mathbb{E}\xi \leq \mathbb{E}\eta$ 。特别的， $|\mathbb{E}\xi| \leq \mathbb{E}|\xi|$;

性质（3）如果 $\xi = a \in \bar{\mathbb{R}}$ a.s., 那么 $\mathbb{E}\xi = a$;

性质（4）（Jensen 不等式）设 g 为 \mathbb{R} 上的凸函数， $\mathbb{E}\xi, \mathbb{E}g(\xi)$ 有定义。那么 $g(\mathbb{E}\xi) \leq \mathbb{E}g(\xi)$;

性质（5）（矩不等式）设 ξ 为随机变量， $0 < s < t$ 。那么 $\|\xi\|_s \leq \|\xi\|_t$;

性质（6）（Hölder 不等式）设 $\frac{1}{p} + \frac{1}{q} = 1, p, q > 1$ ，并记 $\|\xi\|_p := (\mathbb{E}|\xi|^p)^{1/p}$ 。那么 $\|\xi \cdot \eta\|_1 \leq \|\xi\|_p \cdot \|\eta\|_q$;

性质（7）（Minkowski 不等式）设 $p > 0$ 。

(i) $p \geq 1$ 时， $\|\xi + \eta\|_p \leq \|\xi\|_p + \|\eta\|_p$;

(ii) $0 < p < 1$ 时， $\|\xi + \eta\|_p^p \leq \|\xi\|_p^p + \|\eta\|_p^p$ 。

6.3 数学期望的计算公式

在第 4 章中我们已经给出了一个可测函数关于一个参考测度的抽象 Lebesgue 积分的概念；在本章我们也给出了在给定的概率空间中一个随机变量的数学期望的抽象定义。但在实际应用中，我们希望抽象定义的积分、特别是上面抽象定义的数学期望能有更为简单、实用的计算方法。本节就是为此目的而展开讨论的。

定义 6.3.1. 设 $(\Omega_i, \mathcal{F}_i, \mu_i), i = 1, 2$ 是两个测度空间，设映射 $T: \Omega_1 \rightarrow \Omega_2$ 是 $\mathcal{F}_1/\mathcal{F}_2$ 可测的（简称可测的，即 $T^{-1}\mathcal{F}_2 \subset \mathcal{F}_1$ ）。在此基础上，称 T 是保测映射（简称保测的），如果 $\mu_2 = \mu_1 \circ T^{-1}$ 。

设 $(\Omega, \mathcal{F}, \mu)$ 是个测度空间，当 $T: (\Omega, \mathcal{F}, \mu) \rightarrow (\Omega, \mathcal{F}, \mu)$ 保测时，称 μ 是 T -不变测度，简称不变测度，或称 T 保测度 μ ，或称 T 为保测变换。此时 $(\Omega, \mathcal{F}, \mu; T)$ 或 (μ, T) 称为保测系统。

注记 6.1. 在第 5 章介绍随机变量/向量概念时，我们提到过边缘分布测度的说法。事实上，设 μ 为 $(\mathbb{R}^n, \mathcal{B}^n)$ 上的（概率分布）测度，设 $n = p + q$ ， $\mathbb{R}^n = \mathbb{R}^p \times \mathbb{R}^q$ 。我们可以定义自然投影

$$\pi_1: \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}^p, (x, y) \mapsto x$$

及

$$\pi_2: \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}^q, (x, y) \mapsto y.$$

那么 $\mu_i := \mu \circ \pi_i^{-1}, i = 1, 2$ 都可以称为 μ 的边缘（分布）测度。

对于第 4 章介绍的抽象 Lebesgue 积分，我们有下面的积分变换公式，有些文献也称为积分换元公式。

定理 6.3.1. (积分变换公式) 设 $(\Omega_i, \mathcal{F}_i, \mu_i), i = 1, 2$ 是两个测度空间，并且 $T : (\Omega_1, \mathcal{F}_1, \mu_1) \rightarrow (\Omega_2, \mathcal{F}_2, \mu_2)$ 是保测的。那么 $\mu_2 = \mu_1 \circ T^{-1}$ ，且

$$\int_{\Omega_1} f \circ T d\mu_1 = \int_{\Omega_2} f d\mu_2 \quad (6.5)$$

对任意的 $f \in L^1(\Omega_2, \mathcal{F}_2, \mu_2)$ 成立。特别的，上式对非负可测函数 f 总成立。

证明. 注意到 T 是保测映射，因此对任意 $A \in \mathcal{F}_2$ ，

$$\int 1_A \circ T d\mu_1 = \mu_1(T^{-1}A) = \mu_2(A) = \int 1_A d\mu_2.$$

于是 (6.5) 对任意（非负）简单函数 $f \in \mathcal{F}_2$ 成立。类似于我们定义积分的流程，立即知道 (6.5) 对任意非负可测函数 $f \in \mathcal{F}_2$ 成立。进而该式对任意的 $f \in L^1(\Omega_2, \mathcal{F}_2, \mu_2)$ 成立。□

注记 6.2. 上面证明中的方法被称为**典型方法**，它在后续很多场合有用。

当 $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i), i = 1, 2$ 都是概率空间（相应数学期望算符分别记作 $\mathbb{E}_1, \mathbb{E}_2$ ），且 $T : (\Omega_1, \mathcal{F}_1, \mathbb{P}_1) \rightarrow (\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ 是保测映射时，我们可以通过下面的交换图

$$\begin{array}{ccc} (\Omega_1, \mathcal{F}_1, \mathbb{P}_1) & \xrightarrow{T} & (\Omega_2, \mathcal{F}_2, \mathbb{P}_2) \\ & \searrow f \circ T & \downarrow f \\ & & (\mathbb{R}, \mathcal{B}) \end{array}$$

来解释上面定理（积分变换公式）的结论：设 f 和 $f \circ T$ 分别是概率空间 $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ 和 $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ 上的能谈数学期望的随机变量；映射 T 的保测性意味着 f 与 $f \circ T$ 在各自概率空间中数学期望是一致的：

$$\mathbb{E}_2[f] = \mathbb{E}_1[f \circ T].$$

用期望收益来理解上式就是很自然的：同一份真实的随机收益，不同的人的概率建模可能不同，但期望收益应该保持一致（否则就说明其中有的人的概率建模不正确）。

给定概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ ；设 ξ 为其上的随机变量，则

$$\xi : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B}, \mathbb{P} \circ \xi^{-1})$$

就是一个保测映射， $\mathbb{P} \circ \xi^{-1}$ 恰为随机变量 ξ 对应的分布（或称分布测度），对应的**分布函数**为

$$F(x) := \mathbb{P}(\xi \leq x) = \mathbb{P}(\xi^{-1}((-\infty, x])) = \mathbb{P} \circ \xi^{-1}((-\infty, x]).$$

此时，随机变量 ξ 的分布测度 $\mathbb{P} \circ \xi^{-1}$ 本质上就是如下方式（借助 Carathéodory 定理）唯一确定的、定义于 \mathbb{R} 上的 Borel 概率测度 μ_F ：

$$\mu_F((a, b]) := F(b) - F(a), \text{ 其中 } -\infty < a < b < \infty.$$

给定一个“好”的可测函数 $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ ，我们可以借助随机变量 $\xi \sim F$ 构建如下交换图：

$$\begin{array}{ccc} (\Omega, \mathcal{F}, \mathbb{P}) & \xrightarrow{\xi} & (\mathbb{R}, \mathcal{B}, \mu_F) \\ & \searrow \varphi \circ \xi = \varphi(\xi) & \downarrow \varphi \\ & & (\mathbb{R}, \mathcal{B}) \end{array}$$

因此积分变换公式告诉我们， $\varphi(\xi) = \varphi \circ \xi$ 的数学期望与 φ 在概率空间 $(\mathbb{R}, \mathcal{B}, \mu_F)$ 中的数学期望一致。在概率论中，概率空间 $(\mathbb{R}, \mathcal{B}, \mu_F)$ 中的一个可测函数 φ （视作随机变量）的数学期望，也就是 φ 关于测度 μ_F 的积分，它可以写作

$$\int \varphi d\mu_F =: \int \varphi dF.$$

于是如果可测函数 φ 使得 $\mathbb{E}\varphi(\xi)$ 存在，则

$$\mathbb{E}[\varphi(\xi)] := \int \varphi(\xi(\omega)) d\mathbb{P}(\omega) = \int \varphi(x) d\mathbb{P} \circ \xi^{-1}(x) = \int \varphi(x) dF(x).$$

于是我们有下面的定理。

定理 6.3.2. 当随机变量/向量 $\xi \sim F$ 且 φ 为相应的可测函数时，

$$\mathbb{E}[\varphi(\xi)] = \int \varphi(x) dF(x), \quad (6.6)$$

如果上式两端有一者有意义；特别的，当 $\mathbb{E}\xi$ 有意义时，它有如下计算公式

$$\mathbb{E}\xi = \int x dF(x), \quad (6.7)$$

上述定理是初等概率论中可以利用分布函数来计算对应随机变量的数学期望的原因；但在严格的公理化体系下，通常不直接把(6.7)作为数学期望的定义，以免造成不便。

上面的讨论进一步推进，就可以引出概率空间（或测度空间）之间同构的概念；例如，当保测映射实际上是双向可逆保测时，我们就可以认为通过这个可逆保测映射联系的两个概率结构（或可测结构）同构；但由于我们在概率论中通常可以扔掉零概率事件来考虑问题，因而概率结构的同构概念还可以更一般一点。此处详细讨论不再展开，留给读者思考。以下仅举下面几个容易理解的例子来说明有关概念引入后对于说理论证上的便利。

例 6.1. *Galileo Galilei*（伽利略，1564/2/18–1642/1/8；意大利）曾经研究过投掷 3 枚骰子的相关问题。他发现总点数和为 10 与 11 两个事件（分别记作 A_{10} 与 A_{11} ）的频率或概率相同： $\mathbb{P}(A_{10}) = \mathbb{P}(A_{11})$ （均等于 $\frac{1}{8}$ ）。我们来说明这并不是一个纯粹计算上的巧合。

对于投掷单个骰子，我们可以使用 $\Sigma := \{1, \dots, 6\}$ 上的古典概率模型作为它的概率模型，其中可以定义变换 $T(x) := 7 - x$ ，显然变换 $T: \Sigma \rightarrow \Sigma$ 是这个古典概率模型上的同构。对于投掷三枚骰子，我们可以使用 $\Omega_3 := \Sigma^3$ 上的古典概率模型作为它的概率模型，上面可以定义乘积变换 $T: \Omega_3 \rightarrow \Omega_3$ 为 $T(x_1, x_2, x_3) := (T(x_1), T(x_2), T(x_3)) = (7 - x_1, 7 - x_2, 7 - x_3)$ ，仍然是一个同构，从而 $\mathbb{P} \circ T^{-1} = \mathbb{P}$ 。事件 A_n 表示三枚骰子总点数和为 n 这一事件，即

$$A_n := \{(x_1, x_2, x_3) : x_1 + x_2 + x_3 = n\}.$$

于是显然有 $T^{-1}A_{10} = A_{11}$ ，因此 $\mathbb{P}(A_{10}) = \mathbb{P}(A_{11})$ 。 \square

例 6.2. 甲乙二人投掷同一枚均匀的硬币。甲投掷了 $n+1$ 次，得到 X 个正面；乙投掷了 n 次，得到 Y 个正面。请计算甲获得的正面次数多于乙获得的正面次数的概率 $\mathbb{P}(X > Y)$ 。

参考解答. 如果使用组合计数技巧直接进行计算, 当然也能得出问题的解答。这里我们借用上一例题中的观点来给出一个快速而严谨的解答。

首先, 为了说话方便, 我们选择的概率空间为 $\Omega := \{0, 1\}^{2n+1}$ 上的古典概率模型; 其中 1 代表硬币正面, 0 代表硬币反面。对于投掷单个硬币, 我们可以定义 $T(x) = 1 - x$; 它诱导 Ω 上的一个乘积映射: $T(\omega_1, \dots, \omega_{2n+1}) = (1 - \omega_1, \dots, 1 - \omega_{2n+1})$, 它是保测的。这时不妨认为

$$X = \omega_1 + \dots + \omega_{n+1}, Y = \omega_{n+2} + \dots + \omega_{2n+1}.$$

显然 $X \circ T = (n+1) - X, Y \circ T = n - Y$, 从而

$$T^{-1}(\{X > Y\}) = \{X \circ T > Y \circ T\} = \{X \leq Y\} = (\{X > Y\})^c,$$

于是

$$\mathbb{P}(X > Y) = \mathbb{P}(T^{-1}(\{X > Y\})) = \mathbb{P}((\{X > Y\})^c).$$

由此 $\mathbb{P}(X > Y) = \mathbb{P}(X \leq Y) = 1/2$. \square

注意到第 4 章中的 Fubini 定理 (参见定理 4.4.4), 对于非负随机变量, 我们还有如下方式来计算数学期望。

定理 6.3.3. 设 ξ 是非负随机变量, 那么

$$\mathbb{E}\xi = \int_0^\infty \mathbb{P}(\xi > x) dx. \quad (6.8)$$

参考证明. 注意到 $\mathbb{P}(\xi > x) = \mathbb{E}[1_{\{\xi > x\}}]$ 以及数学期望本质上是一种积分, 利用 Fubini 定理就能得到

$$\begin{aligned} \int_0^\infty \mathbb{P}(\xi > x) dx &= \int \mathbb{E}[1_{\{\xi > x\}}] dx \\ &= \mathbb{E}\left[\int_0^\infty 1_{\{\xi > x\}} dx\right] \\ &= \mathbb{E}\left[\int_0^\xi dx\right] = \mathbb{E}\xi. \end{aligned}$$

这里, 我们要指出,

$$1_{\{\xi > x\}} = 1_{(0, \infty)}(\xi - x)$$

实际上是一个非负的二元函数, 自变量为 (ω, x) , 因为 $\xi = \xi(\omega)$; 这个函数的二元可测性是显然的。 \square

6.4 方差、协方差与独立性

以上我们定义了一个随机变量 ξ 的数学期望 $\mathbb{E}\xi$; 如果 $\mathbb{E}\xi$ 是一个有限值, 我们把它称为随机变量 ξ 的**一阶矩**。一般的, 我们把 ξ 的 n 阶矩定义为 $\mathbb{E}[\xi^n]$, 如果它存在; 否则将称 ξ 的 n 阶矩不存在。

进一步, 我们可以定义 ξ 的方差: 当 ξ 的二阶矩存在时, 称

$$\text{Var}(\xi) := \mathbb{E}[(\xi - \mathbb{E}\xi)^2] \quad (6.9)$$

为 ξ 的方差; 当 ξ 的二阶矩不存在时, 我们也称 ξ 的方差不存在。不难知道,

$$\text{Var}(\xi) = \mathbb{E}[\xi^2] - (\mathbb{E}\xi)^2. \quad (6.10)$$

这蕴含了如下矩不等式

$$(\mathbb{E}\xi)^2 \leq \mathbb{E}[\xi^2]. \quad (6.11)$$

现在我们考虑两个随机变量 ξ 与 η 的乘积的数学期望 $\mathbb{E}[\xi \cdot \eta]$ 。通过考察二次多项式函数

$$f(t) := \mathbb{E}[(t\xi - \eta)^2] = t^2\mathbb{E}[\xi^2] - 2t\mathbb{E}[\xi \cdot \eta] + \mathbb{E}[\eta^2],$$

不难得到如下形式的 **Cauchy 不等式**：当 ξ, η 的二阶矩都存在时，

$$|\mathbb{E}[\xi \cdot \eta]|^2 \leq \mathbb{E}[\xi^2] \cdot \mathbb{E}[\eta^2]. \quad (6.12)$$

我们定义 ξ 与 η 的协方差为

$$\text{Cov}(\xi, \eta) := \mathbb{E}[(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)], \quad (6.13)$$

如果上式右端是有限值（否则将称 ξ 与 η 的协方差不存在）。利用前述的 **Cauchy 不等式**，也立即得到有关协方差的 **Cauchy 不等式**：当 ξ, η 的二阶矩都存在时，

$$|\text{Cov}(\xi, \eta)|^2 \leq \text{Var}(\xi) \cdot \text{Var}(\eta). \quad (6.14)$$

当 $\text{Cov}(\xi, \eta) = 0$ 时，我们也称 ξ 与 η **线性不相关**。当 ξ, η 的方差存在（且均不为 0）时，人们进一步定义所谓的 ξ 与 η 的**相关系数**：

$$\text{Corr}(\xi, \eta) := \frac{\text{Cov}(\xi, \eta)}{\sqrt{\text{Var}(\xi) \cdot \text{Var}(\eta)}}. \quad (6.15)$$

此时应有 $\text{Corr}(\xi, \eta) \in [-1, 1]$ 。

关于利用数学期望来探讨随机变量的独立性，我们有如下重要结果。

定理 6.4.1. 设 ξ 与 η 相互独立，并且它们的一阶矩都存在，那么

$$\mathbb{E}[\xi \cdot \eta] = \mathbb{E}\xi \cdot \mathbb{E}\eta. \quad (6.16)$$

亦即此时有 $\text{Cov}(\xi, \eta) = 0$ 。

反之，如果对任意有界连续函数 φ, ψ 都有 $\text{Cov}(\varphi(\xi), \psi(\eta)) = 0$ ，则 ξ 与 η 相互独立。

参考证明：为方便，设 ξ 与 η 的分布测度分别是 μ, ν ，即 $\xi \sim \mu, \eta \sim \nu$ 。

当 ξ 与 η 相互独立时，根据定理 5.1.1， $(\xi, \eta) \sim \mu \times \nu$ ；根据(6.6)以及 Fubini 定理（见定理 4.4.4）

$$\begin{aligned} \mathbb{E}[\xi \cdot \eta] &= \int x \cdot y d\mu \times \nu(x, y) \\ &= \int \left[\int x \cdot y d\mu(x) \right] d\nu(y) \\ &= \int \left[\int x d\mu(x) \right] \cdot y d\nu(y) \\ &= \int [\mathbb{E}\xi] \cdot y d\nu(y) \\ &= [\mathbb{E}\xi] \cdot \int y d\nu(y) = [\mathbb{E}\xi] \cdot [\mathbb{E}\eta]. \end{aligned}$$

反过来， $\text{Cov}(\varphi(\xi), \psi(\eta)) = 0$ 等价于：

$$\mathbb{E}[\varphi(\xi) \cdot \psi(\eta)] = \left[\int \varphi(x) d\mu(x) \right] \left[\int \psi(x) d\nu(x) \right].$$

上式对所有有界连续函数 φ, ψ 成立, 进而对所有有界可测函数 φ, ψ 成立。特别的, 取 $\varphi = 1_A, \psi = 1_B$, 其中 A, B 为可测集, 立即得到

$$\mathbb{P}(\xi \in A, \eta \in B) = \mu(A) \cdot \nu(B).$$

特取 $A = (-\infty, x], B = (-\infty, y]$, 立即得到 ξ 与 η 独立。□

◆ 推论 6.4.1. 设 ξ 与 η 相互独立 (或 $\text{Cov}(\xi, \eta) = 0$), 并且它们的二阶矩都存在, 那么

$$\text{Var}(\xi \pm \eta) = \text{Var}(\xi) + \text{Var}(\eta). \quad (6.17)$$

在统计学中, 给定样本 X_1, \dots, X_n , 设它们的联合分布为

$$(X_1, \dots, X_n) \sim F_\theta,$$

其中 (确定性的) 参数 $\theta \in \Theta$ 未知; 我们将用 \mathbb{E}_θ 和 Var_θ 表示分布律 F_θ 确定的数学期望算子与方差算子。如果我们想利用这些样本构造出了一个统计量 $T = T(X_1, \dots, X_n)$ 来估计 $g(\theta)$, 其中 g 为可测函数, 则当

$$\mathbb{E}_\theta[T] = g(\theta), \forall \theta \in \Theta$$

时, 我们就称 T 是参数 $g(\theta)$ 的无偏估计。如果在所有的无偏估计中, $\text{Var}_\theta(T)$ 达到最小, 则进一步称 T 是参数 $g(\theta)$ 的最小方差无偏估计。另外, 统计学中还有一种矩方法用于构造统计量。记 $\widehat{M}_k := \frac{1}{n}(X_1^k + \dots + X_n^k)$, 称之为 k -阶样本矩。如果通过解方程

$$\mu_k := \mathbb{E}_\theta[\widehat{M}_k] = f_k(\theta), k = 1, \dots, p$$

可以反解出 $g(\theta) = h(\mu_1, \dots, \mu_p)$, 其中 h 是一个不依赖于参数 θ 的 (连续) 函数, μ_1, \dots, μ_p 是任何有意义的常数, 则可以构造统计量

$$T := h(\widehat{M}_1, \dots, \widehat{M}_p).$$

这种构造统计量的办法称为矩方法, 此时 T 称为一个矩估计。统计学中关于如何寻找性质优良的估计量有系统的理论与方法, 本书不展开论述。

习 题 6

习题 6.1. 设 X 服从整数集 $D := (a, b] \cap \mathbb{Z}$ 上的均匀分布, 其中 $a < b$ 是整数。计算 X 的数学期望与方差。

习题 6.2. 设 $\{X_k\}_{k=1}^n \stackrel{\text{i.i.d.}}{\sim} B(1, p)$ 。计算 $X_1 \sim B(1, p)$ 的数学期望与方差, 进而计算 $S_n := X_1 + \dots + X_n \sim B(n, p)$ 的数学期望与方差。此题也表明例 5.10 中估计也可以视作矩估计, 并且是无偏估计。

习题 6.3. 设 $\{\tau_k\}_{k=1}^\infty \stackrel{\text{i.i.d.}}{\sim} \text{Geo}(p)$, 计算 τ_1 的数学期望与方差, 进而计算 $\widetilde{N}_r := \tau_1 + \dots + \tau_r - r \sim NB(r, p)$ 的数学期望与方差, 其中 $r \geq 1$ 是整数。

习题 6.4. 证明例 5.11 中估计量是矩估计, 并且是无偏估计。

习题 6.5. 设 F 是分布函数, $a > 0$, 证明: $\int [F(x+a) - F(x)]dx = a$ 。

习题 6.6. 设随机变量 X 的分布函数 F 满足: 对某常数 $a > 0$,

$$F(x+a) - F(x) = 0 \text{ 或 } 1, \forall x \in \mathbb{R}.$$

试论证: X 是退化的随机变量, 即存在 $b \in \mathbb{R}$ 使得 $\mathbb{P}(X = b) = 1$ 。反过来的结论也成立。【提示: 前一结论用反证法; 后一结论很简单。】

习题 6.7. 设随机变量 X 的分布函数 F 是连续函数，其数学期望存在为 μ 。
证明：下面关于实数 a 的方程

$$\int_{-\infty}^a F(x)dx = \int_a^{\infty} [1 - F(x)]dx$$

具有唯一解 $a = \mu$ 。

习题 6.8. 求证：随机变量 X 可积的充分必要条件是

$$\sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n) < \infty.$$

习题 6.9. 对 $p > 0$ 以及随机变量 X ，求证：

$$\mathbb{E}[|X|^p] = \int_0^{\infty} px^{p-1}\mathbb{P}(|X| > x)dx.$$

习题 6.10. 对可积随机变量 X ，考察函数

$$f(a) := \mathbb{E}|X - a|$$

的最小值点全体所成集合 Med_X ，说明它是单点集或一个非空区间。对任意 $a \in Med_X$ ，证明： $\mathbb{P}(X \leq a) \geq \frac{1}{2}$ ， $\mathbb{P}(X \geq a) \geq \frac{1}{2}$ 。这样的点 a 都称为是 X 的分布对应的中位数（有时仅把最小的 $a = \inf Med_X$ 称为 $\frac{1}{2}$ -分位数）。

习题 6.11. 对 L^2 -可积随机变量 X ，证明函数

$$f(a) := \mathbb{E}[(X - a)^2]$$

的最小值点就是 $a = \mathbb{E}X$ 。

习题 6.12. 设两事件 A, B 都是正概率事件： $\mathbb{P}(A) > 0, \mathbb{P}(B) > 0$ 。称 A 与 B 是正相关的，如果 $\mathbb{P}(B|A) > \mathbb{P}(B)$ 。求证： A 与 B 是正相关的充要条件是 $\text{Cov}(1_A, 1_B) > 0$ 。于是 A 与 B 是正相关的，当且仅当 B 与 A 是正相关的。

§ 7

条件数学期望与条件分布律

在本章，我们将接着上一章的内容继续介绍条件数学期望这一重要概率工具。

条件数学期望是概率论中与数学期望和随机向量的条件分布律等强烈关联的另一套重要工具；它的发展历史与定义流程显得更复杂，但与数学期望的定义的精神相通。数学期望与条件数学期望这两套工具再结合概率论中的独立性概念，衍生出诸多概率论中频繁使用的高级技巧。在本课程中，囿于篇幅及教学体系，我们只能对它们进行初步的介绍。

7.1 条件数学期望的定义

在实际生活中，我们经常遇到类似如下的场景：两家同行业的上市公司，其中一家公司向公众公布了它的财务状况，另一公司没有公布；那么如何合理借助已公布的财务状况等信息，来对另一公司的财务状况做估计？在当前学生成绩作为隐私经常保密的情况下，学生们如何借助一些公开信息结合自己的个人信息合理预测或估算自己及朋友在班级中的排名状况？在商业竞争中，如何利用行业内公开信息以及己方信息，估算竞争对手的（工厂）产能或利润情况等？在风险投资行业中，同样的股份份额，如何根据 A 轮的融资价格以及后续行业发展的有关信息，合理给出 B 轮的参考价格？现代人的生活中，根据已有的一些信息对不确定的量进行估计或预测的需求数不胜数。概率论中对此类预测问题给出的一种比较统一的理论解决方案，就是“条件数学期望”这一工具。

条件数学期望是概率论中继数学期望工具之后的又一重要工具。正如早期的数学期望概念被 Huygens 创造出来的目的是解决赌博中的“平均收益”或“期望收益”这一自然语言中的概念的数学精确化表达，条件数学期望也有它准备解决的数学问题：如何基于已经给定的信息（比如说给定一个随机变量 X 的值），做另一个随机变量的预测？这某种程度上相当于问：自然语言中的“平均条件收益”应当如何给出合理的数学定义？我们将说明，条件数学期望是在二阶损失最小的情况下的最佳预测。基于此，我们最终给出条件数学期望的数学定义，并给出一些典型情况的条件数学期望的计算公式。

7.1.1 数学期望的一个性质

如果把 ξ 解释为赌徒参与某个赌博后的真实收益，那么 $\mathbb{E}\xi$ 就是赌徒参与这个赌博项目的所谓的期望收益（平均收益），或者说这是庄家应向参与这个赌博的赌徒收取的合理“门票价格”：以此价格作为门票价格，长期运营这个赌博项目的话，庄家是“平均角度”保本的（不考虑其他运营成本），略高于此的门票价格将会导致长期稳定盈利，而略低于此的门票价格将会导致亏损，这些结论读者可以结合后续章节中的 Kolmogorov 强大数律来理解。如果 ξ 具有二阶矩，那么其他门票价格 $c \neq \mathbb{E}\xi$ 都将导致庄家与赌徒二者都计算在内（视作一个系统）的平均“二阶损失” $\mathbb{E}[(\xi - c)^2]$ 更大（注意，我们假定赌博是零和博弈， $\xi - c$ 是赌徒的正收益的话，它就也恰好是庄家的损失；反之亦然），即下面的不等式成立

$$\mathbb{E}[(\xi - c)^2] > \mathbb{E}[(\xi - \mathbb{E}\xi)^2] = \text{Var}(\xi), \forall c \neq \mathbb{E}\xi. \quad (7.1)$$

这只要注意到（记 $\mu = \mathbb{E}\xi$ ）

$$\begin{aligned} \mathbb{E}[(\xi - c)^2] &= \mathbb{E}[(\xi - \mu)^2 + 2(\mu - c)(\xi - \mu) + (\mu - c)^2] \\ &= \mathbb{E}[(\xi - \mu)^2] + 2(\mu - c)\mathbb{E}[(\xi - \mu)] + (\mu - c)^2 \\ &= \mathbb{E}[(\xi - \mu)^2] + (\mu - c)^2. \end{aligned}$$

一般而言，平均二阶损失过大的赌博通常难以真正长期进行下去：或者赌徒不乐意参加，或者庄家不乐意主持这样的赌局；除非这两方的参与者至少有一方失去了理智。

7.1.2 经典条件概率到相应的“经典”条件数学期望

在前述章节中，我们已经定义了经典的条件概率（见第 3 章）和经典的（无条件）数学期望（见第 6 章）。对于概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中正概率的事件 $A \in \mathcal{F}$ ，我们有“条件概率空间” $(A, \mathcal{F}_A, \mathbb{P}_A)$ 的定义，其中 $\mathbb{P}_A(\cdot) = \mathbb{P}(\cdot|A)$ 。依据我们定义数学期望的原则，自然可以在“条件概率空间” $(A, \mathcal{F}_A, \mathbb{P}_A)$ 中定义相应的数学期望 \mathbb{E}_A ，我们改记号为 $\mathbb{E}[\cdot|A]$ ，即

$$\mathbb{E}_A[\xi] = \mathbb{E}[\xi|A].$$

注意到 $\xi = 1_B, B \in \mathcal{F}$ 时，应有 $\mathbb{E}_A[1_B] = \mathbb{P}_A(A \cap B) = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$ 。也就是说我们有下面的经典的条件数学期望的出发点：

$$\mathbb{E}[1_B|A] := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad \forall B \in \mathcal{F}. \quad (7.2)$$

按照数学期望的定义流程，很容易知道下面公式成立：

$$\mathbb{E}[\xi|A] := \frac{\mathbb{E}[1_A \cdot \xi]}{\mathbb{P}(A)}. \quad (7.3)$$

至此，我们自然可以考虑两个离散型随机变量 X, Y 的条件数学期望 $\mathbb{E}[Y|X]$ ，它被定义为：当 $X = x$ 发生时，就取值为 $\mathbb{E}[Y|X = x]$ ，即

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X = x], \quad \text{如果 } X = x.$$

上面表达中同样有记号上的偷懒，即把事件 $A = \{X = x\}$ 表达式外面的花括号省略不写了。总之，在 X, Y 都是离散随机变量（且 Y 具有可积性）的

情形下，条件数学期望 $\mathbb{E}[Y|X]$ 被合理地定义为一个新的随机变量 $\varphi(X)$ 了，其中函数 $\varphi(x) := \mathbb{E}[Y|X=x]$ 。

我们设 X, Y 的值域分别为 $\mathcal{R}_X, \mathcal{R}_Y$ ，它们都是可数集。对任意给定的 $x \in \mathcal{R}_X, y \in \mathcal{R}_Y$ ，记

$$\begin{aligned} p_{x,y} &:= \mathbb{P}(X=x, Y=y), \\ p_{x,\cdot} &:= \mathbb{P}(X=x) = \sum_y p_{x,y}, \\ p_{\cdot,y} &:= \mathbb{P}(Y=y) = \sum_x p_{x,y}. \end{aligned}$$

于是 $\mathbb{P}(Y=y|X=x) = \frac{p_{x,y}}{p_{x,\cdot}}, \forall x \in \mathcal{R}_X, y \in \mathcal{R}_Y$ 。当 $\mathbb{E}Y$ 存在时，不难知道

$$\varphi(x) := \mathbb{E}[Y|X=x] = \sum_y y \cdot \frac{p_{x,y}}{p_{x,\cdot}}.$$

基于此，也不难看到条件数学期望的下面两个重要性质。其一称为**全期望公式**：

$$\mathbb{E}[Y] = \mathbb{E}[\varphi(X)] = \mathbb{E}[\mathbb{E}[Y|X]]. \quad (7.4)$$

它是全概率公式的期望版本，很多文献仍然称之为全概率公式。其二是数学期望的性质(7.1)的推广：

$$\mathbb{E}[(Y - \psi(X))^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2], \forall \psi. \quad (7.5)$$

上面方程表明，在所有形如 $\psi(X)$ 的预测中， $\varphi(X) = \mathbb{E}[Y|X]$ 是平均二阶损失最小的对 Y 的预测。

对于同一个概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上任何两个有二阶矩的（实值）随机变量 ξ, η ，我们可以定义“内积”

$$\langle \xi, \eta \rangle := \mathbb{E}[\xi\eta],$$

由此诱导了范数 $\|\eta\|_2 := \sqrt{\langle \eta, \eta \rangle}$ 。给定随机变量 X ，我们可以定义它诱导的一个空间 \mathcal{H}_X

$$\mathcal{H}_X := \{\psi(X) : \psi \text{ 是可测函数, 且 } \mathbb{E}[\psi(X)^2] < \infty\}.$$

当 $\mathbb{E}[Y^2] < \infty$ 且 X, Y 都是离散型随机变量时，(7.5) 表明了两点：第一， $\mathbb{E}[Y|X] \in \mathcal{H}_X$ ；第二， $\|Y - Z\|_2 \geq \|Y - \mathbb{E}[Y|X]\|_2, \forall Z \in \mathcal{H}_X$ 。这说明，条件数学期望 $\mathbb{E}[Y|X]$ 可以视作 Y 向 \mathcal{H}_X 的正交投影：

$$\mathbb{E}[Y|X] = \text{Proj}_{\mathcal{H}_X}(Y). \quad (7.6)$$

7.1.3 抽象的条件数学期望的定义

以上我们已经在具有二阶矩的离散随机变量情形导出了条件数学期望 $\mathbb{E}[Y|X]$ 的一种合理定义：正交投影，见(7.6)。我们把这个定义作为一般的具有二阶矩的随机变量 Y 关于另一个随机变量 X 的条件数学期望的定义。

以下我们准备进一步拓广这个定义。

现在我们给定 $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ ，其含义就是 Y 是概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上有二阶矩的随机变量。再给定一个子 σ -代数 $\mathcal{G} \subset \mathcal{F}$ ，于是我们同样有概率

空间 $(\Omega, \mathcal{G}, \mathbb{P}|_{\mathcal{G}})$ ，进而也有一个 Hilbert 空间

$$\mathcal{H}_{\mathcal{G}} := L^2(\Omega, \mathcal{G}, \mathbb{P}|_{\mathcal{G}}).$$

于是我们仿照上一节，对 $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ ，定义

$$\mathbb{E}[Y|\mathcal{G}] := \text{Proj}_{\mathcal{H}_{\mathcal{G}}}(Y). \quad (7.7)$$

它具有如下优良性质：

$$\mathbb{E}[(Y - \xi)^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y|\mathcal{G}])^2], \forall \xi \in \mathcal{H}_{\mathcal{G}}, \quad (7.8)$$

即 $\mathbb{E}[Y|\mathcal{G}]$ 是 $\mathcal{H}_{\mathcal{G}}$ 中对 Y 的最优预测。

上述定义已经比较宽泛了，但它对 Y 还要求有二阶矩；我们准备进一步把这个条件降低至一阶矩。为此，进一步来分析上述方程(7.7)。它实际上要求：

(1) $\xi := \mathbb{E}[Y|\mathcal{G}]$ 被定义为一个特殊的随机变量 $\xi \in \mathcal{H}_{\mathcal{G}}$ ，即它是 \mathcal{G} -可测的随机变量，并且二阶矩存在；

(2) ξ 进一步满足：对任意 $\eta \in \mathcal{H}_{\mathcal{G}}$ ， $\mathbb{E}[(Y - \xi)\eta] = 0$ ，亦即

$$\mathbb{E}[Y \cdot \eta] = \mathbb{E}[\xi \cdot \eta], \forall \eta \in \mathcal{H}_{\mathcal{G}}.$$

注意到 $\mathcal{H}_{\mathcal{G}}$ 可以视作 $\text{span}\{1_B : B \in \mathcal{G}\}$ 的闭包，我们可以给出下面版本的条件数学期望的定义。

定义 7.1.1. 给定子 σ -代数 $\mathcal{G} \subset \mathcal{F}$ ，当 Y 的数学期望存在时， Y 关于 \mathcal{G} 的**条件数学期望**被定义为具有如下性质的一个随机变量 ξ ：

(1) ξ 是 \mathcal{G} -可测的随机变量；

(2) ξ 进一步满足：

$$\mathbb{E}[Y \cdot 1_B] = \mathbb{E}[\xi \cdot 1_B], \forall B \in \mathcal{G}. \quad (7.9)$$

具有上述性质的随机变量 ξ 是存在且在 \mathbb{P} -几乎处处相等的意义下唯一，我们把它记作 $\mathbb{E}[Y|\mathcal{G}]$ 。

在上面定义中，我们断言具有性质 (1)、(2) 的随机变量 ξ 是存在唯一的，这可以用第 4 章中 R-N 导数的理论来解释。我们定义符号测度 $\mu : \mathcal{G} \rightarrow \mathbb{R}$ 如下

$$\mu(B) := \mathbb{E}[Y \cdot 1_B], \quad B \in \mathcal{G}.$$

显然此时有 $\mathbb{P}(B) = 0$ 蕴含了 $\mu(B) = 0$ 。因而 $\mu \ll \mathbb{P}|_{\mathcal{G}}$ 。于是存在 ρ 为 \mathcal{G} 可测函数，使得

$$\rho = \frac{d\mu}{d\mathbb{P}|_{\mathcal{G}}},$$

即

$$\mu(B) = \int_B \rho d\mathbb{P}|_{\mathcal{G}}, \forall B \in \mathcal{G}.$$

上式改用数学期望表达就是(7.9)。

如果 X 是随机变量或随机向量，我们介绍过 X 生成的 σ -代数 $\sigma(X)$ 的概念。我们可以定义随机变量 Y 关于 $\mathcal{G} = \sigma(X)$ 的**条件数学期望**，把它简单记作 $\mathbb{E}[Y|X]$ ，即有如下符号上的认同：

$$\mathbb{E}[Y|X] := \mathbb{E}[Y|\sigma(X)]. \quad (7.10)$$

下面的结果说明了这种认同的合理性，它在理解和计算条件数学期望等量的过程中经常需要用到：条件数学期望 $\mathbb{E}[Y|\sigma(X)]$ 如果存在，它就可以表达为 X 的可测函数： $\mathbb{E}[Y|\sigma(X)] = \varphi(X)$ ，此时概率论中对于函数 $\varphi(\cdot)$ 的一个偷懒的写法是： $\mathbb{E}[Y|X = x]$ ，亦即此处 $\mathbb{E}[Y|X = x] = \varphi(x)$ ，它在 μ_X -几乎处处相等意义下是唯一确定，其中 μ_X 是 X 的分布测度。

定理 7.1.1. (参见 [20, pp. 41, Corollary 1.97]) 设 $(\Omega_k, \mathcal{F}_k), k = 1, 2$ 是可测空间， $f: \Omega_1 \rightarrow \Omega_2$ 是可测映射。 f 生成的 σ -代数 $\sigma(f)$ 理解为

$$\sigma(f) = f^{-1}(\mathcal{F}_2) \subset \mathcal{F}_1.$$

给定广义实值函数 $g: \Omega_1 \rightarrow \bar{\mathbb{R}}$ 。 g 是 $\sigma(f)/\mathcal{B}(\bar{\mathbb{R}})$ 可测的，当且仅当存在可测函数 $\varphi: (\Omega_2, \mathcal{F}_2) \rightarrow (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ 使得 $g = \varphi \circ f$ 。

下面的交换图解释了上述定理的结论。

$$\begin{array}{ccc} (\Omega_1, \mathcal{F}_1) & \xrightarrow{f} & (\Omega_2, \mathcal{F}_2) \\ & \searrow g & \downarrow \varphi \\ & & (\bar{\mathbb{R}}, \mathcal{B}) \end{array}$$

为了证明上述结论，需要下面的引理。

引理 7.1.1. (参见 [20, pp. 41, Theorem 1.96]) 设 (Ω, \mathcal{F}) 是可测空间， $f: \Omega \rightarrow [0, \infty]$ 是广义非负可测函数。那么以下结论成立：

- (i) 存在非负简单函数列 $\{f_n\}_1^\infty$ ，使得 $f_n \nearrow f$ ；
- (ii) 存在可测集列 $\{A_n\}_1^\infty \subset \mathcal{F}$ 以及非负实数列 $\{\alpha_n \geq 0\}_1^\infty$ 使得

$$f = \sum_{n=1}^{\infty} \alpha_n \cdot 1_{A_n}.$$

证明. 定义 $f_n := \lfloor \frac{2^n \cdot f}{2^n} \rfloor \wedge n$ ，其中 $\lfloor \cdot \rfloor$ 表示取整函数。显然 f_n 可测，并且最多取 $n \cdot 2^n + 1$ 个不同的非负值，因此是非负简单函数。显然 $f_n \nearrow f$ 。此即 (i)。

对于上述 f_n ，有 $f = f_1 + \sum_{n=1}^{\infty} (f_{n+1} - f_n)$ 。显然 $\{f_{n+1} - f_n\}_{n=0}^\infty$ 也是非负简单函数列（约定 $f_0 := 0$ ）。由此知道引理结论 (ii) 成立。 \square

定理 7.1.1 的证明. 定理的充分性是显然的，下面证明必要性。

我们先对非负函数 $g \geq 0$ 给出证明。此时，由上述引理，存在可测集列 $\{A_n\}_1^\infty \subset \sigma(f)$ 及非负实数列 $\{\alpha_n : n \geq 1\}$ 使得

$$g = \sum_{n=1}^{\infty} \alpha_n 1_{A_n}.$$

注意到 $A_n \in \sigma(f)$ ，存在 $B_n \in \mathcal{F}_2$ 使得 $A_n = f^{-1}(B_n)$ ，亦即 $1_{A_n} = 1_{B_n} \circ f$ 。于是可以定义广义非负可测函数

$$\varphi := \sum_{n=1}^{\infty} \alpha_n 1_{B_n},$$

它满足 $g = \varphi \circ f$ 。

对于一般情形, g 是 $\sigma(f)/\mathcal{B}(\mathbb{R})$ 可测的, 因此 $g^+ := g \vee 0, g^- := (-g) \vee 0$ 也是 $\sigma(f)/\mathcal{B}(\mathbb{R})$ 可测的。于是存在

$$\varphi^{(\pm)} = \sum_{n=1}^{\infty} \alpha_n^{(\pm)} \cdot 1_{B_n^{\pm}}$$

使得 $g^{\pm} = \varphi^{(\pm)} \circ f$ 。令 $B := \{\omega_2 : \varphi^{(+)}(\omega_2) = \varphi^{(-)}(\omega_2) = \infty\} \in \mathcal{F}_2$, 显然应有 $f^{-1}(B) = \emptyset, f(\Omega_1) \subset B^c$ (否则与 $g = g^+ - g^-$ 有定义相矛盾)。从而总可以定义 $\varphi(\omega_2) := 1_{B^c}(\omega_2) \cdot (\varphi^{(+)}(\omega_2) - \varphi^{(-)}(\omega_2))$ 。此时显然有 $g = \varphi \circ f$ 。□

7.1.4 由条件数学期望 $\mathbb{E}[\cdot|\mathcal{G}]$ 到条件概率 $\mathbb{P}(\cdot|\mathcal{G})$

以上我们定义了抽象的条件数学期望 $\mathbb{E}[\cdot|\mathcal{G}]$; 在当初数学期望的定义过程中, 我们有下面的对应关系

$$\mathbb{E}1_B = \mathbb{P}(B), \forall B \in \mathcal{F}.$$

现在我们可以模仿这个关系来定义抽象的条件概率 $\mathbb{P}(\cdot|\mathcal{G})$, 即

$$\mathbb{P}(B|\mathcal{G}) := \mathbb{E}[1_B|\mathcal{G}], \forall B \in \mathcal{F}. \quad (7.11)$$

这里需要注意, 与经典的条件概率 $\mathbb{P}(B|A)$ 是一个 $[0, 1]$ 中的数值不同, $\mathbb{P}(B|\mathcal{G})$ 本质上是一个随机变量! 虽然有这个差别, 但是下面例子揭示了二者间的联系, 说明了现在抽象定义的条件概率这一概念是经典条件概率概念的推广。

例 7.1. 设 $A \neq \emptyset, \mathcal{G} = \{\emptyset, A, A^c, \Omega\}$, 它是包含事件 A 的最小 σ -代数。我们假定 $0 < \mathbb{P}(A) < 1$ 。则容易计算出

$$\begin{aligned} \mathbb{P}(B|\mathcal{G})(\omega) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \cdot 1_A(\omega) + \frac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(A^c)} \cdot 1_{A^c}(\omega) \\ &= \mathbb{P}(B|A) \cdot 1_A(\omega) + \mathbb{P}(B|A^c) \cdot 1_{A^c}(\omega). \end{aligned}$$

最后一个等式中出现的条件概率 $\mathbb{P}(B|A), \mathbb{P}(B|A^c)$ 是经典的条件概率; 并且当 A 发生时, 亦即 $\omega \in A$ 时, $\mathbb{P}(B|\mathcal{G})(\omega) = \mathbb{P}(B|A)$ 。这表明, 本节中抽象的条件概率的定义与第 3 章中的经典条件概率定义是相容的。

在有了条件数学期望、条件概率的定义后, 之前的独立性也可以有如下的等价刻画: 任给两事件域 (σ -代数) $\mathcal{G}_1, \mathcal{G}_2 \subset \mathcal{F}$,

$$\mathcal{G}_1, \mathcal{G}_2 \text{ 相互独立} \Leftrightarrow \mathbb{P}(A_2|\mathcal{G}_1) = \mathbb{P}(A_2) \quad \mathbb{P}-a.e., \quad \forall A_2 \in \mathcal{G}_2. \quad (7.12)$$

请读者自己验证该命题, 它是之前经典情形结果(3.14)的推广。在编者看来, 这相当于说: 在现实生活中, 我们基于完全无关的信息所做理性判断, 与没有得到这些信息时所作理性判断应该毫无区别*。

7.1.5 抽象条件数学期望 $\mathbb{E}[Y|\mathcal{G}]$ 的计算

对于抽象条件数学期望 $\mathbb{E}[Y|\mathcal{G}]$ 的计算, 当不知道更多关于 Y, \mathcal{G} 的信息时, 我们也没有太多办法, 只能是按照定义去计算。也就是说, 此时我们只

*但现实生活中的难点在于界定所获得信息与所要考察事件的关联性。在统计学中有一个与此相关的新兴研究邻域: 因果推断。

能去求解满足下面方程（组）的 \mathcal{G} -可测随机变量 ξ :

$$\mathbb{E}[Y \cdot 1_B] = \mathbb{E}[\xi \cdot 1_B], \quad \forall B \in \mathcal{G}.$$

当有更多 Y, \mathcal{G} 的信息时，我们就有其他计算办法。

7.2 条件数学期望的性质

为了减少我们在数学期望和条件数学期望的计算过程中的计算量，我们基于条件数学期望的定义来讨论条件数学期望的性质。

不难验证（细节见附录 B），条件数学期望具有如下一些性质（其中 ξ, η 为随机变量）：

性质（0） $\mathbb{E}[\xi|\mathcal{G}]$ 是 \mathcal{G} -可测随机变量，满足：当 $\mathbb{E}\xi$ 有意义时，下面的全期望公式成立

$$\mathbb{E}\xi = \mathbb{E}[\mathbb{E}[\xi|\mathcal{G}]];$$

性质（1）（线性性质） 设 $a, b \in \mathbb{R}$ 且 $\mathbb{E}\xi, \mathbb{E}\eta, a\mathbb{E}\xi + b\mathbb{E}\eta$ 均有意义，则

$$\mathbb{E}[a\xi + b\eta|\mathcal{G}] = a\mathbb{E}[\xi|\mathcal{G}] + b\mathbb{E}[\eta|\mathcal{G}];$$

性质（2）（ \mathcal{G} 可测线性） 对任意 $A \in \mathcal{G}$ ，有

$$\mathbb{E}[\xi \cdot 1_A|\mathcal{G}](\omega) = 1_A(\omega) \cdot \mathbb{E}[\xi|\mathcal{G}](\omega) \text{ a.s..}$$

更一般的，设 $\mathbb{E}\xi$ 有意义， η 为 \mathcal{G} -可测随机变量，那么

$$\mathbb{E}[\xi\eta|\mathcal{G}] = \eta\mathbb{E}[\xi|\mathcal{G}] \text{ a.s.};$$

性质（3）（正算子/单调性） 如果 $\xi \geq 0$ a.s.，那么 $\mathbb{E}[\xi|\mathcal{G}] \geq 0$ a.s.。更一般的，设 $\mathbb{E}\xi, \mathbb{E}\eta$ 有意义，且 $\xi \leq \eta$ a.s.，则 $\mathbb{E}[\xi|\mathcal{G}] \leq \mathbb{E}[\eta|\mathcal{G}]$ a.s.。特别的， $|\mathbb{E}[\xi|\mathcal{G}]| \leq \mathbb{E}[|\xi||\mathcal{G}]$ a.s.；

性质（4） 对任意常数 $a \in \mathbb{R}$ ， $\mathbb{E}[a|\mathcal{G}] = a$ a.s.；进而如果 ξ 是 \mathcal{G} -可测的，那么 $\mathbb{E}[\xi|\mathcal{G}] = \xi$ ；

性质（5） 设 $\mathbb{E}\xi$ 有意义。又设 $\mathcal{G} \subset \mathcal{G}_1 \subset \mathcal{F}$ 均为 σ -代数，则

$$\mathbb{E}[\mathbb{E}[\xi|\mathcal{G}_1]|\mathcal{G}] = \mathbb{E}[\xi|\mathcal{G}] \text{ a.s.}$$

$$\mathbb{E}[\mathbb{E}[\xi|\mathcal{G}|\mathcal{G}_1] = \mathbb{E}[\xi|\mathcal{G}] \text{ a.s.};$$

性质（6） 设 $\mathbb{E}\xi$ 有意义。如果 ξ 与 \mathcal{G} 独立（其定义为： ξ 生成的 σ -代数与 \mathcal{G} 独立；或者等价的，任给 $A \in \mathcal{G}$ ， ξ 与 A 独立，即 $\mathbb{P}(\{\xi \leq x\} \cap A) = \mathbb{P}(\xi \leq x) \cdot \mathbb{P}(A), \forall x \in \mathbb{R}, A \in \mathcal{G}$ ），那么

$$\mathbb{E}[\xi|\mathcal{G}] = \mathbb{E}\xi \text{ a.s.};$$

性质（7）（条件 Jensen 不等式） 设 g 为 \mathbb{R} 上连续凸函数， $\mathbb{E}[\xi|\mathcal{G}], \mathbb{E}[g(\xi)|\mathcal{G}]$ 有定义。那么

$$g(\mathbb{E}[\xi|\mathcal{G}]) \leq \mathbb{E}[g(\xi)|\mathcal{G}] \text{ a.s.};$$

性质（8）（条件 Hölder 不等式） 设 $\frac{1}{p} + \frac{1}{q} = 1, p, q > 1$ 。那么

$$\mathbb{E}[|\xi\eta||\mathcal{G}] \leq (\mathbb{E}[|\xi|^p|\mathcal{G}])^{1/p} \cdot (\mathbb{E}[|\eta|^q|\mathcal{G}])^{1/q} \text{ a.s.};$$

性质（9）（条件矩不等式）设 $0 < s < t$ 。那么

$$(\mathbb{E}[|\xi|^s|\mathcal{G}])^{1/s} \leq (\mathbb{E}[|\xi|^t|\mathcal{G}])^{1/t} \text{ a.s.};$$

性质（10）（条件 Minkowski 不等式）设 $p > 0$ 。

- (i) $p \geq 1$ 时, $(\mathbb{E}[|\xi + \eta|^p|\mathcal{G}])^{1/p} \leq (\mathbb{E}[|\xi|^p|\mathcal{G}])^{1/p} + (\mathbb{E}[|\eta|^p|\mathcal{G}])^{1/p} \text{ a.s.};$
- (ii) $0 < p < 1$ 时, $\mathbb{E}[|\xi + \eta|^p|\mathcal{G}] \leq \mathbb{E}[|\xi|^p|\mathcal{G}] + \mathbb{E}[|\eta|^p|\mathcal{G}] \text{ a.s.}。$

7.3 条件数学期望的积分变换公式

在讨论数学期望的计算时, 那里有所谓的积分变换公式, 当初我们利用这个公式发展出了利用分布函数来计算数学期望的办法。到了条件数学期望的场合, 同样有类似的积分变换公式。我们在下文中进行讨论。

给定两个概率空间 $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ 和 $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, 其上数学期望算子分别记作 $\mathbb{E}_1, \mathbb{E}_2$ 。假定有可测变换

$$T : (\Omega_1, \mathcal{F}_1) \rightarrow (\Omega_2, \mathcal{F}_2)$$

使得 $\mathbb{P}_2 = \mathbb{P}_1 \circ T^{-1}$ 。假定 $\varphi : (\Omega_2, \mathcal{F}_2) \rightarrow \mathbb{R}$ 是概率空间 $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ 上的可积（或非负）随机变量, 那么 $\varphi \circ T$ 是概率空间 $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ 上的可积（或非负）随机变量, 并且

$$\mathbb{E}_1[\varphi \circ T] = \mathbb{E}_2[\varphi].$$

这是我们之前就讲过的积分变换公式。

现在, 我们进一步假定 $\mathcal{G}_2 \subset \mathcal{F}_2$ 是一个子 σ -代数, 那么 $T^{-1}\mathcal{G}_2 \subset \mathcal{F}_1$ 。在上述假定下, 本节我们将证明下面的条件数学期望版本的积分变换公式:

$$\mathbb{E}_1[\varphi \circ T | T^{-1}\mathcal{G}_2] = \mathbb{E}_2[\varphi | \mathcal{G}_2] \circ T. \quad (7.13)$$

上面公式也可以表述为:

$$\mathbb{E}_{\mathbb{P}_1}[\varphi \circ T | T^{-1}\mathcal{G}_2] = \mathbb{E}_{\mathbb{P}_1 \circ T^{-1}}[\varphi | \mathcal{G}_2] \circ T.$$

请参考下面的交换图:

$$\begin{array}{ccc} (\Omega_1, \mathcal{F}_1, \mathbb{P}_1; T^{-1}\mathcal{G}_2) & \xrightarrow{T} & (\Omega_2, \mathcal{F}_2, \mathbb{P}_1 \circ T^{-1}; \mathcal{G}_2) \\ & \searrow \varphi \circ T & \downarrow \varphi \\ & & \mathbb{R} \end{array}$$

现在我们在可积性条件下来证明(7.13)。事实上, 设 $\xi_2 := \mathbb{E}_2[\varphi | \mathcal{G}_2]$, 显然有 $\xi_1 := \xi_2 \circ T$ 是概率空间 $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ 上的随机变量, 并且它是 $T^{-1}\mathcal{G}_2$ 可测的。于是对任意 $A_1 \in T^{-1}\mathcal{G}_2$, 存在 $A_2 \in \mathcal{G}_2$ 使得 $A_1 = T^{-1}A_2$, 并且

$$\begin{aligned} \mathbb{E}_1[\xi_1 \cdot 1_{A_1}] &= \mathbb{E}_1[\xi_2 \circ T \cdot 1_{A_2} \circ T] \\ &= \mathbb{E}_2[\xi_2 \cdot 1_{A_2}] = \mathbb{E}_2[\varphi \cdot 1_{A_2}] \\ &= \mathbb{E}_1[\varphi \circ T \cdot 1_{A_2} \circ T] = \mathbb{E}_1[\varphi \circ T \cdot 1_{A_1}]. \end{aligned}$$

由 $A_1 \in T^{-1}\mathcal{G}_2$ 的任意性, $\mathbb{E}_1[\varphi \circ T | T^{-1}\mathcal{G}_2] = \xi_2 \circ T$ 几乎处处成立, 即(7.13)几乎处处成立。

7.4 条件分布律及条件数学期望的计算公式

有了一个事件 A 关于一个子 σ -代数 \mathcal{G} 的条件概率 $\mathbb{P}(A|\mathcal{G})$ 的定义，很自然也可以引出一个随机变量 ξ 关于一个子 σ -代数 \mathcal{G} 的条件分布的定义：我们称 $\mathbb{P}(\xi \leq x|\mathcal{G})$ 为随机变量 ξ 关于子 σ -代数 \mathcal{G} 的条件分布。

定义 7.4.1. 假设 $X_i \sim \mu_i, i = 1, 2$ ，其中 μ_1, μ_2 是分布测度。

如果存在一族对 μ_2 -几乎处处 x_2 定义的概率测度族 $\{\mu_{x_2}^{(1)}\}_{x_2}$ ，使得对任意可测集 A_1, A_2

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2) = \int_{A_2} \mu_{x_2}^{(1)}(A_1) d\mu_2(x_2), \quad (7.14)$$

则称这族概率测度族 $\{\mu_{x_2}^{(1)}\}_{x_2}$ 为 X_1 关于 X_2 的条件分布，记作

$$X_1|X_2 = x_2 \sim \mu_{x_2}^{(1)}. \quad (7.15)$$

如果存在二元可测函数

$$(x_1, x_2) \mapsto F_{X_1|X_2}(x_1|x_2),$$

使得对任意固定 x_2 ， $F_{X_1|X_2}(\cdot|x_2)$ 是一个概率分布函数，并且

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = \int_{(-\infty, x_2]} F_{X_1|X_2}(x_1|x_2) d\mu_2(x_2), \quad (7.16)$$

那么我们称概率分布函数族 $\{F_{X_1|X_2}(\cdot|x_2)\}_{x_2}$ 为 X_1 关于 X_2 的条件分布函数（族）。此时，我们记

$$X_1|X_2 = x_2 \sim F_{X_1|X_2}(\cdot|x_2). \quad (7.17)$$

下面例子中的手法本质上能用来证明了随机变量 ξ 关于子 σ -代数 \mathcal{G} 的条件分布

$$F_{\xi|\mathcal{G}}(x|\omega) := \mathbb{P}(\xi \leq x|\mathcal{G})(\omega) \quad (7.18)$$

（它是以 x 为参数的、几乎处处定义的 \mathcal{G} -可测随机变量族）总是有好的版本，使得当样本点 ω 落在一个公共的零测集外时， $F_{\xi|\mathcal{G}}(x|\omega)$ 关于 x 成为一个分布函数。这样的好的条件分布版本 $F_{\xi|\mathcal{G}}(x|\omega)$ 就称为正则条件分布，也称为条件分布律。此时，不难知道，对任意可测函数 φ ，当 $\varphi(\xi)$ 非负或可积时，下面的条件数学期望计算公式成立：

$$\mathbb{E}[\varphi(\xi)|\mathcal{G}](\omega) = \int \varphi(x) dF_{\xi|\mathcal{G}}(x|\omega). \quad (7.19)$$

例 7.2. 给定两个随机变量 X_1, X_2 ，设 μ_i 为 X_i 的分布测度， $i = 1, 2$ 。任意给定 $x \in \mathbb{R}$ ，定义 \mathbb{R} 上测度 μ_x^2

$$\mu_x^2(A) := \mathbb{P}(X_1 \leq x, X_2 \in A), \quad \forall A \in \mathcal{B}.$$

测度族 $\{\mu_x^2\}_{x \in \mathbb{R}}$ 中测度都是有限测度，关于 x 单调递增，且满足： $\mu_x^2 \leq \mu_2$ ，

$$\lim_{x \rightarrow -\infty} \mu_x^2 = 0, \quad \lim_{x \rightarrow \infty} \mu_x^2 = \mu_2$$

及关于 x 的右连续性

$$\lim_{x' \downarrow x} \mu_{x'}^2 = \mu_x^2, \quad \forall x \in \mathbb{R}.$$

它们都关于 μ_2 绝对连续。

考虑上面测度族的子族 $\{\mu_r^2\}_{r \in \mathbb{Q}}$ 。 $\forall r \in \mathbb{Q}$ ，定义可测函数 $F(r|\cdot)$ 如下：

$$F(r|y) := \frac{d\mu_r^2}{d\mu_2}(y) \in [0, 1],$$

亦即 $F(r|y)$ 满足

$$\int_A F(r|y) d\mu_2(y) = \mu_r^2(A) = \mathbb{P}(X_1 \leq r, X_2 \in A), \forall A \in \mathcal{B}. \quad (7.20)$$

以下说明可测函数族 $\{F(r|\cdot)\}_{r \in \mathbb{Q}}$ 在适当缩小定义域后，关于 r 是单调递增的。事实上，对任意 $r_1, r_2 \in \mathbb{Q}, r_1 < r_2$ ，有 $\mu_{r_2}^2 - \mu_{r_1}^2 \geq 0$ 。于是不难知道 $F(r_2|y) - F(r_1|y) \geq 0$ 对 μ_2 -a.e. 的 y 成立，于是我们记此处需要挖去的 μ_2 -零测集为 $A_{r_1, r_2} := \{y : F(r_2|y) - F(r_1|y) < 0\}$ 。取

$$\mathcal{N}_1 := \bigcup_{\substack{r_1 < r_2 \\ r_1, r_2 \in \mathbb{Q}}} A_{r_1, r_2},$$

则 $\mu_2(\mathcal{N}_1) = 0$ 。此时

$$F(\cdot|\cdot) : \mathbb{Q} \times [\mathbb{R} \setminus \mathcal{N}_1] \rightarrow \mathbb{R}, (r, y) \mapsto F(r|y)$$

是良定义的非负可测函数，并且它关于 $r \in \mathbb{Q}$ 是单调递增的。

为了让它进一步满足关于 $r \in \mathbb{Q}$ 的右连续性

$$\lim_{\mathbb{Q} \ni r' \downarrow r} F(r'|y) = F(r|y), \forall r \in \mathbb{Q}$$

以及

$$\lim_{\mathbb{Q} \ni r \downarrow -\infty} F(r|y) = 0, \lim_{\mathbb{Q} \ni r \uparrow \infty} F(r|y) = 1,$$

我们不得不对每个 $r \in \bar{\mathbb{Q}} := \mathbb{Q} \cup \{\pm\infty\}$ 继续挖去一个 μ_2 -零测集 B_r ，使得上面几式对于相应的 $y \notin B_r$ 成立（请读者思考能如此做的原因）。

现在令 $\mathcal{N}_2 := \bigcup_{r \in \bar{\mathbb{Q}}} B_r$ 及 $\mathcal{N} := \mathcal{N}_1 \cup \mathcal{N}_2$ 。于是 $\mu_2(\mathcal{N}) = 0$ ，且

$$F(\cdot|\cdot) : \mathbb{Q} \times [\mathbb{R} \setminus \mathcal{N}] \rightarrow \mathbb{R}, (r, y) \mapsto F(r|y)$$

关于 $r \in \mathbb{Q}$ 具有单调递增、右连续性。之后，对任意 $x \in \mathbb{R} \setminus \mathbb{Q}$ ，令

$$F_{x_1|x_2}(x|y) := \lim_{\mathbb{Q} \ni r \downarrow x} F(r|y).$$

于是我们得到一个在 $\mathbb{R} \times [\mathbb{R} \setminus \mathcal{N}]$ 上定义的二元可测函数（数学专业的同学请思考此处该如何论证它的二元可测性），它满足

$$(i) \quad 0 \leq F_{x_1|x_2}(x|y) \leq 1, \forall x \in \mathbb{R}, y \in \mathbb{R} \setminus \mathcal{N};$$

$$(ii) \quad \text{对任意给定 } y \in \mathbb{R} \setminus \mathcal{N}, F_{x_1|x_2}(\cdot|y) \text{ 是单调递增、右连续函数};$$

$$(iii) \quad \text{对任意给定 } y \in \mathbb{R} \setminus \mathcal{N}, \lim_{x \rightarrow -\infty} F_{x_1|x_2}(x|y) = 0, \lim_{x \rightarrow \infty} F_{x_1|x_2}(x|y) = 1.$$

进一步可定义或修正定义 $y \in \mathcal{N}$ 时 $F_{x_1|x_2}(x|y)$ 的取值：任意取定一个概率分布函数 G ，定义 $F_{x_1|x_2}(x|y) := G(x)$ ， $\forall y \in \mathcal{N}$ 。于是我们得到一个在 $\mathbb{R} \times \mathbb{R}$ 上定义的二元可测函数 $F_{x_1|x_2}(\cdot|\cdot)$ 。在上述条件下，不难证明，对任

意 $x \in \mathbb{R}$

$$\int_A F_{X_1|X_2}(x|y) d\mu_2(y) = \mu_x^2(A) = \mathbb{P}(X_1 \leq x, X_2 \in A), \quad \forall A \in \mathcal{B},$$

于是可以认为总有 $\mathbb{P}(X_1 \leq x|X_2 = y) = F_{X_1|X_2}(x|y)$, 并且定义 7.4.1 中方程(7.16)成立。因此我们称 $F_{X_1|X_2}(\cdot|y)$ 为 X_1 关于 $X_2 = y$ 的条件分布函数(条件分布律), 记作

$$X_1|X_2 = y \sim F_{X_1|X_2}(\cdot|y).$$

本质上, 此处的条件分布律 $F_{X_1|X_2}(\cdot|y)$ 本质上是一族关于参数 y 几乎处处定义的 \mathbb{R} 上的概率分布函数。

在上面记号的基础上, 对任意非负可测或有界可测函数 φ (或使得 $\varphi(\xi, \eta)$ 可积的可测函数 φ), 我们不难证明

$$\mathbb{E}[\varphi(X_1, X_2)|X_2 = y] = \int \varphi(x, y) dF_{X_1|X_2}(x|y). \quad (7.21)$$

这个公式与我们之前得到的(无条件)数学期望的计算公式十分相像, 仅仅是把之前公式中的(无条件)分布函数修改为相应的条件分布函数。

上述讨论显得比较抽象。为获得一些直观, 我们来看看容易理解的情形: ξ, η 均为离散型随机变量。

例 7.3. 设 ξ, η 均为离散型随机变量, 它们的联合分布如下:

$$\mathbb{P}(\xi = a_i, \eta = b_j) = p_{i,j}, \quad i = 1, \dots, M, \quad j = 1, \dots, N,$$

其中 $\sum_{1 \leq i \leq M} \sum_{1 \leq j \leq N} p_{i,j} = 1$ 。令

$$p_{i,\cdot} = \sum_{j=1}^N p_{i,j}, \quad p_{\cdot,j} = \sum_{i=1}^M p_{i,j}.$$

此时, $\mathbb{P}(\xi = a_i) = p_{i,\cdot}$, $\mathbb{P}(\eta = b_j) = p_{\cdot,j}$ 。对应上例中符号体系, $(X_1, X_2) = (\xi, \eta)$, 我们有

$$\mu_x^2(A) = \mathbb{P}(\xi \leq x, \eta \in A) = \sum_{i=1}^M \sum_{j=1}^N p_{i,j} \cdot 1_{[a_i, \infty)}(x) \cdot 1_A(b_j),$$

并且 $\mu_2(A) = \mathbb{P}(\eta \in A)$, $\mu_2(\{b_j\}) = p_{\cdot,j}$ 。因此 μ_x^2, μ_2 都是离散型分布。进而易知, $F(x|y) := \mathbb{P}(\xi \leq x|\eta = y)$ 只对 $y \in \{b_j : 1 \leq j \leq N\}$ 有意义, 并且

$$F(x|b_j) = \frac{d\mu_x^2}{d\mu_2}(b_j) = \frac{\mu_x^2(\{b_j\})}{\mu_2(\{b_j\})} = \sum_{i=1}^M \frac{p_{i,j}}{p_{\cdot,j}} \cdot 1_{[a_i, \infty)}(x).$$

$F(\cdot|b_j)$ 是一个离散型的概率分布函数, 它诱导如下经典的条件概率分布列:

$$\mathbb{P}(\xi = a_i|\eta = b_j) = \frac{p_{i,j}}{p_{\cdot,j}}, \quad i = 1, 2, \dots, M,$$

并且, 此时

$$\mathbb{E}[\varphi(\xi, \eta)|\eta = b_j] = \sum_i \varphi(a_i, b_j) \cdot \frac{p_{i,j}}{p_{\cdot,j}}. \quad (7.22)$$

在下一章，我们将会讨论 (ξ, η) 服从连续型分布时 ξ 关于 η 的条件分布律的问题，那时我们将会得到条件分布律的另一种刻画：**条件密度**。

以上提及的条件分布函数、条件概率分布列、条件密度都是条件分布律的表现形式。

7.5 条件数学期望与独立性

当随机变量/向量 X, Y 相互独立时，关于条件数学期望我们有下面非常实用的定理。

定理 7.5.1. 设 X, Y 相互独立， φ 是使下述讨论有意义的可测函数，那么

$$\mathbb{E}[\varphi(X, Y)|Y] = \mathbb{E}[\varphi(X, y)]|_{y=Y}.$$

特别的，对任意可测集 A, B

$$\mathbb{P}(X \in A, Y \in B|Y) = \mathbb{P}(X \in A) \cdot 1_B(Y).$$

参考证明： 设 X, Y 的分布测度分别为 μ_X, μ_Y ，记 $\psi(y) = \mathbb{E}[\varphi(X, y)]$ 。那么

$$\psi(y) = \int \varphi(x, y) d\mu_X(x),$$

进而对任意可测集 B ,

$$\begin{aligned} & \mathbb{E}[\psi(Y) \cdot 1_B(Y)] \\ &= \int \psi(y) \cdot 1_B(y) d\mu_Y(y) \\ &= \int \left[\int \varphi(x, y) d\mu_X(x) \right] \cdot 1_B(y) d\mu_Y(y) \\ &= \int \left[\int \varphi(x, y) \cdot 1_B(y) d\mu_X(x) \right] d\mu_Y(y) \\ &= \int \left[\varphi(x, y) \cdot 1_B(y) \right] d\mu_X \times \mu_Y(x, y) \quad (\text{Fubini 定理}) \\ &= \mathbb{E}[\varphi(X, Y) \cdot 1_B(Y)]. \end{aligned}$$

由定义立即知道定理的第一个结论成立。取 $\varphi(x, y) = 1_A(x) \cdot 1_B(y)$ 就证明了第二个结论。 \square

尽管有上面的严格证明，我们宁愿把上述定理按照下面更直观的逻辑来理解：

$$\begin{aligned} & \mathbb{E}[\varphi(X, Y)|Y = y] \\ &= \mathbb{E}[\varphi(X, y)|Y = y] \quad (Y = y \text{ 的代换}) \\ &= \mathbb{E}[\varphi(X, y)] \quad (X, Y \text{ 相互独立}). \end{aligned}$$

此外，我们还有下面使用条件分布律来刻画的独立性的重要判据：

定理 7.5.2. 设 Y 的分布测度为 μ_Y 。那么 X, Y 相互独立的充分必要条件是：存在分布函数 F ，使得对 μ_Y -a.e. y ,

$$X|Y = y \sim F \tag{7.23}$$

成立。当上式成立时， $X \sim F$ 。

参考证明: 设 X 的分布函数为 F_X 。当 X, Y 相互独立时, 对任意可测集 B 及任意点 x , 我们有

$$\begin{aligned}\mathbb{P}(X \leq x, Y \in B) &= \mathbb{P}(X \leq x)\mathbb{P}(Y \in B) \\ &= \int_B F_X(x) d\mu_Y(y).\end{aligned}$$

即 X 关于 $Y = y$ 的条件分布函数为

$$F_{X|Y}(x|y) = F_X(x), \quad \mu_Y\text{-a.e. } y.$$

取 $F = F_X$ 即得(7.23)。

反过来, 如果(7.23)成立, 那么对 μ_Y -a.e. y 及任意 x

$$\mathbb{P}(X \leq x|Y = y) = F(x).$$

于是对任意点 x, y

$$\begin{aligned}\mathbb{P}(X \leq x, Y \leq y) &= \mathbb{E}[\mathbb{P}(X \leq x|Y) \cdot 1_{\{Y \leq y\}}] \\ &= \mathbb{E}[\mathbb{P}(X \leq x|Y = t) \Big|_{t=Y} \cdot 1_{\{Y \leq y\}}] \\ &= \mathbb{E}[F(x) \cdot 1_{\{Y \leq y\}}] = F(x) \cdot \mathbb{P}(Y \leq y).\end{aligned}$$

令 $y \rightarrow \infty$, 立即得到 $\mathbb{P}(X \leq x) = F(x)$, 即 $X \sim F$ 。进而

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y), \forall x, y.$$

因此 X, Y 相互独立。 □

习 题 7

习题 7.1. 设随机变量 X, Y 相互独立, 具有共同的分布函数 F 。证明:

$$\mathbb{E}[|X - Y|] = 2 \int F(x)[1 - F(x)]dx.$$

特别的, 如果 m 是分布 F 的中位数 (即 $F(m) \geq \frac{1}{2}$ 且 $F(m-) \leq \frac{1}{2}$), 那么

$$\mathbb{E}|X - m| \leq \mathbb{E}|X - Y|.$$

习题 7.2. 设随机变量 $X \sim F$, m_X 是分布 F 的中位数, 证明:

$$\mathbb{E}|X - m_X| \leq \mathbb{E}|X - y|, \forall y \in \mathbb{R}.$$

利用此性质, 也可以证明: 如果 Y 与 X 相互独立 (不必同分布), 那么

$$\mathbb{E}|X - m_X| \leq \mathbb{E}|X - Y|.$$

【思考: 如果 X, Y 独立, 各自分布函数分别为 F, G , 能否给出 $\mathbb{E}[|X - Y|]$ 的类似上一习题的积分表达式, 能否基于此来论证本习题最后的不等式。回到 Y 退化成常数的情况, 这本质上也给出了本习题的第一个不等式。】

习题 7.3. 设 X, Y 相互独立、同分布且可积。证明: $\mathbb{E}|X - \mathbb{E}X| \leq \mathbb{E}|X - Y|$ 。

习题 7.4. 设 X, Y 相互独立、同分布且可积。证明: $\mathbb{E}|X - Y| \leq \mathbb{E}|X + Y|$ 。

【注记: 离散原型为实数系中的不等式 $\sum_{1 \leq i, j \leq n} |x_i - x_j| \leq \sum_{1 \leq i, j \leq n} |x_i + x_j|$ 。对

于随机情形, 设 X 的分布函数为 F , 可以计算出 $\mathbb{E}(X + y)^+ = \int [1 - F(x)] \cdot$

$1_{\{y > -x\}} \mathrm{d}x$, 进而由此算出 $\mathbb{E}(X + Y)^+ = \int [1 - F(x)][1 - F(-x)] \mathrm{d}x$ 以及 $\mathbb{E}|X + Y| = \int [(1 - F(x))(1 - F(-x)) + F(x)F(-x)] \mathrm{d}x$ 。结合习题 7.1 进一步算出 $\mathbb{E}[|X + Y| - |X - Y|] = \int [1 - F(x) - F(-x)]^2 \mathrm{d}x$ 。】

§ 8

随机变量 (II)

在第 5 章我们介绍了随机变量/向量的概念和离散型随机变量/向量。在本章我们继续介绍一类重要的随机变量/向量：**连续型随机变量/向量**，它们对应的分布称为**连续型分布**。

离散型随机变量与连续型随机变量是初等概率论中最重要的两类随机变量，它们对应的复合随机变量的分布律或其他一些统计特征量（如均值、方差、协方差）的计算（包括基于分布律等的计算论证或否认独立性等）是初等概率论中的典型问题，需要读者认真学习掌握相关的理论与方法。

8.1 连续型分布与密度函数的定义

最容易理解的连续型分布发生在几何概率模型中，那里本质上定义了某区域上的均匀分布。

例 8.1. 设 $D \subset \mathbb{R}^n$ 是 \mathbb{R}^n 中 *Borel* 可测集，且 $0 < |D| < \infty$ 。取 $\mathcal{B}_D = \mathcal{B}(D)$ 为 D 的所有 *Borel* 可测子集全体。于是可以取概率空间 (D, \mathcal{B}_D, m_D) 是样本空间为 D 的几何概率模型，即

$$m_D(A) := \frac{|A \cap D|}{|D|}, \forall A \in \mathcal{F}.$$

定义 $\xi(\omega) := \omega, \forall \omega \in D$ ，取 $\rho_D(x) := 1_D(x)/|D|$ ，则

$$\mathbb{P}(\xi \in A) = m_D(A) = \int_A \rho_D(x) dx.$$

对于任何概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中满足

$$\mathbb{P}(\xi \in A) = \int_A \rho_D(x) dx, \forall A \in \mathcal{B}^n$$

的随机变量 ξ ，我们就称 ξ 服从 D 上的均匀分布，记作 $\xi \sim U(D)$ ，同时也称 ρ_D 为 ξ 的（概率）密度函数。

例 8.2. 对于 $D = (0, 1)$ ，上例中构建了一个概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中的一个服从 $(0, 1)$ 上均匀分布的特殊随机变量 ξ ，我们记 $\xi \sim U(0, 1)$ ；有时我们把它称为**标准均匀分布**。此时它的密度函数为 $\rho_D(x) = 1_{(0,1)}(x)$ 。这个随机变量

的分布函数为

$$F(x) := \mathbb{P}(\xi \leq x) = \int_{(-\infty, x]} 1_{(0,1)}(x) dx = x, \quad \forall x \in [0, 1].$$

容易知道，随机变量 $\eta := 2\xi \pmod{1}$ 也是一个服从 $(0, 1)$ 上均匀分布的随机变量。这表明：随机变量的分布律信息通常不足以决定随机变量的映射法则。

对于 $a < b$ ，我们把 (a, b) 上均匀分布记作 $U(a, b)$ 。容易知道，它的密度函数为 $\frac{1}{b-a} \cdot 1_{(a,b)}(x)$ 。另外，如果 $\xi \sim U(0, 1)$ ，那么 $a + (b-a) \cdot \xi \sim U(a, b)$ 。

类似的，我们可以定义具有密度函数的随机变量，即所谓的连续型随机变量。

定义 8.1.1. 给定可测函数 $\rho: \mathbb{R}^n \rightarrow \mathbb{R}$ ，如果随机向量 ξ 满足

$$\mathbb{P}(\xi \in A) = \int_A \rho(x) dx, \quad \forall A \in \mathcal{B}^n,$$

则称 ρ 是 ξ 的（联合）密度函数；此时我们认为 ξ 是连续型随机向量（当 $n = 1$ 时就称为连续型随机变量）。

根据实变函数中 *Lebesgue* 积分的理论（可以参见本书第 4 章），不难知道上述概率密度函数 ρ 一定是非负可测函数，并且

$$\int_{\mathbb{R}^n} \rho(x) dx = 1. \quad (8.1)$$

很多时候，我们也把具有性质(8.1)的非负可测函数 ρ 直接称为 \mathbb{R}^n 上的（概率）密度函数，简称密度。

以下我们简略地探讨一维随机变量 ξ 的分布函数 F 具有何种性质时就一定是一个连续型随机变量。这个问题的准确陈述如下：给定分布函数 $F(x) := \mathbb{P}(\xi \leq x)$ 。何时存在 $\rho \in L^1(\mathbb{R})$ ，使得

$$F(x) = \int_{-\infty}^x \rho(t) dt, \quad \forall x \in \mathbb{R}. \quad (8.2)$$

定理 4.3.6 告诉我们，这等价于要求 F 是绝对连续函数，并且此时 $\rho = F'$ 几乎处处成立。

在以后，为了节约语言，如果连续型随机变量/向量 X 的（联合）密度函数为 ρ ，我们将简单记作

$$X \sim \rho(x) dx.$$

8.2 连续型随机变量的数学期望计算公式

假设有连续型随机变量/向量 X ，其中

$$X \sim \rho(x) dx.$$

给定可测函数 φ ，假设我们可以定义复合随机变量 $Y := \varphi(X)$ 。根据数学期望的计算公式(6.6)，我们立即得到 Y 的数学期望的计算公式：

$$\mathbb{E}[\varphi(X)] = \int \varphi(x) \rho(x) dx, \quad (8.3)$$

只要上述公式右端的积分有意义。

8.3 边缘密度、条件密度与条件分布函数

设 $d \geq 2$ ，且随机行向量 $X = (X_1, \dots, X_d)$ 以 ρ 为联合密度函数的连续型随机向量。容易知道每个分量 X_i 仍然是连续型随机变量，具有密度 ρ_i ，即 $X_i \sim \rho_i(x_i)dx_i$ ，其中

$$\rho_i(x_i) = \int_{\mathbb{R}^{d-1}} \rho(x) dx_1 \cdots \widehat{dx_i} \cdots dx_d.$$

上式中， $\widehat{dx_i}$ 表示积分表达式中把 dx_i 项去掉。此处，我们称 ρ_i 为随机行向量 $X = (X_1, \dots, X_d)$ （或联合密度 ρ ）的**第 i 个边缘分布密度**，简称**边缘分布密度**。一般的，对任意 $2 \leq n < d$ ，以及 $1 \leq j_1 < \dots < j_n \leq d$ ， $(X_{j_1}, \dots, X_{j_n})$ 仍然是连续型随机向量，它对应的联合分布密度也称为**边缘分布密度**。

设 $(X, Y) \sim \rho(x, y)dx dy$ ，即随机（行）向量 (X, Y) 是连续型的，并且具有联合密度函数 ρ 。这时候， $X \sim \rho_X(x)dx, Y \sim \rho_Y(y)dy$ ，其中

$$\rho_X(x) = \int \rho(x, y) dy, \quad \rho_Y(y) = \int \rho(x, y) dx.$$

我们定义（下式中约定 $\frac{0}{0} = 0$ ）

$$\rho_{X|Y}(x|y) := \frac{\rho(x, y)}{\rho_Y(y)}, \quad (8.4)$$

根据例 7.2 中对条件分布函数 $F_{X|Y}(\cdot|\cdot)$ 的定义，我们知道

$$\begin{aligned} \mathbb{P}(X \leq x, Y \in A) &= \int_A \left[\int_{-\infty}^x \rho(u, v) du \right] dv \\ &= \int_A \left[\int_{-\infty}^x \rho_{X|Y}(u|v) du \right] \rho_Y(v) dv, \end{aligned}$$

即有

$$F_{X|Y}(x|y) = \int_{-\infty}^x \rho_{X|Y}(u|y) du$$

是 X 关于 $Y = y$ 的条件分布函数；固定 y ，关于自变量 x 函数 $F_{X|Y}(x|y)$ 仍然是一个连续型分布函数，对应的密度恰好为 $\rho_{X|Y}(x|y)$ ，因此我们把 $\rho_{X|Y}(x|y)$ 称为 X 关于 $Y = y$ 的**条件分布密度**，记作

$$X|Y = y \sim \rho_{X|Y}(x|y)dx.$$

此时，不难知道条件数学期望有如下计算公式

$$\mathbb{E}[\varphi(X, Y)|Y = y] = \int \varphi(x, y) \rho_{X|Y}(x|y) dx, \quad (8.5)$$

其中 $\varphi(x, y)$ 是使上式右端有意义的可测函数。

对于两个随机变量 X, Y ，如果 $(X, Y) \sim \rho(x, y)dx dy$ ，实变函数的理论告诉我们：对 Lebesgue 几乎处处的 y ，极限

$$F_{X|Y}(x|y) = \lim_{\varepsilon \downarrow 0} \mathbb{P}(X \leq x | Y \in (y - \varepsilon, y + \varepsilon)) \quad (8.6)$$

存在，并且

$$F_{X|Y}(x|y) = \int_{-\infty}^x \rho_{X|Y}(t|y) dt.$$

关于两个连续型随机变量/向量之间的独立性，我们有下面的结果：这是定理 5.1.1 结合密度函数定义的简单推论，因此此处的证明留给读者。

定理 8.3.1. 假设 $X_i \sim \rho_i(x_i)dx_i, i = 1, 2$ 。则 X_1 与 X_2 相互独立的充分必要条件是： (X_1, X_2) 仍然是连续型分布，并且对应的联合密度函数为

$$\rho(x_1, x_2) = \rho_1(x_1) \cdot \rho_2(x_2). \quad (8.7)$$

上述等号是在 Lebesgue 几乎处处相等意义下理解的。上述联合密度函数满足的条件也等价于

$$\rho_{X_1|X_2}(x_1|x_2) = \rho_1(x_1), \quad (8.8)$$

即“条件密度”等于“无条件密度”。

8.4 概率微元法

在初等概率论中，经常要考虑连续型随机变量/向量经过光滑（或分片光滑）映射变成新的随机变量/向量后的分布律的计算问题。本小节的目的就是为此提供有关的方法与理论。

首先做一些记号上的约定。当 $X \sim \rho(x)dx$ 时，我们知道

$$\mathbb{P}(X \in A) = \int_A \rho(x)dx, \quad \forall \text{ 可测集 } A. \quad (8.9)$$

我们认为，上式对应的“概率微元”形式的表达式为

$$\mathbb{P}(X = x + dx) = \rho(x)dx, \quad (8.10)$$

其中 $x + dx$ 理解为 x 处的一个“微分形式”的“平行四边形”（含“矩形”），我们称之为“微元区域”； x 点在这个“微元区域”的边界或内部均可。而关系 $X = x + dx$ 应理解为 $X \in x + dx$ ，即随机变量/向量 X 落入这个“微元区域”中。上述方程的右端项 dx 理解为“微元区域” $x + dx$ 的体积。

基于上述约定，我们来进行前述问题的探讨，即考虑 $Y := \varphi(X)$ 的分布律的计算。我们总假设对某区域 $\mathcal{D} \subset \mathbb{R}^d$

$$\mathbb{P}(X \in \mathcal{D}) = 1.$$

(1) 假定 $\varphi: \mathcal{D} \rightarrow \mathbb{R}^d$ 可逆，并且 $\varphi^{-1}: \varphi(\mathcal{D}) \rightarrow \mathcal{D}$ 光滑。

此时，我们有

$$\begin{aligned} & \mathbb{P}(Y = y + dy) \\ &= \mathbb{P}(\varphi(X) = y + dy) \\ &= \mathbb{P}(X = \varphi^{-1}(y + dy)) \\ &= \mathbb{P}(X = \varphi^{-1}(y) + D\varphi^{-1}(y)dy) \quad (\text{Taylor 公式, } D \text{ 是导算子}) \\ &= \rho(\varphi^{-1}(y)) \cdot |\det(D\varphi^{-1}(y))|dy. \end{aligned}$$

上述形式演算告诉我们， $Y := \varphi(X)$ 的分布密度为

$$\rho_Y(y) = \rho(\varphi^{-1}(y)) \cdot |\det(D\varphi^{-1}(y))|.$$

以上形式演算实际上是可以严谨化的，参考 $X \sim \rho(x)dx$ 时的方程(8.9)，利用对应的积分表达式进行相关概率演算即可。

于是我们有下面的定理。

☞ **定理 8.4.1.** X 是 d 维随机向量, $\mathcal{D} \subset \mathbb{R}^d$ 满足 $\mathbb{P}(X \in \mathcal{D}) = 1$. 又设

$$\varphi: \mathcal{D} \rightarrow \mathbb{R}^d$$

是单射, 并且 $\varphi^{-1}: \varphi(\mathcal{D}) \rightarrow \mathcal{D}$ 光滑, 那么当 $X \sim \rho(x)dx$ 时, $Y := \varphi(X)$ 的密度函数为

$$\rho_Y(y) = \rho(\varphi^{-1}(y)) \cdot |\det(D\varphi^{-1}(y))|. \quad (8.11)$$

(2) 假定 φ 分片可逆, 并且分片的逆映射是光滑的。为方便, 不妨设

$$\varphi: \mathcal{D} = \mathcal{D}_1 \uplus \mathcal{D}_2 \rightarrow \mathbb{R}^d,$$

其中对于 $i = 1, 2$, $\varphi_i := \varphi|_{\mathcal{D}_i}: \mathcal{D}_i \rightarrow \mathbb{R}^d$ 是单射, 并且 $\varphi_i^{-1}: \varphi(\mathcal{D}_i) \rightarrow \mathcal{D}_i$ 是光滑的。

类似于 (1) 中演算, 我们有

$$\begin{aligned} & \mathbb{P}(Y = y + dy) \\ &= \mathbb{P}(\varphi(X) = y + dy) \\ &= \sum_{i=1}^2 \mathbb{P}(\varphi(X) = y + dy, X \in \mathcal{D}_i) \\ &= \sum_{i=1}^2 \mathbb{P}(X = \varphi_i^{-1}(y + dy), X \in \mathcal{D}_i) \\ &= \sum_{i=1}^2 \mathbb{P}(X = \varphi_i^{-1}(y) + D\varphi_i^{-1}(y)dy) \cdot 1_{\varphi(\mathcal{D}_i)}(y) \\ &= \left[\sum_{i=1}^2 \rho(\varphi_i^{-1}(y)) \cdot |\det(D\varphi_i^{-1}(y))| \cdot 1_{\varphi(\mathcal{D}_i)}(y) \right] dy. \end{aligned}$$

上述形式演算告诉我们, $Y := \varphi(X)$ 的分布密度为

$$\rho_Y(y) = \sum_{i=1}^2 \rho(\varphi_i^{-1}(y)) \cdot |\det(D\varphi_i^{-1}(y))| \cdot 1_{\varphi(\mathcal{D}_i)}(y).$$

同样, 我们下面的定理。

☞ **定理 8.4.2.** X 是 d 维随机向量, $\mathcal{D} \subset \mathbb{R}^d$ 满足 $\mathbb{P}(X \in \mathcal{D}) = 1$. 又设映射

$$\varphi: \mathcal{D} = \biguplus_{i=1}^N \mathcal{D}_i \rightarrow \mathbb{R}^d$$

满足: 对任意 $1 \leq i \leq N$, $\varphi_i := \varphi|_{\mathcal{D}_i}$ 是单射并且 $\varphi_i^{-1}: \varphi(\mathcal{D}_i) \rightarrow \mathcal{D}_i$ 光滑, 那么当 $X \sim \rho(x)dx$ 时, $Y := \varphi(X)$ 的密度函数为

$$\rho_Y(y) = \sum_{i=1}^N \rho(\varphi_i^{-1}(y)) \cdot |\det(D\varphi_i^{-1}(y))| \cdot 1_{\varphi(\mathcal{D}_i)}(y). \quad (8.12)$$

下面推论的证明留给读者。

◆ **推论 8.4.1.** 假设 $X_i \sim \rho_i(x_i)dx_i, i = 1, 2$, 并且 X_1 与 X_2 相互独立。令 $Y_1 = X_1 + g(X_2)$, 其中 g 为光滑函数。则 Y_1 关于 $Y_2 := X_2$ 的条件密度为

$$\rho_{Y_1|Y_2}(y_1|y_2) = \rho_1(y_1 - g(y_2)).$$

定义 8.4.1. 假设 $X_i \sim \mu_i, i = 1, 2$ 。如果存在非负（二元）可测函数 $\rho_{X_1|X_2}(x_1|x_2)$ 使得：对任意可测集 A_1, A_2

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2) = \int_{A_2} \left[\int_{A_1} \rho_{X_1|X_2}(x_1|x_2) dx_1 \right] d\mu_2(x_2), \quad (8.13)$$

则称 $\rho_{X_1|X_2}(\cdot|\cdot)$ 为 X_1 关于 X_2 的条件密度，记作

$$X_1|X_2 = x_2 \sim \rho_{X_1|X_2}(x_1|x_2) dx_1. \quad (8.14)$$

从第 7 章的例 7.2 中，我们容易知道，上述条件分布密度的概念是有意义的，它完全由随机向量的联合分布律决定。

在上述推广了的条件密度的概念下，我们其实有比上述推论 8.4.1 的更好结果，它放弃了推论中的函数 g 的光滑性和对 X_2 的分布具有密度的要求，见下面的定理。

定理 8.4.3. 设 $X_1 \sim \rho_1(x)dx$ ， g 是可测函数。如果 X_1 与 X_2 相互独立，那么 $Y_1 := X_1 + g(X_2)$ 关于 $Y_2 := X_2$ 的条件密度为

$$\rho_{Y_1|Y_2}(y_1|y_2) = \rho_1(y_1 - g(y_2)).$$

参考证明： 设 $X_2 \sim \mu_2$ 。那么对任意可测集 A_1, A_2

$$\begin{aligned} \mathbb{P}(Y_1 \in A_1, Y_2 \in A_2) &= \mathbb{P}((X_1 + g(X_2), X_2) \in A_1 \times A_2) \\ &= \int \left[\int 1_{A_1}(x_1 + g(x_2)) 1_{A_2}(x_2) \rho(x_1) dx_1 \right] d\mu_2(x_2) \\ &= \int \left[\int 1_{A_1}(y_1) 1_{A_2}(y_2) \rho(y_1 - g(y_2)) dy_1 \right] d\mu_2(y_2) \\ &= \int_{A_2} \left[\int_{A_1} \rho(y_1 - g(y_2)) dy_1 \right] d\mu_2(y_2). \end{aligned}$$

因此定理中结论成立。 \square

在 (ξ, η) 联合分布是连续型的情形下，现在我们来讨论条件数学期望的计算问题。

例 8.3. 设 (ξ, η) 具有联合密度 ρ ，从而 ξ 与 η 分别具有边缘密度 ρ_ξ, ρ_η ：

$$\rho_\xi(x) := \int \rho(x, y) dy, \quad \rho_\eta(y) := \int \rho(x, y) dx.$$

由此不难知道，对任意非负可测或有界可测函数 φ ，我们有

$$\mathbb{E}[\varphi(\xi, \eta)|\eta = y] = \int \varphi(x, y) \rho_{\xi|\eta}(x|y) dx, \quad (8.15)$$

其中

$$\rho_{\xi|\eta}(x|y) := \frac{\rho(x, y)}{\rho_\eta(y)}$$

恰为我们在第 6 章中介绍的 ξ 关于 η 的条件分布密度函数。上述条件数学期望 $\mathbb{E}[\varphi(\xi, \eta)|\eta = y]$ 的计算公式与无条件数学期望的计算公式在形式上是一致的，只是密度上做了调整。

另外，也有人按如下方式来解释条件数学期望 $\mathbb{E}[\varphi(\xi, \eta)|\eta = y]$ ：

$$\mathbb{E}[\varphi(\xi, \eta)|\eta = y] = \lim_{\delta \downarrow 0} \mathbb{E}[\varphi(\xi, \eta)|\eta \in (y - \delta, y + \delta)]. \quad (8.16)$$

读者可以证明，在本例中，上式是几乎处处成立的；对于一般情况上式（有些文献中把这个式子称为 **Borel Density Lemma**）本质上也对，但要严格证明是有难度的。

8.5 常见连续型分布

在本节的开头，我们就给出了均匀分布的概念。在本小节，我们将介绍更多的常见的连续型分布。

我们先探讨两个连续型随机变量独立和 $X_1 + X_2$ 的分布与和项 X_1, X_2 的分布之间的关系。

例 8.4. (连续型随机变量的独立和的分布律) 给定两个相互独立的连续型随机变量 X_1, X_2 ，假设

$$X_1 \sim \rho_1(x)dx, X_2 \sim \rho_2(x)dx.$$

那么 $Y := X_1 + X_2 \sim \rho(x)dx$ ，其中

$$\rho(x) = \rho_1 * \rho_2(x) := \int \rho_1(t)\rho_2(x-t)dt = \int \rho_1(x-t)\rho_2(t)dt.$$

参考证明：这里，注意到定理 8.3.1，我们可以直接进行计算：

$$\begin{aligned} \mathbb{P}(Y \leq x) &= \mathbb{P}(X_1 + X_2 \leq x) \\ &= \int_{\mathbb{R}^2} 1_{\{x_1+x_2 \leq x\}} \cdot \rho_1(x_1) \cdot \rho_2(x_2) dx_1 dx_2 \\ &= \int \left[\int 1_{\{x_1+x_2 \leq x\}} \cdot \rho_1(x_1) \cdot \rho_2(x_2) dx_2 \right] dx_1 \\ &\stackrel{x_2=s-x_1}{=} \int \left[\int 1_{\{s \leq x\}} \cdot \rho_1(x_1) \cdot \rho_2(s-x_1) ds \right] dx_1 \\ &= \int_{-\infty}^x \left[\int \rho_1(x_1) \cdot \rho_2(s-x_1) dx_1 \right] ds \\ &= \int_{-\infty}^x \rho_1 * \rho_2(s) ds. \end{aligned}$$

上述方程中，第三个等号使用了 Fubini 定理化重积分为累次积分，第四个等号是对内层的积分施行了换元，第五个等号使用了 Fubini 定理实现两重积分的交换次序，最后一个等号是利用了分析学中函数卷积的定义。因此本例中的结论成立。 \square

上面的例子表明，两个连续型随机变量的独立和仍然是连续型的。但实际上，对于随机变量的独立和，只要其中某一个随机变量是连续型的，那么独立和就仍然是连续型的；参见下面的例子。

例 8.5. 设随机变量 X_1, X_2 相互独立，其中 $X_1 \sim \rho_1(x)dx$ ， $X_2 \sim \mu_2$ 。那么 $Y := X_1 + X_2$ 是连续型随机变量，具有密度 ρ_Y ：

$$\rho_Y(x) = \rho_1 * \mu_2(x) := \int \rho_1(x-t)d\mu_2(t).$$

这个结论的证明方法与上一例类似，留给读者思考。

例 8.6. (指数分布 $\mathcal{E}(\lambda)$) 给定 $\lambda > 0$ 。设 $U \sim U(0, 1)$ ，取

$$X := -(\log U)/\lambda,$$

则容易知道 X 的分布函数为

$$F_\lambda(x) := \mathbb{P}(X \leq x) = (1 - e^{-\lambda x})^+, \text{ 其中 } x \in \mathbb{R}.$$

它是一个连续型分布，具有密度函数

$$\rho(x; \lambda) = \lambda e^{-\lambda x} 1_{(0, \infty)}(x).$$

分布 F_λ 称为参数为 λ 的**指数分布**； $1/\lambda$ 视作尺度参数。在本书中我们简单记作： $X \sim \mathcal{E}(\lambda)$ 。容易知道此时

$$\mathbb{P}(X > t) = e^{-\lambda t}, \quad \forall t \geq 0.$$

特别的，我们称 $\mathcal{E}(1)$ 为**标准指数分布**。

例 8.7. (指数分布的无记忆性) 设 ξ 代表了教室里的日光灯的寿命（假定从不熄灯，除非日光灯坏了，那时认为日光灯寿终正寝了）。到 $t > 0$ 时刻日光灯仍然正常工作，这个事件表示为 $\{\xi > t\}$ ；此时我们称 $\xi_t := \xi - t$ 是 t 时刻日光灯的**剩余寿命**。在 $\xi \sim \mathcal{E}(\lambda)$ 的假设下，我们计算在 t 时刻日光灯仍然正常工作的条件下它的剩余寿命的分布如下：对任意 $s > 0$

$$\begin{aligned} & \mathbb{P}(\xi_t > s | \xi > t) \\ &= \mathbb{P}(\xi > s + t | \xi > t) \\ &= \frac{\mathbb{P}(\xi > s + t)}{\mathbb{P}(\xi > t)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\ &= e^{-\lambda s} = \mathbb{P}(\xi_0 > s). \end{aligned}$$

这表明，在 t 时刻日光灯仍然正常工作的条件下它的剩余寿命就如同它没有使用过一样，仍然服从指数分布 $\mathcal{E}(\lambda)$ 。也就是说，指数分布总满足下面的方程

$$\mathbb{P}(\xi > s + t | \xi > t) = \mathbb{P}(\xi > s), \quad \forall s, t \geq 0. \quad (8.17)$$

这个性质就称为指数分布的**无记忆性**。

例 8.8. (Erlang 分布, $\Gamma(n, \lambda)$ 分布) 现在假设

$$\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(\lambda).$$

定义 $S_n := X_1 + \cdots + X_n$ ，则 S_n 的分布密度为

$$\rho_n(x; \lambda) = \frac{\lambda^n x^{n-1}}{(n-1)!} \cdot e^{-\lambda x} \cdot 1_{(0, \infty)}(x).$$

这是下面将要介绍的 *Gamma* 分布族中的 $\Gamma(n, \lambda)$ 分布，在有的文献中也称为 *Erlang* 分布；其中 n 视作自由度参数， $1/\lambda$ 视作尺度参数。

参考证明：本例中的密度 ρ_n 的计算可以利用例 8.4 中结论通过归纳法来完成证明；也可以在后面的 *Gamma* 分布的半生成性的基础上来实现计算。有关细节留给读者。 \square

例 8.9. (Gamma 分布及其半生成性) 对任意给定的 $\alpha > 0, \lambda > 0$

$$\rho(x; \alpha, \lambda) := \frac{\lambda^\alpha \cdot x^{\alpha-1}}{\Gamma(\alpha)} \cdot e^{-\lambda x} \cdot 1_{(0, \infty)}(x)$$

是一个密度函数。我们称这个密度函数对应的分布为 $\Gamma(\alpha, \lambda)$ 分布，简称 *Gamma 分布*；其中 α 视作“自由度”参数， $1/\lambda$ 视作尺度参数。

设 $X_i \sim \Gamma(\alpha_i, \lambda), i = 1, 2$ ，且 X_1 与 X_2 独立，则

$$X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda).$$

这个性质称为 *Gamma 分布的半生成性*。

参考证明：本例中 $\rho(x; \alpha, \lambda)$ 是密度函数的证明不难，细节留给读者。

下面我们论证例中提及的 *Gamma 分布* 的半生成性。根据例 8.4 的结论，记 $\rho_i(x) := \rho(x; \alpha_i, \lambda), i = 1, 2$ ，我们计算卷积如下：不妨设 $x > 0$

$$\begin{aligned} & \rho_1 * \rho_2(x) \\ &= \int \rho_1(t) \rho_2(x-t) dt \\ &= \int \frac{\lambda^{\alpha_1+\alpha_2} t^{\alpha_1-1} (x-t)^{\alpha_2-1}}{\Gamma(\alpha_1) \Gamma(\alpha_2)} \cdot e^{-\lambda x} \cdot 1_{(0,\infty)}(t) \cdot 1_{(0,\infty)}(x-t) dt \\ &\stackrel{t=xu}{=} \frac{\lambda^{\alpha_1+\alpha_2} \cdot x^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1) \Gamma(\alpha_2)} \cdot e^{-\lambda x} \cdot \int_0^1 u^{\alpha_1-1} (1-u)^{\alpha_2-1} du \\ &= \frac{\lambda^{\alpha_1+\alpha_2} \cdot x^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1) \Gamma(\alpha_2)} \cdot e^{-\lambda x} \cdot B(\alpha_1, \alpha_2) \\ &= \frac{\lambda^{\alpha_1+\alpha_2} \cdot x^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1 + \alpha_2)} \cdot e^{-\lambda x} = \rho(x; \alpha_1 + \alpha_2, \lambda). \end{aligned}$$

当 $x \leq 0$ 时 $\rho_1 * \rho_2(x) = 0$ 。因此例中结论成立。 \square

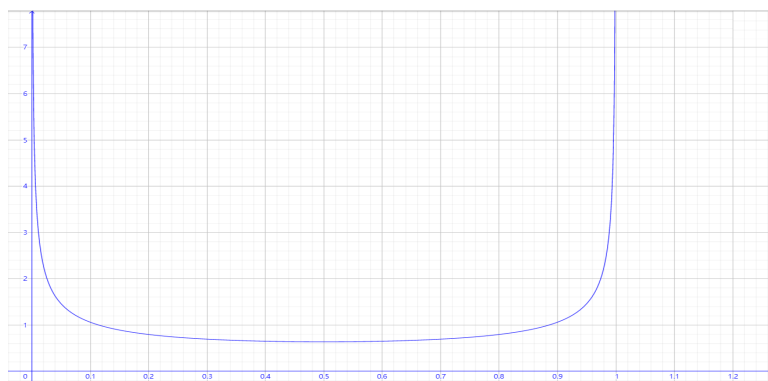


图 8.1: $B(\frac{1}{2}, \frac{1}{2})$ 分布密度函数图

例 8.10. (*Beta 分布*) 对任意给定的 $\alpha > 0, \beta > 0$

$$\rho(x; \alpha, \beta) := \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \cdot 1_{(0,1)}(x)$$

是一个密度函数。我们称这个密度函数对应的分布为 *Beta*(α, β) 分布，简称 *Beta 分布*；其中，由于 *Beta*($\frac{1}{2}, \frac{1}{2}$) 分布的分布函数在 $[0, 1]$ 上为

$$F(x) = \frac{2}{\pi} \arcsin \sqrt{x},$$

这个分布又称为反正弦律。参数 α, β 取自然数的 *Beta* 分布将在讨论均匀分布 $U(0, 1)$ 的次序统计量的分布律问题中出现。

以下我们介绍正态分布及其相关的分布。

我们从一维标准正态分布开始。

例 8.11.（一维标准正态分布） 容易知道

$$\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

是一个密度函数；我们把这个密度对应的分布称为（一维）标准正态分布，记作 $N(0, 1)$ 。它对应的分布函数通常记作 Φ ，即

$$\Phi(x) := \int_{-\infty}^x \varphi(x) dx = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

设 $Z \sim N(0, 1)$ ，容易知道：对任意 $n \in \mathbb{N}$

$$\mathbb{E}[Z^{2n-1}] = 0, \quad \mathbb{E}[Z^{2n}] = (2n-1)!!.$$
 (8.18)

因此， $\mathbb{E}Z = 0$ ， $\text{Var}(Z) = \mathbb{E}[Z^2] = 1$ 。

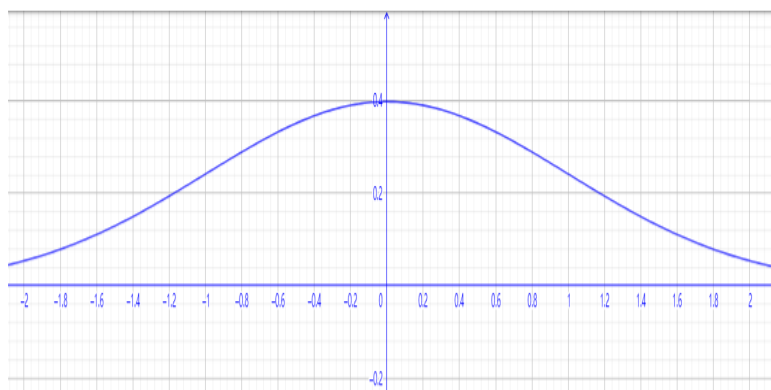


图 8.2: 一维标准正态分布密度函数图

例 8.12.（一维正态分布） 设 $Z \sim N(0, 1)$ 。给定 $\mu \in \mathbb{R}, \sigma > 0$ ，定义

$$X := \mu + \sigma Z.$$

容易知道 X 的密度函数为

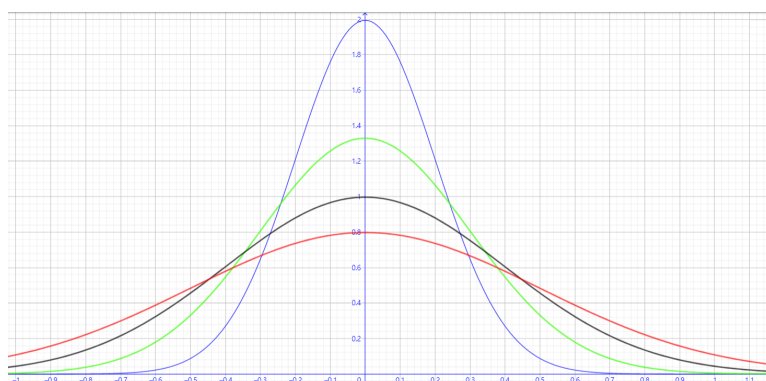
$$\rho(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

我们把这个密度对应的分布称为以 μ 为均值参数，以 σ^2 为方差参数的正态分布（简称正态分布），记作 $N(\mu, \sigma^2)$ 。根据 $\mathbb{E}Z = 0$ ， $\text{Var}(Z) = \mathbb{E}[Z^2] = 1$ ，我们知道

$$\mathbb{E}X = \mu, \quad \text{Var}(X) = \sigma^2.$$

在讲高维正态分布之前，我们先介绍随机向量的协方差矩阵的概念。设 $X = (X_1, \dots, X_n)^T$ 为 n 维随机列向量。如果对任意 $1 \leq i, j \leq n$,

$$\sigma_{i,j} := \text{Cov}(X_i, X_j)$$


 图 8.3: 一维正态分布 $N(0, \sigma^2)$ 密度函数图; $\sigma = 0.2, 0.3, 0.4, 0.5$

都存在, 则我们定义 X 的协方差矩阵为下面的矩阵 Σ :

$$\Sigma := (\sigma_{i,j})_{1 \leq i,j \leq n}.$$

有时仍然把 X 的协方差矩阵记作 $\text{Var}(X) := \Sigma$ 。对随机列向量 $X = (X_1, \dots, X_n)^T$, 记 $\mathbb{E}X := (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^T$, 容易知道

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T]. \quad (8.19)$$

上述数学期望的含义是: 对任意 i, j 指标, 计算出“随机矩阵” $(X - \mathbb{E}X)(X - \mathbb{E}X)^T$ 的 (i, j) 元的数学期望, 作为一个新矩阵的 (i, j) 元, 由此得到的新矩阵就视作矩阵形态随机元的数学期望 $\mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T]$ 。

类似的, 我们还可以定义两个随机列向量 X, Y 之间的协方差矩阵 $\text{Cov}(X, Y)$ (如果方程中的数学期望存在)

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)^T]. \quad (8.20)$$

上述方程中数学期望的理解与(8.19)类似。

不难知道, $X \mapsto \mathbb{E}X$ 是线性的, $(X, Y) \mapsto \text{Cov}(X, Y)$ 是 (共轭) 双线性的。

定理 8.5.1. 设 X, Y 是实数域上的随机列向量, A, B 是常数矩阵, α, β 是常数向量, 使得 $AX + \alpha, BY + \beta$ 有意义, 那么当 $\text{Cov}(X, Y)$ 存在时

$$\mathbb{E}[AX + \alpha] = A(\mathbb{E}X) + \alpha, \quad \mathbb{E}[BY + \beta] = B(\mathbb{E}Y) + \beta, \quad (8.21)$$

$$\text{Cov}(AX + \alpha, BY + \beta) = A\text{Cov}(X, Y)B^T. \quad (8.22)$$

我们稍后将看到, 上述性质在高维正态分布的有关讨论中使用起来非常便捷。

定理 8.5.2. 设矩阵 $\Sigma = (\sigma_{i,j})_{1 \leq i,j \leq n}$ 是 n 维实随机向量 X 的协方差矩阵, 那么 Σ 是对称、非负定矩阵。

参考证明： 取 $a = (a_1, \dots, a_n)^T \in \mathbb{R}^n$ ，则

$$\begin{aligned} a^T \Sigma a &:= \sum_{1 \leq i, j \leq n} \sigma_{i,j} a_i a_j \\ &= \sum_{1 \leq i, j \leq n} \text{Cov}(X_i, X_j) a_i a_j \\ &= \text{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^n a_j X_j\right) \\ &= \text{Var}\left(\sum_{i=1}^n a_i X_i\right) \geq 0. \end{aligned}$$

因此定理结论成立。 \square

例 8.13. (高维标准正态分布) 给定正整数 $n \geq 2$ 。现在设

$$\{Z_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} N(0, 1).$$

我们称随机（列）向量 $Z := (Z_1, \dots, Z_n)^T$ 服从 n 维标准正态分布，记作 $N(0, I_n)$ 。容易知道 Z 的联合分布密度为

$$\varphi_n(x) := \frac{1}{\sqrt{(2\pi)^n}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n x_i^2\right\}.$$

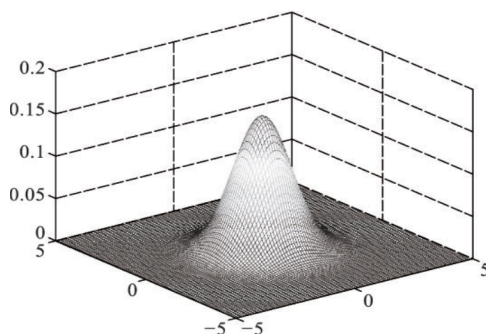


图 8.4: 二维标准正态分布密度函数图

例 8.14. (高维正态分布) 给定正整数 $n \geq 2$ 。现在设 $Z \sim N(0, I_n)$ 。给定 $\mu \in \mathbb{R}^n$ 以及 n 阶可逆方阵 A ，定义

$$X := \mu + AZ.$$

记 $\Sigma := AA^T$ ；它是一个实的对称正定矩阵。容易知道 X 的密度函数为

$$\rho(x; \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left\{-\frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{2}\right\}.$$

我们把这个密度对应的分布称为以 μ 为均值参数，以 Σ 为协方差矩阵参数的高维正态分布（简称正态分布），记作 $N(\mu, \Sigma)$ 。容易知道

$$\mathbb{E}X = \mu, \quad \text{Var}(X) = \Sigma. \quad (8.23)$$

根据上述正态分布的定义方式，我们很容易得到以下的一系列结果。

☞ **定理 8.5.3.** 设 X 是 n 维随机列向量, $X \sim N(\mu, \Sigma)$, 其中 $\mu \in \mathbb{R}^n$, Σ 是 n 阶实对称正定阵。设 A 为 n 阶可逆矩阵, $\beta \in \mathbb{R}^n$, 则

$$Y := \beta + AX \sim N(\tilde{\mu}, \tilde{\Sigma}),$$

其中 $\tilde{\mu} = \beta + A\mu$, $\tilde{\Sigma} = A\Sigma A^T$ 。

◆ **推论 8.5.1.** 设 Z 是 n 维标准正态随机列向量, 即 $Z \sim N(0, I_n)$ 。则对任意单位正交矩阵 $A \in O(n)$,

$$AZ \sim N(0, I_n).$$

根据本书对正态分布的定义, 上述定理和推论的证明都比较简单, 因此留给读者作为练习。

☞ **定理 8.5.4.** 设 $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ 是 $n \geq 2$ 维随机列向量, 其中 X_1 为 n_1 维随机列向量, X_2 为 n_2 维随机列向量, $n_1 + n_2 = n$ 。又设 $X \sim N(\mu, \Sigma)$, 其中 $\mu \in \mathbb{R}^n$, Σ 是 n 阶实对称正定阵, 并且它们根据 $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ 的分块形式有下面的分块表示

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}.$$

那么

(1) $X_1 \sim N(\mu_1, \Sigma_{1,1})$, $X_2 \sim N(\mu_2, \Sigma_{2,2})$, 并且 $\Sigma_{1,2} = \text{Cov}(X_1, X_2)$;

(2) X_1 与 X_2 相互独立的充分必要条件是:

$$\text{Cov}(X_1, X_2) = 0; \quad (8.24)$$

(3) X_1 关于 $X_2 = x_2$ 的条件分布律仍然是正态分布, 具体来说

$$X_1 | X_2 = x_2 \sim N(\tilde{\mu}_1(x_2), \tilde{\Sigma}_{1,1}), \quad (8.25)$$

其中

$$\begin{cases} \tilde{\mu}_1(x_2) &= \mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(x_2 - \mu_2), \\ \tilde{\Sigma}_{1,1} &= \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}. \end{cases}$$

参考证明: 首先, 我们证明 $\Sigma_{1,2} = \text{Cov}(X_1, X_2)$ 。事实上, 这只要注意到

$$X_1 = (I_{n_1}, 0)X, X_2 = (0, I_{n_2})X$$

以及 $\text{Cov}(X, X) = \text{Var}(X) = \Sigma$, 立即利用 $\text{Cov}(X, Y)$ 关于 X, Y 的 (共轭) 双线性性质得到

$$\text{Cov}(X_1, X_2) = (I_{n_1}, 0)\text{Cov}(X, X)(0, I_{n_2})^T = \Sigma_{1,2}.$$

其次, 我们证明: $\Sigma_{1,2} = 0$ 是 X_1 与 X_2 独立的充分必要条件。当 X_1 与 X_2 独立时, 显然有 $\Sigma_{1,2} = \text{Cov}(X_1, X_2) = 0$, 因此我们重点需要证明充分性部分。现在设 $\Sigma_{1,2} = 0$, 从而由于 Σ 是对称矩阵, $\Sigma_{2,1} = \Sigma_{1,2}^T = 0$, 进而 Σ 实际上是分块对角矩阵。根据例 8.14 中密度 $\rho(x; \mu, \Sigma)$ 的表达式, 此时不难验证它满足:

$$\rho(x; \mu, \Sigma) = \rho(x_1; \mu_1, \Sigma_{1,1}) \cdot \rho(x_2; \mu_2, \Sigma_{2,2}).$$

从而 X_1 与 X_2 独立, 并且 $X_1 \sim N(\mu_1, \Sigma_{1,1}), X_2 \sim N(\mu_2, \Sigma_{2,2})$ 。

最后, 我们论证定理中的其他结论。选取合适的矩阵 B , 使得线性变换 $Y = AX$ 具有如下形态

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} I_{n_1} & -B \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

并且 $\text{Cov}(Y_1, Y_2) = 0$ 。这只要取 $B = \Sigma_{1,2}\Sigma_{2,2}^{-1}$ 即可。此时, 显然 Y 仍然服从高维正态分布, 并且 Y_1, Y_2 相互独立 (进而它们也分别服从高维正态分布)。注意到 $Y_2 = X_2$, 并且 $\mathbb{E}X_2 = \mu_2, \text{Var}(X_2) = \Sigma_{2,2}$, 我们立即得到 $X_2 \sim N(\mu_2, \Sigma_{2,2})$; 类似的, 应有 $X_1 \sim N(\mu_1, \Sigma_{1,1})$ 。这完成了定理 (1) 部分的证明。设 $Y_1 \sim N(\hat{\mu}_1, \hat{\Sigma}_{1,1})$, 注意到 $Y_1 = X_1 - BX_2$, 应有:

$$\hat{\mu}_1 = \mathbb{E}Y_1 = \mathbb{E}X_1 - B\mathbb{E}X_2 = \mu_1 - B\mu_2$$

以及

$$\hat{\Sigma}_{1,1} = \text{Var}(Y_1) = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1} =: \tilde{\Sigma}_{1,1}.$$

利用推论 8.4.1, 我们立即得到 $X_1 = Y_1 + BX_2$ 关于 $X_2 = Y_2$ 的条件密度为

$$\rho_{X_1|X_2}(x_1|x_2) = \rho_{Y_1}(x_1 - Bx_2).$$

注意到 $\rho_{Y_1}(y_1) = \rho(y_1; \hat{\mu}_1, \hat{\Sigma}_{1,1})$, 我们有

$$\begin{aligned} \rho_{X_1|X_2}(x_1|x_2) &= \rho_{Y_1}(x_1 - Bx_2) \\ &= \rho(x_1 - Bx_2; \hat{\mu}_1, \hat{\Sigma}_{1,1}) \\ &= \rho(x_1; \hat{\mu}_1 + Bx_2, \hat{\Sigma}_{1,1}) \\ &= \rho(x_1; \tilde{\mu}_1(x_2), \tilde{\Sigma}_{1,1}). \end{aligned}$$

上述函数作为 x_1 的函数 (x_2 视作常数) 是正态分布 $N(\tilde{\mu}_1(x_2), \tilde{\Sigma}_{1,1})$ 的密度函数, 因此定理结论(8.25)成立。□

基于上述结果, 我们进一步有下面的推论; 它的证明就留给读者。

◆ **推论 8.5.2.** 设 X 是 n 维随机列向量, $X \sim N(\mu, \Sigma)$, 其中 $\mu \in \mathbb{R}^n$, Σ 是 n 阶实对称正定阵。设 A 为 $m \times n$ 矩阵且行满秩 (从而 $m = \text{rank}(A) \leq n$), $\beta \in \mathbb{R}^m$, 则

$$Y := \beta + AX \sim N(\tilde{\mu}, \tilde{\Sigma}),$$

其中 $\tilde{\mu} = \beta + A\mu, \tilde{\Sigma} = A\Sigma A^T$ 。

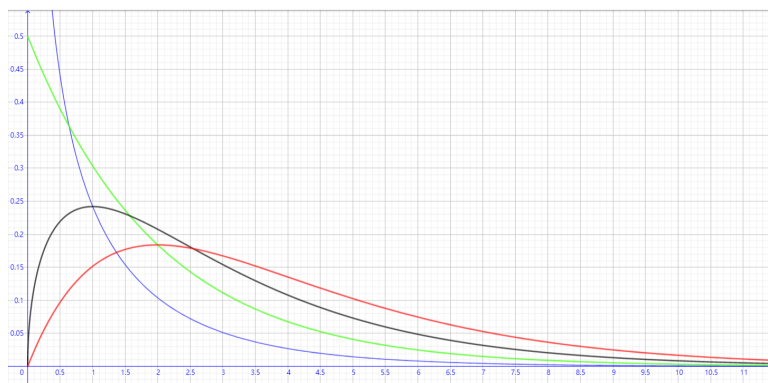
注记 8.1. 在以后, 我们利用特征函数的工具, 可以把正态分布 $N(\mu, \Sigma)$ 中参数 Σ 的正定性要求减弱为非负定, 对应推广的分布称为 *Gauss 分布* 或 *广义正态分布*, 仍然记作 $N(\mu, \Sigma)$ (但需要注意, 当 $\det(\Sigma) = 0$ 时, $N(\mu, \Sigma)$ 不再是连续型分布, 而是一种集中在稍低维度的线性子流形上的分布)。在这种认同下, 上述推论中矩阵 A 的行满秩要求可以去掉。

例 8.15. (卡方分布 $\chi^2(n)$) 现在假设

$$\{Z_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} N(0, 1).$$

定义 $Y_n := Z_1^2 + \cdots + Z_n^2$, 称 Y_n 服从 n 个自由度的卡方分布, 记作 $\chi^2(n)$ 。不难知道,

$$Z_1^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right),$$


 图 8.5: 卡方分布 $\chi^2(n)$ 密度函数图; $n = 1, 2, 3, 4$

从而, 由 *Gamma* 分布的半再生性, $\chi^2(n) = \Gamma(\frac{n}{2}, \frac{1}{2})$ 。

例 8.16. (正态分布 $N(\mu, \sigma^2)$ 的极大似然估计) 设有正态分布 $N(\mu, \sigma^2)$ (其中两个参数均未知) 的 n 个简单样本 $\{X_i\}_{i=1}^n$, 亦即

$$\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2).$$

首先计算 (X_1, \dots, X_n) 的联合密度函数 $L = L(x_1, \dots, x_n; \mu, \sigma^2)$, 得到

$$\log L = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

固定具体样本值 (x_1, \dots, x_n) , 把 L 视作参数 (μ, σ^2) 的函数 (统计学中, 把联合密度函数 L 在视作参数的函数时, 又称之为似然函数), L 的极大值点 $(\hat{\mu}, \hat{\sigma}^2)$ 就是极大似然估计

$$(\hat{\mu}, \hat{\sigma}^2) = \arg \max_{(\mu, \sigma^2)} L(x; \mu, \sigma^2).$$

因此, 我们最终基于这 n 个简单样本 $\{X_i\}_{i=1}^n$ 对参数的似然估计为

$$\begin{cases} \hat{\mu} &:= \bar{X} = \frac{1}{n} [X_1 + \dots + X_n], \\ \hat{\sigma}^2 &:= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \end{cases}$$

其中, \bar{X} 称为样本均值, $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 称为样本方差。不难计算出

$$\mathbb{E}[\bar{X}] = \mu, \quad \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{n-1}{n} \sigma^2 < \sigma^2.$$

也就是说, 上述极大似然估计中, $\hat{\mu}$ 是无偏估计, $\hat{\sigma}^2$ 不是无偏估计; 为此我们可以调整系数, 称 $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 称为样本标准方差, 以此作为真实

分布的方差的估计就是无偏估计了。

下面进一步研究两个参数的极大似然估计的分布律问题。为方便, 对问题进行转化, 令

$$Z_i := \frac{X_i - \mu}{\sigma}.$$

则 $\{Z_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ 。于是

$$\hat{\mu} = \mu + \sigma \bar{Z}, \quad \widehat{\sigma^2} = \frac{\sigma^2}{n} \sum_{i=1}^n (Z_i - \bar{Z})^2.$$

容易知道 $\bar{Z} \sim N(0, \frac{1}{n})$ 。

以下将证明: $\sum_{i=1}^n (Z_i - \bar{Z})^2 \sim \chi^2(n-1)$, 并且它与 \bar{Z} 相互独立。从而

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0, 1), \quad \frac{n\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-1),$$

并且上述两个随机变量相互独立。上述第二个分布律信息可以用来构建 σ^2 的区间估计。

注意到标准正态分布在正交变换下保持不变 (推论 8.5.1), 存在正交矩阵 A , 使得它的第一行元素如下指定:

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}.$$

令 $Y = AZ \sim N(0, I_n)$, 其中 $Z = (Z_1, \dots, Z_n)^T$ 。那么 $\{Y_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, 且

$$\bar{Z} = \sqrt{n}Y_1, \quad \|Z\|^2 = \|Y\|^2.$$

进而

$$\sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 = \sum_{i=2}^n Y_i^2 \sim \chi^2(n-1),$$

且它与 $\bar{Z} = \sqrt{n}Y_1$ 相互独立。

例 8.17. (t -分布) 在上一例中, 为了构建参数 μ 的区间估计, 可以考虑两个相互独立的随机变量 $\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma}$ 与

$$\sqrt{\frac{n\widehat{\sigma^2}}{\sigma^2} / (n-1)}$$

的商

$$W := \frac{\sqrt{n-1}(\hat{\mu} - \mu)}{\sqrt{\widehat{\sigma^2}}}$$

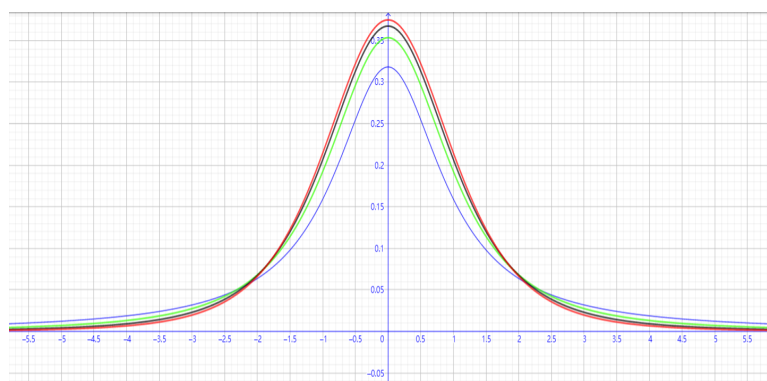
的分布律, 从而把未知参数 σ 消去。 W 的分布律就是本例中将介绍的 $n-1$ 个自由度的 t -分布。

一般而言, 设 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 且 X, Y 相互独立, 则记

$$\frac{X}{\sqrt{Y/n}} \sim t(n),$$

称商 $\frac{X}{\sqrt{Y/n}}$ 服从 n 个自由度的 t -分布。不难算出, $t(n)$ -分布的密度函数为:

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \cdot (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}.$$


 图 8.6: t -分布 $t(n)$ 密度函数图; $n = 1, 2, 3, 4$

注 8.1. t -分布又称为 *Student t -分布*。它的推导是由 *William Sealy Gosset* (威廉·戈塞特, 英国化学家、数学家与统计学家) 于 1908 年首先发表。当时他还在都柏林的健力士酿酒厂工作, 因为不能以他本人的名义发表, 所以论文使用了学生 (*Student*) 这一笔名; *Gosset* 被认为是英国现代统计方法发展的先驱, 小样本理论研究的先驱, 为研究样本分布理论奠定了重要基础, 被统计学家誉为统计推断理论发展史上的里程碑。之后 t 检验以及相关理论经由 *R. Fisher* (罗纳德·费雪) 的工作发扬光大, 而正是他将此分布称为学生分布。

例 8.18. (F -分布) 假定有两族来源于不同分布且互相独立的样本:

$$\{X_i\}_{i=1}^m \sim N(\mu_1, \sigma_1^2), \quad \{Y_j\}_{j=1}^n \sim N(\mu_2, \sigma_2^2).$$

关心的问题是: 是否可以认为 $\sigma_1^2 = \sigma_2^2$? 即两个分布的方差是否相同? 如果 μ_1, μ_2 已知, 则可以知道

$$T_1 := \frac{\sum_{i=1}^m (X_i - \mu_1)^2}{\sigma_1^2} \sim \chi^2(m), \quad T_2 := \frac{\sum_{j=1}^n (Y_j - \mu_2)^2}{\sigma_2^2} \sim \chi^2(n),$$

并且它们相互独立。商 $\frac{T_1/m}{T_2/n}$ 的分布律就称为自由度参数为 (m, n) 的 F -分布, 记作

$$\frac{T_1/m}{T_2/n} \sim F(m, n).$$

基于此就可以构建相应的拒绝域来完成对应的假设检验。不难算出 $F(m, n)$ -分布的密度函数为

$$f_{m,n}(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}} \cdot 1_{(0,\infty)}(x).$$

当 μ_1, μ_2 未知时, 则取

$$T_1 := \frac{\sum_{i=1}^m (X_i - \bar{X})^2}{\sigma_1^2} \sim \chi^2(m-1), \quad T_2 := \frac{\sum_{j=1}^n (Y_j - \bar{Y})^2}{\sigma_2^2} \sim \chi^2(n-1),$$

此时商 $\frac{T_1/(m-1)}{T_2/(n-1)} \sim F(m-1, n-1)$ 。

例 8.19. (Cauchy 分布) 设 $\{Z_i\}_{i=1}^2 \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ 。定义

$$X := \frac{Z_1}{Z_2}.$$

则 X 的密度函数为

$$\rho(x) = \frac{1}{\pi(1+x^2)}.$$

这个密度函数对应的分布律称为**标准 Cauchy 分布**。给定 $\mu \in \mathbb{R}, a > 0$, $Y := \mu + aX$ 的分布称为参数为 (μ, a) 的 *Cauchy 分布*，它的密度函数为

$$\rho_Y(y) = \frac{a}{\pi[a^2 + (y - \mu)^2]}.$$

参考证明：为方，我们记 $X_1 := Z_1/Z_2, X_2 = Z_2$ ，于是 $Z_1 = X_1X_2, Z_2 = X_2$ 。基于此立即得到 (X_1, X_2) 的联合密度函数

$$\rho_{1,2}(x_1, x_2) = \frac{1}{2\pi} \cdot \exp\left\{-\frac{(1+x_1^2)x_2^2}{2}\right\} \cdot |x_2|.$$

进而 $X = X_1$ 的密度函数为

$$\rho_1(x_1) = \int \rho_{1,2}(x_1, x_2) dx_2 = \frac{1}{\pi(1+x_1^2)}.$$

Y 的密度函数的计算留给读者。 □

注记 8.2. 上例中计算方法很典型，问题本来是求 2 维随机向量的一维复合随机变量的分布律，作为变换是从 2 维变成 1 维，肯定是不可逆的。我们提供的方法是补充定义“缺失”的一维复合随机变量，从而把原始的复合映射补全后变成可逆或分片可逆的光滑变换，使用光滑变换的思想求出联合分布密度，再由此去求关心的边缘分布函数。在例 8.16 的计算过程中也本质上也使用了这一方法。

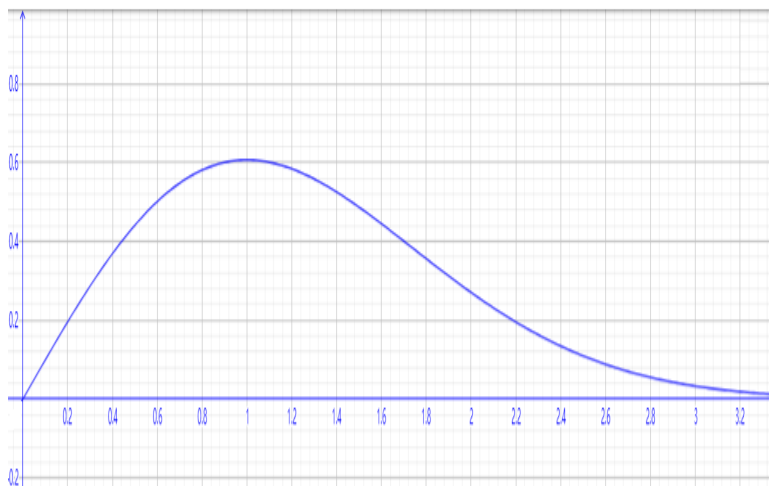


图 8.7: Rayleigh 分布密度函数图

例 8.20. (Rayleigh 分布) 设 $\{Z_i\}_{i=1}^2 \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ 。考虑 (Z_1, Z_2) 的极坐标 (R, Θ) , 即 $R \geq 0, \Theta \in [0, 2\pi)$ 满足:

$$\begin{cases} Z_1 = R \cos \Theta, \\ Z_2 = R \sin \Theta. \end{cases}$$

则 R 与 Θ 相互独立。 $\Theta \sim U(0, 2\pi)$ 。 R 的密度函数为

$$\rho(r) = r \cdot e^{-\frac{r^2}{2}} \cdot 1_{(0, \infty)}(r),$$

这个密度函数对应的分布律称为 Rayleigh 分布, 它的图像见图 8.7。

习 题 8

习题 8.1. 给定 $\alpha, \beta > 0$ 。如果随机变量 X 具有密度函数

$$\rho(x; \alpha, \beta) = \frac{x^{\alpha-1}}{B(\alpha, \beta)(1+x)^{\alpha+\beta}} \cdot 1_{(0, \infty)}(x),$$

则记 $X \sim Z(\alpha, \beta)$, 称 X 服从参数为 (α, β) 的 Z 分布 (或第 II 类 Beta 分布)。求证: 如果 $X_k \sim \Gamma(\alpha_k, \lambda), k = 1, 2$ 相互独立 (其中 $\alpha_1, \alpha_2, \lambda > 0$), 那么:

(1) $Y_1 := X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$, $Y_2 := \frac{X_1}{X_1 + X_2} \sim \text{Beta}(\alpha_1, \alpha_2)$, 且 Y_1, Y_2 相互独立;

(2) $Z_2 := \frac{Y_2}{1-Y_2} = \frac{X_1}{X_2} \sim Z(\alpha_1, \alpha_2)$, 且 Y_1, Z_2 相互独立。

习题 8.2. (1) 设 n 是正整数。求证: $\chi^2(1) = \Gamma(\frac{1}{2}, \frac{1}{2})$, 即当 $X \sim N(0, 1)$ 时, $X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$ 。进而由此推断 $\chi^2(n) = \Gamma(\frac{n}{2}, \frac{1}{2})$ 。

(2) 如果 $X_k \sim N(\mu_k, 1), k = 1, \dots, n$ 相互独立, 那么 $Y := X_1^2 + \dots + X_n^2$ 的分布律由参数 n 及 $\delta := \sqrt{\mu_1^2 + \dots + \mu_n^2}$ 决定, 记作 $Y \sim \chi^2(n; \delta)$, 称之为自由度 n 的非中心 χ^2 -分布;

(3) 设 $\delta > 0, I \sim \text{Poisson}(\frac{\delta^2}{2})$, $Y_k \sim \chi^2(k), k = 1, 2, \dots$ 相互独立且与 I 也相互独立。那么 $Y_{2I+1} \sim \chi^2(1; \delta)$; 特别的, $Y_{2I+n} \sim \chi^2(n; \delta)$ 。

习题 8.3. 对于给定的正整数 n_1, n_2 , 如果 $X_k \sim \chi^2(n_k), k = 1, 2$ 相互独立, 那么就称 $X := \frac{X_1/n_1}{X_2/n_2}$ 服从自由度参数对 (n_1, n_2) 的 F 分布, 记作 $X \sim F(n_1, n_2)$ 。求证: 此时有 $\frac{n_1}{n_2} \cdot X \sim Z(\frac{n_1}{2}, \frac{n_2}{2})$ 。当 $X_1 \sim \chi^2(n_1, \delta), X_2 \sim \chi^2(n_2)$ 且相互独立时, 就称 $X := \frac{X_1/n_1}{X_2/n_2}$ 服从自由度参数对 (n_1, n_2) 、位置参数 $\delta > 0$ 的 F 分布, 记作 $X \sim F(n_1, n_2; \delta)$; 试利用上一习题给出 $F(n_1, n_2; \delta)$ 的类似表示。

习题 8.4. 对给定正整数 n , 如果 $X_1 \sim N(0, 1), X_2 \sim \chi^2(n)$ 且相互独立, 那么就记 $T_n := \frac{X_1}{\sqrt{X_2/n}} \sim t(n)$, 称它服从自由度为 n 的 t 分布。如果 $X_1 \sim N(\mu, 1), X_2 \sim \chi^2(n)$ 且相互独立, 那么就记 $T_n(\mu) := \frac{X_1}{\sqrt{X_2/n}} \sim t(n; \mu)$, 称它服从位置参数为 μ 、自由度为 n 的非中心 t 分布。求证: $T_n(-\mu) \stackrel{d}{=} -T_n(\mu)$ 。

§ 9

随机变量 (III)

本章我们先介绍一个奇异型分布的例子及随机变量的分布的分类理论, 之后介绍分布函数的性质与随机变量的实现。

奇异型 (或奇异连续型) 随机变量的介绍只是为了避免读者在初等概率论的学习过程中僵化了思维, 误以为: 随机变量只有离散型和连续型两种类型, 不是离散型就一定是连续型。事实上, 随机变量的分类上, 有三种所谓的纯型: 离散型、连续型 (又称绝对连续型) 和奇异型 (又称奇异连续型), 一般的随机变量是这三种纯型的凸组合: 任何 \mathbb{R} 上的分布函数 F , 存在 $\alpha_d, \alpha_c, \alpha_s \geq 0$, $\alpha_d + \alpha_c + \alpha_s = 1$, 及离散型分布函数 F_d 、连续型分布函数 F_c 、奇异型分布函数 F_s , 使得

$$F = \alpha_d F_d + \alpha_c F_c + \alpha_s F_s.$$

对此读者只需略作了解即可。

在介绍分布函数的性质与随机变量的实现的章节, 我们讨论了: (1) 分布函数的刻画、基于均匀分布 $U(0,1)$ 随机数发生器如何实现给定分布 F 的随机数发生器以及反过来的问题; (2) 利用之前的理论, 探讨连续型分布的次序统计量的分布律计算; (3) 如何构造标准正态分布的随机数发生器。这些理论对于实现 Monte-Carlo 随机模拟计算大有用处。

9.1 一个奇异型分布的例子及随机变量的分类

定义 9.1.1. 设 μ 是一个 Borel 测度。称 a 是 μ 的一个原子, 如果 $\mu(\{a\}) > 0$ 。现在设 μ 是 \mathbb{R} 上的 Borel 概率测度, 称 μ 为奇异型 (或奇异连续型), 如果测度 μ 无原子且 $\mu \perp \text{Leb}$ 。任给一个分布函数 F , 如果它诱导的分布测度 μ_F 是奇异 (连续) 型, 我们就称 F 是奇异 (连续) 型的, 或是奇异 (连续) 的。

例 9.1. 著名的 Cantor 函数就是一种奇异型分布函数, 它在 $[0,1]$ 上的定义可以如下构造。令 $F_0(x) = x, x \in C_0 := [0,1]$ 。下一步三等分 C_0 并去掉中间的区间, 得到集合 $C_1 = [0, 1/3] \cup [2/3, 1] =: C_{0,0} \cup C_{0,1}$, 产生两个新的分点 $1/3$ 与 $2/3$, 令 $F_1(0) := 0, F_1(1/3) = F_1(2/3) := \frac{F_0(0)+F_0(1)}{2} = \frac{1}{2}, F_1(1) = 1$, 其余取值采用线性插值。假设已经定义 C_n, F_n , 对集合 C_n 的每个连通区间进行三等分并去掉中间的区间, 得到新的集合 C_{n+1} , 产生一系列的新的分点 $\{x_{n+1,k} : k = 1, \dots, 2^{n+1}\}$, 定义函数 F_{n+1} 如下: 在之前得到的分点上

的函数值与 F_n 相同；在新的分点 $x_{n+1,k}$ 上的取值定为在 C_n 中包含 $x_{n+1,k}$ 的连通区间两端点的 F_n -函数值的平均；其余函数值采用线性插值。很容易知道 F_n 一致收敛到某函数 F ，从而 F 仍然为连续函数，并且是一个分布函数。容易验证这样构造的 Cantor 函数 F 是奇异型的。

以下我们简略讨论一维随机变量的分布的分类；高维情形的做法类似。设 $\xi \sim F$ ，其中 F 是分布函数，它诱导了分布测度 μ_F 。

首先我们可以定义 F 的原子点集 \mathcal{D} 如下

$$\mathcal{D} := \{x \in \mathbb{R} : \mu_F(\{x\}) = F(x) - F(x-) > 0\}.$$

进一步定义 α_d

$$\alpha_d := \mu_F(\mathcal{D}) = \sum_x [F(x) - F(x-)].$$

如果点集 \mathcal{D} 非空，则 $\alpha_d > 0$ ，就可以定义 μ_d 和 F_d 如下：

$$\mu_d(A) := \mu_F(A \cap \mathcal{D}) / \alpha_d, \forall A \in \mathcal{B},$$

$$F_d(x) := \mu_d((-\infty, x]), \forall x \in \mathbb{R}.$$

于是 F_d 是离散型概率分布函数。

此时，不难验证

$$G(x) := F(x) - \alpha_d F_d(x) = \mu_F((-\infty, x] \cap \mathcal{D}^c)$$

是一个连续的单调递增函数。因而 G 几乎处处有导函数 G' 。定义 α_c 如下：

$$\alpha_c := \int G'(x) dx.$$

容易知道，一般而言 $0 \leq \alpha_c \leq 1 - \alpha_d$ ，等号成立当且仅当 G 是绝对连续函数。

如果 $\alpha_c = 0$ ，则可任取一连续型分布函数 F_c （比如取 $F_c(x) = x^+ \wedge 1$ ），取 $\alpha_s = 1 - \alpha_d$ ，并取 $F_s(x) = G(x) / \alpha_s$ ，则 F_s 是一个奇异型分布函数。此时有 $\alpha_d + \alpha_c + \alpha_s = 1$ ，

$$F = \alpha_d F_d + \alpha_c F_c + \alpha_s F_s.$$

如果 $\alpha_c = 1 - \alpha_d$ ，则定义 $F_c := G / \alpha_c$ ，此时 F_c 是连续型概率分布函数。取 $\alpha_s = 0$ ，可任取一奇异型分布函数 F_s （比如本节提供的 Cantor 分布函数）。此时有 $\alpha_d + \alpha_c + \alpha_s = 1$ ，

$$F = \alpha_d F_d + \alpha_c F_c + \alpha_s F_s.$$

如果 $0 < \alpha_c < 1 - \alpha_d$ ，则定义

$$F_c(x) := \int_{-\infty}^x G'(x) dx / \alpha_c,$$

它是一个连续型概率分布函数。定义 $\alpha_s := 1 - \alpha_d - \alpha_c$ ，并取

$$F_d = (G - \alpha_c F_c) / \alpha_s,$$


它是一个奇异型概率分布函数。此时有 $\alpha_d + \alpha_c + \alpha_s = 1$ ，

$$F = \alpha_d F_d + \alpha_c F_c + \alpha_s F_s.$$

9.2 分布函数的性质与随机变量的实现

9.2.1 分布函数的性质

关于分布函数，有下面的刻画定理。

 **定理 9.2.1.** 随机变量 ξ 的分布函数 F 具有下面性质：

- (1) F 是单调递增函数；
- (2) F 是右连续的；
- (3) $\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1.$

反之，对于具有上述三条性质的函数 F ，设随机变量 $U \sim U(0, 1)$ ，则 $F^{-1}(U)$ 的分布函数恰好为 F ，即： $F^{-1}(U) \sim F$ 。此处，函数 F^{-1} 称为 F 的广义逆，定义为：

$$F^{-1}(p) := \inf\{t : F(t) \geq p\}, \quad p \in (0, 1). \quad (9.1)$$

因此，我们直接把具有上述三条性质的函数 F 称为分布函数。

证明. 利用概率测度本身的单调性、上连续性、下连续性等，很容易论证出一个随机变量的分布函数具有性质 (1)–(3)，具体细节此处从略。

以下论证定理的第二部分结论。现假设已有某概率空间中随机变量 $U \sim U(0, 1)$ （例 8.2 中已说明这样的概率空间与随机变量存在）以及给定的满足 (1)–(3) 的函数 F 。我们论证 $\xi := F^{-1}(U) \sim F$ 。事实上，我们只需证明广义逆有如下重要性质：

$$F^{-1}(p) \leq x \Leftrightarrow p \leq F(x), \quad \forall p \in (0, 1), x \in \mathbb{R}. \quad (9.2)$$

因为如上述性质成立，那么


$$\mathbb{P}(\xi \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

注意到 $x_p := F^{-1}(p)$ 是通过下确界定义的，因此它满足： $F(x_p) \geq p$ 以及 $F(x_p - \varepsilon) < p, \forall \varepsilon > 0$ 。由此利用 F 的单调性可知： $x_p \leq x$ 时， $p \leq F(x_p) \leq F(x)$ ；反之， $p \leq F(x)$ 时， $F(x_p - \varepsilon) < p \leq F(x)$ ， $x_p - \varepsilon < x, \forall \varepsilon > 0$ ，因此 $x_p \leq x$ 。□

此处，我们提请读者前往习题 9.1 熟悉分布函数的广义逆的更多性质。

注记 9.1. 上述定理理论上给出了从一个均匀随机数发生器出发如何构造服从任意给定分布律的随机数发生器的一种办法，这被称为**反函数方法 (Inverse Transformation Method)**。在本章 §9.2.3 中，我们将会介绍其他构造随机数发生器的办法；更多的构造随机数发生器的办法请参阅其他专业书籍，如 [33]。这些构造随机数发生器的办法对于蒙特卡罗模拟计算来说非常重要。

我们还有以下事实。

 **定理 9.2.2.** 如果 $X \sim F$ ，且 F 是连续函数，那么 $F(X) \sim U(0, 1)$ 。

此处，我们提请读者前往习题 5.6，看看上述定理中分布函数 F 不连续时会发生什么现象。

定理 9.2.2 的证明. 我们证明在定理所给条件下，

$$F(F^{-1}(p)) = p, \forall p \in (0, 1).$$

事实上， $x_p := F^{-1}(p)$ 满足： $F(x_p) \geq p$ 以及 $F(x_p - \varepsilon) < p, \forall \varepsilon > 0$ 。由 F 的连续性立即知道 $F(x_p) = F(x_p -) \leq p \leq F(x_p)$ ，因此 $F(x_p) = p$ 。

由此，对任意 $p \in (0, 1)$

$$\mathbb{P}(F(X) \geq p) = \mathbb{P}(X \geq F^{-1}(p) =: x_p) = 1 - F(x_p -) = 1 - p.$$

由此立即知道 $\mathbb{P}(F(X) \leq p) = p, \forall p \in (0, 1)$ ，亦即 $F(X) \sim U(0, 1)$ 。 \square

注记 9.2. 上述定理也可以不严谨地表述为：当 F 连续时，以下两个概率空间同构

$$(\mathbb{R}, \mathcal{B}, \mu_F) \cong ((0, 1), \mathcal{B}(0, 1), Leb).$$

付出一些努力，可以证明， \mathbb{R} 配上任意 Borel 概率测度 μ ，概率空间 $(\mathbb{R}, \mathcal{B}, \mu)$ 同构于 $(0, 1)$ 上配上 Lebesgue 测度与某离散概率测度的凸组合得到的概率空间。（具有这种性质的概率空间，动力系统中通常称为 **Lebesgue 空间**；本注记说明，一维欧氏空间配上任何 Borel 概率测度，得到的都是 Lebesgue 空间。这一结论在高维欧氏空间中仍然成立。）

9.2.2 次序统计量

本小节我们介绍统计学中的一个重要概念——**次序统计量**。假定我们已经有了来自分布 F 的 n 个简单样本 $\{X_i\}_{i=1}^n$ ，亦即

$$\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} F.$$

我们可以按照它们的取值从小到大排列，记作

$$X_{(1:n)} \leq X_{(2:n)} \leq \cdots \leq X_{(n:n)}.$$

当样本量 n 明确时，也经常简单记 $X_{(i)} := X_{(i:n)}, i = 1, 2, \dots, n$ 。我们称 $\{X_{(i)}\}_{i=1}^n$ 为 $\{X_i\}_{i=1}^n$ 的**次序统计量**，称 $X_{(i)}$ 是第 i 个**次序统计量**。特殊的，我们也称 $X_{(1)}$ 是**最小次序统计量**，称 $X_{(n)}$ 是**最大次序统计量**；此二者是统计学中非常关注的极端值，有时也简称**极值**（extreme value）。

下面我们考虑次序统计量的分布律的计算。

例 9.2. 最小次序统计量 $X_{(1)}$ 的分布律计算如下：

$$\begin{aligned} \mathbb{P}(X_{(1)} > x) &= \mathbb{P}(X_1 > x, \dots, X_n > x) \\ &= \mathbb{P}(X_1 > x) \cdots \mathbb{P}(X_n > x) \\ &= [1 - F(x)]^n, \end{aligned}$$

即 $X_{(1)}$ 的分布函数为

$$F_{(1)}(x) = 1 - [1 - F(x)]^n.$$

同理，最大次序统计量 $X_{(n)}$ 的分布函数为

$$F_{(n)}(x) = F(x)^n.$$

特别的, 如果这些简单样本的公共分布 F 具有密度 ρ , 则最小、最大的次序统计量对应的密度函数分别为

$$\rho_{(1)}(x) = n\rho(x)[1 - F(x)]^{n-1}, \quad \rho_{(n)}(x) = n\rho(x)[F(x)]^{n-1}.$$

例 9.3. 设 $\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} F$, 其中 F 是连续的分布函数。则

$$\mathbb{P}(\exists i, j, 1 \leq i < j \leq n, X_i = X_j) = 0. \quad (9.3)$$

因此我们可以认为次序统计量实际上是按照如下严格大小顺序排列的:

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}.$$

参考证明. 我们只要证明: 对任意 $i, j, 1 \leq i < j \leq n$, $\mathbb{P}(X_i = X_j) = 0$ 。但实际上, 注意到此时 $\{U_i := F(X_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 我们有

$$\begin{aligned} \mathbb{P}(X_i = X_j) &\leq \mathbb{P}(F(X_i) = F(X_j)) \\ &= \mathbb{P}(U_i = U_j) \\ &= \int_{(0,1)^2} 1_{\{x=y\}} dx dy = 0. \end{aligned}$$

因此上例中的结论成立。 \square

以下在 F 具有密度函数 ρ 的条件下讨论次序统计量的分布律。

此时 F 是连续的分布函数。上一小节中的定理 9.2.1 和定理 9.2.2 结合定理 5.1.1 可以应用于此处, 得到

定理 9.2.3. 设 F 是连续的分布函数。

- (i) 如果 $\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} F$, 那么 $\{U_i := F(X_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 并且两者之间的次序统计量有下面的对应关系:

$$U_{(i)} = F(X_{(i)}), i = 1, \cdots, n;$$

- (ii) 如果 $\{U_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 那么 $\{X_i := F^{-1}(U_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} F$, 并且两者之间的次序统计量有下面的对应关系:

$$X_{(i)} = F^{-1}(U_{(i)}), i = 1, \cdots, n.$$

因此, 借助上述定理, 人们经常把统计学中关心的次序统计量的有关问题转化为标准均匀分布的次序统计量问题来考虑。

现在我们基于分布 F 具有密度 ρ 的条件, 用概率微元法来探讨次序统计量的密度函数的计算问题。

首先讨论 $X_{(k)}$ 的密度, 其中 $2 \leq k \leq n-1$ 。此时

$$\begin{aligned} \mathbb{P}(X_{(k)} = x + dx) &:= \mathbb{P}(X_{(k)} \in (x, x + dx]) \\ &= \mathbb{P}(X_1, \cdots, X_n \text{ 中有 } k-1 \text{ 个取值 } \leq x, \\ &\quad \text{有 } 1 \text{ 个取值 } \in (x, x + dx], \\ &\quad \text{有 } n-k \text{ 个取值 } > x + dx) \\ &= \frac{n!}{(k-1)!(n-k)!} \cdot [F(x)]^{k-1} [1 - F(x)]^{n-k} \rho(x) dx. \end{aligned}$$

因此, $X_{(k)}$ 的密度为

$$\rho_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} \cdot [F(x)]^{k-1} [1 - F(x)]^{n-k} \rho(x). \quad (9.4)$$

接着, 我们讨论 $(X_{(k)}, X_{(\ell)})$ 的联合密度, 其中 $1 \leq k < \ell \leq n$. 此时, 设 $x < y$,

$$\begin{aligned} & \mathbb{P}(X_{(k)} = x + dx, X_{(\ell)} = y + dy) := \mathbb{P}(X_{(k)} \in (x, x + dx], X_{(\ell)} \in (y, y + dy]) \\ &= \mathbb{P}(X_1, \dots, X_n \text{ 中有 } k-1 \text{ 个取值 } \leq x, \text{ 有 } 1 \text{ 个取值 } \in (x, x + dx], \\ & \quad \text{有 } \ell - k - 1 \text{ 个取值 } \in (x + dx, y], \text{ 有 } n - \ell \text{ 个取值 } > y + dy) \\ &= \frac{n!}{(k-1)!(\ell-k-1)!(n-k)!} \cdot [F(x)]^{k-1} [F(y) - F(x)]^{\ell-k-1} \\ & \quad \cdot [1 - F(y)]^{n-\ell} \rho(x) \rho(y) dx dy. \end{aligned}$$

因此, $(X_{(k)}, X_{(\ell)})$ 的联合密度为

$$\begin{aligned} \rho_{(k, \ell)}(x, y) &= \frac{n!}{(k-1)!(\ell-k-1)!(n-k)!} \cdot [F(x)]^{k-1} [F(y) - F(x)]^{\ell-k-1} \\ & \quad \cdot [1 - F(y)]^{n-\ell} \rho(x) \rho(y) \cdot 1_{\{x < y\}}. \end{aligned} \quad (9.5)$$

更一般的 $(X_{(k_1)}, \dots, X_{(k_r)})$ 的联合密度 (其中 $r \geq 3, 1 \leq k_1 < k_2 < \dots < k_r \leq n$) 可以类似计算. 我们这里特意指出, $(X_{(1)}, \dots, X_{(n)})$ 的联合密度为

$$\rho_{(1, \dots, n)}(x_1, \dots, x_n) = n! \rho(x_1) \cdots \rho(x_n) \cdot 1_{\{x_1 < \dots < x_n\}}. \quad (9.6)$$

我们特意把均匀分布 $U(0, 1)$ 的次序统计量的分布律总结为如下定理.

☞ **定理 9.2.4.** 设 $\{U_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 那么

(i) $U_{(k)}$ 的密度为

$$\rho_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} \cdot x^{k-1} (1-x)^{n-k} \cdot 1_{(0,1)}(x). \quad (9.7)$$

于是 $U_{(k)} = U_{(k:n)} \sim \text{Beta}(k, n-k+1)$;

(ii) 对 $2 \leq r \leq n$, $1 \leq k_1 < k_2 < \dots < k_r \leq n$, $(U_{(k_1)}, \dots, U_{(k_r)})$ 的联合密度为 (此处约定 $k_0 := 0, k_{r+1} := n+1$ 以及 $x_0 := 0, x_{r+1} = 1$)

$$\begin{aligned} \rho_{(k_1, \dots, k_r)}(x, y) &= \frac{n!}{\prod_{j=1}^{r+1} (k_j - k_{j-1} - 1)!} \cdot \prod_{j=1}^{r+1} (x_j - x_{j-1})^{k_j - k_{j-1} - 1} \\ & \quad \cdot 1_{\{0 < x_1 < \dots < x_r < 1\}}. \end{aligned} \quad (9.8)$$

(iii) $(U_{(1)}, \dots, U_{(n)})$ 的联合密度为

$$\rho_{(1, \dots, n)}(x_1, \dots, x_n) = n! \cdot 1_{\{0 < x_1 < \dots < x_n < 1\}}. \quad (9.9)$$

基于上面的定理, 约定 $U_{(0)} := 0, U_{(n+1)} := 1$, 不难知道下面结论成立.

☞ **定理 9.2.5.** 设 $\{U_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 那么对 $0 \leq k < \ell \leq n+1, \ell - k \geq 2$,

$$\left\{ \frac{U_{(i)} - U_{(k)}}{U_{(\ell)} - U_{(k)}} \right\}_{i=k+1}^{\ell-1}$$

可视作 $U(0, 1)$ 的 $\ell - k - 1$ 个简单样本的次序统计量, 并且它们与 $\{U_{(j)}\}_{j \notin (k, \ell)}$ 相互独立.

9.2.3 随机数发生器的构造

定理 9.2.1 在理论上给出了基于均匀随机数发生器，构造具有给定分布函数 F 的随机数发生器的一般性的办法（反函数方法）：

$$U \sim U(0, 1) \Rightarrow X := F^{-1}(U) \sim F.$$

但在实际应用中，很多连续型分布具有很好计算的密度函数 ρ ，对应的分布函数 F （特别是广义逆 F^{-1} ）并不好计算。因此人们发展出一些新的办法来构造对应的随机数发生器。本小节内容节选自 [33]（部分地方有改动）。

下面的 Rejection Method 就是一种常见办法。

定理 9.2.6. (Rejection Method) 假设 f, g 都是密度函数，并且存在 $C > 0$ ，使得 $h := f/g \leq C$ 总成立（这里约定 $0/0 = 0$ ）。那么按照以下流程将产生随机变量 $X \sim f(x)dx$ ：

步骤 1. 产生 $Y \sim g(x)dx$ ，以及 $U \sim U(0, 1)$ ；

步骤 2. 如果 $U \leq h(Y)/C$ ，那么就输出 $X := Y$ ，否则转上一步。

证明. 设 N 为定理中为了得到随机变量 X 所须进行的循环次数， Y_k 为第 k 次循环中产生随机变量 Y ，显然

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}(Y_N \leq x) = \mathbb{P}(Y \leq x | U \leq h(Y)/C) \\ &= \frac{\mathbb{P}(Y \leq x, U \leq h(Y)/C)}{\mathbb{P}(U \leq h(Y)/C)} \\ &= \frac{1}{K} \mathbb{E} \mathbb{P}(Y \leq x, U \leq h(Y)/C | Y) \quad (K := \mathbb{P}(U \leq h(Y)/C)) \\ &= \frac{1}{K} \int_{-\infty}^x \frac{h(y)}{C} \cdot g(y) dy = \frac{1}{KC} \int_{-\infty}^x f(y) dy. \end{aligned}$$

令 $x \rightarrow \infty$ ，得到 $KC = 1$ 。因此 $X \sim f(x)dx$ 。 \square

容易知道，上面证明中的 $N \sim \text{Geo}(p)$ ，其中 $p = 1/C$ ；从而 $\mathbb{E}N = C$ 。即上述定理中为了得到满足条件的随机数，平均需要进行 C 次循环调用。

例 9.4.（基于标准指数分布的随机数发生器构建标准正态的随机数发生器）注意到 $Z \sim N(0, 1)$ 时， $|Z|$ 具有密度

$$f(x) := \frac{2}{\sqrt{2\pi}} e^{-x^2/2} \cdot 1_{(0, \infty)}(x),$$

我们考虑标准指数分布的密度函数 $g(x) = e^{-x} \cdot 1_{(0, \infty)}(x)$ ，那么容易知道

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2e}{\pi}} \cdot \exp\left\{-\frac{(x-1)^2}{2}\right\} \leq \sqrt{\frac{2e}{\pi}} \approx 1.32.$$

于是可以按照以下流程得到 $|Z|$ 的随机数发生器：

(a) 产生 $Y \sim \mathcal{E}(1)$ 以及 $U \sim U(0, 1)$ ；

(b) 如果 $U \leq \exp\{-(Y-1)^2/2\}$ （或等价地： $-\log U \geq (Y-1)^2/2$ ），那么就令 $X = Y$ ，否则转上一步。

注意到在上述流程中， $-\log U \sim \mathcal{E}(1)$ ，可以进一步改进上面的流程如下：

(a') 产生 $Y_1, Y_2 \sim \mathcal{E}(1)$ ；

(b') 如果 $Y_2 \geq (Y_1-1)^2/2$ ，那么就令 $X = Y_1$ ，否则转上一步。

进一步，我们可以按照以下流程得到 $Z \sim N(0, 1)$ 的随机数发生器：

步骤 1. 产生 $Y_1, Y_2 \sim \mathcal{E}(1)$ ；

步骤 2. 如果 $Y_2 \geq (Y_1 - 1)^2/2$, 那么转下一步, 否则转上一步;

步骤 3. 产生 $Y_3 \sim \mathcal{E}(1)$, 并输出

$$Z := \begin{cases} Y_1, & \text{如果 } Y_3 \leq \log 2 \\ -Y_1, & \text{如果 } Y_3 > \log 2 \end{cases}$$

此时 $Z \sim N(0, 1)$ 。 □

有一种用于构造非负连续型随机变量的随机数发生器的办法称为 **Hazard Rate Method**, 由于其论证需要用到 Poisson 过程, 超出本书的范围, 因此请感兴趣的读者参阅 [33] 中的 Poisson 过程相关的章节。习题 9.9 原则上提供了一种类似的构造非负整数值随机变量的随机数发生器的办法, 可以视作离散版本的 **Hazard Rate Method**。

下面我们讨论基于均匀分布随机数发生器构造标准正态随机数发生器的其他办法。

例 9.5. 设 $X, Y \sim N(0, 1)$ 且相互独立。记 (X, Y) 的极坐标表示为 (R, Θ) (其中 $\Theta \in [0, 2\pi)$), 那么 $R^2 \sim \mathcal{E}(\frac{1}{2})$, $\Theta \sim U(0, 2\pi)$ 且相互独立。由此可以有下面基于均匀分布随机数发生器得到标准正态分布随机数发生器的办法 (称为 **Box-Muller Approach**):

设 $U_1, U_2 \sim U(0, 1)$ 且相互独立, 那么下面方程

$$X := \sqrt{-2 \log U_1} \cos(2\pi U_2), \quad Y := \sqrt{-2 \log U_1} \sin(2\pi U_2)$$

给出了两个独立同分布的标准正态随机变量 X, Y 。 □

上述办法中, 计算常数 π 、调用函数 \cos, \sin 都是比较耗时的。注意到, 如果我们能得到 (V_1, V_2) 在单位圆内均匀分布, 则 (V_1, V_2) 的极坐标表示 (R, Θ) 就满足 $R := \sqrt{V_1^2 + V_2^2} \sim U(0, 1)$, $\Theta \sim U(0, 2\pi)$, 且二者相互独立。此时, $\cos \Theta = \frac{V_1}{\sqrt{V_1^2 + V_2^2}}$, $\sin \Theta = \frac{V_2}{\sqrt{V_1^2 + V_2^2}}$ 。这样就可以使用下例中的流程来构造标准正态随机数发生器 (规避出现计算 π 、调用函数 \cos, \sin):

例 9.6. 以下得到一对独立的标准正态随机变量的方法称为 **Polar Method**:

步骤 1. 产生 $U_1, U_2 \sim U(0, 1)$;

步骤 2. 令 $V_1 := 2U_1 - 1, V_2 := 2U_2 - 1, S := V_1^2 + V_2^2$, 如果 $S \leq 1$, 那么转下一步, 否则转上一步;

步骤 3. 输出 $X := \sqrt{\frac{-\log S}{S}} V_1, Y := \sqrt{\frac{-\log S}{S}} V_2$ 。 □

更多构造特定随机数发生器的办法, 请参见 [33, Chapter 11]。

习 题 9

习题 9.1. 设 F 是分布函数, F^{-1} 是它的广义逆。试证明:

- (1) F^{-1} 是单调递增、左连续的;
- (2) $F(F^{-1}(p)) \geq p, \forall p \in (0, 1)$;
- (3) 对 $p \in (0, 1)$, $F^{-1}(p) \leq x$ 当且仅当 $p \leq F(x)$;

(4) 如果 $F(a) > F(a-)$, 那么 $F^{-1}(p) \equiv a, \forall p \in (F(a-), F(a)]$;

(5) 如果 $a < b$ 且 $F(x) \equiv p, \forall x \in (a, b)$, 那么 $F^{-1}(p+) > F^{-1}(p)$;

(6) 如果 $p \in (0, 1)$ 是 F^{-1} 的连续点, 那么 $F^{-1}(p) < x$ 蕴含了 $p < F(x)$ 。

习题 9.2. 设有两个分布函数 F_0, F_1 以及一个常数 $\alpha \in (0, 1)$ 。再设有随机变量 $X_0 \sim F_0, X_1 \sim F_1, I \sim B(1, \alpha)$, 并且此三个随机变量相互独立。试求证: $X_I \sim (1 - \alpha)F_0 + \alpha F_1$ 。

习题 9.3. 设有随机变量 X_1, X_2 , 分别是离散型、连续型随机变量。求证: 对任意 $C \in \mathbb{R}, \mathbb{P}(X_1 = X_2 + C) = 0$ 。

习题 9.4. 设有相互独立随机变量 $X_1 \sim F_1, X_2 \sim F_2$, 其中 F_1, F_2 为分布函数。求证: F_1 是连续函数蕴含了 $\mathbb{P}(X_1 = X_2) = 0$ 。

习题 9.5. 请给出基于 $U(0, 1)$ 的随机数发生器构建标准指数随机数发生器的办法。【提示: *Inverse Transformation Method*。】

习题 9.6. 设 $U \sim U(0, 1)$, 定义 X_1 为 $2U$ 的整数部分; 一般的, 设 $2^n U$ 的整数部分为 N_n , 如果它是奇数则令 $X_n = 1$, 否则 $X_n = 0$ 。求证: $\{X_n\}_1^\infty$ 是 i.i.d. 的, 服从等概率的 0-1 两点分布, 并且 $U = \sum_{n=1}^\infty \frac{X_n}{2^n}$ 。

习题 9.7. 本习题想要说明: $((0, 1), \mathcal{B}_{(0,1)}, \text{Leb})$ 已经是一个非常丰富的概率空间了。事实上, 对任意 $x \in (0, 1)$, 设它的二进制表示为 $x = \sum_{n=1}^\infty \frac{x_n}{2^n}$, 其中 $x_n \in \{0, 1\}$ (此处不允许序列 $\{x_n\}_1^\infty$ 从某个位置开始后续全部为 1, 以保证表示的唯一性)。我们可以把序列 $\{x_n\}_1^\infty$ 排列成如下无穷矩阵形式:

$$\begin{array}{ccccccc} x_1 & x_2 & x_9 & x_{10} & \cdots \\ x_4 & x_3 & x_8 & x_{11} & \cdots \\ x_5 & x_6 & x_7 & x_{12} & \cdots \\ x_{16} & x_{15} & x_{14} & x_{13} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \end{array}$$

其中第 i 行、第 j 列的元素记作 $x_{i,j}$ 。我们可以如下定义随机变量

$$U_i := \sum_{j=1}^\infty \frac{x_{i,j}}{2^j}.$$

请证明: $\{U_i\}_{i=1}^\infty$ 是 i.i.d. 的随机变量序列, 共同分布为 $(0, 1)$ 上均匀分布。

习题 9.8. 设 $\{X_n\}_{n=1}^\infty \stackrel{\text{i.i.d.}}{\sim} B(1, \frac{1}{2})$, 令 $Y := \sum_{n=1}^\infty \frac{2X_n}{3^n}$ 。求证: Y 是良定义的随机变量, 它的分布函数恰为例 9.1 中介绍的 *Cantor* 分布函数。

习题 9.9. 设 X 是非负整数值随机变量, 定义 $h(k) := \mathbb{P}(X = k | X \geq k)$, $k \geq 0$ 。若 $\{U_n\}_{n=0}^\infty \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 证明: $Z := \min\{n : U_n \leq h(n)\}$ 与 X 同分布。(注: 此处构建随机数发生器的办法可以称为离散版本的 *Hazard Rate Method*; 用 *Inverse Transformation Method* 及此处方法都可以构建几何分布 $\text{Geo}(\frac{1}{2})$ 的随机数发生器; 请读者思考两种方法在计算机上实现的优劣。)

习题 9.10. 设有相互独立随机变量 $X_1 \sim F_1, X_2 \sim F_2$, 其中 F_1, F_2 为分布函数。求证: F_1 是连续函数蕴含了 $X_1 + X_2$ 的分布函数也是连续函数; X_1 是连续型的蕴含了 $X_1 + X_2$ 也是连续型的。


§ 10

随机变量列的收敛与大数律

在本章我们将先介绍 Chebyshev 不等式，之后介绍有关随机变量列的各种收敛的定义：几乎处处收敛/几乎必然收敛*、依概率收敛、 L^p -收敛与依分布收敛；在讨论了几种收敛之间的关系后，我们将进一步介绍历史上几个著名的大数律，其中也将介绍证明几乎处处收敛的重要概率工具：Borel-Cantelli 引理。最后，我们给出本章的理论的一些应用实例。

10.1 Chebyshev 不等式

在概率论中，有一个简单、但又非常重要的不等式，这就是下面的 Chebyshev 不等式；有些文献称之为 Markov 不等式，此处我们不作区分。

 **定理 10.1.1.** (Chebyshev 不等式) 设 ξ 是一个随机变量，则对任意 $\varepsilon > 0$

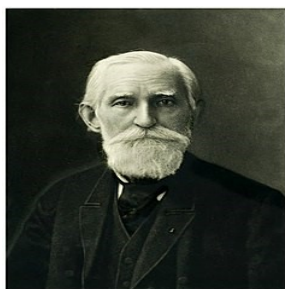
$$\mathbb{P}(|\xi| \geq \varepsilon) \leq \frac{1}{\varepsilon} \cdot \mathbb{E}|\xi|.$$

上面定理的证明很简单，只需在天然的不等式 $1_{\{|\xi| \geq \varepsilon\}} \leq \frac{|\xi|}{\varepsilon}$ 两边取数学期望即可；但这个定理及其思想在概率论中非常重要，它实现了通过数学期望的计算来估算概率，为估算概率带来了极大的便利。历史上 Bernoulli 得到他的大数律是通过复杂的概率估算实现的；而基于 Chebyshev 不等式，今天的我们对 Bernoulli 大数律的证明不过短短一行；请参见后文相关论述。

注 10.1. 原始的 Chebyshev 不等式是用二阶矩来估计概率的，是俄国数学家、力学家 P. L. Chebyshev (切比雪夫, 1821/5/26–1894/12/8) 首创的方法，用来简洁地论证 Bernoulli 的弱大数律，进而得到更一般的 Chebyshev 弱大数律。Chebyshev 生于奥卡多沃，卒于彼得堡。他早年接受家庭教育，1841 年毕业于莫斯科大学并获银质奖章，1847 年开始任彼得堡大学副教授，1850 年升为教授，1859 年被选为彼得堡科学院院士。他终身未娶，日常生活十分简朴，他的一点积蓄全部用来买书和造机器，最大乐趣是与年轻人讨论数学问题。以 Chebyshev 为首，19 世纪下半叶俄国开始形成自己的第一个有国际影响的数学学派，称为圣彼得堡学派或 Chebyshev 学派。

Chebyshev 培养的知名学生有：A. Korkin (1837/3/3–1908/9/1；俄罗斯数学家，在偏微分方程方向有贡献，是当时的圣彼得堡学派的第二把交椅)，K. A. Posse (1847/9/29–1928/8/24；俄罗斯数学家，在数学分析、特别是逼近论方面有贡献)，A. A. Markov (1856/6/14–1922/7/20；俄国)，A. Lyapunov (1857/6/6–1918/11/3；俄罗斯数学家、力学家、物理学家，以发展了动力系

* “几乎处处收敛”或“几乎必然收敛”在有些文献中也称为“以概率 1 收敛”。



П. Л. Чебышев

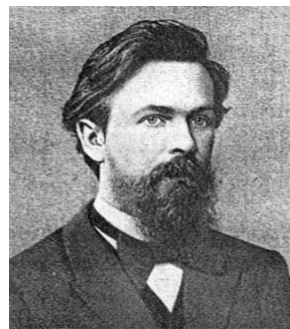


图 10.1: P. L. Chebyshev(1821-1894) 图 10.2: A. A. Markov(1856-1922)

统的稳定性理论闻名，对数学物理和概率论也有很多贡献），*D. Grave* (1863/9/6–1939/12/19；俄罗斯-苏联数学家，被认为是代数学 *Kyiv* 大学学派创始人，这个大学后来成为苏联代数学的中心；他的另一个博士论文导师是 *A. Korkin*)，*V. A. Markov* (1871/5/8–1897/1/18；俄罗斯数学家，*A. A. Markov* 的弟弟，与他哥哥一块证明了 *Markov* 不等式)。

A. A. Markov (马尔可夫, 1856/6/14–1922/7/20；俄国) 是 *Chebyshev* 最亲密的学生。他 1874 年进入彼得堡大学物理数学系学习，1878 年毕业于圣彼得堡大学并获得金质奖章；1880 年获得硕士学位，并在彼得堡大学任教，1884 年获得物理数学博士学位，1886 年成为教授。他在 1896 年当选为圣彼得堡科学院院士；上面的不等式是他整理提出的。*Markov* 被 *Kolmogorov* 评价为“*Chebyshev* 思想的优秀表达者”[49, 序言]；他同时也是马氏链 (*Markov Chain*) 理论的开创者。他的儿子也叫 *A. A. Markov* (小马尔可夫, 1903/9/22–1979/10/11)，也是数学家，但专长是数理逻辑。

Markov 培养的知名学生有：*G. Voronoy* (1868/4/28–1908/11/20；俄罗斯数学家，以 *Voronoi* 图、*Voronoi* 迭代/算法、*Voronoi* 公式知名)，*V. Kagan* (1869/3/10–1953/5/8；俄罗斯-苏联数学家，专长是几何学，他是 *Y. Sinai* 和 *G. Barenblatt* 的外祖父，他的另一个博士论文导师是 *Konstantin Posse*)，*N. Günther* (1871/12/17–1941/5/4；俄罗斯数学家，以位势理论、积分和偏微分方程方面的工作而知名，他是连续三届国际数学家大会的邀请报告人：1924 年多伦多、1928 年博洛尼亚、1932 年苏黎世)，*V. I. Romanovsky* (1879/12/4–1954/12/6；俄罗斯-苏联-乌兹别克数学家，塔什干数学学派创始人)，*J. V. Uspensky* (1883/4/29–1947/1/27；俄罗斯-美国数学家，以《*Theory of Equations*》而知名)，*J. Tamarkin* (1888/7/11–1945/11/18；俄罗斯-美国数学家，以数学分析方向的工作知名，美国女数学家 *D. L. Bernstein*、美国数学家 *N. Dunford*、斯坦福计算机系创始系主任 *G. Forsythe* 等都是他的学生)，*A. Besicovitch* (1891/1/23–1970/11/2；俄罗斯数学家，以 *Hausdorff-Besicovitch* 维数、*Besicovitch* 函数、*Besicovitch* 覆盖定理等闻名于世) 等。

Chebyshev 不等式有下面一个简单但又非常实用的推论，它给出了第 5 章随机变量定义的补丁程序中条件 $\mathbb{P}(|\xi| = \infty) = 0$ 的一种论证手段：

◆ 推论 10.1.1. 对于广义实值随机变量 ξ ， $\mathbb{E}|\xi| < \infty$ 蕴含了 $|\xi| < \infty$ a.s.。

另外 *Chebyshev* 不等式还有如下一些重要分布上的应用。

例 10.1. 设 $Z \sim N(0, 1)$ ，则对 $a > 0$

$$\mathbb{P}(Z \geq a) \leq e^{-a^2/2}.$$

参考证明：考虑引进带参数 $t > 0$ 的函数 $g(x, t) = e^{tx}$ 。于是 $Z \geq a$ 当且仅当 $g(Z, t) \geq g(a, t) > 0$ ，由 *Chebyshev* 不等式

$$\mathbb{P}(Z \geq a) \leq \frac{\mathbb{E}g(Z, t)}{g(a, t)} = \exp\left\{\frac{t^2}{2} - at\right\}.$$

上式中取 $t = a$ ，立即得到欲证明的不等式。 \square

例 10.2. 设 $X \sim \text{Poisson}(\lambda)$ ，则

$$\mathbb{P}(|X - \lambda| \geq \delta) \leq \frac{\lambda}{\delta^2}, \quad \forall \delta \geq \sqrt{\lambda}, \quad (10.1)$$

$$\mathbb{P}(X \geq \lambda + \delta) \leq \exp\left\{-\frac{\delta^2}{2(\lambda + \delta)}\right\}, \quad \forall \delta > 0 \quad (10.2)$$

$$\mathbb{P}(X \leq \lambda - \delta) \leq \exp\left\{-\frac{\lambda - \frac{5}{3}\delta}{2(\lambda - \delta)^2} \cdot \delta^2\right\}, \quad \forall \delta \in (0, \frac{\lambda}{2}]. \quad (10.3)$$

参考证明：注意到 $\mathbb{E}X = \lambda, \text{Var}(X) = \lambda$ ，(10.1)是 Chebyshev 不等式的简单应用；

注意到 $\mathbb{E}[e^{tX}] = \exp\{\lambda(e^t - 1)\}$ ，先考虑 $t > 0$ ，有

$$\mathbb{P}(X \geq \lambda + \delta) \leq \exp\{-t(\lambda + \delta)\} \mathbb{E}[e^{tX}] = \exp\{\lambda(e^t - 1) - t(\lambda + \delta)\}.$$

在上式中取 $t = \log(1 + \frac{\delta}{\lambda})$ ，得到

$$\mathbb{P}(X \geq \lambda + \delta) \leq \exp\left\{\delta + (\lambda + \delta) \log\left(1 - \frac{\delta}{\lambda + \delta}\right)\right\}.$$

再结合不等式 $\log(1 - \frac{\delta}{\lambda + \delta}) \leq -\frac{\delta}{\lambda + \delta} - \frac{\delta^2}{2(\lambda + \delta)}$ 立即得到(10.2)。

现在，对 $t > 0$ ，考虑 $\mathbb{E}[e^{-tX}] = \exp\{-\lambda(1 - e^{-t})\}$ ，类似有

$$\mathbb{P}(X \leq \lambda - \delta) \leq \exp\{t(\lambda - \delta)\} \mathbb{E}[e^{-tX}] = \exp\{-\lambda(1 - e^{-t}) + t(\lambda - \delta)\}.$$

在上式中取 $t = -\log(1 - \frac{\delta}{\lambda})$ ，得到

$$\mathbb{P}(X \leq \lambda - \delta) \leq \exp\left\{-\delta + (\lambda - \delta) \log\left(1 + \frac{\delta}{\lambda - \delta}\right)\right\}.$$

当 $\delta \in (0, \frac{\lambda}{2}]$ 时， $\frac{\delta}{\lambda - \delta} \in (0, 1]$ ，此时注意到

$$\log\left(1 + \frac{\delta}{\lambda - \delta}\right) \leq \frac{\delta}{\lambda - \delta} - \frac{\delta^2}{2(\lambda - \delta)^2} + \frac{\delta^3}{3(\lambda - \delta)^3},$$

立即可以整理得到(10.3)。 \square

10.2 各种收敛性的定义

概率论中，关于随机变量列的收敛，主要有下面几种不同的收敛性。

定义 10.2.1. 设 $\{\xi_n\}_{n=1}^{\infty}$ 为一个随机变量序列， ξ 为一个随机变量。

- (1) 称 $\{\xi_n\}_{n=1}^{\infty}$ 几乎处处收敛/几乎必然收敛到 ξ ，记作 $\xi_n \xrightarrow{a.e.} \xi$ 或 $\xi_n \xrightarrow{a.s.} \xi$ ，或 $\lim_{n \rightarrow \infty} \xi_n \stackrel{a.e.}{=} \xi$ ，或 $\lim_{n \rightarrow \infty} \xi_n \stackrel{a.s.}{=} \xi$ ，如果

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \xi_n \neq \xi\right) = 0;$$

- (2) 称 $\{\xi_n\}_{n=1}^{\infty}$ 依概率收敛到 ξ ，记作 $\xi_n \xrightarrow{\mathbb{P}} \xi$ ，如果

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\xi_n - \xi| \geq \varepsilon) = 0, \quad \forall \varepsilon > 0;$$

(3) 给定 $p > 0$, $\{\xi_n\}_{n=1}^\infty$ 称为 L^p 收敛到 ξ , 记作 $\xi_n \xrightarrow{L^p} \xi$, 如果

$$\lim_{n \rightarrow \infty} \mathbb{E}[|\xi_n - \xi|^p] = 0;$$

(4) $\{\xi_n\}_{n=1}^\infty$ 称为依分布收敛到 ξ , 记作 $\xi_n \xrightarrow{d} \xi$, 如果

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

在 F 的所有连续点 x 上成立; 此处 F_n, F 分别是 ξ_n, ξ 的分布函数; 有时也记 $F_n \xrightarrow{w} F$ 或 $\xi_n \xrightarrow{d} F$.*

需要指出的是, 在上述定义中, 与其他收敛性不同, 依分布收敛的随机变量列本身可以定义在不同的概率空间中, 而不必定义在同一个概率空间; 本质上它是分布函数列的一种收敛性质。关于依分布收敛 (及更一般的测度的弱收敛), 第 11 章将进一步详细论述。本章重点讨论前三种收敛性质之间的关系。

10.3 几种收敛之间的关系

10.3.1 L^p -收敛、依概率收敛与几乎处处收敛之间的关系

一般而言, L^p -收敛、依概率收敛与几乎处处收敛之间的关系如下。

定理 10.3.1. 设 $\{\xi_n : n \geq 1\}$ 为一个随机变量序列, ξ 为一个随机变量。

(1) 给定 $p > 0$, $\xi_n \xrightarrow{L^p} \xi$ 蕴含了 $\xi_n \xrightarrow{\mathbb{P}} \xi$;

(2) $\xi_n \xrightarrow{a.s.} \xi$ 蕴含了 $\xi_n \xrightarrow{\mathbb{P}} \xi$;

(3) 若 $\xi_n \xrightarrow{\mathbb{P}} \xi$, 则存在子序列 $\{\xi_{n_k} : k \geq 1\}$ 使得 $\xi_{n_k} \xrightarrow{a.s.} \xi$ 。

证明. 结论 (1) 由 Chebyshev 不等式保证。

结论 (2) 的证明如下。容易验证

$$\{\omega : \lim \xi_n(\omega) \neq \xi(\omega)\} = \bigcup_{k=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{\omega : |\xi_n(\omega) - \xi(\omega)| \geq \frac{1}{k}\},$$

从而 $\xi_n \xrightarrow{a.s.} \xi$ 时,

$$0 = \mathbb{P}(\lim \xi_n(\omega) \neq \xi(\omega)) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{\omega : |\xi_n(\omega) - \xi(\omega)| \geq \frac{1}{k}\}\right).$$

*当 μ_n, μ 分别是 ξ_n, ξ 的分布测度时, 也记 $\mu_n \xrightarrow{w} \mu$, 参见第 11 章。

进而对任意 $k \geq 1$

$$\begin{aligned} 0 &= \mathbb{P} \left(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{\omega : |\xi_n(\omega) - \xi(\omega)| \geq \frac{1}{k}\} \right) \\ &= \lim_{N \rightarrow \infty} \mathbb{P} \left(\bigcup_{n=N}^{\infty} \{\omega : |\xi_n(\omega) - \xi(\omega)| \geq \frac{1}{k}\} \right) \\ &\geq \overline{\lim}_{n \rightarrow \infty} \mathbb{P} \left(\{\omega : |\xi_n(\omega) - \xi(\omega)| \geq \frac{1}{k}\} \right), \end{aligned}$$

即知 $\lim_{n \rightarrow \infty} \mathbb{P}(|\xi_n - \xi| \geq \varepsilon) = 0, \forall \varepsilon > 0$ 。

下面证明结论 (3)。由于 $\xi_n \xrightarrow{\mathbb{P}} \xi$ ，对任意整数 $k \geq 1$ ，必存在 n_k 使得

$$\mathbb{P}(|\xi_{n_k} - \xi| \geq \frac{1}{k}) \leq \frac{1}{2^k}.$$

由 B-C 引理，集合 $\mathcal{N} := \bigcap_{K=1}^{\infty} \bigcup_{k=K}^{\infty} \{\omega : |\xi_{n_k}(\omega) - \xi(\omega)| \geq \frac{1}{k}\}$ 对应的概率为 0；

显然 $\forall \omega \notin \mathcal{N}$ ， $\lim_{k \rightarrow \infty} \xi_{n_k}(\omega) = \xi(\omega)$ 。因此 $\xi_{n_k} \xrightarrow{\text{a.s.}} \xi$ 。 \square

为了进一步阐述各种收敛之间的关系，我们引入一致可积的概念；这个概念本身也是概率论中的重要概念之一。

定义 10.3.1. 设 $\{\xi_\alpha\}_{\alpha \in I}$ 为（可积的）随机变量族，称它是一致可积的，如果

$$\lim_{N \rightarrow \infty} \sup_{\alpha \in I} \mathbb{E} \left[|\xi_\alpha| \cdot 1_{\{|\xi_\alpha| \geq N\}} \right] = 0.$$

显然，如果随机变量族被同一个可积的随机变量所控制，则这族随机变量是一致可积的；反之未必成立。

下面定理给出了一致可积的一个等价刻画，它形式上与泛函分析中的 Ascoli-Azela 定理非常相似。

定理 10.3.2. 设 $\{\xi_\alpha\}_{\alpha \in I}$ 为一个（可积的）随机变量族，则它一致可积的充分必要条件是：

(1) 等度绝对连续*：对任意 $\varepsilon > 0$ ，存在 $\delta > 0$ 使得当 $A \in \mathcal{F}$ 满足 $\mathbb{P}(A) < \delta$ 时就有 $\mathbb{E} \left[|\xi_\alpha| \cdot 1_A \right] < \varepsilon, \forall \alpha \in I$ ；

(2) 一致 L^1 有界： $\sup_{\alpha \in I} \mathbb{E} |\xi_\alpha| < \infty$ 。

证明. 必要性。对任意 $A \in \mathcal{F}, N > 0$ ，有

$$\begin{aligned} \mathbb{E} \left[|\xi_\alpha| \cdot 1_A \right] &= \mathbb{E} \left[|\xi_\alpha| \cdot 1_{A \cap \{|\xi_\alpha| \geq N\}} \right] + \mathbb{E} \left[|\xi_\alpha| \cdot 1_{A \cap \{|\xi_\alpha| < N\}} \right] \\ &\leq \mathbb{E} \left[|\xi_\alpha| \cdot 1_{\{|\xi_\alpha| \geq N\}} \right] + N \mathbb{P}(A). \end{aligned}$$

*此条也可等价地表述为：对任意 $\varepsilon > 0$ ，存在 $\delta > 0$ 使得当 $A_\alpha \in \mathcal{F}$ 满足 $\mathbb{P}(A_\alpha) < \delta$ 时就有 $\mathbb{E} \left[|\xi_\alpha| \cdot 1_{A_\alpha} \right] < \varepsilon, \forall \alpha \in I$ ；这样的表述从应用上更方便，但书中的表述形式上显得更弱。请读者思考：为何这两种表述其实是等价的？

由一致可积性，推出 (1) 中的等度绝对连续性。在上式中取 $A = \Omega$ ，得

$$\mathbb{E}[|\xi_\alpha|] \leq \mathbb{E}\left[|\xi_\alpha| \cdot 1_{\{|\xi_\alpha| \geq N\}}\right] + N,$$

立得 (2) 中的一致 L^1 有界性。

充分性。利用一致 L^1 有界性，由 Chebyshev 不等式，当 $N \rightarrow \infty$ 时

$$\sup_{\alpha \in I} \mathbb{P}(|\xi_\alpha| \geq N) \leq \frac{1}{N} \sup_{\alpha \in I} \mathbb{E}|\xi_\alpha| \rightarrow 0.$$

进而由等度绝对连续性， $\sup_{\alpha \in I} \mathbb{E}\left[|\xi_\alpha| \cdot 1_{\{|\xi_\alpha| \geq N\}}\right] \rightarrow 0$ 。 \square

下面给出一种比较实用的检验一致可积性的方法。证明留给读者。

定理 10.3.3. 给定随机变量族 $\{\xi_\alpha\}_{\alpha \in I}$ ，它一致可积的一个充分条件是：存在非负函数 $G: [0, \infty) \rightarrow \mathbb{R}$ ，满足 $\lim_{t \rightarrow \infty} \frac{G(t)}{t} = \infty$ 及 $\sup_{\alpha \in I} \mathbb{E}[G(|\xi_\alpha|)] < \infty$ 。

之前我们已经说明， L^1 收敛蕴含了依概率收敛。反过来，在适当条件下，依概率收敛也能导出 L^1 收敛。

定理 10.3.4. 可积随机变量列 $\{\xi_n\}_{n=1}^\infty$ L^1 收敛于 ξ 的充要条件是：

(1) $\{\xi_n\}_{n=1}^\infty$ 一致可积；

(2) $\xi_n \xrightarrow{\mathbb{P}} \xi$ 。

证明. (A) **必要性：** 首先，当随机变量列 $\{\xi_n\}_{n=1}^\infty$ 在 L^1 意义下收敛于 ξ 时， $\xi_n \xrightarrow{\mathbb{P}} \xi$ 是显然的。此即定理结论 (2)。

其次，我们有 $\mathbb{E}|\xi| \leq \mathbb{E}|\xi_n - \xi| + \mathbb{E}|\xi_n| < \infty$ 以及

$$\sup_n \mathbb{E}|\xi_n| \leq \sup_n \mathbb{E}|\xi_n - \xi| + \mathbb{E}|\xi| < \infty.$$

这表明 $\{\xi_n\}_{n=1}^\infty$ 一致 L^1 -有界。

对任意 $A \in \mathcal{F}$ ，由 $\mathbb{E}\left[|\xi_n| \cdot 1_A\right] \leq \mathbb{E}\left[|\xi| \cdot 1_A\right] + \mathbb{E}\left[|\xi_n - \xi|\right]$ 立得（此处约定 $\xi_0 := \xi$ ）

$$\sup_n \mathbb{E}\left[|\xi_n| \cdot 1_A\right] \leq \max_{0 \leq k \leq N} \mathbb{E}\left[|\xi_k| \cdot 1_A\right] + \sup_{n \geq N} \mathbb{E}\left[|\xi_n - \xi|\right].$$

由 $\mathbb{E}\left[|\xi_n - \xi|\right] \rightarrow 0$ 及 $\xi_k \in L^1$ ，立得 $\{\xi_n\}_{n=1}^\infty$ 的等度（绝对）连续性。

由定理 10.3.2， $\{\xi_n\}_{n=1}^\infty$ 一致可积。此即定理结论 (1)。

(B) **充分性：** 对任意 $\varepsilon > 0$ ，

$$\begin{aligned} \mathbb{E}\left[|\xi_n - \xi|\right] &\leq \varepsilon + \mathbb{E}\left[|\xi_n - \xi| \cdot 1_{\{|\xi_n - \xi| \geq \varepsilon\}}\right] \\ &\leq \varepsilon + \mathbb{E}\left[|\xi_n| \cdot 1_{\{|\xi_n - \xi| \geq \varepsilon\}}\right] + \mathbb{E}\left[|\xi| \cdot 1_{\{|\xi_n - \xi| \geq \varepsilon\}}\right]. \end{aligned}$$

因为 $\lim_{n \rightarrow \infty} \mathbb{P}(|\xi_n - \xi| \geq \varepsilon) = 0$ ，由 $\{\xi_n\}_{n=1}^\infty$ 的一致可积与 ξ 的可积性，上面

不等式右端可以任意小，即 $\xi_n \xrightarrow{L^1} \xi$ 。 \square

讨论 L^1 收敛与其他收敛的关系的一个目的是研究极限与数学期望交换

$$\mathbb{E}\left[\lim_{n \rightarrow \infty} \xi_n\right] = \lim_{n \rightarrow \infty} \mathbb{E}[\xi_n]$$

的条件。定理 10.3.4 告诉我们，若 $\{\xi_n\}_{n=1}^{\infty}$ 一致可积，且 $\xi_n \xrightarrow{\mathbb{P}} \xi$ ，那么上面极限与数学期望就可以交换；这是比 Lebesgue 控制收敛定理稍弱的条件。

10.3.2 阅读材料：随机序与随机控制收敛定理

两个实数总是可以比较大小、排一个顺序。但很多场合也希望对两个或多个随机的数（随机变量）排“大小”顺序，也就是所谓的随机序。我们这里只介绍如下在经济、金融及统计中广为应用的一种随机序。更多的随机序及相关应用，请读者参考相关文献。

定义 10.3.2. 两个随机变量 X, Y （未必定义在同一个概率空间中），称 X 比 Y 随机小（或称 X 被 Y 随机控制），如果 $\mathbb{P}(X \leq x) \geq \mathbb{P}(Y \leq x), \forall x$ 。此时我们记作 $X \preceq Y$ 。

在一些文献中，这种随机序也被称为一阶随机占优。显然它只是一种偏序关系。

给定两个分布函数 F_1, F_2 ，如果 $F_1 \geq F_2$ ，那么它们的广义逆就显然满足 $F_1^{-1} \leq F_2^{-1}$ 。注意到定理 9.2.1，我们容易论证下面的结果。

定理 10.3.5. 以下命题相互等价：

- (1) $X \preceq Y$;
- (2) 存在定义在同一个概率空间中的随机变量 $\tilde{X} \stackrel{d}{=} X, \tilde{Y} \stackrel{d}{=} Y$ 使得 $\tilde{X} \leq \tilde{Y}$ 几乎处处成立;
- (3) 对于任意单调非降函数 g , $g(X) \preceq g(Y)$;
- (4) 对于任意单调非降函数 g , 当 $\mathbb{E}g(X), \mathbb{E}g(Y)$ 都有意义时, 总有 $\mathbb{E}g(X) \leq \mathbb{E}g(Y)$ 。

给定非负随机变量 $X \sim F$ ，设 $0 < \mathbb{E}X < \infty$ ，定义

$$F^*(x) := \frac{\mathbb{E}[X \cdot 1_{\{X \leq x\}}]}{\mathbb{E}X}.$$

那么 F^* 也是一个分布函数，它称为 F 的 **size-biased** 变换。取 $f(y) := y, g(y) := 1_{(x, \infty)}(y)$ ，那么它们都是单调递增函数。设 \tilde{X} 与 X 同分布且独立，那么 $[f(\tilde{X}) - f(X)] \cdot [g(\tilde{X}) - g(X)] \geq 0$ 几乎处处成立。两边取数学期望，并整理就得到了 $F^*(x) \leq F(x), \forall x$ 。于是我们有下面的结果。

命题 10.3.1. 设 $X \sim F$ 为非负随机变量，满足 $0 < \mathbb{E}X < \infty$ 。设 F^* 是 F 的 **size-biased** 变换， $X^* \sim F^*$ 。那么 $X \preceq X^*$ 。

上面定义的随机序的另一个重要应用是关于一致可积性以及控制收敛定理的如下讨论；这给出了比控制收敛定理稍弱条件的极限与数学期望交换的条件，相应结果可以称为随机控制收敛定理。

定理 10.3.6.（随机控制收敛定理）设 $\{X_\alpha : \alpha \in I\}$ 、 X 满足：

- (1) $|X_\alpha| \preceq X, \forall \alpha \in I$;

(2) $\mathbb{E}[X] < \infty$.

那么 $\{X_\alpha : \alpha \in I\}$ 是一致可积的。特别的, 如果此时还有: $\mathbb{N} \subset I$, 且 $X_n \rightarrow X_\infty$ 几乎处处 (或依概率、依分布) 成立, 那么

$$\lim_{n \rightarrow \infty} \mathbb{E}X_n = \mathbb{E}X_\infty.$$

证明. 在所给条件下, 存在 $\{\tilde{X}_\alpha : \alpha \in I\}$ 、 \tilde{X} 满足: (1) $X_\alpha \stackrel{d}{=} \tilde{X}_\alpha, \forall \alpha \in I, X \stackrel{d}{=} \tilde{X}$; (2) $|\tilde{X}_\alpha| \leq \tilde{X}$ 几乎处处。此时, 对任意 $M > 0$

$$|\tilde{X}_\alpha| \cdot 1_{\{|\tilde{X}_\alpha| \geq M\}} \leq \tilde{X} \cdot 1_{\{\tilde{X} \geq M\}},$$

进而

$$\sup_{\alpha \in I} \mathbb{E}[|X_\alpha| \cdot 1_{\{|X_\alpha| \geq M\}}] = \sup_{\alpha \in I} \mathbb{E}[|\tilde{X}_\alpha| \cdot 1_{\{|\tilde{X}_\alpha| \geq M\}}] \leq \mathbb{E}[\tilde{X} \cdot 1_{\{\tilde{X} \geq M\}}].$$

结合 $\mathbb{E}X = \mathbb{E}[\tilde{X}] < \infty$, 即知 $\{X_\alpha : \alpha \in I\}$ 是一致可积的。

定理的后一结论是一致可积性的简单推论。 \square

注记 10.1. 定理 10.3.4 结合第 11 章的 Skorokhod 嵌入定理告诉我们, 在已经有随机变量序列的 (几乎处处/依概率/依分布) 收敛性的前提下, 一致可积性差不多是极限与数学期望交换的充分必要条件。现在, 我们又得到了随机控制方法下的一致可积性的判别方法, 进而得到了随机控制收敛定理。很自然要问一个反过来的问题: 是否一致可积的 (非负) 随机变量族总能找到可积的随机控制? 这个问题答案是否定的, 如下反例是赵敏智老师提供的: 设 $\{X_n\}_{n=1}^\infty$ 满足 $\frac{X_n}{n+1} \sim B(1, p_n), p_n = \frac{1}{(n+1)\log(n+1)}$ 。显然有 $X_n \xrightarrow{L^1} 0$ 。但不存在 $Y \in L^1$, 使得 $X_n \preceq Y$, 否则

$$\mathbb{P}(Y > n) \geq \mathbb{P}(X_n > n) = \frac{1}{(n+1)\log(n+1)}, \sum_{n=0}^\infty \mathbb{P}(Y > n) = \infty,$$

与 $Y \in L^1$ 相矛盾。

注记 10.2. 一般的, 我们可以按照如下办法定义更多的随机序: 给定一个特定的“效用函数集” \mathcal{U} , 对任意两个给定的随机变量 X, Y (本质上是谈它们对应的分布律), 我们称 Y 是 \mathcal{U} -随机占优于 X , 记作 $X \preceq_{\mathcal{U}} Y$, 如果

$$\mathbb{E}[u(X)] \leq \mathbb{E}[u(Y)], \forall u \in \mathcal{U}$$

(假定涉及的数学期望有意义)。上面定义的一阶随机占优, 使用的效用函数集合为

$$\mathcal{U}_1 := \{g : g \text{ 是单调不减函数}\} \supset \{1_{(a, \infty)}(x) : a \in \mathbb{R}\}.$$

我们把 $X \preceq_{\mathcal{U}_1} Y$ 简记为 $X \preceq_1 Y$ 。

在文献中, 有所谓的二阶随机占优, $X \preceq_2 Y$, 它的原始定义为:

$$\int_{-\infty}^x [\mathbb{P}(X \leq t) - \mathbb{P}(Y \leq t)] dt \geq 0, \forall x \in \mathbb{R}.$$

当 $X, Y \in L^1$ 时, 这也等价于 $\mathbb{E}[X \wedge x] \leq \mathbb{E}[Y \wedge x], \forall x \in \mathbb{R}$ (也等价于 $\mathbb{E}[(x - X)^+] \geq \mathbb{E}[(x - Y)^+], \forall x \in \mathbb{R}$)。可以证明, 在可积随机变量类中, 二阶随机占优也等价于 \mathcal{U}_2 -随机占优, 其中

$$\mathcal{U}_2 := \{g : g \text{ 是单调不减的凹函数}\} \supset \{x \wedge a : a \in \mathbb{R}\}.$$

另外，还有所谓的三阶随机占优， $X \preceq_3 Y$ ，它的原始定义为：

$$\int_{-\infty}^x \int_{-\infty}^y [\mathbb{P}(X \leq t) - \mathbb{P}(Y \leq t)] dt dy \geq 0, \forall x \in \mathbb{R}.$$

易知，在 L^2 机变量类中，三阶随机占优也等价于 \mathcal{U}_3 -随机占优，其中

$$\mathcal{U}_3 := \{-(x-a)^2 \cdot 1_{[a, \infty)}(x) : a \in \mathbb{R}\}.$$

为了刻画一致可积性，我们也可以定义如下的随机占优序 \preceq' ：取

$$\mathcal{U}' := \{xg(x) : g \text{ 单调不减, 且 } g(-\infty) = 0\} \supset \{x \cdot 1_{(a, \infty)}(x) : a \in \mathbb{R}\},$$

我们定义 $X \preceq' Y$ ，如果 $X \preceq_{\mathcal{U}'} Y$ 。可以证明，随机变量族 $\{X_\alpha\}_{\alpha \in I}$ 是一致可积的，当且仅当存在 $Y \in L^1$ ，使得 $|X_\alpha| \preceq' Y, \forall \alpha \in I$ 。

10.3.3 依概率收敛与依分布收敛之间的关系

显然，依概率收敛可以导出依分布收敛。

定理 10.3.7. 设 $\{\xi_n\}_{n=1}^\infty$ 依概率收敛到 ξ ，那么 $\{\xi_n\}_{n=1}^\infty$ 依分布收敛到 ξ 。

证明. 留作习题。 \square

在特定条件下，依概率收敛与依分布收敛之间可以做到互相等价。

定理 10.3.8. 设 $\{\xi_n\}_{n=1}^\infty$ 依分布收敛到某常数 c 的充分必要条件是 $\{\xi_n\}_{n=1}^\infty$ 依概率收敛到这个常数 c 。

证明. 留作习题。 \square

上面的定理 10.3.8 再结合第 11 章中的特征函数理论，就给出了论证弱大数律的另一种特殊技巧：**特征函数方法**。Khinchine 弱大数律就是使用这种方法建立起来的。我们提请读者参阅第 11 章的相关内容。

10.4 大数律简介

10.4.1 Borel-Cantelli 引理

下面的 Borel-Cantelli 引理本身的证明非常简单，但它是概率论中用来证明有关几乎处处收敛性命题的重要工具。

定理 10.4.1. (Borel-Cantelli 引理) 设 $\{A_n\}_{n=1}^\infty \subset \mathcal{F}$ 。如果

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

那么 $\mathbb{P}(A_n \text{ i.o.}) = 0$ ，其中 $\{A_n \text{ i.o.}\} = \overline{\lim}_{n \rightarrow \infty} A_n = \{\sum_{n=1}^{\infty} 1_{A_n} = \infty\}$ 。

证明. 定义广义非负随机变量 $\xi := \sum_{n=1}^{\infty} 1_{A_n}$ 。显然 $\mathbb{E}\xi = \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ 。于是 $\xi < \infty$ a.s.，此即 $\mathbb{P}(A_n \text{ i.o.}) = 0$ 。 \square

上面的 Borel-Cantelli 引理，有时也称为 Borel-Cantelli 第一引理。利用上述 Borel-Cantelli 引理，我们可以证明如下结果。

◆ 推论 10.4.1. * 设 $S_n := \sum_{k=1}^n X_k$ 为非负随机变量列 $\{X_n\}_{n=1}^\infty$ 的部分和。假定 $\mathbb{E}S_n \rightarrow \infty$, 且 $\sup\{\mathbb{E}X_n : n \geq 1\} < \infty$ 。如果存在正数 $C, \delta > 0$ 使得: 对一切正整数 n , 我们有下面估计

$$\text{Var}(S_n) \leq C \cdot (\mathbb{E}S_n)^{2-\delta} \quad (10.4)$$

或者更弱的估计

$$\text{Var}(S_n) \leq C \cdot (\mathbb{E}S_n)^2 / (\log \mathbb{E}S_n)^{1+\delta}, \quad (10.5)$$

那么

$$\lim_{n \rightarrow \infty} \frac{S_n}{\mathbb{E}S_n} \stackrel{\text{a.s.}}{=} 1. \quad (10.6)$$

证明. 不妨设 $0 < M := \sup\{\mathbb{E}X_n : n \geq 1\} \leq 1$ 。注意到 $0 \leq \mathbb{E}X_n \leq 1$, 容易知道数列 $\{E(n) := \mathbb{E}S_n\}_{n=1}^\infty$ 的整数部分能够取到全体自然数。简单起见, 我们设估计 (10.4) 成立。取 $n_k := \inf\{n : E(n) \geq k^{2/\delta}\}$ 。那么

$$k^{2/\delta} \leq E(n_k) \leq k^{2/\delta} + 1, \quad \forall k \geq 1.$$

此时由 Chebyshev 不等式及估计 (10.4)

$$\mathbb{P}(|\frac{S_{n_k}}{E(n_k)} - 1| \geq \varepsilon) \leq \frac{\text{Var}(S_{n_k})}{\varepsilon^2 \cdot E(n_k)^2} \leq \frac{C}{\varepsilon^2 \cdot k^2}, \quad \forall k \geq 1, \varepsilon > 0.$$

由 Borel-Cantelli 引理, $\lim_{k \rightarrow \infty} \frac{S_{n_k}}{E(n_k)} \stackrel{\text{a.s.}}{=} 1$ 。

对于充分大的 n , 存在充分大的 k 使得 $n \in [n_k, n_{k+1})$ 。此时利用 $S_n, E(n)$ 的单调性,

$$\frac{E(n_k)}{E(n_{k+1})} \cdot \frac{S_{n_k}}{E(n_k)} \leq \frac{S_n}{E(n)} \leq \frac{E(n_{k+1})}{E(n_k)} \cdot \frac{S_{n_{k+1}}}{E(n_{k+1})}.$$

注意到 $\frac{E(n_k)}{E(n_{k+1})} \rightarrow 1$, 我们很容易看到 $\frac{S_n}{E(n)} \rightarrow 1$ 几乎处处成立。□

由上述证明可以知道以下更一般的结论。

◆ 推论 10.4.2. 设 $\{S_n\}_{n=1}^\infty$ 非负、单调上升, 且 $E(n) := \mathbb{E}S_n \rightarrow \infty$ 。记 $V(n) := \text{Var}(S_n)$ 。如果存在子列 n_k 满足

$$\lim_{k \rightarrow \infty} \frac{E(n_k)}{E(n_{k+1})} = 1 \text{ 且 } \sum_{k=1}^\infty \frac{V(n_k)}{E(n_k)^2} < \infty, \quad (10.7)$$

则推论 10.4.1 中的结论 (10.6) 仍然成立。

某种程度上, Borel-Cantelli 引理的逆命题成立, 具体陈述为如下定理; 在一些文献中这被称为 Borel-Cantelli 第二引理[†]。

☞ 定理 10.4.2. 设 $\{A_n\}_{n=1}^\infty \subset \mathcal{F}$, 满足下面条件中的任意一条:

(1) $\{A_n\}_{n=1}^\infty$ 相互独立;

*本推论本质上隐含于 Dvoretzky 与 Erdős 在 1950 年的一项合作工作中; 编者发掘整理成此处的形式 (见 [8])。

[†]部分文献把条件 (1) 下的结果单独叫做 Borel-Cantelli 第二引理, 或者与之前的 Borel Cantelli 引理的结果放在一块陈述, 合称 Borel-Cantelli 引理; 有些文献也把条件 (3)-(4') 下的结果称为 Borel-Cantelli 第二引理。在本书中我们把几种条件下的结果统一称为 Borel-Cantelli 第二引理。

(2) $\{A_n\}_{n=1}^\infty$ 成对独立;

(3) 存在 $\rho = \{\rho_n\}_1^\infty \in \ell^1$ 使得对任意 $i \neq j$

$$\mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i) \cdot \mathbb{P}(A_j) \leq \rho_{|i-j|} \sqrt{\mathbb{P}(A_i) \cdot \mathbb{P}(A_j)}.$$

(3') 存在 $\rho = \{\rho_n\}_1^\infty \in \ell^1$ 使得对任意 $i \neq j$

$$\mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i) \cdot \mathbb{P}(A_j) \leq \rho_{|j-i|} (\mathbb{P}(A_i) + \mathbb{P}(A_j)),$$

(4) 存在 $a = (a_i : i \geq 1) \in \ell^2$ 使得对任意 $i \neq j$

$$\mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i) \cdot \mathbb{P}(A_j) \leq a_i a_j \sqrt{\mathbb{P}(A_i) \cdot \mathbb{P}(A_j)};$$

(4') 存在无穷矩阵 $a = (a_{i,j} : i \geq 1, j \geq 1) : \ell^2 \rightarrow \ell^2, x \mapsto a \cdot x$ 为有界线性算子, 使得对任意 $i \neq j$

$$\mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i) \cdot \mathbb{P}(A_j) \leq a_{i,j} \sqrt{\mathbb{P}(A_i) \cdot \mathbb{P}(A_j)};$$

那么 $\sum_{n=1}^\infty \mathbb{P}(A_n) = \infty$ 蕴含了 $\mathbb{P}(A_n \text{ i.o.}) = 1$ 。

利用上述推论 10.4.1, 我们有下面相对统一且简单的证明。

定理 10.4.2 的证明. 简单起见, 我们只在条件 (3) 下给出证明, 其余情况类似可证 (留作习题)。

令 $S_n := \sum_{k=1}^n 1_{A_k}$ 。显然 S_n 满足推论 10.4.1 前半部分条件。再考察方差:

$$\begin{aligned} \text{Var}(S_n) &= \sum_{k=1}^n \text{Var}(1_{A_k}) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(1_{A_i}, 1_{A_j}) \\ &\leq \sum_{k=1}^n \mathbb{P}(A_k) + 2 \sum_{1 \leq i < j \leq n} \rho_{|i-j|} \sqrt{\mathbb{P}(A_i) \cdot \mathbb{P}(A_j)} \\ &\leq \mathbb{E}S_n + \sum_{1 \leq i < j \leq n} \rho_{|i-j|} \cdot (\mathbb{P}(A_i) + \mathbb{P}(A_j)) \\ &\leq (1 + 2\|\rho\|_1) \cdot \mathbb{E}S_n. \end{aligned}$$

于是推论 10.4.1 的所有条件都满足, 由此 $\lim_{n \rightarrow \infty} \frac{S_n}{\mathbb{E}S_n} = 1$ 几乎处处成立。由于 $\mathbb{E}S_n \rightarrow \infty, S_n \rightarrow \infty$ 几乎处处成立, 亦即 $\mathbb{P}(A_n \text{ i.o.}) = 1$ 。□

下面介绍另一个证明 [38, pp. 18]。

定理 10.4.2 在条件 (3) 下的另一证明. 令 $S_n := \sum_{k=1}^n 1_{A_k}$ 。同上导出: $\exists C > 0$

使得 $\text{Var}(S_n) \leq C\mathbb{E}S_n, \forall n \geq 1$ 。现在, 对任意 $a > 0$, 取 $N \geq 1$ 充分大, 使得 $\mathbb{E}S_N > a$ 。于是对任意 $n \geq N$

$$\begin{aligned} \mathbb{P}(S_\infty \leq a) &\leq \mathbb{P}(S_n \leq a) = \mathbb{P}(-(S_n - \mathbb{E}S_n) \geq \mathbb{E}S_n - a) \\ &\leq \frac{\mathbb{E}[|S_n - \mathbb{E}S_n|^2]}{[\mathbb{E}S_n - a]^2} = \frac{\text{Var}(S_n)}{[\mathbb{E}S_n - a]^2} \leq \frac{C \cdot \mathbb{E}S_n}{[\mathbb{E}S_n - a]^2}. \end{aligned}$$

令 $n \rightarrow \infty$, 注意到 $\mathbb{E}S_n \rightarrow \infty$, 立即得到 $\mathbb{P}(S_\infty \leq a) = 0, \forall a > 0$. 于是定理结论成立. \square

为便于读者参考、比较, 此处我们也写下传统教科书中在条件 (1) 下对定理 10.4.2 的一个证明:

定理 10.4.2 在条件 (1) 下的证明. 注意到 $\{A_n \text{ i.o.}\} = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n$, 结合集合运算的 De Morgan 律、概率测度的上下连续性、条件 (1) 等, 有

$$\begin{aligned} 1 - \mathbb{P}(A_n \text{ i.o.}) &= \mathbb{P}\left(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n\right)^c = \mathbb{P}\left(\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n^c\right) \\ &= \lim_{N \rightarrow \infty} \mathbb{P}\left(\bigcap_{n=N}^{\infty} A_n^c\right) = \lim_{N \rightarrow \infty} \lim_{M \rightarrow \infty} \mathbb{P}\left(\bigcap_{n=N}^M A_n^c\right) \\ &= \lim_{N \rightarrow \infty} \lim_{M \rightarrow \infty} \prod_{n=N}^M [1 - \mathbb{P}(A_n)] \\ &\leq \lim_{N \rightarrow \infty} \lim_{M \rightarrow \infty} \prod_{n=N}^M e^{-\mathbb{P}(A_n)} = \lim_{N \rightarrow \infty} \exp\left\{-\sum_{n=N}^{\infty} \mathbb{P}(A_n)\right\}. \end{aligned}$$

注意到 $\sum_{n=N}^{\infty} \mathbb{P}(A_n) = \infty$, 因此 $1 - \mathbb{P}(A_n \text{ i.o.}) \leq 0$, 即 $\mathbb{P}(A_n \text{ i.o.}) = 1$. \square

习题 10.29 则告诉我们下面的关于 Borel-Cantelli 第二引理的更一般结果:

定理 10.4.3. (Kochen-Stone) 设 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$, 令 $S_n := \sum_{k=1}^n 1_{A_k}$. 如果

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{(\mathbb{E}S_n)^2}{\mathbb{E}[S_n^2]} &= 1, \text{ 或者等价的,} \\ \lim_{n \rightarrow \infty} \frac{\text{Var}(S_n)}{(\mathbb{E}S_n)^2} &= 0, \end{aligned} \quad (10.8)$$

那么 $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ 蕴含了 $\mathbb{P}(A_n \text{ i.o.}) = 1$.

以后 Borel-Cantelli (第一/二) 引理也时常简写为 B-C (第一/二) 引理。

注 10.2. *Émile Borel* (波莱尔, 1871/1/7–1956/2/3; 法国) 生于圣阿夫里克 (Saint-Affrique), 卒于巴黎, 父亲是一个新教牧师。他以第一名的身份同时通过巴黎高等师范学校和巴黎综合理工的考试, 在 1889 年入学巴黎高等师范学校, 1892 年毕业, 1893 年发表博士论文 (导师是 *G. Darboux*), 并在里尔大学任教。1897 年他回到巴黎高等师范学校并任函数论方向的主席到 1941 年。*E. Borel* 与 *R.-L. Baire*、*H. Lebesgue* 三人是当时的测度论方向的前驱; 他还在数论 (正规数)、双曲几何与狭义相对论等方面有重要贡献。1922 年他筹建了巴黎统计研究所 (*Paris Institute of Statistics*), 1928 年他与 *Birkhoff* 一起筹建了 *Poincaré* 研究所 (*Institut Henri Poincaré*)。

Borel 培养的知名学生有: *P. Dienes* (1882/11/24–1952/3/23; 匈牙利数学家、诗人), *H. Lebesgue* (1875/6/28–1941/7/26; 法国数学家, 以 *Lebesgue* 积分闻名于世), *P. Montel* (1876/4/29–1975/1/22; 法国数学家, 研究复分析中的全纯函数, 以 *Montel* 定理、*Montel* 空间、*Normal* 族闻名; 据说他同时也是 *Lebesgue* 的学生), *G. Valiron* (1884/9/7–1955/3/17; 法国数学家, 以 *Valiron* 定理知名, 1932 年苏黎世国际数学家大会的大会报告人) 等。



图 10.3: Borel(1871--1956)



图 10.4: Cantelli(1875-1966)

Francesco Paolo Cantelli (坎泰利, 1875/12/20–1966/7/21; 意大利) 生于帕勒莫 (Palermo), 卒于罗马。他 1899 年博士毕业于帕勒莫大学, 早期从事天文学和天体力学方面的工作。1903–1923 年他转而研究数理金融学和精算科学以及概率论, 由此才逐渐成名。他在概率论方面的著名工作有 Borel-Cantelli 引理、Glivenko-Cantelli 定理等。

下面纯分析的引理在证明随机变量列的几乎处处收敛时非常有用; 证明留给读者。

♠ 引理 10.4.1. 设 $\{X_n\}_{n=1}^\infty$ 满足: 存在 $\{n_k\}_{k=1}^\infty \subset \mathbb{N}$ 使得 $X_{n_k} \xrightarrow{a.s.} X$, 并且

$$\max_{n_{k-1} < n \leq n_k} |X_n - X_{n_{k-1}}| \xrightarrow{a.s.} 0 \quad (\text{或} \quad \max_{n_{k-1} \leq n < n_k} |X_n - X_{n_k}| \xrightarrow{a.s.} 0),$$

那么 $X_n \xrightarrow{a.s.} X$ 。

下面的定理给出了通过 p 阶绝对矩的估算来论证几乎处处收敛的一种比较粗糙、但非常实用的办法; 证明同样留给读者。

☞ 定理 10.4.4. 设 $\{X_n\}_{n=1}^\infty$ 满足: $\exists p > 0, \sum_{n=1}^\infty \mathbb{E}[|X_n|^p] < \infty$ 。那么 $X_n \xrightarrow{a.s.} 0$ 。

10.4.2 从 Bernoulli 弱大数律到 Borel 强大数律

概率论发展史上最早的大数定律由 Jacob Bernoulli 给出 (也被认为是概率论历史上第一个极限定理), 陈述如下。

☞ 定理 10.4.5. (Bernoulli 弱大数律) 设 $\{X_n\}_{n=1}^\infty$ 为独立同分布的 Bernoulli 两点分布的随机变量序列, $X_1 \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$, 其中 $q = 1 - p$, $p \in (0, 1)$ 。那么弱大数律成立: $\bar{X}_n \xrightarrow{\mathbb{P}} p$, 其中 $\bar{X}_n := \frac{1}{n}(X_1 + \cdots + X_n)$ 。

通常把上述定理中的 X_n 解释为第 n 次独立重复某项实验得到的结果, $X_n = 1$ 代表成功, 概率为 p ; $X_n = 0$ 代表失败, 概率为 q ; 从而可以把 \bar{X}_n 解释为前 n 次重复实验中成功的频率。而上述定理的结论是, 当重复实验充分多次后, 成功的频率就稳定在成功的概率附近。这被认为是古典概率论对概率的一种直观的解释, 这也是我们第 1 章就介绍了的“频率稳定性”的精

确数学陈述。

定理 10.4.5 的证明： 使用二阶矩对应的 Chebyshev 不等式得到： $\forall \varepsilon > 0$

$$\mathbb{P}(|\bar{X}_n - p| \geq \varepsilon) \leq \frac{\mathbb{E}[|\bar{X}_n - p|^2]}{\varepsilon^2} = \frac{pq}{n\varepsilon^2} \rightarrow 0,$$

其中 $q := 1 - p$ 。 □

基于上面的论证技巧，Chebyshev 建立了如下更一般的弱大数律：

定理 10.4.6. (Chebyshev 弱大数律) 设 $\{X_n\}_1^\infty$ 两两无关，且它们的方差存在并具有共同有限上界

$$\sup_n \text{Var}(X_n) < \infty.$$

则弱大数律成立：

$$\frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{\mathbb{P}} 0, \quad (10.9)$$

其中 $S_n = X_1 + \cdots + X_n$ 。

Markov 指出， $\frac{\text{Var}(S_n)}{n^2} \rightarrow 0$ 时就有 (10.9)，这被称为 **Markov 弱大数律**。

重新回到原始的 Bernoulli 弱大数律，Borel 进一步考虑使用四阶矩对应的 Chebyshev 不等式得到：

$$\mathbb{P}(|\bar{X}_n - p| \geq \varepsilon) \leq \frac{\mathbb{E}[|\bar{X}_n - p|^4]}{\varepsilon^4} = \frac{3n(n-1)(pq)^2 + npq(p^3 + q^3)}{n^4\varepsilon^4} \leq \frac{3pq}{n^2\varepsilon^4}.$$

于是用 B-C 引理得出结论： $\lim_{n \rightarrow \infty} \bar{X}_n \stackrel{\text{a.s.}}{=} p$ 。此处也可以通过论证

$$\sum_{n=1}^{\infty} \mathbb{E}[|\bar{X}_n - p|^4] < \infty,$$

一定程度上绕开 B-C 引理而得出结论。上述方法运用到一般的 i.i.d. 随机变量序列上就得到

定理 10.4.7. (Borel 强大数律) 设 $\{X_n\}_{n=1}^\infty$ 为独立同分布的随机变量序列，满足 $\mathbb{E}|X_1|^4 < \infty$ 。那么 $\lim_{n \rightarrow \infty} \bar{X}_n = \mathbb{E}X_1$ a.s.。

钟开莱 [61] 告诉我们，1932 年 A. Rajchman (雷奇曼, 1890/11/13–1940/?/?; 波兰犹太裔数学家) 在 Chebyshev 弱大数律的条件下得到了如下结果：

定理 10.4.8. (Rajchman 强大数律) 设 $\{X_n\}_1^\infty$ 两两无关，且它们的方差存在并具有共同有限上界

$$\sup_n \text{Var}(X_n) < \infty.$$

则强大数律成立： $\frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{\text{a.s.}} 0$ ，其中 $S_n = X_1 + \cdots + X_n$ 。

证明。 不妨设 $\mathbb{E}X_n = 0, \forall n \geq 1$ 及 $M := \sup_n \text{Var}(X_n)$ ，我们论证 $\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$ 。

由 $\{X_n\}_1^\infty$ 两两无关，

$$\mathbb{E}[S_n^2] = \text{Var}(S_n) = \sum_{k=1}^n \text{Var}(X_k) \leq nM.$$

由此容易知道 $\sum_{n=1}^{\infty} \mathbb{E}[(\frac{S_{n^2}}{n^2})^2] \leq \sum_{n=1}^{\infty} \frac{M}{n^2} < \infty$, 进而 $\frac{S_{n^2}}{n^2} \xrightarrow{\text{a.s.}} 0$.

以下考察 $D_n := \sup\{|S_k - S_{n^2}| : n^2 \leq k < (n+1)^2\}$. 利用 Cauchy 不等式, 得到

$$D_n^2 \leq 2n \sum_{n^2 < k < (n+1)^2} X_k^2.$$

于是同理可得 $\sum_{n=1}^{\infty} \mathbb{E}[(\frac{D_n}{n^2})^2] \leq \sum_{n=1}^{\infty} \frac{4M}{n^2} < \infty$, 进而 $\frac{D_n}{n^2} \xrightarrow{\text{a.s.}} 0$.

现在, 对任意 k 充分大, 存在 n 充分大, 使得 $n^2 \leq k < (n+1)^2$. 此时

$$\frac{|S_k|}{k} \leq \frac{|S_{n^2}| + D_n}{n^2}.$$

根据之前已经论证的结论, 立即知道定理结论成立。 \square

注记 10.3. 历史上, *J. Bernoulli* 对他的弱大数律的证明是通过对二项分布的“尾部”概率的复杂估计而得出的; 后来 *De Moivre* (棣莫弗, 1667/5/26–1754/11/27; 法国-英国) 和 *Laplace* (拉普拉斯, 1749/3/23–1827/3/5; 法国) 直接对二项分布的概率 $C_n^k p^k (1-p)^{n-k}$ (前者对 $p = \frac{1}{2}$, 后者对一般 p) 给出渐近公式, 基于这种方法, 不仅重新论证了 *Bernoulli* 弱大数定律, 而且得到了 *Bernoulli* 分布对应的“局部中心极限定理”和“中心极限定理”; 参见第 11 章习题 11.32. 现在我们论证弱大数律所用的 *Chebyshev* 不等式估计概率的思想、论证强大数律所用的基于 *B-C* 引理结合 *Chebyshev* 不等式 (及其他推广的概率不等式) 进行概率估计的方法以及以后将介绍的论证中心极限定理的特征函数方法等经典思想与方法都是后人在前辈经典工作的基础上总结提高、反思创新而陆续发展出来的。

10.4.3 从 Khintchine 弱大数律到 Kolmogorov 强大数律

回到 i.i.d. 情形, *Chebyshev* 弱大数律要求二阶矩条件; 在一阶矩条件下, *Khintchine* 证明了他的弱大数律, 陈述如下 (证明请参见第 11 章):

定理 10.4.9. (*Khintchine* 弱大数律) 设 $\{X_n\}_{n=1}^{\infty}$ 为独立同分布的随机变量序列, 满足 $\mathbb{E}|X_1| < \infty$. 记 $\mu := \mathbb{E}X_1$, 那么 $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$.

后来, *Kolmogorov* 进一步得到了冠以他的名字的一个强大数律; 它是独立同分布随机变量列情形下的最优结果. 命题陈述如下:

定理 10.4.10. (*Kolmogorov* 强大数律) 设 $\{X_n\}_{n=1}^{\infty}$ 为独立同分布的随机变量序列. 则存在随机变量 μ , 使得 $\lim_{n \rightarrow \infty} \overline{X}_n = \mu$ a.s. 成立的充分必要条件是: $\mathbb{E}|X_1| < \infty$. 此时, $\mu = \mathbb{E}X_1$ 是常数.

在 *Kolmogorov* 强大数律的证明过程中, 数学分析中著名的 *Kronecker* 引理发挥了重要的作用, 它把证明平均值的极限存在性问题转化为相应级数的收敛性问题. 我们不加证明的陈述这一重要引理如下.

♠ 引理 10.4.2. (Kronecker 引理) 设 $\{x_n\}_{n=1}^{\infty}$ 为实数列。如果 $\sum_{n=1}^{\infty} \frac{x_n}{n} < \infty$, 那么 $\lim_{n \rightarrow \infty} \frac{x_1 + \cdots + x_n}{n} = 0$ 。更一般的, 设 $0 < a_n \nearrow \infty$, 且 $\sum_{n=1}^{\infty} \frac{x_n}{a_n} < \infty$, 那么 $\lim_{n \rightarrow \infty} \frac{x_1 + \cdots + x_n}{a_n} = 0$ 。

为了证明 Kolmogorov 强大数律, 我们还需要一些预备结果。

在处理独立和的几乎处处收敛问题中, 一个重要工具就是下面的 Kolmogorov 不等式, 它某种程度上可以视作 Chebyshev 不等式的推广。

♠ 引理 10.4.3. (Kolmogorov 不等式) 设随机变量列 $\{X_k\}_{k=1}^n$ 相互独立, 且 $\mathbb{E}X_k = 0, \mathbb{E}X_k^2 < \infty, k = 1, \dots, n$ 。那么对任意 $\varepsilon > 0$

$$\mathbb{P}(\max_{1 \leq k \leq n} |S_k| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \mathbb{E}[|S_n|^2] = \frac{1}{\varepsilon^2} \sum_{k=1}^n \text{Var}(X_k).$$

证明. 不妨设 $n \geq 2$, 令 $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$,

$$A_k := \{|S_i| < \varepsilon, 1 \leq i < k\} \cap \{|S_k| \geq \varepsilon\},$$

及 $A := \bigcup_{k=1}^n A_k$ 。于是

$$\begin{aligned} \mathbb{E}[|S_n|^2 \cdot 1_{A_k}] &= \mathbb{E}[\mathbb{E}[|S_n|^2 \cdot 1_{A_k} | \mathcal{F}_k]] = \mathbb{E}[\mathbb{E}[|S_n|^2 | \mathcal{F}_k] \cdot 1_{A_k}] \\ &= \mathbb{E}[\mathbb{E}[|S_n - S_k + S_k|^2 | \mathcal{F}_k] \cdot 1_{A_k}] \\ &\geq \mathbb{E}[|S_k|^2 \cdot 1_{A_k}] \geq \varepsilon^2 \mathbb{P}(A_k), \end{aligned}$$

进而 $\mathbb{E}[|S_n|^2 \cdot 1_A] \geq \varepsilon^2 \sum_{k=1}^n \mathbb{P}(A_k) = \varepsilon^2 \mathbb{P}(A)$ 。注意到

$$\mathbb{E}[|S_n|^2 \cdot 1_A] \leq \mathbb{E}[|S_n|^2] = \sum_{k=1}^n \text{Var}(X_k),$$

引理结论成立。 □

利用 Kolmogorov 不等式, 我们很容易证明下面的结论。

♠ 引理 10.4.4. 设随机变量列 $\{X_n\}_{n=1}^{\infty}$ 相互独立, 且 $\mathbb{E}X_k = 0, \forall k$ 。如果

$$\sum_{n=1}^{\infty} \mathbb{E}X_n^2 < \infty,$$

那么 $\sum_{n=1}^{\infty} X_n$ 几乎处处收敛。特别的, 如果 $\sum_{n=1}^{\infty} \frac{\text{Var}(X_n)}{n^2} < \infty$, 那么 $\sum_{n=1}^{\infty} \frac{X_n}{n}$ 几乎处处收敛。

证明. 后一推论是显然的。我们来证明引理的前一个结论。令 $S_n := \sum_{k=1}^n X_k$ 。

由 $\{X_n : n \geq 1\}$ 的相互独立性及零均值性, 容易知道存在随机变量 S_{∞} 使得 $S_n \xrightarrow{L^2} S_{\infty}$, 进而存在子列 $\{n_k : k \geq 1\} \subset \mathbb{N}$ 使得 $S_{n_k} \xrightarrow{\text{a.s.}} S_{\infty}$ 。

又由 Kolmogorov 不等式, 对任意 $k \geq 0$ (此处约定 $n_0 := 0$ 及 $S_0 := 0$)

$$\mathbb{P}\left(\max_{n_k < p \leq n_{k+1}} |S_p - S_{n_k}| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \cdot \mathbb{E}|S_{n_{k+1}} - S_{n_k}|^2,$$

且 $\sum_{k=0}^{\infty} \mathbb{E}|S_{n_{k+1}} - S_{n_k}|^2 = \sum_{n=1}^{\infty} \mathbb{E}X_n^2 < \infty$. 由 B-C 引理, $\max_{n_k < p \leq n_{k+1}} |S_p - S_{n_k}| \rightarrow 0$ 几乎处处成立, 进而由引理 10.4.1, $S_n \rightarrow S_{\infty}$ 几乎处处成立. \square

定理 10.4.10 的证明. (1) 一阶矩有限条件的必要性: 由于 $\{X_k\}_{k=1}^{\infty}$ 相互独立, $A_k := \{\omega : |X_k(\omega)| \geq k\}, k = 1, 2, \dots$ 也相互独立. 而 $\lim_{n \rightarrow \infty} \bar{X}_n = \mu$ a.s.

无疑蕴含了 $\lim_{n \rightarrow \infty} \frac{X_n}{n} = 0$ a.s. 成立, 进而应有 $\mathbb{P}(A_n \text{ i.o.}) = 0$. 由定理 10.4.2,

这意味着 $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. 由同分布性, 此即 $\sum_{n=1}^{\infty} \mathbb{P}(|X_1| \geq n) < \infty$, 上述结论亦等价于 $\mathbb{E}|X_1| < \infty$ (为什么? 留作习题).

(2) 以下证明一阶矩有限条件是充分的. 此时不妨设 $\mathbb{E}X_1 = 0$. 令 $Y_n := X_n \cdot 1_{\{|X_n| \leq n\}}, n \geq 1$. 我们断言:

断言 1. $\frac{1}{n} \sum_{k=1}^n \mathbb{E}Y_k \rightarrow 0$ 成立;

断言 2. $\frac{1}{n} \sum_{k=1}^n (X_k - Y_k) \rightarrow 0$ 几乎处处成立;

断言 3. $\frac{1}{n} \sum_{k=1}^n (Y_k - \mathbb{E}Y_k) \rightarrow 0$ 几乎处处成立.

显然, 我们只需要证明上述 3 个断言即可.

注意到 $\mathbb{E}Y_n = \mathbb{E}X_1 \cdot 1_{\{|X_1| \leq n\}} \rightarrow \mathbb{E}X_1 = 0$, 断言 1 显然成立.

由于 $\mathbb{P}(X_n - Y_n \neq 0) = \mathbb{P}(|X_n| > n) = \mathbb{P}(|X_1| > n)$, 而 $\mathbb{E}|X_1| < \infty$ 保证了 $\sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) < \infty$. 由 B-C 引理, $\mathbb{P}(X_n - Y_n \neq 0 \text{ i.o.}) = 0$, 即对于几乎处处的样本点, 序列 $\{X_n - Y_n\}_{n=1}^{\infty}$ 中非零元素只有有限个. 从而必定有断言 2 成立.

下面证明断言 3 成立. 我们计算

$$\begin{aligned} \sum_n \frac{\text{Var}(Y_n - \mathbb{E}Y_n)}{n^2} &\leq \sum_n \frac{\mathbb{E}Y_n^2}{n^2} = \sum_n \frac{\mathbb{E}X_1^2 \cdot 1_{\{|X_1| \leq n\}}}{n^2} \\ &= \mathbb{E}[X_1^2 \cdot \sum_{n \geq |X_1| \vee 1} \frac{1}{n^2}] \\ &\leq \mathbb{E}[X_1^2 \cdot \frac{2}{|X_1| \vee 1}] \leq 2\mathbb{E}|X_1| < \infty. \end{aligned}$$

由引理 10.4.4 及引理 10.4.2, 断言 3 成立. \square

在 Kolmogorov 本人证明其强大数律的过程中, 他是利用 Kronecker 引理结合其三级数定理来完成证明的. 著名的华人概率学家钟开莱给出了一个更为初等的证明, 感兴趣的读者可以参考其著作 [9]. 读者也可以尝试按照如下路径 (不使用引理 10.4.4) 证明断言 3: 记 $X_n := Y_n - \mathbb{E}Y_n$,

$\tilde{S}_n := \tilde{X}_1 + \cdots + \tilde{X}_n$, 基于估计 $\sum_n \frac{\mathbb{E}[\tilde{X}_n^2]}{n^2} < \infty$, 首先论证 $\frac{\tilde{S}_{2^n}}{2^n} \rightarrow 0$ 几乎处处成立; 之后借助 Kolmogorov 不等式论证 $\max_{k \in (2^n, 2^{n+1}]} \frac{|\tilde{S}_k - \tilde{S}_{2^n}|}{k} \rightarrow 0$ 几乎处处成立; 最后在前两步基础上就能论证 $\frac{\tilde{S}_n}{n} \rightarrow 0$ 几乎处处成立。

某种程度上, 我们有 Kolmogorov 强大数律的如下逆命题。

定理 10.4.11. 设 $\{X_n\}_1^\infty$ 独立同分布, 如果 $\mathbb{E}|X_1| = \infty$, 那么对于任意数列 $\{b_n\}_1^\infty$, $\overline{\lim}_{n \rightarrow \infty} \frac{|S_n - b_n|}{n} = \infty$ 几乎处处成立。

证明. 令 $A_k := \{\frac{|X_k|}{k} \geq a\}$, 其中 $a > 0, k \geq 1$ 。由于 $\mathbb{E}|X_1| = \infty$, 我们知道 $\sum_k \mathbb{P}(A_k) = \infty$, 进而由 B-C 第二引理, $\mathbb{P}(\frac{|X_n|}{n} \geq a \text{ i. o. }) = 1$ 对任意 $a > 0$ 成立。于是 $\{\frac{|X_n|}{n}\}_1^\infty$ 几乎处处无界。但 $\frac{S_n}{n} - \frac{n-1}{n} \cdot \frac{S_{n-1}}{n-1} = \frac{X_n}{n}$, 因此必有随机数列 $\{\frac{S_n}{n}\}_1^\infty$ 几乎处处为无界的。这相当于对于 $b_n = 0$ 的情况证明了定理结论。

对于一般情况, 记 X_n^s 为 X_n 的对称化, 于是 $\mathbb{E}|X_1^s| = \infty$, 进而 $\{\frac{S_n^s}{n}\}_1^\infty$ 几乎处处为无界的随机数列。但 S_n^s 可以由 $S_n - b_n$ 对称化得到, 因此随机数列 $\{\frac{S_n - b_n}{n}\}_1^\infty$ 有界的概率为 0。□

在 $\{X_n\}_{n=1}^\infty$ 同分布, 而独立性减弱的前提下, 也有一些最优的结果。称随机变量 X, Y 是正象限相关的, 如果

$$\mathbb{P}(X \geq x, Y \geq y) \geq \mathbb{P}(X \geq x) \cdot \mathbb{P}(Y \geq y), \forall x, y;$$

称它们是负象限相关的, 如果

$$\mathbb{P}(X \geq x, Y \geq y) \leq \mathbb{P}(X \geq x) \cdot \mathbb{P}(Y \geq y), \forall x, y.$$

定理 10.4.12. (见 [16] 及 [27]。) 设 $\{X_n\}_1^\infty$ 同分布, 并且成对独立 (或成对负象限相关)。则存在随机变量 μ , 使得 $\lim_{n \rightarrow \infty} \overline{X}_n = \mu$ a.s. 成立的充分必要条件是: $\mathbb{E}|X_1| < \infty$ 。此时, $\mu = \mathbb{E}X_1$ 是常数。

证明. 我们只对“成对独立”的情形移植 [16] 中给出的证明, 另一情形的证明请参见 [27]。

必要性部分的证明与 Kolmogorov 强大数律完全类似。此处从略。

充分性。由于此时 $\{X_n^+ : n \geq 1\}$ 与 $\{X_n^- : n \geq 1\}$ 都满足定理假设条件, 且 $X_n = X_n^+ - X_n^-$, 故不妨直接假设 $X_n \geq 0$ 。

令 $Y_n := X_n \cdot 1_{\{X_n \leq n\}}$, 并记 $S_n^* = Y_1 + \cdots + Y_n$ 。现任取 $\varepsilon > 0$ 及 $\alpha > 1$,

令 $n_k := \lfloor \alpha^k \rfloor$, 利用 Chebyshev 不等式计算:

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{n_k}^* - \mathbb{E}S_{n_k}^*}{n_k}\right| \geq \varepsilon\right) &\leq \sum_{k=1}^{\infty} \frac{\text{Var}(S_{n_k}^*)}{\varepsilon^2 \cdot n_k^2} = \sum_{k=1}^{\infty} \frac{1}{\varepsilon^2 \cdot n_k^2} \sum_{i=1}^{n_k} \text{Var}(Y_i) \\ &\leq O(1) \cdot \sum_{i=1}^{\infty} \frac{\mathbb{E}Y_i^2}{i^2} = O(1) \cdot \mathbb{E}\left[X_1^2 \cdot \sum_{i \geq X_1} \frac{1}{i^2}\right] \\ &\leq O(1) \cdot \mathbb{E}\left[X_1^2 \cdot \frac{1}{1 \vee X_1}\right] \leq O(1) \cdot \mathbb{E}X_1 < \infty. \end{aligned}$$

由此 $\lim_{n \rightarrow \infty} \frac{S_{n_k}^* - \mathbb{E}S_{n_k}^*}{n_k} = 0$ a.s.。但 $\mathbb{E}Y_n \rightarrow \mathbb{E}X_1$, 从而 $\frac{\mathbb{E}S_n^*}{n} \rightarrow \mathbb{E}X_1$ 。因此

$$\lim_{n \rightarrow \infty} \frac{S_{n_k}^*}{n_k} = \mathbb{E}X_1 \text{ a.s.}$$

另外, $\sum_n \mathbb{P}(X_n \neq Y_n) = \sum_n \mathbb{P}(X_1 > n) \leq \mathbb{E}X_1 < \infty$, 由 B-C 引理,

$\mathbb{P}(X_n \neq Y_n \text{ f.o.}) = 1$ 。于是 $\lim_{k \rightarrow \infty} \frac{S_{n_k}}{n_k} = \mathbb{E}X_1$ 。

利用 S_n 的单调性, 我们可以证明: 几乎处处有

$$\frac{1}{\alpha} \cdot \mathbb{E}X_1 \leq \liminf_n \frac{S_n}{n} \leq \limsup_n \frac{S_n}{n} \leq \alpha \cdot \mathbb{E}X_1.$$

由 $\alpha > 1$ 的任意性。上下极限必定几乎处处相等, 且等于 $\mathbb{E}X_1$ 。 \square

10.5 应用举例

这里, 我们基于 Borel 强大数律, 给出一个极限与数学期望不能交换的例子。

例 10.3. 设 $\{\xi_n\}_1^\infty$ 独立同分布, 满足 $\mathbb{P}(\xi_1 = \frac{2}{3}) = \frac{2}{3}, \mathbb{P}(\xi_1 = -\frac{2}{3}) = \frac{1}{3}$, 从而 $\mathbb{E}\xi_1 = \frac{2}{9} > 0$; 令 $X_n := 1 + \xi_n$ 。考虑乘积 $M_n := X_1 \cdots X_n$, 有 $\mathbb{E}[M_n] = (\frac{11}{9})^n$ 。但是 $\mathbb{E}[(\log X_1)^4] < \infty$ 且

$$\mathbb{E}[\log X_1] = \frac{2}{3} \log \frac{5}{3} + \frac{1}{3} \log \frac{1}{3} = \frac{1}{3} \log \frac{25}{27} < 0.$$

根据 Borel 强大数律, 应有 $M_n = \exp\{\sum_{k=1}^n \log X_k\} \xrightarrow{\text{a.s.}} 0$ 。因此此时

$$\mathbb{E}[\lim_{n \rightarrow \infty} M_n] = 0 \neq \infty = \lim_{n \rightarrow \infty} \mathbb{E}[M_n]. \quad \square$$

对于上例, 有人把它解释如下: 假定某人每次投资都有 $\frac{2}{3}$ 的概率获得 $\frac{2}{3}$ 的利润率, $\frac{1}{3}$ 的概率获得 $-\frac{1}{3}$ 的利润率 (即亏损 $\frac{1}{3}$ 的本金); 那么平均而言, 他单次投资的期望利润率为 $\frac{2}{9} > 0$ 。如果他每次都使用全部资金进行投资, 看似平均而言单次的投资都应该平均角度赚钱的, 然而长期而言他必然会破产。如果单次投资的收益率同分布于 ξ (其中 $\xi > -1$), 且各次投资之间相互独立, 那么真正高明的专业投资者应当做到如下两个要求

$$\mathbb{E}\xi > 0, \mathbb{E}[\log(1 + \xi)] > 0.$$

只有这样他才能真正做到通过复利的方式长期、稳定地获利。当然要注意，例 10.3 中研究的投资方式是“*All-In*”的梭哈式赌博，在现实生活中应该极力避免；在风险资产的投资与管理领域，人们通常提议通过合理的仓位管理的模式来避免大亏损、积累单次的小盈利，以期做到资产的稳定增值。

例 10.4.（赠券收集问题） 任给 $n \geq 2$ ，设 $\{X_k^{(n)}\}_{k=1}^\infty i.i.d.$ ，共同分布是 $\{1, 2, \dots, n\}$ 上均匀分布。记 $\tau_k^n := \inf\{m : \#\{X_j^{(n)} : 1 \leq j \leq m\} = k\}$ ，并记 $T_n := \tau_n^n$ 。在现实生活中，商家在客户消费满一定额度（比如 100 元）后赠送一张奖券（假设一套共 n 张），集齐一套就可以换取特定奖品，因而我们关心 $\mathbb{E}T_n$ 的大小。这个问题也可以解释成 *Polyá* 坛子模型：假定在坛子里面有 n 个编号不同的球。每次有放回的摸取一球，并记录下编号。 T_n 就是首次全部编号记录齐全时的取球次数。现在的问题是： $\mathbb{E}T_n = ?$ 当 n 充分大时，我们能对 T_n 本身的大小规模说些什么？

参考解答. 令 $Y_{n,k} := \tau_k^n - \tau_{k-1}^n$ ，约定 $\tau_0^n := 0$ ； $Y_{n,k}$ 表示在已获得 $k-1$ 个编号的基础上，要获得新编号额外所需的取球次数。显然它与 $\{Y_{n,j}\}_{j=1}^{k-1}$ 相互独立，并且 $Y_{n,k} \sim \text{Geo}(1 - \frac{k-1}{n})$ 。于是 $T_n = Y_{n,1} + Y_{n,2} + \dots + Y_{n,n}$ 。

注意到当 $Y \sim \text{Geo}(p)$ 时， $\mathbb{E}Y = \frac{1}{p}$ ， $\text{Var}(Y) = \frac{1-p}{p^2}$ ，我们有

$$\begin{aligned}\mathbb{E}T_n &= \sum_{k=1}^n (1 - \frac{k-1}{n})^{-1} = n \sum_{k=1}^n \frac{1}{k} \sim n \log n, \\ \text{Var}(T_n) &= \sum_{k=1}^n (1 - \frac{k-1}{n})^{-2} \frac{k-1}{n} \leq n^2 \sum_{k=1}^\infty \frac{1}{k^2} = Cn^2.\end{aligned}$$

于是 $\text{Var}(\frac{T_n}{\mathbb{E}T_n}) \leq \frac{C}{(\log n)^2} \rightarrow 0$ ，这表明 $\frac{T_n}{\mathbb{E}T_n} \xrightarrow{\mathbb{P}} 1$ ，亦即： $\frac{T_n}{n \log n}$ 依概率收敛到 1。□

在上面例子中，敏锐的同学或许会联系起推论 10.4.1 或推论 10.4.2 试图得到几乎处处收敛。如果随机变量列 $\{T_n\}_{n=1}^\infty$ 具有单调上升性质，这个企图就能实现。但不幸的是， $\{T_n\}_{n=1}^\infty$ 某种意义上可以认为定义在不同的概率空间中，一般来说无法在几乎处处意义下比较 T_n 与 T_{n+1} 的大小关系，因而无法单纯利用此处的方差分析的方法推断几乎处处收敛。进一步敏锐的同学也许会想到后文将提及的中心极限定理（见第 8 章的 *Lindeberg-Feller* 中心极限定理），但同样遗憾的是，此处 *Lindeberg* 条件不成立；有兴趣的同学可以去验证这一点。

为了介绍下一个例子，我们回忆一下超几何分布 $H(b, r; n)$ ，它可以用 *Polyá* 的罐子模型来解释：在有 b 个黑球、 r 个红球的罐子里无放回的取 $n \leq b+r$ 个球所获得的蓝球数量记作 X ，则 $X \sim H(b, r; n)$ ，概率分布列为

$$\mathbb{P}(X = k) = \frac{C_b^k C_r^{n-k}}{C_{b+r}^n}, \ell_1 := (n-r)^+ \leq k \leq \ell_2 := b \wedge n. \quad (10.10)$$

不难通过组合数的技巧论证出：

$$\mathbb{E}X = \frac{nb}{b+r}, \quad \mathbb{E}[X(X-1)] = \frac{n(n-1)b(b-1)}{(b+r)(b+r-1)}. \quad (10.11)$$

由此进一步算出

$$\text{Var}(X) = \frac{nbr(b+r-n)}{(b+r)^2(b+r-1)}. \quad (10.12)$$

例 10.5. (估算池塘里的鱼) 在例 5.11, 考虑了池塘里的鱼数的估计问题呢。估计方法如下: 首先从池塘中捕捞出 M 条成鱼, 都做好标记, 并重新放回池塘。过几天后 (假定期间成鱼数量不改变, 且之前作的标记不会脱落), 再次进行捕捞, 捞出了 n 条成鱼, 其中有 $X = m$ 条是之前做过标记的成鱼。那么池塘里的鱼数 N 的极大似然估计为

$$\hat{N} := \frac{nM}{X}.$$

可以证明, 它也可以认为是矩估计, 因为 $\mathbb{E}X = \frac{nM}{N}$ 。

历史上, 上述估计方法最早是 Laplace 在 1786 年为了估计法国人口时提出的。从逻辑上来说, 这个估计方法很巧妙。我们这里的目的是想进一步了解这个估计的效果, 以及得到好的估计效果时参数 M, n 应当满足的要求。

仅供参考的一个解答. 我们定义 $Y := \frac{N}{\hat{N}} = \frac{NX}{nM}$, 它是真实值与估计值的比值。如果这个比值靠近 1, 我们就认为估计效果好。根据关于超几何分布的理论, 我们得到 $\mathbb{E}Y = 1$ 以及

$$\text{Var}(Y) = \frac{(N-n)(N-M)}{nM(N-1)} = \frac{(1-\frac{n}{N})(1-\frac{M}{N})}{1-\frac{1}{N}} \cdot \frac{N}{nM}.$$

直观上, 我们在做这个估计方案时, 应该取 $M \ll N, n \ll N$, 否则从成本等方面来说就不合适。但为了得到好的估计效果 (确切说, $\text{Var}(Y) \rightarrow 0$, 从而 $Y \xrightarrow{\mathbb{P}} 1$), 我们应该要求 $N \ll n \cdot M$ 。

总结以上论述, 为了得到好的估计效果, 我们对估计方案中参数 M, n 的安排应该是: $M \ll N, n \ll N$, 但 $n \cdot M \gg N$ 。□

对于上面的例子, 读者也可以尝试直接计算 \hat{N} 的数学期望与方差来进行有关讨论。

习 题 10

习题 10.1. 证明: 当 $n \rightarrow \infty$ 时, 数列 $\{a_n\}_{n=1}^{\infty}$ 以 a 为极限当且仅当:

$$\sum_{n=1}^{\infty} 1_{\{|a_n - a| \geq \varepsilon\}} < \infty, \forall \varepsilon > 0.$$

这是概率论中能利用本章的 B-C 引理证明几乎处处收敛的原因。请类似书写 $\overline{\lim}_n a_n \leq a$, $\overline{\lim}_n a_n \geq a$, $\underline{\lim}_n a_n \leq a$, $\underline{\lim}_n a_n \geq a$ 的等价刻画。

习题 10.2. 证明随机变量列 $X_n \xrightarrow{\mathbb{P}} 0$ 当且仅当 $\mathbb{E}[\frac{|X_n|}{1+|X_n|}] \rightarrow 0$ 。

习题 10.3. 当 $\xi_n \xrightarrow{\mathbb{P}} \xi$ 时, 我们记为 $\xi_n = \xi + o_{\mathbb{P}}(1)$; 当任给 $\varepsilon > 0$ 总有 $M > 0$ 使得 $\sup_n \mathbb{P}(|\xi_n| \geq M) \leq \varepsilon$, 则记作 $\xi_n = O_{\mathbb{P}}(1)$ 。试证明:

(i) $o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$;

(ii) $o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$, $O_{\mathbb{P}}(1) + O_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$;

(iii) $o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$, $O_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$;

(iv) 设 $\xi_n \xrightarrow{d} \xi$, 那么 $\xi_n = O_{\mathbb{P}}(1)$;

(v) 对于连续函数 f , 总有 $f(\xi + o_{\mathbb{P}}(1)) = f(\xi) + o_{\mathbb{P}}(1)$ 。

习题 10.4. 设 X 为非负非退化随机变量。求证: $\lim_{x \rightarrow \infty} x \mathbb{E}[\frac{1}{X} \cdot 1_{\{X > x\}}] = 0$ 。

习题 10.5. 试证明: 当存在某 $p > 1$ 使得 $\sup\{\mathbb{E}[|X_{\alpha}|^p] : \alpha \in I\} < \infty$ 时, 随机变量族 $\{X_{\alpha}\}_{\alpha \in I}$ 是一致可积的。

习题 10.6. 给出定理 10.3.3 的证明。它本质上是上一习题的推广。

习题 10.7. 设 $\{X_n\}_{n=1}^{\infty}$ 是一列非负的单调递减的随机变量列, $X_n \rightarrow 0$ 依概率收敛, 那么 $X_n \rightarrow 0$ 几乎处处收敛。

习题 10.8. 设 $\{U_k\}_{k=1}^n$ 为 *i.i.d.* 的 $U(0, 1)$ 分布样本; 假设它们从小到大排列为:

$$U_n^{(1)} < U_n^{(2)} < \dots < U_n^{(n)}.$$

试证明: $U_n^{(1)} \rightarrow 0$ 几乎处处成立, 且 $nU_n^{(1)} \xrightarrow{d} \mathcal{E}(1)$ 。

习题 10.9. 沿用习题 10.8 中假定与记号。对于固定的正整数 m , 类似讨论 $U_n^{(m)}$ 及 $nU_n^{(m)}$ 的极限行为。

习题 10.10. 沿用习题 10.8 中假定与记号。设正整数 K_n 满足 $1 \leq K_n \leq n$ 且 $K_n \rightarrow \infty$, 求证: $Y_n := nU_n^{(K_n)} \rightarrow \infty$ 依概率成立, 即

$$\lim_{n \rightarrow \infty} \mathbb{P}(Y_n \geq M) = 1, \forall M \in \mathbb{R}.$$

习题 10.11. 设 $\{U_n\}_{n=1}^{\infty}$ 是一列 *i.i.d.* 的服从 $U(0, 1)$ 分布的随机变量列, 设 Y_n 为 U_1, \dots, U_{2n+1} 中的样本中位数。求证: $Y_n \rightarrow \frac{1}{2}$ 依概率收敛。【提示: 计算 $\mathbb{E}[(Y_n - \frac{1}{2})^2]$ 。通过计算 $\mathbb{E}[(Y_n - \frac{1}{2})^4]$ 可以证明这个收敛还是几乎处处成立的。】

习题 10.12. 沿用习题 10.8 中假定与记号。给定 $p \in (0, 1)$, 如果正整数列 $\{K_n\}$ 满足 $1 \leq K_n \leq n$ 且 $\frac{K_n}{n} \rightarrow p$, 求证: $U_n^{(K_n)} \rightarrow p$ 几乎处处成立。

习题 10.13. 设 $\{X_k\}_{k=1}^n$ 为 *i.i.d.* 的标准指数分布样本。 $X_{(n)}$ 是这些样本中的最大值。试证明: $X_{(n)} \rightarrow \infty$ 几乎处处成立, 且 $X_{(n)} - \log n$ 依分布收敛, 极限的分布函数为 $F(x) = e^{-e^{-x}}$ (这是第一类极值分布, 也称为 *Gumbel* 分布)。由此容易知道 $\frac{X_{(n)}}{\log n} \rightarrow 1$ 依概率收敛。

习题 10.14. 设 $X_n \sim B(n, p_n), n = 1, 2, \dots$ 。如果 $n \cdot p_n \rightarrow \lambda > 0$, 试论证:

$$X_n \xrightarrow{d} \text{Poisson}(\lambda).$$

习题 10.15. 考虑 n 个人戴着 n 顶外观相同、帽内带标记区分的帽子参加聚会; 入场脱帽、离场取帽, 假定帽子均匀混合, 取帽时随机抓取。记 N_n 为离场时拿对了帽子的实际人数。求证: $N_n \xrightarrow{d} \text{Poisson}(1)$ 。

习题 10.16. 设 $\{X_n\}_{n=1}^\infty$ 是一列非负的随机变量列, 正数列 $a_n \uparrow \infty$ 。那么几乎处处意义下,

$$\lim_{n \rightarrow \infty} \frac{X_n}{a_n} = 0 \Leftrightarrow \lim_{n \rightarrow \infty} \frac{\max\{X_1, \dots, X_n\}}{a_n} = 0.$$

习题 10.17. 设 $\{X_n\}_{n=1}^\infty$ 是一列非负的随机变量列, 正数列 $a_n \uparrow \infty$ 。那么几乎处处意义下,

$$\overline{\lim}_{n \rightarrow \infty} \frac{X_n}{a_n} = 1 \Leftrightarrow \overline{\lim}_{n \rightarrow \infty} \frac{\max\{X_1, \dots, X_n\}}{a_n} = 1.$$

习题 10.18. 设 $\{X_n\}_1^\infty$ 独立同分布, 共同分布为参数为 $p \in (0, 1)$ 的几何分布, 即 $\mathbb{P}(X_1 = n) = pq^{n-1}, n \geq 1$, 其中 $q := 1 - p$ 。证明:

$$\lim_n \frac{\max\{X_k : 1 \leq k \leq n\}}{-\log_q n} = 1$$

几乎处处成立。

习题 10.19. 对于 *i.i.d.* 的标准指数分布随机变量 $\{X_n\}_1^\infty$, 请按下面步骤证明

$$\lim_{n \rightarrow \infty} \frac{\max\{X_1, \dots, X_n\}}{\log n} = 1.$$

(1) 先使用 *B-C* 第一、第二引理论证 $\overline{\lim}_{n \rightarrow \infty} \frac{X_n}{\log n} = 1$;

(2) 根据习题 10.17, 这等价于论证了

$$\overline{\lim}_{n \rightarrow \infty} \frac{\max\{X_1, \dots, X_n\}}{\log n} = 1.$$

再使用 *B-C* 第一引理论证

$$\underline{\lim}_{n \rightarrow \infty} \frac{\max\{X_1, \dots, X_n\}}{\log n} \geq 1.$$

二者结合就完成了结论的证明。请完成有关论证的技术细节。

习题 10.20. 给定 *i.i.d.* 标准 *Cauchy* 分布的随机变量列 $\{X_n\}_1^\infty$ 及正实数序列 $a_n \nearrow \infty$ 。那么

$$\begin{aligned} \overline{\lim}_{n \rightarrow \infty} \frac{|X_n|}{a_n} = 0, \text{ a.s.} &\Leftrightarrow \sum_{n=1}^{\infty} \frac{1}{a_n} < \infty, \\ \overline{\lim}_{n \rightarrow \infty} \frac{|X_n|}{a_n} = \infty, \text{ a.s.} &\Leftrightarrow \sum_{n=1}^{\infty} \frac{1}{a_n} = \infty. \end{aligned}$$

因此, $\overline{\lim}_{n \rightarrow \infty} \frac{|X_n|}{a_n} \in \{0, \infty\}$, 不可能出现 $\overline{\lim}_{n \rightarrow \infty} \frac{|X_n|}{a_n} = 1$ 。

【注: 在厚尾分布中, 具有类似现象的例子还有很多。】

习题 10.21. 沿用习题 10.8 中假定与记号。设 $a_n \nearrow \infty$ 。求证:

$$\underline{\lim}_{n \rightarrow \infty} a_n U_n^{(1)} = \infty, \text{ a.s.} \Leftrightarrow \sum_{n=1}^{\infty} \frac{1}{a_n} < \infty,$$

并且,

$$\lim_{n \rightarrow \infty} a_n U_n^{(1)} = 0, a.s. \Leftrightarrow \sum_{n=1}^{\infty} \frac{1}{a_n} = \infty.$$

因此, $\lim_{n \rightarrow \infty} [a_n U_n^{(1)}] \in \{0, \infty\}$, 不可能出现 $\lim_{n \rightarrow \infty} [a_n U_n^{(1)}] = 1$ 。

【注: 本题与上一题有对偶的关系。】

习题 10.22. 证明定理 10.4.2 在 (3) 以外的其他条件下的结论。

习题 10.23. 设 X, Y 为两随机变量, 均值为 0, 方差为 1, 协方差 ρ 。试证明:

$$\mathbb{E}[\max(X^2, Y^2)] \leq 1 + \sqrt{1 - \rho^2}.$$

【提示: 利用 $\max(a, b) = \frac{a+b+|a-b|}{2}$ 及 Cauchy 不等式。】

习题 10.24. (1) 给定 $a > 0, p \in (0, 1)$ 。求证:

$$\sup\left\{\frac{a^2 \mathbb{P}(|X| \geq a)}{\mathbb{E}[X^2]} : \mathbb{E}[X^2] = a^2 \cdot p\right\} = 1.$$

也就是说, 在满足 $\mathbb{E}[X^2] = a^2 \cdot p$ 的随机变量 X 中, Chebyshev 不等式 $\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[X^2]}{a^2}$ 能达到最佳界。【提示: 考虑两点分布。】

(2) 设 $0 < \mathbb{E}[X^2] < \infty$ 。当 $a \rightarrow \infty$ 时, $\frac{a^2 \mathbb{P}(|X| \geq a)}{\mathbb{E}[X^2]} \rightarrow 0$ 。

习题 10.25. 设 $a > 0$ 。求证:

$$\inf\{\mathbb{P}(|X| > a) : \mathbb{E}X = 0, \text{Var}(X) = 1\} = 0.$$

这表明在 $\mathbb{E}X = 0, \text{Var}(X) = 1$ 条件下, $\mathbb{P}(|X| > a)$ 没有正下界。

习题 10.26. 设 $a \geq 1, \sigma^2 > 0$ 。求证:

$$\inf\{\mathbb{P}(|X| > a) : \mathbb{E}X = 1, \text{Var}(X) = \sigma^2\} = 0.$$

这表明在 $\mathbb{E}X = 1, \text{Var}(X) = \sigma^2$ 条件下, $\mathbb{P}(|X| > a)$ 没有正下界。

习题 10.27. 设随机变量 X 满足: $\mathbb{E}X = 0, \text{Var}(X) = 1$ 。证明: 对任意 $x > 0$,

$$\mathbb{P}(X \geq x) \leq \frac{1}{1 + x^2}.$$

上述不等式无法改进: 取两点分布的随机变量 Y 满足

$$\mathbb{P}(Y = -\frac{1}{x}) = \frac{x^2}{1 + x^2}, \quad \mathbb{P}(Y = x) = \frac{1}{1 + x^2}.$$

则显然有 $\mathbb{E}Y = 0, \mathbb{E}Y^2 = 1$ 。

习题 10.28. 设随机变量 ξ 满足 $\mathbb{E}\xi > 0, \mathbb{E}\xi^2 < \infty, \beta \in (0, 1)$ 。求证:

$$\mathbb{P}(\xi > \beta \mathbb{E}\xi) \geq \frac{(1 - \beta)^2 (\mathbb{E}\xi)^2}{(1 - \beta)^2 (\mathbb{E}\xi)^2 + \text{Var}(\xi)}.$$

【提示: 利用不等式 $1_{\{x > 0\}} \geq x(2 - x)$, 取 $x = (\xi - \beta \mathbb{E}\xi)/a$, 寻找最佳的正常数 a 。也可直接利用习题 10.27 的结论。】

习题 10.29. (Kochen-Stone Lemma) 假设 $\sum_n \mathbb{P}(A_n) = \infty$ 。求证:

$$\overline{\lim}_{n \rightarrow \infty} \frac{[\sum_{k=1}^n \mathbb{P}(A_k)]^2}{\sum_{1 \leq i, j \leq n} \mathbb{P}(A_i \cap A_j)} = \alpha > 0$$

蕴含了 $\mathbb{P}(\{A_n\} i.o.) \geq \alpha$ 。【提示: 利用上一题结论。】

习题 10.30. 利用分析方法证明下面不等式 (参见 [13, Theorem 1.4]) *:

$$\frac{a \cdot e^{-\frac{a^2}{2}}}{1+a^2} \leq \int_a^\infty e^{-x^2/2} dx \leq \frac{e^{-\frac{a^2}{2}}}{a}, \quad \forall a > 0.$$

这给出了标准正态随机变量 $Z \sim N(0, 1)$ 的尾部概率的精细估计:

$$\frac{a \cdot e^{-\frac{a^2}{2}}}{\sqrt{2\pi}(1+a^2)} \leq \mathbb{P}(Z \geq a) = 1 - \Phi(a) \leq \frac{e^{-\frac{a^2}{2}}}{\sqrt{2\pi}a}, \quad \forall a > 0.$$

习题 10.31. 设 $X = (X_1, \dots, X_n)$ 为 n 维平方可积的随机向量, 均值为 μ , 协方差矩阵为 Σ . 设 Σ 非退化. 证明: 对于 \mathbb{R}^n 中任意的紧凸集 S ,

$$\mathbb{P}(X \in S) \leq \frac{1}{1 + \inf\{(x - \mu)\Sigma^{-1}(x - \mu)^T : x \in S\}}.$$

【提示: 先看一维情形, 利用习题 10.27 的结论。】

习题 10.32. 设非负随机变量 ξ 具有正的二阶矩. 求证: $\mathbb{P}(\xi = 0) \leq \frac{\text{Var}(\xi)}{\mathbb{E}(\xi^2)}$.

习题 10.33. 证明定理 10.3.7.

习题 10.34. 证明定理 10.3.8.

习题 10.35. 设 $X_n \sim B(1, p_n)$ 相互独立, 令 $S_n := X_1 + \dots + X_n$. 求证:

(1) 当 $\sum_n p_n < \infty$ 时, $S_\infty < \infty$;

(2) 当 $\sum_n p_n = \infty$ 时, $S_\infty = \infty$, 更准确的说, $\lim_{n \rightarrow \infty} \frac{S_n}{\mathbb{E}S_n} = 1$ a.e.

习题 10.36. 设 $X_n \sim \text{Poisson}(\lambda_n)$ 相互独立. 记 $S_n := X_1 + \dots + X_n$. 试证:

(1) 当 $\sum_n \lambda_n < \infty$ 时, $S_\infty < \infty$;

(2) 当 $\sum_n \lambda_n = \infty$ 时, $S_\infty = \infty$;

(3) 当 $\sum_n \lambda_n = \infty$ 且 $\sup_n \lambda_n < \infty$ 时, $\lim_n \frac{S_n}{\mathbb{E}S_n} = 1$ 几乎处处成立.

习题 10.37. 设 $X_n \sim \mathcal{E}(\lambda_n)$ 相互独立. 记 $S_n := X_1 + \dots + X_n$. 试证:

(1) 当 $\sum_n \frac{1}{\lambda_n} < \infty$ 时, $S_\infty < \infty$;

(2) 当 $\sum_n \frac{1}{\lambda_n} = \infty$ 时, $S_\infty = \infty$;

(3) 当 $\sum_n \frac{1}{\lambda_n} = \infty$ 且 $\sup_n \frac{1}{\lambda_n} < \infty$ 时, $\lim_n \frac{S_n}{\mathbb{E}S_n} = 1$ 几乎处处成立.

习题 10.38. 设 $\alpha > 0, \beta \geq 0$. 试证明:

$$\lim_{x \rightarrow \infty} x^{\alpha+\beta} \cdot \mathbb{P}(|X| > x) = 0 \Leftrightarrow \lim_{x \rightarrow \infty} x^\alpha \cdot \mathbb{E}[|X|^\beta \cdot 1_{\{|X|>x\}}] = 0.$$

对于必要性, $\alpha > 0$ 是必须的.

*此处, 我们基于 [13, Theorem 1.4] 的证明思想, 对不等式的左端做了改进. 请读者自行对比. 估计式右端因子 $\frac{1}{a}$ 可以替换为 $\frac{2}{a+\sqrt{1+a^2}}$.

习题 10.39. (控制收敛定理的推广) 设有可积的随机变量序列 $\{X_n\}_{n=1}^{\infty}, \{Y_n\}_{n=1}^{\infty}$ 及可积的随机变量 X, Y 。假设 $X_n \rightarrow X, Y_n \rightarrow Y$ 几乎处处收敛 (或依概率收敛, 或更弱的依分布收敛), 并且 $|X_n| \leq Y_n, \mathbb{E}Y_n \rightarrow \mathbb{E}Y$ 。试证明: $\mathbb{E}X_n \rightarrow \mathbb{E}X$ 。【提示: *Fatou* 引理。】

习题 10.40. 设 $X_n \xrightarrow{d} X$ 。又设连续函数 g, h 满足

(i) $g \geq 0$, 且存在 $M > 0$ 使得 $g(x) > 0, \forall |x| \geq M$;

(ii) $\lim_{|x| \rightarrow \infty} \frac{h(x)}{g(x)} = 0$;

(iii) $\sup_n \mathbb{E}g(X_n) < \infty$ 。

那么 $\mathbb{E}h(X_n) \rightarrow \mathbb{E}h(X)$ 。【本题来源: [13, Theorem 3.8]。】

习题 10.41. 设 $X_n \rightarrow X$ 依概率收敛, g 为连续函数, 那么 $g(X_n) \rightarrow g(X)$ 依概率收敛。

习题 10.42. 依概率收敛是可以度量化。对于随机变量 X, Y , 定义 *Ky Fan* 度量

$$\alpha(X, Y) := \inf\{\delta \geq 0 : \mathbb{P}(|X - Y| > \delta) \leq \delta\}.$$

试证明: $\alpha(\cdot, \cdot)$ 是度量, 且诱导了依概率收敛。

习题 10.43. 设 $\mathbb{E}X_1^- < \infty$ 且 $X_n \nearrow X$ 。求证: $\mathbb{E}X_n \nearrow \mathbb{E}X$ 。

习题 10.44. 设 $S_n \sim B(n, p)$, 其中 $p \in (0, 1)$ 。任给 $a \in (p, 1)$, 记

$$r := r(p, a) = \left(\frac{p}{a}\right)^a \cdot \left(\frac{1-p}{1-a}\right)^{1-a},$$

求证: $0 < r < 1$, 且 $\mathbb{P}\left(\frac{S_n}{n} \geq a\right) \leq r^n$ 。

习题 10.45. 设 $\{X_n\}_{n=1}^{\infty}$ 是一列 *i.i.d.* 的离散型随机变量。求证: $\frac{R_n}{n} \rightarrow 0$ 几乎处处成立。更强的结论是:

$$\lim_{n \rightarrow \infty} \frac{R_n}{\mathbb{E}R_n} = 1, \quad \lim_{n \rightarrow \infty} \frac{\mathbb{E}R_n}{n} = 0.$$

习题 10.46. 基于例 10.1 中估计, 证明: 如果 $\{X_n\}_{n=1}^{\infty} \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, 则 $S_n := X_1 + \cdots + X_n$ 满足

$$\frac{S_n - n\lambda}{n^\alpha} \xrightarrow{\text{a.s.}} 0, \forall \alpha > \frac{1}{2}.$$

习题 10.47. 基于例 10.2 中估计, 证明: 如果 $\{X_n\}_{n=1}^{\infty} \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, 则 $S_n := X_1 + \cdots + X_n$ 满足

$$\frac{S_n - n\lambda}{n^\alpha} \xrightarrow{\text{a.s.}} 0, \forall \alpha > \frac{1}{2}.$$

习题 10.48. 本习题考虑例题 3.4 的推广。设有无穷大的空坛子以及带自然数标记的无穷多的球。给定一个严格单调的函数 $f: \mathbb{N} \rightarrow \mathbb{N}$ (从而 $f(n) \geq n$)。第一步时, 先在坛内放入标号 1 到 $f(1)$ 的球, 再从坛内随机的取一球; 一般的, 在第 $n \geq 2$ 步时, 在坛内放入标号 $f(n-1)+1$ 到 $f(n)$ 的球, 之后从坛内随机的取一球。记事件 E 为“坛子最终是空的”。求证:

$$\sum_{n=1}^{\infty} \frac{1}{f(n)} < \infty \implies \mathbb{P}(E) = 0,$$

$$\sum_{n=1}^{\infty} \frac{1}{f(n)} = \infty \implies \mathbb{P}(E) = 1.$$

§ 11

随机变量的特征函数与中心极限定理

在第 5 章我们就引入了分布函数的概念；本章将继续介绍特征函数的概念，之后介绍特征函数的一个经典且重要的应用：中心极限定理。我们的论述以一维情形为主，因为大部分场合高维的对应命题、对应性质等的表述是类似的。

11.1 特征函数与分布函数

特征函数的概念发源于 Laplace 的工作。后来 A. M. Lyapunov（李雅普诺夫，1857/6/6–1918/11/3；俄国）在 1900、1901 年各发表了一篇文章，首创使用特征函数方法，得到了关于独立和的第一个比较一般的中心极限定理（基于 Lyapunov 条件），由此特征函数方法开始了广泛的应用。但这个方法的奠基性工作—特征函数的连续性定理，是在这个方法已经使用较多年后才由 P. Lévy（莱维，1886/9/15–1971/12/15）完成严格论证的。



图 11.1: A. M. Lyapunov(1857-1918)

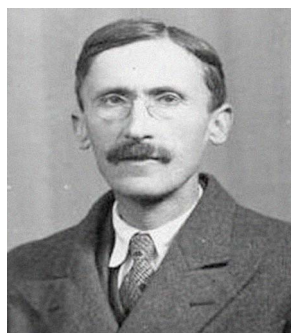


图 11.2: P. Lévy(1886-1971)

注 11.1. A. M. Lyapunov（李雅普诺夫，1857/6/6–1918/11/3；俄国）是 P. Chebyshev 的学生，以动力系统方向的稳定性理论知名，在数学物理和概率论中也有诸多贡献。他 1876 年入读圣彼得堡大学数学物理系，一个月后转系到数学系。他先是在当时的力学教授 D. K. Bobylev

的指导下完成第一个独立科学研究工作;之后成为 *Chebyshev* 的学生。1880 年因流体静力学方面的工作获得一枚金质奖章,并于同年毕业。他培养的学生有: *Nikola Saltikov* 和 *Vladimir Steklov* (1864/1/9–1926/5/30; 俄罗斯-苏联数学家、力学家、物理学家,以 *Poincaré-Steklov* 知名) 等。

法国数学家 *Paul Pierre Lévy* (莱维, 1886/9/15–1971/12/15) 生于巴黎、卒于巴黎,是 *Hadamard* (阿达马, 1865/12/8–1963/10/17; 法国) 和 *Vito Volterra* (伏特拉, 1860/5/3–1940/10/11; 意大利) 的学生。他 1904 年考入巴黎高等师范学校, 后在巴黎综合理工学院毕业, 1913 年在巴黎综合理工学院任教, 1920 年升任教授; 1950 年起在美国各大学任教, 1962 年返回法国, 1964 年 *Lévy* 在 78 岁时当选为巴黎科学院院士。*Lévy* 在泛函分析、函数论、拓扑学、力学等多方面都有贡献; 但他最主要的贡献在概率论研究, 他重新发现并完善了特征函数理论, 发展了中心极限定理定理, 提出“分布律”, 独创从样本函数角度研究随机过程, 对 *Brown* 运动和鞅的研究具有高度建设性, 其影响遍及整个概率论。

Lévy 培养的知名学生有: *Wolfgang Doeblin* (1915/3/17–1940/6/21; 德国-法国数学家, 犹太裔, 另一个指导老师是 *Maurice René Fréchet*; 马氏链的 *Doeblin* 比值极限定理就是以他的姓氏命名的), *Michel Loève* (1907/1/22–1979/2/17; 法国-美国概率学家、数理统计学家, 犹太裔; 以 *Karhunen-Loève* 定理知名), *Benoît Mandelbrot* (1924/11/20–2010/10/14; 波兰出生的法国-美国数学家, 在分形几何方向有突出贡献), *Georges Matheron* (1930/1/2–2000/8/7; 法国数学家、工程师, 与 *Jean Serra* 一起创立了数学形态学) 等。

给定分布函数 F , 它对应的特征函数 (记作 $\hat{\mu}_F$) 定义为

$$\hat{\mu}_F(\lambda) := \int \exp(i\lambda \cdot x) dF(x), \quad \lambda \in \mathbb{R},$$

其中 $i = \sqrt{-1}$ 为虚单位。设随机变量 $X \sim F$ 。上面定义也等价于

$$\hat{\mu}_F(\lambda) = \mathbb{E} \exp(i\lambda \cdot X).$$

对于高维随机变量或高维随机变量的联合分布函数, 也可类似定义它们的特征函数, 只是相应的变元 λ, x 应视作相应欧氏空间中的向量值变元, 而乘积 $\lambda \cdot x$ 视作相应的内积。调和分析中的 Fourier 变换 $\rho \mapsto \mathcal{F}(\rho)$ 定义为*:

$$\mathcal{F}(\rho)(\lambda) := \int \rho(x) \cdot e^{-i\lambda \cdot x} dx.$$

容易知道, 当分布 F 具有密度 ρ 时, 相应的特征函数恰好是:

$$\hat{\mu}_F(\lambda) = \hat{\rho}(\lambda) := \int \rho(x) \cdot e^{i\lambda \cdot x} dx, \text{ 如果 } dF(x) = \rho(x) dx.$$

对应于上面的符号, 则 $\hat{\rho}(\lambda) = \mathcal{F}(\rho)(-\lambda)$; 因此, 通过适当调整 Fourier 变换定义 (在本章将把 Fourier 变换定义为 $\mathcal{F}(\rho)(\lambda) := \hat{\rho}(\lambda)$), 此处的特征函数可以视作对应的密度函数的 Fourier 变换, 在更一般情形则可视作分布测度的“Fourier 变换”。

有关特征函数的基本性质, 我们提请读者参看课后习题 11.2、11.4、11.15 等习题中的结论。简而言之, 随机变量的特征函数具有如下一些分析性质: (1) 0 处的函数值为 1; (2) 它们都是复值、有界、一致连续函数; (3) 当对应分布是连续型分布时, 特征函数在无穷远处的极限是 0。关于特征函数更精确的刻画, 读者可以参见本章 §11.4 的 Bochner-Khintchine 定理。

这里, 我们特别提请读者熟悉正态分布等常见分布的特征函数。

例 11.1. (1) 设 $Z \sim N(0, 1)$, 那么 Z 的特征函数为 $\mathbb{E}[e^{itZ}] = \exp\{-\frac{t^2}{2}\}$;

*也有很多文献把 Fourier 变换定义为 $\mathcal{F}(\rho)(\lambda) := \int \rho(x) \cdot e^{-2\pi i \lambda \cdot x} dx$, 这样定义的好处是 Fourier 变换成为了一个 L^2 -等距变换。

- (2) 设 $X \sim N(\mu, \sigma^2)$, 那么 X 的特征函数为 $\mathbb{E}[e^{itX}] = \exp\{i\mu t - \frac{\sigma^2 t^2}{2}\}$;
 (3) 设 $Y \sim N(\mu, \Sigma)$, 其中 $\mu \in \mathbb{R}^m, \Sigma = (\sigma_{i,j})_{1 \leq i,j \leq m}$, 那么 Y 的特征函数为 $\mathbb{E}[e^{it \cdot Y}] = \exp\{i\mu \cdot t - \frac{t\Sigma \cdot t}{2}\}$.

证明. 第一条结论最为关键, 其他结论可以在此基础上通过线性变换及借助独立性 (高维情形先考虑高维的标准正态分布) 的办法来完成对相应特征函数的计算; 对应细节留给读者。

当 $\lambda \in \mathbb{R}$ 时, 配方法结合换元容易计算出

$$\mathbb{E}e^{\lambda Z} = \int \frac{e^{\lambda x}}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = e^{\frac{\lambda^2}{2}}.$$

因此 (1) 中结论在形式演算下来理解并不难。但数学上严格证明要一定的功夫。借助留数定理来完成计算与论证, 是一种方法, 细节留给读者。此处我们使用微分的方法。事实上, 令

$$\psi(t) := \mathbb{E}[e^{itZ}] = \int \frac{e^{itx}}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

注意到被积函数的虚部是一个奇函数, 对应积分为 0, 因此

$$\psi(t) = \int \frac{\cos(tx)}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

我们可以在积分号下关于 t 求导:

$$\begin{aligned} \psi'(t) &= \int \frac{-x \sin(tx)}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int \frac{\sin(tx)}{\sqrt{2\pi}} d e^{-\frac{x^2}{2}} \\ &= - \int \frac{t \cos(tx)}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = -t\psi(t). \end{aligned}$$

这表明 $\frac{d}{dt}[\psi(t)e^{\frac{t^2}{2}}] = 0$, 从而 $\psi(t)e^{\frac{t^2}{2}} = \psi(0) = 1$, 即 $\psi(t) = e^{-\frac{t^2}{2}}$. \square

11.2 特征函数的唯一性定理

从特征函数到分布函数有下面的一些逆转公式/反演公式。

定理 11.2.1. (逆转公式) 设 F 是分布函数, $\hat{\mu}_F$ 是其对应的特征函数。

- (1) 对于 F 的任意两个连续点 $a < b$, 有

$$F(b) - F(a) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iat} - e^{-ibt}}{it} \cdot \hat{\mu}_F(t) dt; \quad (11.1)$$

- (2) 对任意 $\varphi \in C_0(\mathbb{R})$,

$$\int \varphi dF = \lim_{t \downarrow 0} \frac{1}{2\pi} \int \varphi(x) dx \int e^{-ixz} \cdot \hat{\mu}_F(z) \cdot e^{-\frac{t}{2}|z|^2} dz; \quad (11.2)$$

- (3) 如果 $\|\hat{\mu}_F\|_1 := \int |\hat{\mu}_F| dt < \infty$, 那么 F 具有密度函数 $\rho(x)$, 且

$$\rho(x) = \frac{1}{2\pi} \int e^{-itx} \cdot \hat{\mu}_F(t) dt, \quad (11.3)$$

并且 $\sup_x \rho(x) \leq \frac{\|\hat{\mu}_F\|_1}{2\pi} < \infty$ (从而 F 是 Lipschitz 连续函数)。

(4) * 如果 $\|\hat{\mu}_F\|_2 := \sqrt{\int |\hat{\mu}_F|^2 dt} < \infty$, 那么 F 是 $\frac{1}{2}$ -Hölder 连续的, 并且具有密度函数 $\rho(x)$ 。此时

$$\rho(x) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T e^{-itx} \cdot \hat{\mu}_F(t) dt \quad (11.4)$$

在 $L^2(\mathbb{R})$ 中的 L^2 -收敛意义下成立。

证明. 先证明 (2)。假设 $X \sim F, \xi \sim N(0, 1)$ 且二者独立, 那么对任意紧支集连续函数 φ 及 $t > 0$

$$\begin{aligned} \mathbb{E}[\varphi(X + \sqrt{t}\xi)] &= \mathbb{E}\left[\int \varphi(X + \sqrt{t}z) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz\right] \\ &= \mathbb{E}\left[\int \varphi(X + \sqrt{t}z) \cdot \frac{1}{\sqrt{2\pi}} \mathbb{E}[e^{-iz\xi}] dz\right] \\ &= \mathbb{E}\left[\frac{1}{\sqrt{2\pi}} \int \varphi(X + \sqrt{t}z) \cdot e^{-iz\xi} dz\right] \\ &= \mathbb{E}\left[\frac{1}{\sqrt{2\pi t}} \int \varphi(x) \cdot e^{-i\xi \cdot \frac{x-X}{\sqrt{t}}} dx\right] = \frac{1}{\sqrt{2\pi t}} \int \varphi(x) \cdot \mathbb{E}\left[e^{-i\xi \cdot \frac{x-X}{\sqrt{t}}}\right] dx \\ &= \frac{1}{\sqrt{2\pi t}} \int \varphi(x) \cdot \mathbb{E}\left[e^{-i\xi \cdot \frac{x}{\sqrt{t}}} \cdot \hat{\mu}_F\left(\frac{\xi}{\sqrt{t}}\right)\right] dx \\ &= \frac{1}{2\pi\sqrt{t}} \int \varphi(x) \cdot \left[\int e^{-iy \cdot \frac{x}{\sqrt{t}}} \cdot \hat{\mu}_F\left(\frac{y}{\sqrt{t}}\right) e^{-\frac{y^2}{2}} dy\right] dx \\ &= \frac{1}{2\pi} \int \varphi(x) \cdot \left[\int e^{-iz \cdot x} \cdot \hat{\mu}_F(z) e^{-\frac{t}{2} \cdot z^2} dz\right] dx. \end{aligned}$$

再令 $t \downarrow 0$, 注意到 φ 连续有界, 由控制收敛定理立即得到结论 (2)。

再证明 (1)。事实上,

$$\begin{aligned} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iat} - e^{-ibt}}{it} \cdot \hat{\mu}_F(t) dt &= \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iat} - e^{-ibt}}{it} \cdot \mathbb{E}[e^{itX}] dt \\ &= \mathbb{E}\left[\frac{1}{2\pi} \int_{-T}^T \frac{e^{-iat} - e^{-ibt}}{it} \cdot e^{itX} dt\right] \\ &= \mathbb{E}\left[\frac{1}{2\pi} \int_{-T}^T \frac{\sin[(X-a)t] - \sin[(X-b)t]}{t} dt\right]. \end{aligned}$$

利用 $\int \frac{\sin \alpha t}{t} dt = \pi \operatorname{sgn}(\alpha)$, 其中 $\operatorname{sgn}(\alpha) := 1_{(0, \infty)}(\alpha) - 1_{(-\infty, 0)}(\alpha)$, 结合控

*此条结论在 [20, pp. 299] 有非常简短的描述, 后来作者发现在钟开莱的中文译书 [61] 的习题 6.2.11 中也有出现; 较多概率论教科书中缺失这一漂亮结果的论述。此处, 也感谢李洪全教授介绍参考文献 [19]。

制收敛定理*, 立即得到:

$$\lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iat} - e^{-ibt}}{it} \cdot \hat{\mu}_F(t) dt = \frac{1}{2} \mathbb{E}[\text{sgn}(X - a) - \text{sgn}(X - b)].$$

简单计算表明

$$\mathbb{E}[\text{sgn}(X - a) - \text{sgn}(X - b)] = F(b) + F(b-) - (F(a) + F(a-)).$$

因此当 a, b 为 F 的连续点时定理的结论 (1) 成立。对于一般情况 $a < b$, 有

$$\lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iat} - e^{-ibt}}{it} \cdot \hat{\mu}_F(t) dt = \mu_F((a, b)) + \frac{\mu_F(\{a\}) + \mu_F(\{b\})}{2}. \quad (11.5)$$

现在来证明 (3)。我们先论证 μ_F 是绝对连续的。事实上, 假定 $a < b$ 都是 F 的连续点, 那么由 (11.1)

$$[F(b) - F(a)] \leq \frac{1}{2\pi} \int | \frac{e^{-iat} - e^{-ibt}}{it} | \cdot |\hat{\mu}_F(t)| dt.$$

计算表明, $| \frac{e^{-iat} - e^{-ibt}}{it} | \leq |b - a|$, 从而

$$|F(b) - F(a)| \leq \frac{\|\hat{\mu}_F\|_1}{2\pi} \cdot |b - a|.$$

由于上述估计对于 F 的所有连续点 a, b 都成立, 很容易推断它实际上对所有 a, b 成立。由此 F 是绝对连续的, 具有密度 ρ 。现在对任意 $a < b$ 都有:

$$\frac{F(b) - F(a)}{b - a} = \frac{1}{2\pi} \int \frac{e^{-iat} - e^{-ibt}}{i(b-a)t} \cdot \hat{\mu}_F(t) dt.$$

令 $b \downarrow a$, 由控制收敛定理

$$\rho(a) = \frac{1}{2\pi} \int e^{-iat} \cdot \hat{\mu}_F(t) dt.$$

显然 ρ 具有定理中的上界估计。

最后证明 (4)。对 $L^2(\mathbb{R})$, 我们记其上内积为 $\langle \cdot, \cdot \rangle$, 即

$$\langle f, g \rangle := \int \bar{f}g dx, \quad \forall f, g \in L^2(\mathbb{R}).$$

特别的, Fourier 变换与其逆变换在 $L^2(\mathbb{R})$ 中有定义 (见 [19, Section 2.2.3–2.2.4]), 且满足

$$\langle \hat{f}, \hat{g} \rangle = 2\pi \langle f, g \rangle, \quad \forall f, g \in L^2(\mathbb{R}).$$

假定 $a < b$ 都是 F 的连续点。注意到

$$\frac{e^{-iat} - e^{-ibt}}{it} = \int e^{-ixt} 1_{(a,b]}(x) dx = \bar{1}_{(a,b]}(t),$$

有 $\|\bar{1}_{(a,b]}\|_2 = \sqrt{2\pi} \|1_{(a,b]}\|_2 = \sqrt{2\pi|b-a|}$ 。

*提示: 根据本书第 4 章例 4.6 的结论 (4.8), 我们可以粗略地取控制函数为常数 $M := \sup\{|\frac{1}{\pi} \int_{-T}^T \frac{\sin x}{x} dx| : T > 0\} < \infty$ 。[61] 指出, $0 \leq \int_0^T \frac{\sin x}{x} dx \leq \int_0^\pi \frac{\sin x}{x} dx, \forall T > 0$, 因而此处控制函数可以取为常数 $M := \frac{2}{\pi} \int_0^\pi \frac{\sin t}{t} dt \leq 2$ 。[13, Exercise 1.7.5] 指出, 对 $T > 0$, $\int_0^T \frac{\sin x}{x} dx = \frac{\pi}{2} - \int_0^\infty \frac{(\cos T + y \sin T)e^{-Ty}}{1+y^2} dy$, 进而 $|\int_0^T \frac{\sin x}{x} dx - \frac{\pi}{2}| \leq \frac{1}{T}, \forall T > 0$ 。

由 (11.1) 以及 Cauchy 不等式,

$$F(b) - F(a) = \frac{1}{2\pi} \langle \hat{1}_{(a,b]}, \hat{\mu}_F \rangle \leq \sqrt{\frac{|b-a|}{2\pi}} \cdot \|\hat{\mu}_F\|_2.$$

因此 F 是 $\frac{1}{2}$ -Hölder 连续的。

由于 $\hat{\mu}_F \in L^2(\mathbb{R})$, 存在 $\rho \in L^2(\mathbb{R})$ 使得 $\hat{\mu}_F = \hat{\rho}$ 。以下我们论证 ρ 就是 F 的密度函数。事实上, 对任意 $a < b$, 有

$$\begin{aligned} F(b) - F(a) &= \frac{1}{2\pi} \langle \hat{1}_{(a,b]}, \hat{\mu}_F \rangle = \frac{1}{2\pi} \langle \hat{1}_{(a,b]}, \hat{\rho} \rangle \\ &= \langle 1_{(a,b]}, \rho \rangle = \int_{(a,b]} \rho(x) dx. \end{aligned}$$

由此, ρ 应为实函数, 且非负, 并且它是分布 F 的密度函数。□

上面的逆转公式是一维情形的结论。需要指出的是, 高维情形下上述定理中 (3) 的类似结果仍然成立。基于这个结果容易知道, 随机变量 (或向量) 的分布函数与它的特征函数之间是一一对应的; 这个结论又称为特征函数的唯一性定理。

在承认特征函数的唯一性定理后, 使用特征函数的方法来刻画随机变量的独立性就是很简单的事情了。

定理 11.2.2. 设 X, Y 的分布测度分别为 μ_X, μ_Y , 联合分布测度记作 $\mu_{(X,Y)}$ 。那么 X 与 Y 独立的充分必要条件是

$$\hat{\mu}_{(X,Y)}(u, v) = \hat{\mu}_X(u) \hat{\mu}_Y(v).$$

上式也可以简单记为: $\hat{\mu}_{(X,Y)} = \hat{\mu}_X \otimes \hat{\mu}_Y$, 右端 \otimes 表示函数的张量积。这正好和我们用分布测度描述的独立性对应: X 与 Y 独立的充分必要条件是 $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$ 。

11.3 特征函数的连续性定理

通过特征函数, 我们可以估计分布函数的尾部特征, 具体而言我们有下面的引理:

引理 11.3.1. 设 f 是随机变量 X 的特征函数。那么对任意 $\varepsilon, \delta > 0$, 存在绝对常数 $K = K_\delta > 0$ (仅依赖于 δ), 使得

$$\mathbb{P}(|X| \geq \varepsilon) \leq K \cdot \int_0^1 [1 - \operatorname{Re}(f(t\delta/\varepsilon))] dt \quad (11.6)$$

$$\mathbb{E}[X^2 \cdot 1_{\{|X| \leq \varepsilon\}}] \leq K \cdot \varepsilon^2 \cdot [1 - \operatorname{Re}(f(\frac{\delta}{\varepsilon}))]. \quad (11.7)$$

证明. 补充定义函数 $\frac{\sin x}{x}$ 在 0 处的取值为 1, 则此函数成为 \mathbb{R} 上有界一致连续函数; 类似的, 补充定义 $\frac{1-\cos x}{x^2}$ 在 0 处的取值为 $\frac{1}{2}$, 此函数也成为 \mathbb{R} 上有界一致连续函数。于是对任意 $\delta \in (0, 1)$, 存在 $K = K_\delta > 0$ 使得

$$1 - \frac{\sin x}{x} \geq \frac{1}{K}, \quad \forall |x| \geq \delta,$$

并且

$$\frac{1 - \cos x}{x^2} \geq \frac{1}{K}, \quad \forall |x| \leq \delta.$$

记 F 为 X 的分布函数, 有 (此处 $M := \delta/\varepsilon$)

$$\begin{aligned}\mathbb{P}(|X| \geq \varepsilon) &= \mathbb{E}[1_{\{|MX| \geq \delta\}}] \leq K\mathbb{E}[1 - \frac{\sin MX}{MX}] \\ &= K \int_0^1 [1 - \operatorname{Re} f(Mt)] dt.\end{aligned}$$

此即引理的第一个不等式。

类似的,

$$\begin{aligned}\mathbb{E}[|X|^2 \cdot 1_{\{|X| \leq \varepsilon\}}] &= \frac{1}{M^2} \mathbb{E}[|MX|^2 \cdot 1_{\{|MX| \leq \delta\}}] \\ &\leq \frac{K}{M^2} \mathbb{E}(1 - \cos MX) \\ &= \frac{K}{M^2} \cdot [1 - \operatorname{Re}(f(M))].\end{aligned}$$

此即引理的第二个不等式。 \square

在第 2 章中, 我们已经介绍过依分布收敛的概念, 再次陈述如下。

定义 11.3.1. 设 $\{X_n\}_{n=0}^\infty$ 为随机变量列。称 $\{X_n\}_{n=1}^\infty$ 依分布收敛到 X_0 , 记作 $X_n \xrightarrow{d} X_0$, 如果 $F_n(x), F_0(x)$ 分别是 X_n, X_0 的分布函数, 且对于 F_0 的任意连续点 x , 总有 $\lim_{n \rightarrow \infty} F_n(x) = F_0(x)$; 有时也记 $F_n \xrightarrow{w} F_0$, 或 $X_n \xrightarrow{d} F_0$ 。注意, 此处诸随机变量 X_0, X_1, \dots 可以定义在不同的概率空间中, 而不必定义在同一个概率空间中。

更一般的, 设 $\{F_n\}_{n=0}^\infty$ 是递增右连续函数的序列。如果在 F_0 的连续点上 F_n 收敛到 F_0 , 就称 F_n 弱收敛到 F_0 , 记作 $F_n \xrightarrow{w} F_0$ 。当 F_n, F_0 都是概率分布函数时, 我们也把 $F_n \xrightarrow{w} F_0$ 称为 F_n 正常的弱收敛到 F_0 。

利用分布函数广义逆的性质, 很容易论证如下 Skorokhod 嵌入定理。

定理 11.3.1. (Skorokhod 嵌入定理) 设 $\{F_n\}_{n=1}^\infty$ 是一列分布函数。如果 $F_n \xrightarrow{w} F_0$, 其中 F_0 是某一分布函数, 那么存在某个概率空间上的随机变量列 $\{X_n\}_{n=0}^\infty$, 使得:

(i) $X_n \sim F_n, \forall n \geq 0$;

(ii) $X_n \rightarrow X_0$ 几乎处处成立。

证明. 取 $U \sim U(0, 1)$ 为均匀分布随机变量。记 F_0^{-1} 为 F_0 的广义逆, 即

$$F_0^{-1}(y) := \inf\{x : F_0(x) \geq y\}.$$

令 $X_n = F_n^{-1}(U) \sim F_n, X = F_0^{-1}(U) \sim F_0$ 。可以论证 (反证法; 留作习题): 在定理的条件下,

$$\lim_n F_n^{-1}(p) = F_0^{-1}(p)$$

对于 F_0^{-1} 的连续点 $p \in (0, 1)$ 成立。于是定理的两个结论都成立。 \square

由上述 Skorokhod 嵌入定理, 容易知道下面结果成立。

定理 11.3.2. (Helly-Bray 定理) 设 $X_n \xrightarrow{d} X$, 那么

$$\lim_{n \rightarrow \infty} \mathbb{E}f(X_n) = \mathbb{E}f(X)$$

对任意有界连续函数 f 成立。

下面的定理则称为特征函数的连续性定理，通常也称为 **Lévy 连续性定理**。它本质上说明了从分布函数到特征函数这个一一映射的双向连续性。这是实际应用中利用特征函数来论证依分布收敛（特别是中心极限定理）的理论依据。

定理 11.3.3.（特征函数的连续性定理） 设 $\{F_n\}_{n=1}^\infty$ 是一列分布函数， $\{f_n\}_{n=1}^\infty$ 是其对应的特征函数列。

- (1) 如果 $F_n \xrightarrow{w} F$ ，其中 F 是某一分布函数，那么 $\hat{\mu}_{F_n} \rightarrow \hat{\mu}_F$ 点点成立；
- (2) 如果对每个 $t \in \mathbb{R}$ 点点极限 $\lim_{n \rightarrow \infty} \hat{\mu}_{F_n}(t) = g(t)$ 存在，并且 $g(t)$ 在 $t=0$ 处连续，那么 g 是某一分布 F 的特征函数，即 $g = \hat{\mu}_F$ ，并且 $F_n \xrightarrow{w} F$ 。

为了论证上述定理，我们需要如下结论（其中 $m \geq 1$ 为整数）。

定理 11.3.4.（Helly 选择定理） \mathbb{R}^m 上任何分布函数列 $\{F_n\}_{n=1}^\infty$ 都有弱收敛的子列 $\{F_{n_k}\}$ 。

证明. 此定理实际上可以看成点集拓扑理论中 Tychonov 定理的一个推论。这里我们给出另外一个简单的证明。

先证明 $m=1$ 的情形。任取 \mathbb{R} 的一个可数稠密子集 $D = \{x_n : n \geq 1\}$ （比如可以取 $D = \mathbb{Q}$ ）。利用对角线法则容易验证，存在子列 $\{n_k\}_{k=1}^\infty$ 使得对任意给定 $i \geq 1$ ， $\{F_{n_k}(x_i)\}_{k=1}^\infty$ 收敛。令

$$F(r) := \lim_{k \rightarrow \infty} F_{n_k}(r), r \in D.$$

容易知道 $F : D \rightarrow \mathbb{R}$ 是单调不减函数。补充定义

$$F(x) := \lim_{D \ni r \downarrow x} F(r), x \in \mathbb{R} \setminus D.$$

容易知道，此时 F 成为 \mathbb{R} 上一个递增函数，并且在 D^c 上右连续（注意！未必是概率分布函数）。记 F 的连续点集为 \mathcal{C}_F 。并记

$$\bar{F}(x) := \overline{\lim}_{k \rightarrow \infty} F_{n_k}(x), \quad \underline{F}(x) := \underline{\lim}_{k \rightarrow \infty} F_{n_k}(x).$$

那么对任意 $x \in \mathcal{C}_F$ ，存在 D 中点列 $r_n \uparrow x, r'_n \downarrow x$ ，于是 $r_n < x < r'_n$ ，

$$\bar{F}(x) \leq \overline{\lim}_{k \rightarrow \infty} F_{n_k}(r'_n) = F(r'_n) \rightarrow F(x),$$

$$\underline{F}(x) \geq \underline{\lim}_{k \rightarrow \infty} F_{n_k}(r_n) = F(r_n) \rightarrow F(x)$$

成立，这表明 $\lim_{k \rightarrow \infty} F_{n_k}(x) = F(x), \forall x \in \mathcal{C}_F$ 。注意， F 未必具有右连左极性；但由于 F 是一个单调函数，我们可以对它进行右连续修正，这并不会改变它的连续点上的取值；我们把修正后的函数记作 \tilde{F} ，显然 F 与 \tilde{F} 具有相同的连续点集。那么 $F_{n_k} \xrightarrow{w} \tilde{F}$ 。因此定理结论成立。

对于 $m \geq 2$ 的情形类似可证。 \square

定理 11.3.3 的证明. (1) 我们借助 Skorokhod 嵌入定理来论证。不妨假设 $X_n \rightarrow X$ 是几乎处处的，其中 $X_n \sim F_n, X \sim F$ 。那么对于任意 $\lambda \in \mathbb{R}$ ， $e^{i\lambda X_n} \rightarrow e^{i\lambda X}$ 几乎处处成立。由 Lebesgue 控制收敛定理（或有界收敛定理），

$$\hat{\mu}_{F_n}(\lambda) = \mathbb{E}[e^{i\lambda X_n}] \rightarrow \mathbb{E}[e^{i\lambda X}] = \hat{\mu}_F(\lambda).$$

(2) 此处我们假设 $X_n \sim F_n$ 。由 Helly 定理, $\{F_n\}_{n=1}^\infty$ 有子列 $\{F_{n_k}\}_{k=1}^\infty$ 弱收敛于某个递增右连续的函数 G , 并且容易知道 $0 \leq G(x) \leq 1, \forall x \in \mathbb{R}$ 。我们先论证 G 是分布函数。任取 $T > 0$, 由引理 11.3.1 (取 $\varepsilon = T, \delta = 2$, 此时可取 $K = 2$),

$$\mathbb{P}(|X_{n_k}| \geq T) \leq 2 \int_0^1 [1 - \operatorname{Re}(\hat{\mu}_{F_{n_k}}(2\lambda/T))] d\lambda.$$

对固定的 T , 由控制收敛定理,

$$\lim_{k \rightarrow \infty} \int_0^1 [1 - \operatorname{Re}(\hat{\mu}_{F_{n_k}}(2\lambda/T))] d\lambda = \int_0^1 [1 - \operatorname{Re}(g(2\lambda/T))] d\lambda.$$

注意到 g 在 0 处的连续性以及 $g(0) = 1$,

$$\lim_{T \rightarrow \infty} \int_0^1 [1 - \operatorname{Re}(g(2\lambda/T))] d\lambda = 0.$$

注意到当 $\pm T$ 均为 G 的连续点时

$$\mathbb{P}(X_{n_k} \in (-T, T]) = F_{n_k}(T) - F_{n_k}(-T) \rightarrow G(T) - G(-T).$$

记 $\mathcal{D} := \{T : \pm T \text{ 为 } G \text{ 的连续点}\}$, 则 \mathcal{D} 是 \mathbb{R} 中 Lebesgue 满测集, 因而

$$\lim_{\mathcal{D} \ni T \rightarrow \infty} [G(T) - G(-T)] \geq \overline{\lim}_{\mathcal{D} \ni T \rightarrow \infty} \lim_{k \rightarrow \infty} [1 - \mathbb{P}(|X_{n_k}| \geq T)] \geq 1,$$

由此, G 是概率分布函数。

现在我们论证 $g = \hat{\mu}_G$ 。事实上, 我们已经论证 $F_{n_k} \xrightarrow{w} G$ 。由上面的 (1) 立即知道 $\hat{\mu}_{F_{n_k}} \rightarrow \hat{\mu}_G$ 点点成立, 于是 $g = \hat{\mu}_G$ 。由唯一性定理, G 由 g 唯一决定。

现在我们知道, G 是一个概率分布, 且 $\hat{\mu}_{F_n} \rightarrow \hat{\mu}_G$ 点点成立。我们要证明 $F_n \xrightarrow{w} G$ 。但是, 由上述证明容易知道, $\{F_n\}_{n=1}^\infty$ 任意子列的弱极限都必定是由 g 所唯一决定的分布函数 $F = G$, 于是上述结论成立。□

注 11.2. *Eduard Helly* (1884/6/1–1943/? /?; 奥地利) 生于维也纳, 卒于美国芝加哥。他 1907 年博士毕业于维也纳大学, 之后到哥廷根大学进修 1 年。他在第一次世界大战中应召入伍, 1915 年中弹被俘, 起初关押在西伯利亚的布列左夫卡集中营; 狱中他组织数学讨论班 (当时的工程师狱友 *Tibor Radó* (1895/6/2–1965/12/29; 匈牙利) 受此影响而对纯粹数学感兴趣, 并在之后成功逃回匈牙利, 1923 年获得博士学位, 成为数学家, 1929 年移居美国); 在转到西伯利亚的尼科利斯克-乌苏里斯克集中营中, 他继续举办讨论班, 并在泛函分析方面做出重要贡献。他最后在 1920 年重返维也纳, 次年结婚; 1938 年奥地利被纳粹德国占领, 他逃亡到了美国, 在爱因斯坦帮助下, 在新泽西的中学教书, 之后 1941 年与妻子移居芝加哥, 并为美军的通讯部队服务。他在数学中的重要贡献有: *Helly* 定理、*Helly* 族、*Helly* 选择定理、*Helly-Bray* 定理、*Helly* 度量等。

11.4 特征函数的等价刻画

下面的 Bochner-Khintchine 定理回答了什么样的函数可以作为概率测度的特征函数。

定理 11.4.1. (*Bochner-Khintchine 定理*) 设 g 是 \mathbb{R} 上复值连续函数, 且 $g(0) = 1$ 。那么 g 是某个概率测度的特征函数的充要条件是: g 是非负定的, 即对任意 $n \geq 2$ 及实数 $\lambda_1, \dots, \lambda_n$, 矩阵 $M_n := (g(\lambda_i - \lambda_j))_{1 \leq i, j \leq n}$ 非负定。

证明. 必要性的证明比较简单，留给读者。

充分性的一个证明思路如下（细节留给读者补充）：对任意 $m > 0$ 定义

$$p_m(x) := \frac{1}{2\pi m} \int_{(0,m)^2} g(t-s) \cdot e^{-i(t-s)x} dt ds.$$

容易验证 p_m 非负，且它是一个密度函数。

定义 $g_m(\lambda) := g(\lambda) \cdot (1 - \frac{|\lambda|}{m}) \cdot 1_{(-m,m)}(\lambda)$ 。容易验证 g_m 是 p_m 的 Fourier 变换（或者说特征函数），但是 g_m 点点收敛于 g ，由连续性定理， g 是某个随机变量的特征函数。 \square

11.5 中心极限定理

最早的中心极限定理是 18 世纪中期 De Moivre 和 Laplace 对 i.i.d. 的 Bernoulli 序列，基于相应概率分布列的纯粹分析方法建立起来的；见课后习题 11.32。

Lyapunov 是使用特征函数方法来证明中心极限定理的第一人；他是 Chebyshev 的学生，是 Markov 的学弟。在他之前，Chebyshev 和 Markov 是运用“矩方法”来证明一些特殊情形的中心极限定理的。概率理论后来的发展证明，“特征函数是证明最多种多样的极限定理的强有力工具”（见 [49, 序言]）。

下面这个常见的经典中心极限定理通常也被称为 Lindeberg-Lévy 定理。

定理 11.5.1. 设随机变量列 $\{X_n\}_1^\infty$ i.i.d., $0 < \sigma^2 := \text{Var}(X_1) < \infty$ 。记 $\mu := \mathbb{E}X_1$, $\bar{X}_n := \frac{X_1 + \dots + X_n}{n}$, 那么 $Z_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{d} N(0, 1)$, 即

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x) := \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du, \quad \forall x \in \mathbb{R}. \quad (11.8)$$

证明. 标准正态的特征函数为 $f(t) := e^{-t^2/2}$ 。

不妨设 $\mu = 0, \sigma^2 = 1$, 即 $\mathbb{E}X_1 = 0, \mathbb{E}[X_1^2] = 1$ 。设 X_1 的特征函数为 $g(t) := \mathbb{E}e^{itX_1}$ 。那么 $Z_n = \sqrt{n}\bar{X}_n$ 的特征函数为

$$f_n(t) := \mathbb{E}[e^{itZ_n}] = g\left(\frac{t}{\sqrt{n}}\right)^n.$$

容易知道，

$$|f_n(t) - f(t)| = \left| g\left(\frac{t}{\sqrt{n}}\right)^n - f\left(\frac{t}{\sqrt{n}}\right)^n \right| \leq n \left| g\left(\frac{t}{\sqrt{n}}\right) - f\left(\frac{t}{\sqrt{n}}\right) \right|.$$

根据 $\mathbb{E}X_1 = 0, \mathbb{E}[X_1^2] = 1$, 有

$$g\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right).$$

又显然有 $f\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$, 因此

$$n \left| g\left(\frac{t}{\sqrt{n}}\right) - f\left(\frac{t}{\sqrt{n}}\right) \right| = n \cdot o\left(\frac{1}{n}\right) = o(1) \rightarrow 0,$$

即 $f_n \rightarrow f$ 点点成立。由特征函数的连续性定理， $Z_n \xrightarrow{d} N(0, 1)$ 成立。 \square

对于 i.i.d. 序列的弱大数律，无论是最初的 Bernoulli 弱大数律，还是后来的 Chebyshev 弱大数律和 Markov 弱大数律，都要求分布具有二阶矩。1929 年，Khinchine (1894–1959) 在只要求分布具有一阶矩的条件下证明了下面的 Khinchine 弱大数律。它是 Kolmogorov 强大数律的先声。

定理 11.5.2. (*Khinchine 弱大数律*) 设 $\{X_n\}_{n=1}^{\infty}$ 为独立同分布的随机变量序列，满足 $\mathbb{E}|X_1| < \infty$ 。记 $\mu := \mathbb{E}X_1$ ，那么 $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$ 。

证明. 不妨设 $\mu = 0$ 。记 f 为 X_1 的特征函数。考察 \bar{X}_n 的特征函数 f_n ：

$$f_n(t) := \mathbb{E}[e^{it\bar{X}_n}] = f\left(\frac{t}{n}\right)^n.$$

因为 $\mathbb{E}X_1 = 0$ ，容易知道在 0 附近 $f(t) = 1 + o(t)$ 。于是

$$|f_n(t) - 1| \leq n|f\left(\frac{t}{n}\right) - 1| \rightarrow 0,$$

即 $\bar{X}_n \xrightarrow{d} 0$ 。由定理 10.3.8 立即知道 $\bar{X}_n \xrightarrow{\mathbb{P}} 0$ 。 □



图 11.3: A. Khinchine (1894-1959)

注 11.3. A. Khinchine (辛钦, 1894/7/19–1959/11/18) 是前苏联数学家、教育学家。1916 年他毕业于莫斯科大学，并留校从事教学工作，1919 年成为教授。1935 年获得数学物理博士学位，导师是 Nikolai Luzin (1883/1/29–1950/1/28; 苏联-俄罗斯)。1939 年他被选为苏联科学院通讯院士，1944 年被选为俄罗斯联邦教育科学院院士。他是前苏联概率论学派的创始人之一，在概率论的极限定理、平稳随机过程、分析学中的 Denjoy 积分、数论（特别是连分数）等方面都有建树。本书中的 Markov 过程，最初 Kolmogorov 起名为无后效过程，也是 Khinchine 建议改为 Markov 过程。

Khinchine 培养的知名学生有：A. Buchstab (1905/10/4–1990/2/27; 苏联数学家，数论专家，以 Buchstab 函数知名)，A. Gelfond (1906/10/24–1968/11/7; 苏联数学家，在数论、解析函数、积分方程等方面都有贡献，以 Gelfond 定理知名，1934 年彻底解决 Hilbert 第 7 问题) 等。

下面给出独立随机变量阵列形式的 Lindeberg-Feller 中心极限定理：定理的充分性部分归功于 Lindeberg，必要性部分则归功于 Feller。本书不提供这个定理的证明，感兴趣的读者可以参见 [18, 42, 61] 等著作。

定理 11.5.3. (*Lindeberg-Feller 中心极限定理*) 设 $\{X_{n,k} : 1 \leq k \leq k_n, n \geq 1\}$ 为一给定的独立随机变量阵列，亦即对每个 $n \geq 1$, $X_{n,1}, \dots, X_{n,k_n}$ 是相互

独立的。假设对每个 $n \geq 1$ 及 $k = 1, \dots, k_n$, $\mathbb{E}X_{n,k} = 0$, $\mathbb{E}X_{n,k}^2 =: \sigma_{n,k}^2 < \infty$, 并且 $\sigma_n^2 := \sigma_{n,1}^2 + \dots + \sigma_{n,k_n}^2 > 0$ 。那么下面两式

$$\lim_{n \rightarrow \infty} \frac{\max_{1 \leq k \leq k_n} \sigma_{n,k}^2}{\sigma_n^2} = 0, \quad (11.9)$$

$$\frac{1}{\sigma_n} \cdot \sum_{k=1}^{k_n} X_{n,k} \xrightarrow{d} N(0, 1) \quad (11.10)$$

成立的充分必要条件是下面的 *Lindeberg* 条件成立

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma_n^2} \sum_{k=1}^{k_n} \mathbb{E}X_{n,k}^2 \cdot 1_{\{|X_{n,k}| \geq \varepsilon \sigma_n\}} = 0, \quad \forall \varepsilon > 0. \quad (11.11)$$

把上述定理应用于独立随机变量列, 得到下面的中心极限定理。

定理 11.5.4. 设 $\{X_n\}_{n=1}^\infty$ 为一给定的独立随机变量列。假设对每个 $n \geq 1$, $\mathbb{E}X_n = 0$, $0 < \mathbb{E}X_n^2 =: \sigma_n^2 < \infty$, 并且 $B_n^2 := \sigma_1^2 + \dots + \sigma_n^2 > 0$ 。那么下面两式

$$\lim_{n \rightarrow \infty} \frac{\max_{1 \leq k \leq n} \sigma_k^2}{B_n^2} = 0, \quad (11.12)$$

$$\frac{1}{B_n} \cdot \sum_{k=1}^n X_k \xrightarrow{d} N(0, 1) \quad (11.13)$$

成立的充分必要条件是下面的 *Lindeberg* 条件成立

$$\lim_{n \rightarrow \infty} \frac{1}{B_n^2} \sum_{k=1}^n \mathbb{E}X_k^2 \cdot 1_{\{|X_k| \geq \varepsilon B_n\}} = 0, \quad \forall \varepsilon > 0. \quad (11.14)$$

在实际使用中, *Lindeberg* 条件验证起来不是那么容易。很多场合, 下面的 *Lyapunov* 中心极限定理更方便使用。

定理 11.5.5. (*Lyapunov* 中心极限定理) 设 $\{X_n\}_{n=1}^\infty$ 为一给定的独立随机变量列。假设对每个 $n \geq 1$, $\mathbb{E}X_n = 0$, $0 < \mathbb{E}X_n^2 =: \sigma_n^2 < \infty$, 并且 $B_n^2 := \sigma_1^2 + \dots + \sigma_n^2 > 0$ 。如果存在 $\delta > 0$ 使得

$$\lim_{n \rightarrow \infty} \frac{1}{B_n^{2+\delta}} \sum_{k=1}^n \mathbb{E}[|X_k|^{2+\delta}] = 0, \quad (11.15)$$

那么

$$\frac{1}{B_n} \cdot \sum_{k=1}^n X_k \xrightarrow{d} N(0, 1).$$

上述定理中的条件(11.15)被称为 *Lyapunov* 条件。

注 11.4. *Jarl Waldemar Lindeberg* (1876/8/4–1932/12/24; 芬兰) 生于赫尔辛基, 卒于赫尔辛基。他是赫尔辛基大学教授, 以中心极限定理方面的工作闻名于世。*Lindeberg* 是赫尔辛基综合理工学院的教师的儿子, 小时候就显露出数学上的天赋与兴趣。他早期的兴趣是偏微分方程和变分, 但 1920 年后兴趣转向了概率与统计。1920 年他发表了他在中心极限定理方面的第一篇论文, 其结果类似于 *Lyapunov* 的工作 (但那时他并不知情), 但所用方法是卷

积的方法，而不是 *Lyapunov* 使用的特征函数方法。两年后他用他的方法得到了更强的结果，*Lindeberg* 条件下的中心极限定理。1935 年，*Alan Turing*（图灵，1912/6/23–1954/6/7；英国数学家、计算机科学家、逻辑学家、密码学家、哲学家、理论生物学家）在不知道 *Lindeberg* 的工作的情况下，在博士论文中也证明了这个中心极限定理。

William Feller（费勒，1906/7/7–1970/1/14）是美籍克罗地亚裔数学家，他的父亲是波兰籍犹太人，母亲是奥地利人；他的博士论文导师是 *Richard Courant*（1888/1/8–1972/1/27；美籍德国数学家）。*Feller* 原本 1928 年在 *Kiel* 大学拥有临时职位，到 1933 年他因为拒绝向纳粹宣誓效忠而逃亡到丹麦的哥本哈根。后来他到瑞典的斯德哥尔摩讲学。据 *Feller* 记载，那时大学里面的法西斯主义持续扩散；那时的瑞典数学家、斯德哥尔摩大学教授 *Torsten Carleman*（1892/7/8–1949/1/11；以 *Carleman* 条件、*Carleman* 不等式、*Denjoy-Carleman* 定理、平均遍历定理、*Carleman* 核、*Carleman* 公式等闻名）甚至表示，犹太人和外国人应当被处死。1939 年 *Feller* 来到美国，先后在 *Brown* 大学、*Cornell* 大学任职。1950 年他成为普林斯顿大学教授。

Feller 是二十世纪最伟大的概率学家之一。根据瑞典统计学家 *Harald Cramér*（1893/9/25–1985/10/5）的说法，在二十世纪中叶，概率论研究在法国和俄罗斯、数理统计研究在英国和美国各自盛行。*Gian-Carlo Rota*（1932/4/27–1999/4/18；美籍意大利数学家、哲学家）评价 *Feller* 的两卷本概率论教材是“概率论方面最成功的专著”。他也培养了很多学生，如 *Patrick Billingsley*（1925/5/3–2011/4/22；美国数学家、演员，美国艺术与科学院院士），*George Forsythe*（1917/1/8–1972/4/9；斯坦福计算机系创立者、系主任，曾任计算机协会理事），*Henry McKean*（1930–；美国数学家，美国科学院院士，纽约大学教授），*Lawrence Shepp*（1936/9/9–2013/4/23；美国数学家，专长是统计和计算射线成像技术），*Hale Trotter*（1931/5/30–；美籍加拿大数学家，以 *Lie-Trotter* 乘积公式、*Steinhaus-Johnson-Trotter* 算法、*Lang-Trotter* 猜测闻名），*Benjamin Weiss*（1941–；美籍以色列数学家，在遍历论、拓扑动力系统、概率论、博弈论、描述集合论等方向有贡献），*David A. Freedman*（1938/3/5–2008/10/17；著名数理统计学家，*UC Berkeley* 的统计学教授）等。

习 题 11

习题 11.1. 本题来自 [3, Lemma 8.0.1]。设以下依分布收敛成立：

$$(X_n - b_n)/a_n \xrightarrow{d} X,$$

其中 $a_n > 0, b_n \in \mathbb{R}$ 。则存在 $\alpha_n > 0, \beta_n \in \mathbb{R}$ 及随机变量 Y 使得

$$(X_n - \beta_n)/\alpha_n \xrightarrow{d} Y$$

的充要条件是：

$$a_n/\alpha_n \rightarrow a \in [0, \infty), \quad (b_n - \beta_n)/\alpha_n \rightarrow b \in \mathbb{R}.$$

此时，如果 $X \sim F$ ，则 $Y \stackrel{d}{=} aX + b$ 。当 $a > 0$ 时，我们称 X, Y 是同一分布类型，即认为

$$\mathcal{D}_F := \{F((\cdot - b)/a) : a > 0, b \in \mathbb{R}\}$$

是由 F 确定的一个分布族类型， X, Y 的分布函数都落入此分布族。

习题 11.2. 本习题讨论特征函数的连续性、可导性（或可微性）与对应随机变量的各阶矩之间的关系。设 X 是随机变量， $f(\lambda) := \mathbb{E}[e^{i\lambda X}]$ 为其特征函数。试证明：

(i) $f : \mathbb{R} \rightarrow \mathbb{C}$ 是一个有界、一致连续的复值函数；【提示：利用不等式 $|e^{i\theta} - 1| = 2|\sin \frac{\theta}{2}| \leq 2$ 以及控制收敛定理。】

(ii) 当对某正整数 $n \geq 1, \mathbb{E}[|X|^n] < \infty$ 时， f 具有从一阶直到 n 阶的导函数，且它们都是一致连续的，并且具有表达式：

$$f^{(k)}(\lambda) = \mathbb{E}[(iX)^k \cdot e^{i\lambda X}], \quad k = 1, \dots, n.$$

特别的, 我们有 $f^{(k)}(0) = i^k \cdot \mathbb{E}[X^k], k = 1, \dots, n$ 。【提示: 我们有估计式

$$|e^{i\theta} - \sum_{k=0}^n \frac{(i\theta)^k}{k!}| \leq \min\left(\frac{2|\theta|^n}{n!}, \frac{|\theta|^{n+1}}{(n+1)!}\right).$$

在此基础上, 论证方法类似 (i)。】

- (iii) f 在 0 点具有 2 阶导数的充要条件是 $\mathbb{E}[X^2] < \infty$; 【提示: 充分性见 (ii), 只需论证必要性。容易借助 $\varepsilon - N$ 语言论证, 当 f 在 0 点具有 2 阶导数时

$$f^{(2)}(0) = \lim_{\lambda \downarrow 0} \frac{f(\lambda) + f(-\lambda) - 2f(0)}{\lambda^2}.$$

之后借助不等式 $1 - \cos \theta \geq K\theta^2$ 在 0 点附近的小邻域成立, K 为某对应正常数。】

- (iv) f 在 0 点具有 $2n$ 阶导数的充要条件是 $\mathbb{E}X^{2n} < \infty$; 【提示: 方法基本同 (iii)。】

习题 11.3. 设有相互独立随机变量 $X_1 \sim F_1, X_2 \sim F_2$, 其中 F_1, F_2 为分布函数。求证: F_1 是连续函数蕴含了 $X_1 + X_2$ 的分布函数也是连续函数; X_1 是连续型的蕴含了 $X_1 + X_2$ 也是连续型的。

习题 11.4. 设 f, g 是特征函数。求证: $f \cdot g, pf + (1-p)g$ (其中 $p \in (0, 1)$)、 \bar{f} 、 $\operatorname{Re}(f)$ 、 $|f|^2$ 都是特征函数。本书中, $\operatorname{Re}(z), \operatorname{Im}(z)$ 分别表示复数 z 的实部与虚部。

习题 11.5. 随机变量 ξ 服从格分布或格点分布, 如果存在 $a \neq 0, b \in \mathbb{R}$, 使得 $\mathbb{P}(\xi \in a\mathbb{Z} + b) = 1$ 。请证明: ξ 服从格分布当且仅当其特征函数 f 满足: $\exists \lambda_0 \neq 0, s.t. |f(\lambda_0)| = 1$ 。特别的, 如果 ξ 服从格分布且非退化, 那么存在 $\lambda_0 \neq 0$, 使得 $\{\lambda \in \mathbb{R} : |f(\lambda)| = 1\} = \lambda_0\mathbb{Z}$ 。最后这个结论如果改成 $\exists \lambda_0 \neq 0, \{\lambda \in \mathbb{R} : |f(\lambda)| = 1\} \subset \lambda_0\mathbb{Q}$, 证明是简单的。本处这个形式的结论的证明需要用到数论中有关最大公约数的知识。

习题 11.6. 设随机变量 ξ 非退化, 其特征函数为 f 。求证: $\{\lambda \in \mathbb{R} : |f(\lambda)| = 1\}$ 是 Lebesgue 零测集 (实际上是可数集)。【提示: 上一习题结论。】

习题 11.7. 求证: $\forall a > 0, f_a(\lambda) := (1 - \frac{|\lambda|}{a})_+$ 是一个特征函数。

【提示: 它对应的密度函数为 $\frac{1 - \cos(ax)}{\pi ax^2}$, 它被称为参数 a 的 Polyá 分布。】

习题 11.8. (Polyá) 设 $f: \mathbb{R} \rightarrow \mathbb{R}$ 是偶函数, $f(0) = 1$, 且在 $[0, \infty)$ 上它是非负连续递减的凸函数。求证: f 是特征函数。

习题 11.9. 利用上一题, 证明: $\forall \alpha \in (0, 1], f(\lambda) := e^{-|\lambda|^\alpha}$ 是特征函数。

习题 11.10. 求证: $\forall \alpha \in (0, 2], f(\lambda) := e^{-|\lambda|^\alpha}$ 是特征函数。

【提示 (F. Spitzer 的论证方法): 对于 $\alpha \in (0, 2]$, 论证 $\psi(t) := 1 - (1 - \cos t)^{\alpha/2}$ 是特征函数, 并且 $e^{-|t|^\alpha} = \lim_{n \rightarrow \infty} [\psi(\frac{\sqrt{2}t}{n^{1/\alpha}})]^n$ 。】

习题 11.11. 本习题的目的是想说明, 如果把独立同分布条件修改为两两独立、同分布条件, 则经典的中心极限定理中结论不再成立。设 $\{\xi_n\}_{n=0}^\infty$ 是 *i.i.d.* 的随机变量列, 满足 $\mathbb{P}(\xi_0 = 1) = \mathbb{P}(\xi_0 = -1) = \frac{1}{2}$ 。令 $X_1 := \xi_0, X_2 := \xi_0 \cdot \xi_1$; 对 $m = 2^{n-1} + j, 1 \leq j \leq 2^{n-1}$, 递归的定义 $X_m := \xi_n X_j$ 。求证:
(1) $\{X_n\}_{n=1}^\infty$ 两两独立且同分布, $\mathbb{E}X_1 = 0, \text{Var}(X_1) = 1$;
(2) $S_{2^n} = \xi_0(1 + \xi_1) \cdots (1 + \xi_n)$ 的分布律是

$$\mathbb{P}(S_{2^n} = 0) = 1 - \frac{1}{2^n}, \quad \mathbb{P}(S_{2^n} = \pm 2^n) = \frac{1}{2^{n+1}},$$

从而 $\sqrt{n}\bar{X}_n$ 不可能依分布收敛到正态分布。

习题 11.12. 给定 $\{X_n\}_1^\infty \stackrel{\text{i.i.d.}}{\sim} \text{Cauchy}(0, 1)$ 。假定 $b_n \nearrow \infty$ 。定义 $X_{n,k} := X_k \cdot 1_{\{|X_k| \leq n/b_n\}}$ 以及 $S'_n = \sum_{k=1}^n X_{n,k}$ 。求证: $Z_n := \sqrt{\frac{\pi b_n}{2}} \cdot \frac{S'_n}{n} \xrightarrow{d} N(0, 1)$ 。

【提示: 特征函数方法; 也可以考虑使用 *Lindeberg-Feller* 中心极限定理。】

习题 11.13. 设 ξ 的分布函数是 F , 求证:

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T e^{-ia\lambda} \hat{\mu}_F(\lambda) d\lambda = \mathbb{P}(\xi = a),$$

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |\hat{\mu}_F(\lambda)|^2 d\lambda = \sum_{a \in \mathbb{R}} [\mathbb{P}(\xi = a)]^2.$$

习题 11.14. 如果随机变量 ξ 的密度函数 ρ 是两次连续可微的紧支撑函数, 那么 $\int |\hat{\rho}| d\lambda < \infty$ 。【提示: *Fourier* 变换的结论结合 *Cauchy* 不等式。此处“二次连续可微”条件可以减弱为“连续可微”, 紧支撑条件仍然保留。】

习题 11.15. 如果随机变量 ξ 具有密度函数 ρ , $\hat{\rho}$ 在无穷远处趋于 0。【提示: 数学分析中的 *Riemann-Lebesgue* 引理。】

习题 11.16. 证明正态分布 $N(\mu, \sigma^2)$ 的各阶矩唯一确定其分布。

习题 11.17. 证明 *Poisson* 分布的各阶矩唯一确定其分布。

习题 11.18. 证明指数分布的各阶矩唯一确定其分布。

习题 11.19. 证明对数标准正态分布的各阶矩不能唯一确定其分布。

习题 11.20. 对有界连续函数 f , 令 $I_f(x) := \int_0^x f(t) dt$ 。如果 f 是特征函数, 求证:

$$|f(x + \delta) - f(x)|^2 \leq 2(1 - \text{Re}(f(\delta))),$$

$$\left| \frac{I_f(x + \delta) - I_f(x - \delta)}{2\delta} \right|^2 \leq \frac{1 + \text{Re}(f(\delta))}{2}.$$

习题 11.21. 设 $\{f_n\}_{n=1}^\infty$ 为特征函数列, 满足 $f_n \rightarrow g_\delta$ 在区间 $(-\delta, \delta)$ 上收敛, 并且 g_δ 在 0 处连续。求证:

(1) $\{f_n\}_{n=1}^\infty$ 等度连续, 并且 $f_n \rightarrow g_\delta$ 在区间 $(-\delta, \delta)$ 上的收敛是一致的; 【提示: 反证法。】

(2) 函数 g_δ 可以延拓为一个特征函数 g , 并且当延拓唯一时, 我们有 $f_n \rightarrow g$ 在 \mathbb{R} 上成立。【提示: 此处需要使用 *Bochner-Khintchine* 定理。更多细节请参考 [25]。】

习题 11.22. 设 g_δ 是特征函数 g 在区间 $(-\delta, \delta)$ 上的限制, 其中 $\delta > 0$. 如果 g_δ 是解析函数或解析函数的边界函数, 那么 g_δ 唯一决定了 g . 【注: 正态分布的特征函数就是一个满足本习题条件的典型例子. 结合上一习题, 在使用特征函数方法论证中心极限定理时, 只需论证收敛在包含 0 的一个有界区间内成立即可.】

习题 11.23. 设随机变量列 $\{\xi_n : n \geq 1\}$ 的特征函数列 $\{f_n : n \geq 1\}$ 满足: $f_n(t) \rightarrow 1$ 在区间 $[-\varepsilon, \varepsilon]$ 上成立, 其中 $\varepsilon > 0$ 为某正常数. 试证明: ξ_n 依概率收敛到 0.

【提示: 利用(11.6)来论证. 习题 11.20–11.22 取材于 [25] 中的一些命题. 本题也可借用习题 11.20 中的前一估计式来证明. 后一方法的好处是可以把本题中的区间 $[-\varepsilon, \varepsilon]$ 替换为一般的正 Lebesgue 测度集 A 而保持结论不变.】

习题 11.24. 本习题讨论特征函数与非负定性, 由此给出 Böckner-Khintchine 定理的另一种证明 (充分性部分). 此处只考虑一维情况, 高维类似. 以下设 $f: \mathbb{R} \rightarrow \mathbb{C}$ 是一个非负定函数, 那么

(i) $f(0) \geq 0$, $f(-t) = \bar{f}(t)$, 且 $|f(t)| \leq f(0)$;

(ii) 给定 $c > 0$. 存在支撑在 $[-\frac{\pi}{c}, \frac{\pi}{c}]$ 上的有限测度 μ , 使得它的特征函数 $\hat{\mu}(t) := \int e^{it} d\mu$ 满足: $\hat{\mu}|_{c\mathbb{Z}} = f|_{c\mathbb{Z}}$; 【提示: 考察 $\mu_m(dx) := \frac{c}{2\pi} G_m(x) \cdot 1_{[-\frac{\pi}{c}, \frac{\pi}{c}]}(x)$, 其中 $G_m(x) := \frac{1}{m} \sum_{0 \leq k, \ell < m} f(c(k-\ell)) e^{-ic(k-\ell)x}$ 非负, 可算出 $f(0) = \mu_m(\mathbb{R}) = \mu_m([-\frac{\pi}{c}, \frac{\pi}{c}])$. 于是再取子列 $\mu_{m'}$ 弱收敛到某 μ , 那么不难证明 μ 就满足此处的要求. 此处可以使用 Helly 定理及特征函数的连续性定理等来完成结论 (ii) 的论证.】

(iii) 基于上面结论, 立即知道, 如果 f 是在 0 处连续的非负定函数, 那么它是某个有限测度的特征函数. 特别的, 当 $f(0) = 1$ 时, 这也给出了 Böckner-Khintchine 定理的另一种证明.

习题 11.25. 设特征函数 f 在 0 处有展开: $f(x) = 1 + o(x^2)$, 求证: $f(x) = 1$.

习题 11.26. 求证: 对于 $\alpha > 2$, $f(\lambda) := e^{-|\lambda|^\alpha}$ 不是特征函数. 【注: 此结论本质上源于 J. Marcinkiewicz 的工作, 他证明了: 如果 $P(t)$ 是多项式, 且 $e^{P(t)}$ 是特征函数, 那么 P 至多是二次多项式.】

习题 11.27. 求证: $f(\lambda) := |\cos \lambda|$ 不是特征函数 (但 $\cos \lambda$ 是特征函数).

习题 11.28. 给出定理 11.4.1 必要性部分的证明.

习题 11.29. 设有密度函数 ρ , 它的特征函数 $\hat{\rho}$ 非负可积, 那么 ρ 在 0 处连续且严格正, 并且 $\frac{\rho(t)}{\rho(0)}$ 也是一个特征函数.

【注: 这是一个有趣的现象. 建议读者自行研究几组经典案例, 比如说: (1) 标准 Cauchy 分布与所谓的标准双侧指数分布, 后者的密度函数为 $\rho(x) := \frac{1}{2} e^{-|x|}$; (2) Polya 分布 (见习题 11.7) 与所谓的三角形分布或帐篷分布, 后者的密度函数形态为: $\rho_a(x) = \frac{1}{a} \cdot (1 - \frac{|x|}{a})_+$.】

习题 11.30. 在统计学中有一个著名的 Cramer-Wold Device. 它可以有两层含义, 陈述如下:

- (1) 设 X, Y 是随机变量。假设对任意 $\alpha, \beta \in \mathbb{R}$, 有已知的分布函数 $F_{\alpha, \beta}$, 满足 $\alpha X + \beta Y \sim F_{\alpha, \beta}$, 那么 (X, Y) 的联合分布 F 是由分布函数族 $\{F_{\alpha, \beta} : \alpha, \beta \in \mathbb{R}\}$ 唯一确定的;
- (2) 设 X_n, Y_n 是随机变量序列。假设对任意 $\alpha, \beta \in \mathbb{R}$, 有已知的分布函数 $F_{\alpha, \beta}$, 满足 $\alpha X_n + \beta Y_n \xrightarrow{d} F_{\alpha, \beta}$, 那么 (X_n, Y_n) 是依分布收敛到由 (1) 中确定的分布 F 。

你能说出 *Cramer-Wold Device* 成立的原因吗?

习题 11.31. 对随机变量 X 及任意 $\lambda \geq 0$, 记

$$Q_X(\lambda) := \sup_x \mathbb{P}(x \leq X \leq x + \lambda) = \sup_x \mathbb{P}(|X - x| \leq \lambda/2),$$

在 [31, Chapter III] 中称之为 *concentration function*, 可以翻译为集中度函数; 显然它关于 λ 是单调不减的函数。试证明:

- (1) 对独立随机变量 X, Y , $Q_{X+Y}(\lambda) \leq \min\{Q_X(\lambda), Q_Y(\lambda)\}$;
- (2) $Q_X(n\lambda) \leq nQ_X(\lambda)$, $n \geq 1$ 为自然数;
- (3) 设 f_X 为 X 的特征函数, 那么对任意 $a, \lambda \geq 0$

$$Q_X(\lambda) \leq \frac{1}{aK(\frac{a\lambda}{4})^2} \int_{-a}^a |f_X(t)| dt,$$

其中函数 K 为 $K(\lambda) := \min\{|\frac{\sin x}{x}| : |x| \leq \lambda\}$ 。特别的,

$$Q_X(\lambda) \leq (\frac{96}{95})^2 \max(\lambda, \frac{1}{a}) \int_{-a}^a |f_X(t)| dt.$$

【提示: 考虑密度 $\rho(x) := (1 - |x|)^+$ 的特征函数 $\hat{\rho}(t) = \frac{4 \sin^2 \frac{t}{2}}{t^2}$, 设 $Y \sim \rho(x)dx$ 与 X 独立, 计算 $\mathbb{E}[e^{ia(X-Y)Y}]$ 得到

$$\int \hat{\rho}(a(x - \gamma)) dF_X(x) = \frac{1}{a} \int_{-a}^a e^{-i\gamma t} \rho(\frac{t}{a}) f_X(t) dt,$$

并基于此来估计 Q_X 。】

- (4) 对任意 $a, \lambda \geq 0$ 满足 $a\lambda \leq \frac{\pi}{4}$, 有 $Q_X(\lambda) \geq \frac{3}{8\pi} \cdot K(\frac{a\lambda}{2}) \int_{-a}^a |f_X(t)|^2 dt$ 。

特别的, 对任意 $a, \lambda \geq 0$ $Q_X(\lambda) \geq \frac{95\lambda}{256\pi(1+2a\lambda)} \cdot \int_{-a}^a |f_X(t)|^2 dt$ 。【提示: 考虑 X 的对称化 $X^s := X - Y$, 再取独立随机变量 U 使其特征函数为 $\rho(\frac{t}{4a})$, 令 $V = X^s + U$ 。于是 $Q_V(\lambda) \leq Q_X(\lambda)$, 但

$$\mathbb{P}(|V| \leq b) = \frac{1}{\pi} \int_{-4a}^{4a} |f_X(t)|^2 \cdot \frac{\sin bt}{t} \cdot (1 - \frac{|t|}{4a})^+ dt, \quad \forall b > 0.$$

基于此来进行有关估计。】

习题 11.32. 本习题的目的是引导读者温习 *De Moivre* 和 *Laplace* 关于 *Bernoulli* 分布对应的中心极限定理的有关讨论; 这里提供了主要思想框架, 细节就是作为习题留给读者补充的。以下 $\{X_n\}_1^\infty$ 是 *i.i.d.* 的 $B(1, p)$ 分布随机变量列, 则 $S_n := X_1 + \cdots + X_n \sim B(n, p)$ 。此处总假定 $p \in (0, 1), q := 1 - p$,

定义 $P_n(k) := C_n^k p^k q^{n-k}$ 。我们首先论证所谓的“局部极限定理”：对满足 $|k - np| = o((npq)^{2/3})$ 的所有 k ，一致有

$$P_n(k) \sim \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}} =: \bar{P}_n(k),$$

其含义是：当 $n \rightarrow \infty$ 时（此处一定程度上容许 $p = p_n, q = q_n$ 与 n 有关），对任意满足 $\varphi(n)/(npq)^{2/3} \rightarrow 0$ 的非负函数 φ 成立下式

$$\sup\left\{\left|\frac{P_n(k)}{\bar{P}_n(k)} - 1\right| : |k - np| \leq \varphi(n)\right\} \rightarrow 0.$$

这可以按照以下步骤来实现论证：

(a) 记 $\hat{p} := k/n$ ，则根据 *Stirling* 公式，当 $n \rightarrow \infty, k \rightarrow \infty, n - k \rightarrow \infty$ 时，

$$\varepsilon_n := \varepsilon_n(k, n - k) = C_n^k \cdot \sqrt{2\pi n \hat{p}(1 - \hat{p})} \hat{p}^k (1 - \hat{p})^{n-k} - 1 \rightarrow 0;$$

(b) $P_n(k) = \frac{1 + \varepsilon_n}{\sqrt{2\pi n \hat{p}(1 - \hat{p})}} e^{-nH_p(\hat{p})}$ ，其中 $H_p(x) := x \log \frac{x}{p} + (1 - x) \log \frac{1-x}{q}$ ；

(c) 当 n 充分大时， $H_p(\hat{p}) = \frac{1}{2pq}(\hat{p} - p)^2 + O(|\hat{p} - p|^3)$ ；

(d) $P_n(k) = \bar{P}_n(k) \cdot [1 + \bar{\varepsilon}_n]$ ，其中

$$1 + \bar{\varepsilon}_n(k, n - k) = [1 + \varepsilon_n(k, n - k)] e^{nO(|\hat{p} - p|^3)} \sqrt{\frac{pq}{\hat{p}(1 - \hat{p})}};$$

(e) 对于前述 φ ， $\sup\{\bar{\varepsilon}_n(k, n - k) : |k - np| \leq \varphi(n)\} \rightarrow 0$ 。

上述“局部极限定理”结果也可以简单的表述为

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{npq}} = x\right) \sim \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2}}, \quad x = o((npq)^{1/6}).$$

当然上面的 x 要求能使得 $np + x\sqrt{npq}$ 为非负整数。

基于“局部极限定理”就可以进一步论证中心极限定理；此时 $p \in (0, 1)$ 为常数。论证方法如下：记 $t_k := t_k^{(n)} = \frac{k - np}{\sqrt{npq}}$ ， $\Delta t_k := t_{k+1} - t_k = \frac{1}{\sqrt{npq}}$ 。那么“局部极限定理”结果也可以表述为

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{npq}} = t_k\right) \sim \frac{\Delta t_k}{\sqrt{2\pi}} e^{-\frac{t_k^2}{2}}, \quad t_k = o((npq)^{1/6}).$$

确切的说，

$$P_n(np + t_k \sqrt{npq}) = \frac{\Delta t_k}{\sqrt{2\pi}} e^{-\frac{t_k^2}{2}} \cdot [1 + \bar{\varepsilon}_n(t_k)],$$

其中对任意 $T > 0$ ， $\sup\{|\bar{\varepsilon}_n(t_k)| : |t_k| \leq T\} \rightarrow 0$ 。记

$$P_n(a, b] := \sum_{t_k \in (a, b]} P_n(np + t_k \sqrt{npq}), \quad \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du.$$

在上述结果基础上，进一步论证： $n \rightarrow \infty$ 时

(f) $\sup\{|P_n(a, b] - [\Phi(b) - \Phi(a)]| : -T \leq a \leq b \leq T\} \rightarrow 0$ ；

(g) 上式对 $T = \infty$ 也成立。

由此就证明了所谓的“*De Moivre-Laplace* 积分定理”（即中心极限定理）：

$$\sup\{|P_n(a, b] - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du| : -\infty \leq a \leq b \leq \infty\} \rightarrow 0.$$

原则上做的更精细一点，应能估计出如下逼近的速度（这是 *Bernoulli* 分布情形的所谓 *Berry-Esseen* 估计）

$$\sup_x |P_n(-\infty, x] - \Phi(x)| \leq \frac{p^2 + q^2}{\sqrt{npq}}.$$

为了得到 *Bernoulli* 的弱大数律，只需注意到，在上述结果的基础上，

$$\mathbb{P}(|\frac{S_n}{n} - p| \leq \varepsilon) = \mathbb{P}(|\frac{S_n - np}{\sqrt{npq}}| \leq \varepsilon \sqrt{\frac{n}{pq}}) \approx 2\Phi(\varepsilon \sqrt{\frac{n}{pq}}) - 1 \rightarrow 1.$$

§ 12

检验我们的概率建模：步入《统计学》！

马克思主义理论中强调，“实践是检验真理的（唯一）标准”。对于我们的概率建模来说，我们的模型是否准确可以通过收集大量来自随机实验的数据进行检验，这意味着我们将进入现代意义上的《统计学》的有关领域。

事实上，在本书的若干章节已经涉及到了《统计学》的部分初步内容，提及了《统计学》中的若干理论思想与方法（如小概率事件原理、极大似然原理等思想，极大似然估计、矩估计、Bayes 方法等方法）。例如，第 5 章的 §5.2.3 介绍了两个离散分布的参数的极大似然估计；第八章的例 8.16 和例 8.17 探讨了基于单个正态分布总体的样本如何给出均值参数、方差参数的极大似然估计，并探寻出有关这些估计的精确分布律（这同样可以用来检验我们的模型假设），例 8.18 基于两个不同的正态分布总体的样本数据，探讨了如何构建有关方差参数的统计量并得到它的精确分布律用于检验这两个正态分布总体是否具有相同的方差。这些参数估计和假设检验的问题都是现代《统计学》中的典型问题；由于课程安排的原因，我们在本书中并没有仔细展开讨论。

本章我们将主要基于前面两章给出的极限定理，构建大样本下的对有关模型假设的检验。本章中，我们关心的核心问题是以下两个典型问题：

- (1) 给定（一维）大样本数据 $\{x_k\}_{k=1}^n$ ，设它们来自随机变量 $X \sim F$ （其中 F 已知）的独立同分布样本 $\{X_k\}_{k=1}^n$ （即此处承认 $\{X_k\}_{k=1}^n$ 的相互独立性），如何检验这些数据来源于同一个分布总体 F ，即要对假设 $\{X_k\}_{k=1}^n \stackrel{\text{i.i.d.}}{\sim} F$ 中“具有共同的分布函数 F ”的这部分假设做检验（此处对独立性不做检验）；
- (2) 给定（二维）大样本数据 $\{(x_k, y_k)\}$ ，设它们来自随机向量 $(X, Y) \sim F$ （此处 F 未知）的独立同分布样本 $\{(X_k, Y_k)\}$ （即此处承认 $\{(X_k, Y_k)\} \stackrel{\text{i.i.d.}}{\sim} (X, Y)$ ），如何检验 X, Y 之间的相互独立性？

上述问题（1）是对总体分布的分布律的检验；问题（2）是对独立性的检验。这两个检验问题正好覆盖了我们概率论课程中有关随机变量的概念中最重要两个方面：分布律与独立性。

12.1 对总体分布的分布律的检验

12.1.1 定性检验：P-P 图或 Q-Q 图

强大数律在统计学中的一个更重要应用是对样本经验函数（作为分布函数的估计）的讨论。设 $\{X_k\}_{k=1}^{\infty} \stackrel{\text{i.i.d.}}{\sim} F$ ，其中 F 是一个概率分布函数。定义

$$\hat{F}_n(x) := \frac{1}{n} \sum_{k=1}^n 1_{\{X_k \leq x\}}.$$

它称为前 n 个样本对应的**经验分布函数**。根据强大数律，容易知道

$$\hat{F}_n(x) \xrightarrow{\text{a.s.}} F(x)$$

对任意固定的 $x \in \mathbb{R}$ 成立。令

$$D_n := \|\hat{F}_n - F\|_{\infty} = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|, \quad (12.1)$$

它被称为 **Kolmogorov-Smirnov 统计量**。下面的 Glivenko-Cantelli 定理告诉我们，上述收敛 $\hat{F}_n(x) \xrightarrow{\text{a.s.}} F(x)$ 关于 x 是一致收敛。

定理 12.1.1. (Glivenko-Cantelli 定理) 对于任意给定分布函数 F 以及样本 $\{X_k\}_{k=1}^{\infty} \stackrel{\text{i.i.d.}}{\sim} F$ 的 Kolmogorov-Smirnov 统计量 D_n ，我们有

$$D_n \xrightarrow{\text{a.s.}} 0.$$

上述定理被统计学文献称为“**统计基本定理**”（fundamental statistical theorem），最早是 Glivenko 在 1933 年对连续的分布函数 F 、Cantelli 在 1933 年对一般的分布函数 F 给出的。统计学中用来检验样本是否服从给定分布 F 的 P-P 图（或 Q-Q 图）的方法本质上就依赖此定理：当 F 是连续的分布函数时，只需注意到对于样本 $\{X_k\}_{k=1}^n \stackrel{\text{i.i.d.}}{\sim} F$ 对应的次序统计量 $\{X_n^{(k)}\}_{k=1}^n$ ， $\hat{F}(X_n^{(k)}) = \frac{k}{n}$ ，因此为检验这些样本确实来自于分布 F ，只需在二维平面上作图检验点列 $\{(\frac{k}{n}, F(X_n^{(k)}))\}_{k=1}^n$ （或点列 $\{(F^{-1}(\frac{k}{n}), X_n^{(k)})\}_{k=1}^n$ ，其中 F^{-1} 表示广义逆）近似位于直线 $y = x$ 上*。P-P 图（或 Q-Q 图）的方法对于离散型分布也可以使用，只是需要略作调整。

定理 12.1.1 的证明。 根据强大数律，容易知道

$$\hat{F}_n(x) \xrightarrow{\text{a.s.}} F(x), \quad \hat{F}_n(x-) \xrightarrow{\text{a.s.}} F(x-), \quad \forall x \in \mathbb{R}.$$

当 F 是连续函数时，定理的结论 $\sup\{|\hat{F}_n(x) - F(x)| : x \in \mathbb{R}\} \rightarrow 0$ 证明可以更简单，留给读者；但我们现在的分布函数 F 是一般的，有可能带有跳跃点。为此，对任意给定正整数 $m \geq 2$ ，对任意 $1 \leq j < m$ ，定义 $x_m^j := F^{-1}(\frac{j}{m})$ 。那么存在 $N_m : \Omega \rightarrow \mathbb{N}$ 使得：对于几乎处处的 ω 及 $n \geq N_m$ ，我们有

$$|\hat{F}_n(x_m^j) - F(x_m^j)| < \frac{1}{m}, \quad |\hat{F}_n(x_m^j-) - F(x_m^j-)| < \frac{1}{m},$$

其中 $j = 1, \dots, m-1$ 。我们约定 $x_m^0 := -\infty, x_m^m = +\infty$ ，于是上式中 j 可以取到 0 和 m 。现在，对任意 $x \in \mathbb{R}$ ，存在唯一 $j \in [1, m]$ ，使得 $x \in (x_m^{j-1}, x_m^j]$ ，

*在实际应用中，对应考虑的点列略作了调整，做 P-P 图时使用点列 $\{(\frac{2k-1}{2n}, F(X_n^{(k)}))\}_{k=1}^n$ ；做 Q-Q 图时使用点列 $\{(F^{-1}(\frac{2k-1}{2n}), X_n^{(k)})\}_{k=1}^n$ 。

此时利用 $F(x_m^j-) - F(x_m^{j-1}) \leq \frac{1}{m}$, 我们有

$$\begin{aligned}\hat{F}_n(x) - F(x) &\leq \hat{F}_n(x_m^j) - F(x) \\ &\leq F(x_m^j-) + \frac{1}{m} - F(x) \\ &\leq F(x_m^j-) - F(x_m^{j-1}) + \frac{1}{m} \leq \frac{2}{m}.\end{aligned}$$

同理, $F(x) - \hat{F}_n(x) \leq \frac{2}{m}$. 因此

$$D_n = \sup_x |\hat{F}_n(x) - F(x)| \leq \frac{2}{m}, \quad \forall n \geq N_m.$$

令 $m \rightarrow \infty$ 就证明了定理的结论. \square

容易知道, 对于任意固定的 x , 经验分布函数 $\hat{F}_n(x)$ 满足中心极限定理:

$$G_n(x) := \sqrt{n}[\hat{F}_n(x) - F(x)] \xrightarrow{d} N(0, F(x)(1 - F(x))). \quad (12.2)$$

此处, $F(x) = 0$ 或 1 时, 上述极限分布理解为退化到 0 的单点分布。

注意到用于检验总体分布律的 P-P 图或 Q-Q 图的方法只是一种形象的、定性的检验方法, 我们通常还需要更定量的检验方法, 使得我们有一定的把握 (比如 95% 的 “概率”) 确认或否认总体的分布律是已知的分布律 F 。这时候我们需要对分布律 F 分情况来进行讨论。

12.1.2 定量检验: 分布函数 F 连续的情形

在分布函数 F 连续的情形, Kolmogorov 证明了下面定理中的依分布收敛部分, Smirnov 计算出了对应的极限分布律 (它被称为 Kolmogorov 分布); 有关结果的证明可以参见 [58, §4.2.3]。

定理 12.1.2. (Kolmogorov 定理) 设分布函数 F 连续。对样本 $\{X_k\}_{k=1}^\infty \stackrel{\text{i.i.d.}}{\sim} X \sim F$ 的 Kolmogorov-Smirnov 统计量 D_n , 我们有

$$\sqrt{n}D_n \xrightarrow{d} K := \sup_{t \in [0,1]} |B_t|,$$

其中 $\{B_t : t \in [0, 1]\}$ 表示标准 Brown 运动。从而*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\sqrt{n}D_n \leq x) = 1 + 2 \sum_{j=1}^{\infty} (-1)^j \cdot e^{-2(jx)^2}, \quad \forall x > 0. \quad (12.3)$$


当 n 很大时 ($n \geq 100$), 利用上面的定理, 我们可以构建 Kolmogorov 检验: 对于指定的 $\alpha \in (0, 1)$, 取 K 的分布的 $1-\alpha$ 分位数 K_α , 即 $\mathbb{P}(K \leq K_\alpha) = 1-\alpha$, 当 $\sqrt{n}D_n \leq K_\alpha$ 时, 就差不多有 $1-\alpha$ 的把握认为原始假设 $X \sim F$ 成立; 否则拒绝原始假设。

当 n 不算太大时 ($n \leq 100$), 注意到 D_n 的精确分布律其实与 F 无关, 因此可以借助 Owen 在 1962 年制作的表格 [29], 同样可对于指定的 $\alpha \in (0, 1)$, 取 $\sqrt{n}D_n$ 的分布的 $1-\alpha$ 分位数 $x_{n,\alpha}$, 即 $\mathbb{P}(\sqrt{n}D_n \leq x_{n,\alpha}) = 1-\alpha$, 当 $\sqrt{n}D_n \leq x_{n,\alpha}$ 时, 就有 $1-\alpha$ 的把握认为原始假设 $X \sim F$ 成立; 否则拒绝原始假设。

* K 的分布函数的另一种表达式是 $\mathbb{P}(K \leq x) = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} \exp\{-(2k-1)^2\pi^2/(8x^2)\}$ 。

12.1.3 定量检验：分布函数 F 离散的情形

下面的 Pearson 定理可以视作 Kolmogorov 定理在离散型分布情形的对应。在统计学中，Pearson 定理通常用来进行拟合优度检验。

 **定理 12.1.3. (Pearson 定理)** 设 $\{X_k\}_{k=1}^n \stackrel{\text{i.i.d.}}{\sim} \begin{pmatrix} 1 & \cdots & m \\ p_1 & \cdots & p_m \end{pmatrix}$, 其中 $2 \leq m < \infty$ 。记

$$n_j := \sum_{k=1}^n 1_{\{X_k=j\}}, \quad \hat{p}_j := \frac{n_j}{n}.$$

那么

$$Q := \sum_{j=1}^m \frac{n}{p_j} \cdot (\hat{p}_j - p_j)^2 \xrightarrow{d} \chi^2(m-1). \quad (12.4)$$

证明. 首先, 注意到

$$\hat{p}_j = \frac{1}{n} \sum_{k=1}^n 1_{\{X_k=j\}}$$

以及 $\mathbb{E}[1_{\{X_k=j\}}] = p_j$, $\text{Var}(1_{\{X_k=j\}}) = p_j(1-p_j)$, 根据中心极限定理

$$W_j := \sqrt{\frac{n}{p_j}} \cdot (\hat{p}_j - p_j) \xrightarrow{d} N(0, 1-p_j).$$

先讨论 $W := (W_1, \dots, W_m)^T$ 的极限分布律。这里可以使用 Cramer-Wold Device (见本章习题 11.30): 对任意 $a = (a_1, \dots, a_m)^T$,

$$\begin{aligned} a^T W &= \sum_{j=1}^m a_j \cdot W_j \\ &= \sum_{j=1}^m a_j \cdot \sqrt{\frac{n}{p_j}} \cdot \frac{1}{n} \sum_{k=1}^n (1_{\{X_k=j\}} - p_j) \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^n \sum_{j=1}^m \frac{a_j}{\sqrt{p_j}} \cdot (1_{\{X_k=j\}} - p_j) \xrightarrow{d} N(0, a^T B a), \end{aligned}$$

其中 $B = (B_{i,j})_{1 \leq i,j \leq m}$ 是一个矩阵, 此时就可以认为 $W \xrightarrow{d} N(0, B)$ 。可以算出

$$\begin{aligned} a^T B a &= \text{Var}\left(\sum_{j=1}^m \frac{a_j}{\sqrt{p_j}} \cdot (1_{\{X_1=j\}} - p_j)\right) \\ &= \sum_{i=1}^m a_i^2 (1-p_j) - 2 \sum_{1 \leq i < j \leq m} a_i a_j \cdot \sqrt{p_i p_j}. \end{aligned}$$

即 $B = I - \alpha \alpha^T$, 其中 $\alpha = (\sqrt{p_1}, \dots, \sqrt{p_m})^T$ 。

定理中要讨论的量为

$$Q := \sum_{j=1}^m \frac{n}{p_j} \cdot (\hat{p}_j - p_j)^2 = \sum_{j=1}^m W_j^2 = \|W\|^2.$$

注意到总有

$$\sum_{j=1}^m \sqrt{p_j} W_j = 0,$$

存在正交矩阵 A , 使得它的第一行元素如下指定:

$$A = \begin{pmatrix} \sqrt{p_1} & \cdots & \sqrt{p_m} \\ * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}.$$

此时 $A\alpha = (1, 0, \dots, 0)^T, ABA^T = A(I - \alpha\alpha^T)A^T = \text{diag}\{0, I_{m-1}\}$. 令

$$Z = AW.$$

此时 $Z_1 = 0$; $W \xrightarrow{d} N(0, B)$ 意味着 $Z \xrightarrow{d} N(0, ABA^T)$. 于是

$$(Z_2, \dots, Z_m)^T \xrightarrow{d} N(0, I_{m-1}).$$

现在 $Q = \|W\|^2 = \|Z\|^2 = Z_2^2 + \dots + Z_{m-1}^2 \xrightarrow{d} \chi^2(m-1)$. \square

12.2 对独立性的检验

12.2.1 技术准备：关于卡方分布的讨论

利用特征函数这个工具, 我们可以进一步讨论 χ^2 -分布. 此处, 若 $\{X_k\}_{k=1}^n \stackrel{\text{i.i.d.}}{\sim} N(\mu, 1)$, 则记 $Y := X_1^2 + \dots + X_n^2 \sim \chi^2(n; \delta)$ (其中 $\delta := \sqrt{n}|\mu|$), 并称 Y 服从 n 个自由度的卡方分布, 中心参数为 δ ; 当 $\delta = 0$ (即 $\mu = 0$) 时就简单记作 $Y \sim \chi^2(n)$, 称 Y 服从 n 个自由度的 (中心) 卡方分布.

定理 12.2.1. 以下二次型中涉及的矩阵 A, A_1, A_2 都是实对称矩阵. 关于 χ^2 -分布, 我们有下面的结果:

- (1) 设 $X_k \sim N(a_k, 1), k = 1, 2, \dots, n$, 且它们相互独立, 记 $X = (X_1, \dots, X_n)^T$. 给定 n 阶实对称矩阵 A

$$Y := X^T A X$$

服从 χ^2 -分布的充分必要条件是 A 是幂等矩阵, 即 $A^2 = A$. 此时 $Y \sim \chi^2(r; \delta)$, 其中 $r = \text{rank}(A)$, $\delta^2 = a^T A a$.

- (2) 设 $Y = X^T A X, Y_1 = X^T A_1 X$, 其中 $X \sim N(a, I_n)$. 如果 Y, Y_1 分别服从自由度为 m, m_1 的 χ^2 -分布 (不必是中心的), 并且 $A_2 := A - A_1 \geq 0$ (即半正定), 且 $A_2 \neq 0$, 那么

$$Y_2 := X^T A_2 X$$

服从自由度为 $m_2 = m - m_1$ 的 χ^2 -分布, 并且 Y_1, Y_2 相互独立.

- (3) 设 $Y_i = X^T A_i X, i = 1, 2$ 都服从 χ^2 -分布*, 其中 $X \sim N(a, I_n)$. 则 Y_1, Y_2 相互独立的充要条件是 $A_1 A_2 = 0$.

*此处, Y_1, Y_2 都服从 χ^2 -分布这一条件可以去掉. 证明参见 Plackett [30].

☞ **定理 12.2.2.** (Cochran 定理) 设 $X \sim N(a, I_n)$, 并且

$$X^T X = \sum_{i=1}^m X^T A_i X,$$

其中 A_i 均为实对称矩阵。以下两条之间相互等价:

(1) 存在 $\{(n_i, \delta_i)\}_{i=1}^m$, 使得 $X^T A_i X \sim \chi^2(n_i; \delta_i), i = 1, \dots, m$, 并且它们相互独立;

(2) $\{A_i\}_{i=1}^m$ 的秩满足

$$\sum_{k=1}^m \text{rank}(A_k) = n. \quad (12.5)$$

当(12.5)成立时, $n_i = \text{rank}(A_i), \delta_i^2 = a^T A_i a, i = 1, \dots, m$.

以上定理的证明并不太难, 留给读者作为练习; 亦可参见 [41]。这几个结果在回归分析课程的方差分析理论中有重要应用价值。

12.2.2 Pearson-Fisher 定理

下面的定理在实务上用来进行独立性的列联表检验。

☞ **定理 12.2.3.** (Pearson-Fisher 定理) 设 $\{(X_k, Y_k)\}_{k=1}^n$ 是来自离散型分布 (X, Y) 的样本, 其中

$$\mathbb{P}(X = i, Y = j) = p_{i,j}, \quad 1 \leq i \leq K, 1 \leq j \leq L,$$

$2 \leq K, L < \infty$ 。记 $\mathbb{P}(X = i) = p_i, \mathbb{P}(Y = j) = q_j$, 并令

$$\hat{p}_{i,j} := \frac{1}{n} \sum_{k=1}^n 1_{\{X_k=i, Y_k=j\}}, \quad \hat{p}_i := \frac{1}{n} \sum_{k=1}^n 1_{\{X_k=i\}}, \quad \hat{q}_j := \frac{1}{n} \sum_{k=1}^n 1_{\{Y_k=j\}}.$$

那么当 X, Y 相互独立 (即 $p_{i,j} = p_i q_j, \forall i, j$) 时

$$V := \sum_{i=1}^K \sum_{j=1}^L \frac{n}{\hat{p}_i \hat{q}_j} \cdot (\hat{p}_{i,j} - \hat{p}_i \hat{q}_j)^2 \xrightarrow{d} \chi^2((K-1)(L-1)). \quad (12.6)$$

证明. 我们基于 Pearson 定理来证明此处的定理。这里我们将使用上一章习题 10.3 中的 $O_{\mathbb{P}}(1), o_{\mathbb{P}}(1)$ 记号及其性质, 并结合之前的 Cochran 定理来进行论证。

同上, 定义

$$W_{i,j} := \sqrt{\frac{n}{p_{i,j}}} \cdot (\hat{p}_{i,j} - p_{i,j}) \xrightarrow{d} N(0, 1 - p_{i,j}).$$

我们记

$$Q := \sum_{i,j} \frac{n}{\hat{p}_i \hat{q}_j} \cdot (\hat{p}_{i,j} - \hat{p}_i \hat{q}_j)^2$$

$$Q_* := \sum_{i,j} \frac{n}{p_{i,j}} \cdot (\hat{p}_{i,j} - \hat{p}_i \hat{q}_j)^2.$$

注意到 $n \cdot (\hat{p}_{i,j} - \hat{p}_i \hat{q}_j)^2 = O_{\mathbb{P}}(1)$, $\hat{p}_i = p_i + o_{\mathbb{P}}(1)$, $\hat{q}_j = q_j + o_{\mathbb{P}}(1)$, 容易知道

$$Q = Q_* + o_{\mathbb{P}}(1).$$

现在定义

$$\begin{aligned} Q_{X,Y} &:= \sum_{i,j} \frac{n}{p_{i,j}} \cdot (\hat{p}_{i,j} - p_{i,j})^2, \\ Q_X &:= \sum_i \frac{n}{p_i} \cdot (\hat{p}_i - p_i)^2, \\ Q_Y &:= \sum_j \frac{n}{q_j} \cdot (\hat{q}_j - q_j)^2. \end{aligned}$$

在假设 $p_{i,j} = p_i q_j, \forall i, j$ 下, (1) 中结果告诉我们,

$$Q_{X,Y} \xrightarrow{d} \chi^2(KL-1), Q_X \xrightarrow{d} \chi^2(K-1), Q_Y \xrightarrow{d} \chi^2(L-1).$$

现在不难计算出

$$Q_{X,Y} = Q_X + Q_Y + Q_*.$$

其中 Pearson 定理的证明过程中说明了存在线性的一一变换

$$\{W_{i,j} : 1 \leq i \leq K, 1 \leq j \leq L\} \mapsto \{Z_{\alpha}\}_{1 \leq \alpha \leq KL}$$

使得 $Z_1 \equiv 0, (Z_2, \dots, Z_{KL})^T \xrightarrow{d} N(0, I_{KL-1})$, 并且

$$Q_{X,Y} = Z_2^2 + \dots + Z_{KL}^2.$$

于是不难知道 Q_X, Q_Y, Q_* 均可以表达为 Z_2, \dots, Z_{KL} 的二次型。结合 Cochran 定理 (或定理 12.2.1 的 (3)), 立即知道 $Q_* \xrightarrow{d} \chi^2((K-1)(L-1))$ 。进而 $Q \xrightarrow{d} \chi^2((K-1)(L-1))$ 。□

读者在《统计学》的相关课程中将学习到更多假设检验的方法与理论。

如编者在“前言与导读”中所说, 本章中 Kolmogorov 定理这一重要结果的表述与证明都需要随机过程的语言与泛函中心极限定理等高端的概率理论, 而本书中多处 (特别是第 4 章) 内容限于教学安排无法细致展开, 因而对概率论真正感兴趣读者需要进一步阅读、学习初等概率论的后续高级理论的教材资料。

习 题 12

习题 12.1. (经验的总结与结论的外推?) 考虑反复抛掷一枚有可能不均匀的硬币得到正反面序列, 为方便表示为 1 与 0 的序列, 也就是说假设 $\{X_k\}_{k=1}^{n+1} \stackrel{\text{i.i.d.}}{\sim} B(1, p)$, 其中 $p \in (0, 1)$ 未知。现在已知 $S_n := X_1 + X_2 + \dots + X_n$ 的取值。

(1) 计算 $\mathbb{P}(X_i = 1 | S_n = k)$, 其中 $1 \leq i \leq n, 0 \leq k \leq n$;

(2) 计算 $\mathbb{P}(X_{i_1} = 1, \dots, X_{i_r} = 1 | S_n = k)$, 其中 $2 \leq r \leq k \leq n, 1 \leq i_1 < \dots < i_r \leq n$;

(3) 计算 $\mathbb{P}(X_{n+1} = 1 | S_n = k)$ 。

【容易知道, (1) 中 $\mathbb{P}(X_i = 1 | S_n = k) = \frac{k}{n}$; (1)、(2) 中结果均可以用 Polyá 无放回取球模型来解释。但根据模型假设, (3) 中 $\mathbb{P}(X_{n+1} = 1 | S_n = k) = p$,

与 k 值无关, 也就是说, 如果把 (1)、(2) 中的结果视作经验的总结, 把 (3) 中结果视作类比的外推, 那么这个外推在此处的模型假设下就是无效的。第一章中的赌徒佯谬、热手效应如果用这个模型来看可以给出一种解释。】

习题 12.2. (掷硬币实验的 *Bayes* 模型) 类似上述模型。只不过我们说现在已知有 $n+1$ 枚材质相同的均匀硬币, 但这个抛掷硬币的随机实验是“风洞实验”, 其中假定 P 是一个与“风洞”环境有关、与硬币无关的随机变量, 它使得抛掷单枚硬币后获得正面 (对应记录为 1, 否则记录为 0) 的实际概率值就是 $P \sim U(0, 1)$, 而 $\{X_k\}_{k=1}^{n+1}$ 就是在“风洞”中同时抛掷这 $n+1$ 枚硬币所获得的实验记录。数学上认为:

$$\{X_k\}_{k=1}^{n+1} | P = p \stackrel{\text{i.i.d.}}{\sim} B(1, p).$$

我们现在同样已知 $S_n := X_1 + \cdots + X_n = k$ 。

(1) 计算 $\mathbb{P}(X_i = 1 | S_n = k)$, 其中 $1 \leq i \leq n, 0 \leq k \leq n$;

(2) 计算 $\mathbb{P}(X_{i_1} = 1, \cdots, X_{i_r} = 1 | S_n = k)$, 其中 $2 \leq r \leq k \leq n, 1 \leq i_1 < \cdots < i_r \leq n$;

(3) 计算 $\mathbb{P}(X_{n+1} = 1 | S_n = k)$ 。

习题 12.3. (掷硬币实验的 *Bayes* 模型的 *Bayes* 估计问题) 考虑上一习题中的模型。在已知数据 X_1, \cdots, X_n 的条件下, 对 (随机) 参数 P 的极大似然估计方法如下: 首先写出条件分布律

$$P | (X_1 = x_1, \cdots, X_n = x_n).$$

如果它是一个连续型分布, 特别地如果它的密度函数是单峰的, 则对应密度函数的极大值点就作为对应的极大似然 *Bayes* 估计 \hat{P} ; 如果上述条件分布律是一个离散型分布, 则概率分布列中取值概率最大的点就作为对应的极大似然 *Bayes* 估计 \hat{P} 。试给出本模型下的极大似然 *Bayes* 估计 \hat{P} 。

附录 A

单调类定理及其在概率论中的应用

A.1 集合类简介

下面给出一些集合类的定义，它们通常在某些集合运算下具有封闭性。以下设 Ω 为非空集合， $\mathcal{E} \subset 2^\Omega$ 为其上一集合系/集合类/集族。

在第 2 章已经给出了 σ -代数的概念，在第 4 章也给出了 π -系、半环、环、代数等概念（见定义 4.2.4-4.2.6）；这里重新陈述如下。

定义 A.1.1. \mathcal{E} 称为 σ -代数 (σ -algebra) 或 σ -域 (σ -field)，如果它满足：

(i) $\Omega \in \mathcal{E}$ ；（此条类似于代数学中群等结构的“么元”的要求；）

(ii) 补运算封闭：如果 $A \in \mathcal{E}$ ，则 $A^c \in \mathcal{E}$ ；

(iii) 可列并运算封闭：如果 $A_n \in \mathcal{E}, n \in \mathbb{N}$ ，则 $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{E}$ 。

不难知道， σ -代数 \mathcal{E} 也关于可列交运算封闭。

把上述 σ -代数定义中的第三条提及的“可列并运算封闭”减弱为“有限并运算封闭”，就得到代数的概念，叙述如下：

定义 A.1.2. \mathcal{E} 称为代数 (Algebra) 或域 (Field)，如果它满足：

(i) $\Omega \in \mathcal{E}$ ；

(ii) 如果 $A \in \mathcal{E}$ ，则 $A^c \in \mathcal{E}$ ；

(iii) 如果 $A, B \in \mathcal{E}$ ，则 $A \cup B \in \mathcal{E}$ 。

显然代数对集合并、交、补、差运算有限封闭。此外上述 (i)–(iii) 也可换为

(i') $\Omega \in \mathcal{E}$ ；

(ii') 如果 $A, B \in \mathcal{E}$ ，则 $A \setminus B \in \mathcal{E}$ 。

事实上, $A^c = \Omega \setminus A$, $A \cup B = (A^c \cap B^c)^c = \Omega \setminus [(\Omega \setminus A) \setminus B]$ 。

很多场合, 我们关心的概率/测度只是在一些简单的事件/集合上能知道其取值或有计算办法; 复杂的事件/集合上的概率/测度值需要复杂的手续才能求出。一个经典的例子是我们经常遇到的 \mathbb{R} 上的 Lebesgue 测度 Leb , 它在区间 $(a, b]$ (其中 $a \leq b$) 上的测度值就是区间长度值 $b - a$; 立足于这些特殊集合上的测度取值就能在一定意义下唯一确定 Lebesgue 测度。而区间的全体 $\mathcal{E} := \{(a, b] : a \leq b\}$ 所具有的性质, 可以抽象为下文所谓的“ π 系”及“半环”概念:

定义 A.1.3. \mathcal{E} 称为 π -系 (π -system), 如果 $A, B \in \mathcal{E}$ 蕴含了 $A \cap B \in \mathcal{E}$ 。在此基础上, π -系 \mathcal{E} 进一步称为半环 (Semi-ring)*, 如果 $A, B \in \mathcal{E}$ 蕴含了: 存在 \mathcal{E} 中两两不交的有限集合列 $\{A_k \in \mathcal{E}\}_{k=1}^n$, 使得 $A \setminus B = \biguplus_{k=1}^n A_k$ 。

注意: 对半环 \mathcal{E} , $A, B \in \mathcal{E}$ 时, 未必有 $A \setminus B \in \mathcal{E}$! 即, 对于半环, 集合的差运算未必封闭。

如果半环定义中的后一条修改成“集合的差运算封闭”, 再补充上“集合的并运算封闭”就得到了环的概念, 陈述如下:

定义 A.1.4. \mathcal{E} 称为环 (Ring), 如果它满足:

- (i) $A, B \in \mathcal{E}$, 则 $A \cap B \in \mathcal{E}$;
- (ii) $A, B \in \mathcal{E}$, 则 $A \cup B \in \mathcal{E}$;
- (iii) $A, B \in \mathcal{E}$, 则 $A \setminus B \in \mathcal{E}$ 。

容易知道, 代数就是包含了全空间 Ω 的环; 因此有些文献也称环为预代数 (pre-algebra)。

在求复杂事件/集合的概率/测度过程中, 经常要使用取极限的办法。为此目的, 我们还需引入如下几类集合系的概念。

定义 A.1.5. 称 \mathcal{E} 为单调系 (Monotone System), 如果它满足:

- (i) 如果 $A_n \in \mathcal{E}, n \in \mathbb{Z}^+$, 且 $A_n \uparrow A$, 则 $A \in \mathcal{E}$;
- (ii) 如果 $A_n \in \mathcal{E}, n \in \mathbb{Z}^+$, 且 $A_n \downarrow A$, 则 $A \in \mathcal{E}$ 。

定义 A.1.6. 称 \mathcal{E} 为 λ -系或 D -系 (λ -System/Dynkin System), 如果它满足:

- (i) $\Omega \in \mathcal{E}$;
- (ii) 如果 $A \in \mathcal{E}$, 则 $A^c \in \mathcal{E}$;
- (iii) 如果 $A_n \in \mathcal{E}, n \in \mathbb{Z}^+$, 且它们互不相交, 则 $\biguplus_{n=1}^{\infty} A \in \mathcal{E}$ 。

*在 [13] 中有比半环定义略苛刻的 Semialgebra 的定义, 其中后一条件替换成了: 集合系中元素的补集可以表达成集合系内部有限个元素的不交并; 这隐含了集合系中存在全空间的有限覆盖的要求。因此, 本书例??中集合类不满足 [13] 中的 Semialgebra 的要求。

注记 A.1. 也有人把 λ -系定义为具有下面三条性质的集合类 \mathcal{E} :

- (i) $\Omega \in \mathcal{E}$;
- (ii) 如果 $A, B \in \mathcal{E}$, 且 $B \subset A$, 则 $A \setminus B \in \mathcal{E}$;
- (iii) 如果 $A_n \in \mathcal{E}, n \in \mathbb{Z}^+$, 且 $A_n \uparrow A$, 则 $A \in \mathcal{E}$.

不难证明, 这两种定义方式是等价的。

容易知道, λ -系一定是单调系。

注记 A.2. 设 Ω 是非空集合, \mathcal{E} 是 Ω 上集合系。

- (1) 所有包含 \mathcal{E} 的环的交集仍然是一个环, 记作 $R(\mathcal{E})$, 称为由 \mathcal{E} 生成的环;
- (2) 所有包含 \mathcal{E} 的代数的交集仍然是一个代数, 记作 $\mathcal{A}(\mathcal{E})$, 称为由 \mathcal{E} 生成的代数;
- (3) 所有包含 \mathcal{E} 的 σ -代数的交集仍然是一个 σ -代数, 记作 $\sigma(\mathcal{E})$, 称为由 \mathcal{E} 生成的 σ -代数;
- (4) 所有包含 \mathcal{E} 的 λ -系的交集仍然是一个 λ -系, 记作 $\lambda(\mathcal{E})$, 称为由 \mathcal{E} 生成的 λ -系;
- (5) 所有包含 \mathcal{E} 的单调系的交集仍然是一个单调系, 记作 $m(\mathcal{E})$, 称为由 \mathcal{E} 生成的单调系。

有时为强调上述“生成”运算是在全空间 Ω 中进行的, 我们在对应算符 R 、 \mathcal{A} 、 σ 、 λ 、 m 上附加下标 Ω , 即对应有: $R_\Omega(\mathcal{E})$ 、 $\mathcal{A}_\Omega(\mathcal{E})$ 、 $\sigma_\Omega(\mathcal{E})$ 、 $\lambda_\Omega(\mathcal{E})$ 、 $m_\Omega(\mathcal{E})$ 等记号。

例 A.1. 设 Ω 为非空集合, $\mathcal{E} := 2^\Omega$, 它是 σ -代数。

例 A.2. (同例 4.4) $\Omega = \mathbb{R}$, $\mathcal{E}_1 := \{(a, b] : -\infty < a \leq b < \infty\}$,

$\mathcal{E}_2 := \{E = \bigcup_{k=1}^n A_k : n \in \mathbb{N}, \{A_k\}_{k=1}^n \subset \mathcal{E}_1 \text{ 是两两不交的集合列}\}.$

则 \mathcal{E}_1 是 π -系、半环, 但不是环; \mathcal{E}_2 是环, 但不是代数。 \mathcal{E}_2 恰好是 \mathcal{E}_1 生成的环: $\mathcal{E}_2 = R(\mathcal{E}_1)$ 。

例 A.3. $\Omega = \mathbb{R}$, $\mathcal{E}_1 := \mathbb{R}$ 上 Borel 可测集全体, $\mathcal{E}_2 := \mathbb{R}$ 上 Lebesgue 可测集全体。 $\mathcal{E}_1, \mathcal{E}_2$ 都是 σ -代数, 且 $\mathcal{E}_1 \subset \mathcal{E}_2$ 。

例 A.4. 设 Ω 为一个无穷集合 (可列或不可数),

$$\mathcal{E} := \{A \subset \Omega : A \text{ 可数或 } A^c \text{ 可数}\}.$$

\mathcal{E} 是 σ -代数。(相关习题: 习题 2.6、2.7.)

例 A.5. $\Omega = \mathbb{R}$, $\mathcal{E} := \mathbb{R}$ 上开集全体,

$$\mathcal{E}_0 := \{(a, b) : -\infty < a < b < \infty\},$$

$$\mathcal{E}'_0 := \{(a, b) : a < b, a, b \in \mathbb{Q}\}.$$

那么 $\sigma(\mathcal{E}) = \sigma(\mathcal{E}_0) = \sigma(\mathcal{E}'_0)$ 恰好是 \mathbb{R} 上的 Borel σ -代数。

注意到例 A.5 告诉我们， \mathbb{R} 上的 Borel σ -代数可以通过一个可数个元素的集族生成，为此我们称 \mathbb{R} 上的 Borel σ -代数是可数生成的。我们可以进一步引入一般的 σ -代数的可数生成概念。

定义 A.1.7. 设 \mathcal{F} 是 Ω 上 σ -代数。 \mathcal{F} 称为是**可数生成的**，如果存在 \mathcal{F} 的一个可数子集 $\mathcal{E} = \{A_n : n \in \mathbb{N}\}$ ，使得 $\mathcal{F} = \sigma(\mathcal{E})$ 。有文献把 σ -代数的可数生成也称为**可分**。

例 A.6. 设 Ω 为不可数集， $\mathcal{F} := \{A \subset \Omega : A \text{ 可数或 } A^c \text{ 可数}\}$ 。 \mathcal{F} 是 σ -代数（见例 A.4），但它不是可数生成的。

定义 A.1.8. 设 Ω 为一拓扑空间。由 Ω 上所有开集生成的 σ -代数，称为 Ω 上 **Borel σ -代数**，记作 $\mathcal{B}(\Omega)$ 。

注记 A.3. 当拓扑空间 Ω 有可数邻域基时，其上 Borel σ -代数 $\mathcal{B}(\Omega)$ 就是可数生成的。

例 A.7. 设 $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ 为广义实数集（把 $\pm\infty$ 视作区别于普通实数的两个不同点），具有相应的诱导拓扑。记 $\bar{\mathcal{B}} := \mathcal{B}(\bar{\mathbb{R}})$ 为 $\bar{\mathbb{R}}$ 上 Borel σ -代数，它仍然是可数生成的。我们称 $(\bar{\mathbb{R}}, \bar{\mathcal{B}})$ 为**广义实数空间**。

A.2 单调类定理

下面我们探讨刚刚定义的几个集合系与更早定义的 σ -代数之间的关系。

定理 A.2.1. 设 \mathcal{F} 是空间 Ω 上一集合系。

- (1) \mathcal{F} 是单调系，又是代数 $\iff \mathcal{F}$ 是 σ -代数；
- (2) \mathcal{F} 是 λ -系，又是 π -系 $\iff \mathcal{F}$ 是 σ -代数。

证明. 从定义出发即可证明。（留作习题 A.1。） □

定理 A.2.2.（单调类定理） 设 \mathcal{E}, \mathcal{F} 是空间 Ω 上集合系，且 $\mathcal{E} \subset \mathcal{F}$ 。

- (1) 如果 \mathcal{E} 是代数， \mathcal{F} 是单调系，那么 $\sigma(\mathcal{E}) \subset \mathcal{F}$ ；
- (2) 如果 \mathcal{E} 是 π -系， \mathcal{F} 是 λ -系，那么 $\sigma(\mathcal{E}) \subset \mathcal{F}$ 。

证明. 此处只给出 (1) 的证明。(2) 的证明留作习题 A.2。

本质上我们只需要证明： $\sigma(\mathcal{E}) = m(\mathcal{E})$ 。显然 $\sigma(\mathcal{E}) \supset \mathcal{E}$ ，由定理 A.2.1，应有 $\sigma(\mathcal{E}) \supset m(\mathcal{E})$ 。下面证明 $\sigma(\mathcal{E}) \subset m(\mathcal{E})$ 。注意到 $\mathcal{E} \subset m(\mathcal{E})$ ，同样由定理 A.2.1，只需要证明 $m(\mathcal{E})$ 是一个代数即可。

为此，令

$$\mathcal{E}_1 := \{A \in m(\mathcal{E}) : A^c \in m(\mathcal{E}), A \cup B \in m(\mathcal{E}), \forall B \in \mathcal{E}\}.$$

以下验证 \mathcal{E}_1 是一个单调类。设 $A_n \in \mathcal{E}_1$ 且 $A_n \nearrow A$ （或 $A_n \searrow A$ ）。那么 $A_n, A_n^c \in m(\mathcal{E})$ ，此时 $A, A^c \in m(\mathcal{E})$ 。进而对任意 $B \in \mathcal{E}$ ， $A_n \cup B \in m(\mathcal{E})$ ， $A_n \cup B \nearrow A \cup B \in m(\mathcal{E})$ （或 $A_n \cup B \searrow A \cup B \in m(\mathcal{E})$ ），于是 $A \cup B \in m(\mathcal{E})$ ，即 $A \in \mathcal{E}_1$ 。从而 \mathcal{E}_1 是单调类，注意到 $\mathcal{E} \subset \mathcal{E}_1 \subset m(\mathcal{E})$ ，有 $\mathcal{E}_1 = m(\mathcal{E})$ 。这表

明：(i) $m(\mathcal{E})$ 关于集合的补运算封闭；(ii) 对任意 $A \in m(\mathcal{E}), B \in \mathcal{E}$ ，总有 $A \cup B \in m(\mathcal{E})$ 。

进一步令

$$\mathcal{E}_2 := \{A \in m(\mathcal{E}) : A \cup B \in m(\mathcal{E}), \forall B \in m(\mathcal{E})\}.$$

通过上一步骤，我们已经证明了： $\mathcal{E} \subset \mathcal{E}_2$ 。很容易证明 \mathcal{E}_2 也是一个单调类，从而类似的推出 $\mathcal{E}_2 = m(\mathcal{E})$ 。这表明 $m(\mathcal{E})$ 关于集合并运算封闭。

于是 $m(\mathcal{E})$ 是一个代数，进而是一个 σ -代数。 \square

注记 A.4. 上面的论证方法称为**范畴法**，是数学中论证某些集族具有某种特殊结构或性质的常用方法。另外，上面的单调类定理非常重要，在后续有关积分或数学期望的讨论中将看到它的威力。

A.3 乘积概率测度的存在唯一性

在第3章，对两个概率空间 $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k), k = 1, 2$ ，我们定义它们的乘积空间时，定义了

$$\mathbb{P}_1 \times \mathbb{P}_2(A_1 \times A_2) := \mathbb{P}_1(A_1) \cdot \mathbb{P}_2(A_2), A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2.$$

在那里我们断言 $\mathbb{P}_1 \times \mathbb{P}_2$ 可以唯一延拓为 $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ 上的概率测度，其中 $\mathcal{F}_1 \otimes \mathcal{F}_2 := \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ 。这里我们需要使用 Carathéodory 扩张定理和它后续的扩张唯一性定理；在第4章中我们已经看到，在这两个定理的证明中都需要使用单调类定理。

下面我们重点检查：(1) $\mathcal{F}_1 \times \mathcal{F}_2$ 是一个半环；(2) $\mathbb{P}_1 \times \mathbb{P}_2$ 在 $\mathcal{F}_1 \times \mathcal{F}_2$ 上是一个预测度，进而它可以视作 $\mathcal{F}_1 \times \mathcal{F}_2$ 生成的环上的预测度；(3) 这个预测度是有限预测度（从而自然是 σ -有限的）。以上三点就确保了可以应用 Carathéodory 扩张定理和扩张唯一性定理来论证乘积概率测度的唯一确定性。

对(1)的验证如下。首先，对任意 $A = A_1 \times A_2, B = B_1 \times B_2 \in \mathcal{F}_1 \times \mathcal{F}_2$ ，容易知道 $A_1 \cap B_1 \in \mathcal{F}_1, A_2 \cap B_2 \in \mathcal{F}_2$ ，进而

$$A \cap B = (A_1 \times A_2) \cap (B_1 \times B_2) = (A_1 \cap B_1) \times (A_2 \cap B_2) \in \mathcal{F}_1 \times \mathcal{F}_2.$$

这表明 $\mathcal{F}_1 \times \mathcal{F}_2$ 是 π -系。进一步，注意到 $(B_1 \times B_2)^c = (B_1^c \times \Omega_2) \dot{\cup} (B_1 \times B_2^c)$ ，我们有

$$A \setminus B = [(A_1 \cap B_1^c) \times A_2] \dot{\cup} [(A_1 \cap B_1) \times (A_2 \cap B_2^c)],$$

显然 $(A_1 \cap B_1^c) \times A_2, (A_1 \cap B_1) \times (A_2 \cap B_2^c)$ 都是 $\mathcal{F}_1 \times \mathcal{F}_2$ 中元素，因此 $\mathcal{F}_1 \times \mathcal{F}_2$ 是半环。

现在来验证(2)。 $\mathbb{P}_1 \times \mathbb{P}_2$ 的非负性是显然的；注意到 $\emptyset = \emptyset \times \emptyset$ ，

$$\mathbb{P}_1 \times \mathbb{P}_2(\emptyset) = 0;$$

如果对 $A \in \mathcal{F}_1, B \in \mathcal{F}_2$ ，存在 $\{A_k\}_{k=1}^N \subset \mathcal{F}_1, \{B_k\}_{k=1}^N \subset \mathcal{F}_2$ （为方便，我们可以要求 A_k, B_k 都不是空集，其中 $2 \leq N \leq \infty$ ）使得

$$A \times B = \biguplus_{k=1}^N (A_k \times B_k),$$

那么, 应有 $A_k \subset A, B_k \subset B, \forall k$, 并且 $\{A_k\}_{k=1}^N, \{B_k\}_{k=1}^N$ 分别是 A 与 B 的覆盖。考察

$$\mathcal{E}_A := \sigma_A(\{A_k : 1 \leq k \leq N\}), \quad \mathcal{E}_B := \sigma_B(\{A_k : 1 \leq k \leq N\}).$$

不难知道: \mathcal{E}_A 中存在 A 的分割 $\{C_i \neq \emptyset : 1 \leq i \leq m\}$ 生成 σ -代数 \mathcal{E}_A , \mathcal{E}_B 中存在 B 的分割 $\{D_j \neq \emptyset : 1 \leq j \leq n\}$ 生成 σ -代数 \mathcal{E}_B , 其中 $m, n \leq N$ 。对任意 $1 \leq k \leq N$ 于是存在非空指标集 I_k, J_k , 使得

$$A_k = \biguplus_{i \in I_k} C_i, \quad B_k = \biguplus_{j \in J_k} D_j.$$

于是

$$A_k \times B_k = \biguplus_{(i,j) \in I_k \times J_k} (C_i \times D_j), \quad A \times B = \biguplus_{\substack{(i,j) \in I_k \times J_k \\ 1 \leq k \leq N}} (C_i \times D_j).$$

注意到

$$A \times B = \biguplus_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} (C_i \times D_j),$$

且 $C_i \times D_j$ 均非空, 实际上 $\{I_k \times J_k : 1 \leq k \leq N\}$ 形成 $I \times J$ 的分割, 其中 $I := \{i : 1 \leq i \leq m\}, J := \{j : 1 \leq j \leq n\}$ 。于是自然有

$$\begin{aligned} & \sum_{k=1}^N \mathbb{P}_1 \times \mathbb{P}_2(A_k \times B_k) \\ &= \sum_{k=1}^N \sum_{(i,j) \in I_k \times J_k} \mathbb{P}_1 \times \mathbb{P}_2(C_i \times D_j) \\ &= \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \mathbb{P}_1 \times \mathbb{P}_2(C_i \times D_j) \\ &= \left[\sum_{i=1}^m \mathbb{P}_1(C_i) \right] \cdot \left[\sum_{j=1}^n \mathbb{P}_2(D_j) \right] \\ &= \mathbb{P}_1(A) \cdot \mathbb{P}_2(B) = \mathbb{P}_1 \times \mathbb{P}_2(A \times B). \end{aligned}$$

最后验证 (3), 这是显然的, 因为 $\mathbb{P}_1 \times \mathbb{P}_2(\Omega_1 \times \Omega_2) = 1 < \infty$ 。

A.4 定理 5.1.1 的证明: (5.3) \Rightarrow (5.5) 部分

在这里, 我们拟提供 (5.3) \Rightarrow (5.5) 部分的证明; 在第 5 章处只是简单提到用单调类定理来证明这一点。

为方便, 我们只考虑 X, Y 都是一维随机变量的情况。此时, 取

$$\mathcal{E} := \{(A, B) : A, B \in \mathcal{B}, \text{ 使得 } \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\},$$

及 $\mathcal{E}_A = \{B : (A, B) \in \mathcal{E}\}, \mathcal{E}^B := \{A : (A, B) \in \mathcal{E}\}$ 。

X, Y 相互独立, 即 (5.3) 成立, 意味着对任意 $x, y \in \mathbb{R}$, 取

$$A := (-\infty, x], B := (-\infty, y]$$

时, $(A, B) \in \mathcal{E}$ 。

对任意固定的 $A \in \mathcal{B}$, 不难验证 \mathcal{E}_A 是 \mathbb{R} 上的 σ -代数。当 $A := (-\infty, x]$ 时, 我们已知

$$\mathcal{E}_0 := \{(-\infty, y] : y \in \mathbb{R}\} \subset \mathcal{E}_A.$$

注意到 \mathcal{E}_0 是一个 π -系, 并且 $\mathcal{B} = \sigma(\mathcal{E}_0)$, 由单调类定理我们立即得到 $\mathcal{E}_A = \mathcal{B}$ 。

现在, 对任意固定的 $B \in \mathcal{B}$, 同样不难验证 \mathcal{E}^B 是 \mathbb{R} 上的 σ -代数。由上一步骤, 我们已知

$$\mathcal{E}_0 \subset \mathcal{E}^B.$$

同上逻辑立即得到 $\mathcal{E}^B = \mathcal{B}$ 。由此 $\mathcal{E} = \{(A, B) : A, B \in \mathcal{B}\}$, 即(5.5) 成立。

习 题 A

习题 A.1. 证明定理 A.2.1。

习题 A.2. 证明定理 A.2.2 的 (2)。

习题 A.3. 设 $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ 为概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中三个相互独立的子 σ -代数。记 $\mathcal{G}_1 \vee \mathcal{G}_2$ 为 $\mathcal{G}_1, \mathcal{G}_2$ 生成的 σ -代数, 求证: $\mathcal{G}_1 \vee \mathcal{G}_2$ 与 \mathcal{G}_3 相互独立。【提示: 令 $\mathcal{G}_{1,2}$ 为 $\mathcal{G}_1, \mathcal{G}_2$ 生成的 π 系, 记

$$\mathcal{E} := \{A \in \mathcal{G}_1 \vee \mathcal{G}_2 : \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B), \forall B \in \mathcal{G}_3\}.$$

论证: $\mathcal{G}_1 \cup \mathcal{G}_2 \subset \mathcal{G}_{1,2} \subset \mathcal{E}$, 且 \mathcal{E} 是 λ -系, 再利用单调类定理完成证明。】

附录 B

数学期望、条件数学期望的一些性质的证明

在这个附录里，我们准备补全数学期望、条件数学期望的一些性质的证明。

B.1 数学期望的一些性质的证明

B.1.1 数学期望的线性性质

如第 6 章所述，在数学期望的定义 **Step 1**. 完成后， $\mathbb{E} : \mathcal{S}_+ \rightarrow \mathbb{R}$ 的线性性质是不难知道成立的，即以下结论成立： $\forall \xi, \eta \in \mathcal{S}_+, c \in \mathbb{R}_+$

$$\mathbb{E}[c\xi] = c\mathbb{E}\xi, \quad \mathbb{E}[\xi + \eta] = \mathbb{E}\xi + \mathbb{E}\eta. \quad (\text{B.1})$$

在此基础上，不难知道 $\mathbb{E} : \mathcal{S}_+ \rightarrow \mathbb{R}$ 具有单调性： $\xi, \eta \in \mathcal{S}_+$ 且 $\xi \leq \eta$ ，那么 $\mathbb{E}\xi \leq \mathbb{E}\eta$ 。

我们先证明在数学期望的定义 **Step 2**. 完成后，在非负可测随机变量空间中具有线性性质。事实上，对任意非负可测随机变量 ξ 及 $c \in \mathbb{R}_+$ （此处不妨设 $c > 0$ ）

$$\begin{aligned} \mathbb{E}[c\xi] &:= \sup\{\mathbb{E}\eta : 0 \leq \eta \leq c\xi, \eta \in \mathcal{S}_+\} \\ &= \sup\{\mathbb{E}[c\eta] : 0 \leq \eta \leq \xi, \eta \in \mathcal{S}_+\} \\ &= \sup\{c \cdot \mathbb{E}[\eta] : 0 \leq \eta \leq \xi, \eta \in \mathcal{S}_+\} \\ &= c \cdot \sup\{\mathbb{E}[\eta] : 0 \leq \eta \leq \xi, \eta \in \mathcal{S}_+\} = c \cdot \mathbb{E}\xi. \end{aligned}$$

由此，不难知道，单调收敛定理/Levi 定理成立：设 $\{\xi_n\}_{n=1}^\infty$ 为一列非负可测随机变量，且存在随机变量 ξ 使得 $\xi_n \nearrow \xi$ 。那么

$$\lim_{n \rightarrow \infty} \mathbb{E}[\xi_n] = \mathbb{E}[\xi].$$

它的证明见第 4 章对应定理的证明。于是对任意 $\xi, \eta \geq 0$ 为非负可测随机变量，存在 $\xi_n, \eta_n \in \mathcal{S}_+$ 使得 $\xi_n \nearrow \xi, \eta_n \nearrow \eta$ ，进而 $\xi_n + \eta_n \nearrow \xi + \eta$ ，由单调

收敛定理/Levi 定理,

$$\begin{aligned}\mathbb{E}[\xi + \eta] &= \lim_{n \rightarrow \infty} \mathbb{E}[\xi_n + \eta_n] \quad (\text{Levi 定理}) \\ &= \lim_{n \rightarrow \infty} [\mathbb{E}\xi_n + \mathbb{E}\eta_n] \quad (\mathbb{E} : \mathcal{S}_+ \rightarrow \mathbb{R} \text{ 的线性性质}) \\ &= \mathbb{E}\xi + \mathbb{E}\eta. \quad (\text{Levi 定理})\end{aligned}$$

现在, 我们证明在数学期望的定义 **Step 3**. 完成后, 在随机变量空间中具有线性性质。对于 $c \in \mathbb{R}$ 及使 $\mathbb{E}\xi$ 有意义的 ξ , $\mathbb{E}[c\xi] = c\mathbb{E}\xi$ 的证明是简单的, 留给读者。以下我们证明, 当 ξ, η 均可积时, 总有

$$\mathbb{E}[\xi + \eta] = \mathbb{E}\xi + \mathbb{E}\eta. \quad (\text{B.2})$$

此时, 由于 $|\xi + \eta| \leq |\xi| + |\eta|$, $\xi + \eta$ 也可积。注意到

$$(\xi + \eta)^+ - (\xi + \eta)^- = \xi + \eta = \xi^+ - \xi^- + \eta^+ - \eta^-,$$

我们有

$$(\xi + \eta)^+ + \xi^- + \eta^- = (\xi + \eta)^- + \xi^+ + \eta^+.$$

由上一步骤的结论, 我们得到

$$\mathbb{E}[(\xi + \eta)^+] + \mathbb{E}[\xi^-] + \mathbb{E}[\eta^-] = \mathbb{E}[(\xi + \eta)^-] + \mathbb{E}[\xi^+] + \mathbb{E}[\eta^+].$$

移项整理即得 $\mathbb{E}[\xi + \eta] = \mathbb{E}\xi + \mathbb{E}\eta$ 。更一般情形下(B.2) 结果的证明留给读者作为练习。

B.1.2 Jensen 不等式与矩不等式的证明

为了证明 Jensen 不等式, 这里我们需要使用凸函数的一个等价性质: g 是区间 I 上的凸函数, 当且仅当: $\forall x_0 \in I^\circ$ (此处 I° 表示 I 的内点的全体), $\exists \alpha \in \mathbb{R}$, 使得

$$g(x) \geq g(x_0) + \alpha(x - x_0), \forall x \in I.$$

因此, 当 g 为 \mathbb{R} 上的凸函数时, 对于可积随机变量 ξ , 取 $x_0 := \mathbb{E}\xi$, 则存在常数 $\alpha \in \mathbb{R}$

$$g(\xi) \geq g(\mathbb{E}\xi) + \alpha(\xi - \mathbb{E}\xi).$$

于是对上式两边取数学期望, 立即得到

$$\mathbb{E}[g(\xi)] \geq g(\mathbb{E}\xi).$$

对于 ξ 未必可积, 但 $g(\mathbb{E}\xi)$ 和 $\mathbb{E}[g(\xi)]$ 有意义的情形的 Jensen 不等式的证明, 留给读者思考。

现在, 我们可以来证明矩不等式。事实上, 对任意 $p > 1$, $g(x) := x^p$ 为 \mathbb{R}_+ 上的凸函数。于是对任意随机变量 ξ , $|\xi|$ 是非负随机变量, 并且由 Jensen 不等式

$$\mathbb{E}[g(|\xi|)] \geq g(\mathbb{E}|\xi|),$$

此即 $\|\xi\|_1 \leq \|\xi\|_p$ 。特别的, 对 $0 < s < t$, 取 $p = t/s > 1$, $\| |\xi|^s \|_1 \leq \| |\xi|^s \|_p$, 化简即得 $\|\xi\|_s \leq \|\xi\|_t$ 。

B.1.3 Hölder 不等式与 Minkowski 不等式的证明

关于数学期望版本的 Hölder 不等式，其论证基础是数学分析中如下版本的 Hölder 不等式：

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab, \forall a, b \geq 0,$$

其中 $p > 1, q > 1$ 满足 $\frac{1}{p} + \frac{1}{q} = 1$ 。在承认上述 Hölder 不等式的基础上，假设 $\xi \in L^p, \eta \in L^q$ ，并且 $\|\xi\|_p \cdot \|\eta\|_q > 0$ ，那么取 $a = |\xi|, b = t \cdot |\eta|$ ，其中 $t > 0$ ，立即得到

$$\frac{|\xi|^p}{p} + \frac{|\eta|^q}{q} \cdot t^q \geq t|\xi\eta|, \forall t > 0.$$

对上式取数学期望，整理得到

$$\|\xi\eta\|_1 \leq \frac{\|\xi\|_p^p}{pt} + \frac{\|\eta\|_q^q}{q} \cdot t^{q-1}, \forall t > 0.$$

特取 $t = \|\xi\|_p^{p/q} \cdot \|\eta\|_q^{-1}$ ，代入上式整理即得 $\|\xi\eta\|_1 \leq \|\xi\|_p \cdot \|\eta\|_q$ 。

现在我们来证明数学期望版本的 Minkowski 不等式。

我们先证明 $\|\xi + \eta\|_1 \leq \|\xi\|_1 + \|\eta\|_1$ 。这只要注意到三角不等式

$$|\xi + \eta| \leq |\xi| + |\eta|,$$

对上式两边取数学期望即得到。

下面设 $p > 1$ 。此时不妨设 $\|\xi\|_p < \infty, \|\eta\|_p < \infty$ ，则

$$\begin{aligned} \|\xi + \eta\|_p^p &= \mathbb{E}[|\xi + \eta|^p] \leq \mathbb{E}[|\xi + \eta|^{p-1} \cdot (|\xi| + |\eta|)] \quad (\text{三角不等式}) \\ &\leq [\|\xi\|_p + \|\eta\|_p] \cdot \|\xi + \eta\|_p^{p-1} \quad (\text{Hölder 不等式}) \\ &\leq [\|\xi\|_p + \|\eta\|_p] \cdot \|\xi + \eta\|_p^{p-1}, \quad (\|\xi + \eta\|_p^{p-1} \leq \|\xi + \eta\|_p^{p-1}) \end{aligned}$$

于是当 $\|\xi + \eta\|_p > 0$ 时， $\|\xi + \eta\|_p \leq \|\xi\|_p + \|\eta\|_p$ ；相反情况下这个不等式是显然的。

现在设 $0 < p < 1$ ，我们证明 $\|\xi + \eta\|_p^p \leq \|\xi\|_p^p + \|\eta\|_p^p$ ，即

$$\mathbb{E}[|\xi + \eta|^p] \leq \mathbb{E}[|\xi|^p] + \mathbb{E}[|\eta|^p].$$

注意到此时当 $a, b \geq 0$ 且 $a + b = 1$ 时， $a \leq a^p, b \leq b^p$ ，从而

$$1 = a + b \leq a^p + b^p.$$

这进一步意味着：当 $a, b > 0$ 时总有

$$(a + b)^p \leq a^p + b^p.$$

现在

$$|\xi + \eta|^p \leq [|\xi| + |\eta|]^p \leq |\xi|^p + |\eta|^p.$$

对上式两边取数学期望即完成此情形下 Minkowski 不等式的证明。

B.2 条件数学期望的一些性质的证明

对于给定概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ ，给定的子 σ -代数 $\mathcal{G} \subset \mathcal{F}$ 以及随机变量 $\xi \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ ，由于 $\xi := \mathbb{E}[\xi|\mathcal{G}]$ 被定义为满足下面方程组的 \mathcal{G} -可测随机

变量

$$\mathbb{E}[\xi \cdot 1_B] = \mathbb{E}[\tilde{\xi} \cdot 1_B], \forall B \in \mathcal{G}, \quad (\text{B.3})$$

因此取 $B = \Omega$ 就得到了全期望公式。

B.2.1 条件数学期望 $\mathbb{E}[\cdot|\mathcal{G}]$ 的单调性与 \mathcal{G} -线性性质

注意到 \mathbb{E} 是正算子，容易知道， $\xi \geq 0$ 几乎处处成立时必定有 $\eta \geq 0$ 几乎处处成立，进而 $\mathbb{E}[\cdot|\mathcal{G}]$ 是正算子，即它具有单调性；同理，由 \mathbb{E} 是线性算子，也容易得到 $\mathbb{E}[\cdot|\mathcal{G}]$ 的线性性质。于是注意到 $\pm\xi \leq |\xi|$ ，我们有 $\pm\eta := \mathbb{E}[\pm\xi|\mathcal{G}] \leq \mathbb{E}[|\xi||\mathcal{G}]$ 几乎处处成立。此即

$$|\mathbb{E}[\xi|\mathcal{G}]| \leq \mathbb{E}[|\xi||\mathcal{G}].$$

当 $\xi \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ 时，上式结合全概率公式立即得到

$$\|\mathbb{E}[\xi|\mathcal{G}]\|_1 \leq \|\xi\|_1.$$

由此，可以把 $\mathbb{E}[\cdot|\mathcal{G}]$ 视作算子：

$$\mathbb{E}[\cdot|\mathcal{G}] : L^1(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow L^1(\Omega, \mathcal{G}, \mathbb{P}),$$

它是正算子、线性算子。

以下我们证明它实际上是 \mathcal{G} -线性算子。事实上，由定义容易证明：对任意 $A \in \mathcal{G}$ ， $\xi \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ ，

$$\mathbb{E}[\xi \cdot 1_A|\mathcal{G}] = 1_A \cdot \mathbb{E}[\xi|\mathcal{G}],$$

亦即对任意 $B \in \mathcal{G}$

$$\mathbb{E}[(\xi \cdot 1_A) \cdot 1_B] = \mathbb{E}[(1_A \cdot \mathbb{E}[\xi|\mathcal{G}]) \cdot 1_B].$$

由于 A 的任意性，立即根据数学期望的定义流程得到

$$\mathbb{E}[(\xi \cdot \eta) \cdot 1_B] = \mathbb{E}[(\eta \cdot \mathbb{E}[\xi|\mathcal{G}]) \cdot 1_B]$$

对任意 \mathcal{G} -可测随机变量 η 成立，只要它使得 $\xi \cdot \eta \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ 仍然成立。对上式，进一步由 $B \in \mathcal{G}$ 的任意性及条件数学期望的定义（注意到此时 $\eta \cdot \mathbb{E}[\xi|\mathcal{G}]$ 是 \mathcal{G} -可测随机变量），立即得到

$$\mathbb{E}[(\xi \cdot \eta)|\mathcal{G}] = \eta \cdot \mathbb{E}[\xi|\mathcal{G}].$$

因此 $\mathbb{E}[\cdot|\mathcal{G}]$ 是 \mathcal{G} -线性算子。

B.2.2 条件数学期望 $\mathbb{E}[\cdot|\mathcal{G}]$ 的性质（6）的证明

此小节我们要证明：当 ξ 与 \mathcal{G} 相互独立时（假定 ξ 可积）

$$\mathbb{E}[\xi|\mathcal{G}] = \mathbb{E}\xi \text{ a.s. }$$

这只要注意到 ξ 与 \mathcal{G} 的充要条件是：对任意 Borel 可测集 A 及可测集 $B \in \mathcal{G}$

$$\mathbb{P}(\{\xi \in A\} \cap B) = \mathbb{P}(\xi \in A) \cdot \mathbb{P}(B).$$

把上式写成数学期望形式，立即得到

$$\mathbb{E}[1_A(\xi) \cdot 1_B] = \mathbb{E}[1_A(\xi)] \cdot \mathbb{E}[1_B].$$

注意到上式中 A 的任意性，我们立即得到

$$\mathbb{E}[\varphi(\xi) \cdot 1_B] = \mathbb{E}[\varphi(\xi)] \cdot \mathbb{E}[1_B], \forall B \in \mathcal{G},$$

其中 φ 是使 $\varphi(\xi)$ 可积的可测函数。由条件数学期望定义，立即得到

$$\mathbb{E}[\varphi(\xi)|\mathcal{G}] = \mathbb{E}[\varphi(\xi)] \text{ a.s. },$$

特取 $\varphi(x) = x$ 即得 $\mathbb{E}[\xi|\mathcal{G}] = \mathbb{E}\xi$ 几乎处处成立。

B.2.3 条件 Jensen 不等式的证明

简单起见，设 g 是 \mathbb{R} 上连续的凸函数，设随机变量 ξ 可积且 $g(\xi)$ 可积。我们证明

$$g(\mathbb{E}[\xi|\mathcal{G}]) \leq \mathbb{E}[g(\xi)|\mathcal{G}] \text{ a.s..}$$

同样利用凸函数的等价刻画，对 $\tilde{\xi} := \mathbb{E}[\xi|\mathcal{G}]$ 存在 $\alpha(\tilde{\xi})$ （其中 $\alpha(\cdot)$ 可测），使得

$$g(x) \geq g(\tilde{\xi}) + \alpha(\tilde{\xi}) \cdot (x - \tilde{\xi}), \forall x.$$

取 $x = \xi$ 代入上式，并取条件数学期望，根据条件数学期望的单调性、 \mathcal{G} -线性性质以及 $\mathbb{E}[\xi - \tilde{\xi}|\mathcal{G}] = 0$ ，立即得到

$$\mathbb{E}[g(\xi)|\mathcal{G}] \geq g(\tilde{\xi}),$$

从而完成条件 Jensen 不等式的证明。

其他一些不等式，如条件 Hölder 不等式、条件矩不等式、条件 Minkowski 不等式就留作习题。

习 题 B

习题 B.1. 证明条件 Hölder 不等式。

习题 B.2. 证明条件矩不等式。

习题 B.3. 证明条件 Minkowski 不等式。

习题 B.4. 设 $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ 为概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中三个子 σ -代数。 $\mathcal{G}_1 \vee \mathcal{G}_2$ 总表示由 $\mathcal{G}_1 \cup \mathcal{G}_2$ 生成的 σ -代数。假设 $\mathcal{G}_1 \vee \mathcal{G}_2$ 与 \mathcal{G}_3 独立，试证明：对任意 \mathcal{G}_1 可测的随机变量 f ，总有（当下面等式的左右两边条件数学期望有意义时）

$$\mathbb{E}[f|\mathcal{G}_2 \vee \mathcal{G}_3] = \mathbb{E}[f|\mathcal{G}_2].$$

【提示：先论证对于任意 $A_1 \in \mathcal{G}_1$ 有 $\mathbb{P}(A_1|\mathcal{G}_2 \vee \mathcal{G}_3) = \mathbb{P}(A_1|\mathcal{G}_2)$ 。这个过程中需要论证

$$\mathcal{E} := \{A \in \mathcal{G}_2 \vee \mathcal{G}_3 : \mathbb{P}(A_1 \cap A) = \int_A \mathbb{P}(A_1|\mathcal{G}_2)d\mathbb{P}\}$$

是 λ -系，它包含了一个 π -系 $\mathcal{E}_0 := \{A_2 \cap A_3 : A_2 \in \mathcal{G}_2, A_3 \in \mathcal{G}_3\}$ 。后者包含了 $\mathcal{G}_2 \cup \mathcal{G}_3$ 。】

习题 B.5. 上一题也可以继续推广（本质上都是在推广： $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}X$ ，如果 X 与 \mathcal{G} 独立）：设 $\mathcal{G}_1, \mathcal{G}_2$ 为概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 中两个子 σ -代数。任给可积的（或非负的）随机变量 X ，假设 X 与 \mathcal{G}_2 关于 \mathcal{G}_1 条件独立：

$$\mathbb{P}(\{X \leq x\} \cap A|\mathcal{G}_1) = \mathbb{P}(X \leq x|\mathcal{G}_1) \cdot \mathbb{P}(A|\mathcal{G}_1), \quad \forall x \in \mathbb{R}, A \in \mathcal{G}_2,$$

那么 $\mathbb{E}[X|\mathcal{G}_1 \vee \mathcal{G}_2] = \mathbb{E}[X|\mathcal{G}_1]$ 。

附录 C

矩问题与 Laplace 变换

在本书中，我们已经告诉过大家，分布函数是最主要的刻画随机变量分布律的方式；此外分布测度、特征函数也是与分布函数等价的刻画分布律的重要工具。在一些特殊情况下，还有例如概率分布列（针对离散型随机变量）、概率密度函数（针对连续型随机变量）等刻画分布律的工具。

在这个附录里，我们准备介绍其他一些特殊情况下刻画分布律的工具，为此我们要介绍矩问题与 Laplace 变换。

C.1 矩问题

矩问题 (Moment Problem) 是一个古老的问题；该问题的研究起源于人们对于分布的各阶矩能否确定分布的思考与探索。Moment Problem 这一术语最早出现在 Stieltjes 在 1894-1895 年间的工作中；但实际上，早在 1873 年 Chebyshev 就提出并解决了这一重要问题的一种特殊情形，之后他的学生 Markov 也在这个问题上有突出的研究成果；他们考虑这一问题的目的在于用矩的方法来探讨中心极限定理。Stieltjes、Hamburger、Nevanlinna、Riesz、Hausdorff、Carleman、Stone 等人在矩问题上也有系统的研究成果。本节只介绍一些基本的矩问题及相应的一般性结果。对这方面更深入的结果感兴趣的读者请参见 [37]。

注 C.1. *T. J. Stieltjes* (1856/12/29–1894/12/31) 是荷兰数学家，以 *Lebesgue-Stieltjes* 积分知名。

H. L. Hamburger (1889/8/5–1956/8/14) 是德国数学家，以 *Hamburger* 矩问题知名。

R. H. Nevanlinna (1895/10/22–1980/5/28) 是芬兰数学家，以复变函数中的 *Nevanlinna* 理论知名；家族多位成员是数学家。

M. Riesz (1886/11/16–1969/9/4) 是匈牙利数学家，以 *Riesz-Thorin* 定理、*M. Riesz* 扩张定理、*Riesz* 兄弟定理、*Riesz* 位势、*Riesz* 函数、*Riesz* 变换、*Riesz* 平均等知名；*F. Riesz* (1880/1/22–1956/2/28) 是他的哥哥，在泛函分析中有许多重要贡献，以 *Riesz* 表示定理知名。

F. Hausdorff (1868/11/8–1942/1/26) 是德国数学家，现代拓扑学奠基人之一，以 *Hausdorff* 测度、*Hausdorff* 维数、*Hausdorff* 空间、*Hausdorff* 极大原理、*Hausdorff* 距离、*Hausdorff* 悖论、*Hausdorff* 矩问题、*Hausdorff-Young* 不等式等知名。

T. Carleman (1892/7/8–1949/1/11) 是瑞典数学家，以 *Carleman* 条件、*Carleman* 不等式、*Denjoy-Carleman* 定理、*Carleman* 核、*Carleman* 公式等知名。

Marshall Harvey Stone (1903/4/8–1989/1/9) 是美国数学家，*G. D. Birkhoff* (1884/3/21–1944/11/12; 美国) 的学生，曾任教于哈佛大学和芝加哥大学，在实分析、泛函分析、拓扑、

Bool 代数等方面有贡献，以 *Stone-von Neumann* 定理、*Stone-Čech* 紧化、*Stone-Weierstrass* 定理、*Stone* 表示定理知名。另一个容易与之混淆的数学家 *Stone* 是英国数学家 *Arthur Harold Stone* (1916/9/30–2000/8/6)，他是 *Solomon Lefschetz* (1884/9/3–1972/10/5；俄罗斯出生的美国数学家) 的学生，研究领域是拓扑，曾任教于 *Manchester* 大学和 *Rochester* 大学，与 *P. Erdős* (1913/3/26–1996/9/20) 合作证明了 *Erdős-Stone* 定理。

C.1.1 Hausdorff 矩问题

下面形式的矩问题被称为 **Hausdorff 矩问题**：给定数列 $\{\mu_n : n \geq 0\}$ （其中 $\mu_0 := 1$ ），求解 $[0, 1]$ 上概率分布函数 F ，使得

$$\mu_k = \int_0^1 x^k dF(x), \quad k = 0, 1, \dots \quad (\text{C.1})$$

我们不加证明的引述如下结果：对有关证明感兴趣的读者，请参见 [39, Chapter III] 或习题 C.8。

定理 C.1.1. (Hausdorff 定理) 数列 $\{\mu_n\}_{n=0}^\infty$ （其中 $\mu_0 := 1$ ）可以实现为某个集中在 $[0, 1]$ 上的概率分布 F 的各阶矩的充要条件是该数列是完全单调序列，即： $\mu_n \geq 0, \forall n \geq 1$ ，且

$$(-1)^k \Delta^k \mu_n \geq 0, \quad \forall n, k \geq 0, \quad (\text{C.2})$$

其中

$$\Delta^k \mu_n := \sum_{\ell=0}^k (-1)^\ell C_k^\ell \cdot \mu_{n+k-\ell}.$$

如果上述条件成立，那么(C.1)中的 F 存在且唯一，并且

$$(-1)^k \Delta^k \mu_n = \int_0^1 x^n (1-x)^k dF(x), \quad \forall n, k \geq 0. \quad (\text{C.3})$$

C.1.2 Hamburger 矩问题：可解的条件

下面形式的矩问题被称为 **Hamburger 矩问题**：给定数列 $\{m_n : n \geq 0\}$ （其中 $m_0 := 1$ ），求解 \mathbb{R} 上概率测度 μ ，使得

$$m_k = \int x^k d\mu, \quad k = 0, 1, \dots \quad (\text{C.4})$$

(C.4)并不总能求解，因而需要研究可解条件；在有解时要进一步研究解的唯一性条件、何时有多解（此时甚至可进一步研究解空间的结构）。

定理 C.1.2. (Hamburger 定理) (1) 数列 $\{m_n\}_{n=1}^\infty$ 可以实现为某个不集中在有限个原子上的分布 μ 的各阶矩的充要条件是：对任意 $n \geq 1$ ，矩阵 $H_n := (m_{i+j})_{0 \leq i, j \leq n}$ （称为 *Hankel* 矩阵）是正定的，其中总约定 $m_0 := 1$ 。
(2) 数列 $\{m_n\}_{n=1}^\infty$ 可以实现为某个集中在恰好 $N \geq 1$ 个原子上分布 μ 的各阶矩的充要条件是：对任意 $k = 0, 1, \dots, N-1$ ，*Hankel* 矩阵 H_k 是正定的；而对任意 $k \geq N$ ，*Hankel* 矩阵 H_k 是退化的非负定矩阵。

证明. 我们只给出 (1) 中的充分必要性；(2) 的证明类似。

必要性: 设随机变量 X 的分布为 μ 。对于固定的 $n \geq 1$, 考察数学期望

$$\mathbb{E}_\mu \left[\left(\sum_{k=0}^n x_k \cdot X^k \right)^2 \right] = \sum_{0 \leq k, \ell \leq n} m_{k+\ell} x_k x_\ell \geq 0.$$

如果对某 n 及非零 $x = (x_0, \dots, x_n)$ 上面数学期望取值为零, 那么随机变量 X 几乎处处满足下面的代数方程

$$\sum_{k=0}^n x_k \cdot X^k = 0.$$

这样, μ 只能是集中在有限个原子上的离散分布。矛盾。因此由各阶矩构造的 Hankel 矩阵应该是正定的。

充分性: 对任意固定的 $n \geq 1$, 考虑“截断矩问题”如下:

$$m_k = \int x^k d\mu, \quad k = 0, 1, \dots, 2n-1. \quad (\text{C.5})$$

在所给条件下, 它总是有解的 (具体证明细节见 [1], 此处从略; 也可直接证明, 截断矩问题有一个支撑在至多 $2n$ 个点上的分布测度作为解。), 设 μ_n 就是一个解。由 Helly 定理, 存在测度 μ (现在不知道是否是概率测度), 使得 μ 是 $\{\mu_n : n \geq 1\}$ 的某个子列的弱极限; 不妨设这个子列就是它本身。下面论证这个 μ 是概率测度, 且是矩问题的解。任取 $-A < 0, B > 0$ 为 μ 的连续点, 那么对任意 $k \geq 0$,

$$\int_{-A}^B x^k d\mu(x) = \lim_{n \rightarrow \infty} \int_{-A}^B x^k d\mu_n(x).$$

取整数 r 使得 $2r > k$, 对 $n > r$, 有

$$m_k = \int x^k d\mu_n = \left[\int_{-A}^B + \int_{-\infty}^{-A} + \int_B^{\infty} \right] x^k d\mu_n.$$

记 $C = \min(A, B)$, 有

$$\begin{aligned} \left| \left[\int_{-\infty}^{-A} + \int_B^{\infty} \right] x^k d\mu_n \right| &= \left| \left[\int_{-\infty}^{-A} + \int_B^{\infty} \right] \frac{x^{2r}}{x^{2r-k}} d\mu_n \right| \\ &\leq \frac{1}{C^{2r-k}} \left[\int_{-\infty}^{-A} + \int_B^{\infty} \right] x^{2r} d\mu_n \leq \frac{m_{2r}}{C^{2r-k}}, \end{aligned}$$

进而 $A, B \rightarrow \infty$ 时

$$\left| \int_{-A}^B x^k d\mu(x) - m_k \right| \leq \frac{m_{2r}}{C^{2r-k}} \rightarrow 0.$$

这证明了: $\int x^k d\mu(x) = m_k, k = 0, 1, \dots$ 。容易知道, 在所给条件下分布 μ 不集中在有限个原子上。□

C.1.3 Hamburger 矩问题: 解的唯一性与不唯一性

上面的 Hamburger 定理给出了矩问题的解存在的充分必要条件。在补充某些特殊条件的情况下, 给定了各阶矩就能唯一确定分布。例如, 假设分布是集中在有限区间的, 那么它的各阶矩唯一确定分布。一般而言, 我们有下面的结果。

☞ **定理 C.1.3.** (Carleman 准则; 证明见 [1]) 设数列 $\{m_n\}_{n=0}^{\infty}$ 是分布 F 的各阶矩。当

$$\sum_{n=0}^{\infty} \frac{1}{m_{2n}^{1/(2n)}} = \infty$$

时, $\{m_n\}_{n=0}^{\infty}$ 唯一决定分布 F 。

设数列 $\{m_n\}_{n=0}^{\infty}$ 是分布 F 的各阶矩, 且 F 是集中在 $[0, \infty)$ 上的分布函数 (此时的矩问题称为 *Stieltjes 矩问题*)。当

$$\sum_{n=0}^{\infty} \frac{1}{m_n^{1/(2n)}} = \infty$$

时, $\{m_n\}_{n=0}^{\infty}$ 唯一决定分布 F 。

下面的结果及证明均引自 [13, 5th Edition, Theorem 3.3.25], 也可参见 [49]; 它给出的条件略强于 Carleman 准则, 但更方便使用, 证明也不难。

☞ **定理 C.1.4.** 设数列 $\{m_n\}_{n=0}^{\infty}$ 是分布 F 的各阶矩。当

$$\overline{\lim}_{n \rightarrow \infty} \frac{m_{2n}^{1/(2n)}}{2n} = r < \infty$$

时, $\{m_n\}_{n=0}^{\infty}$ 唯一决定分布 F 。

证明. 设 $X \sim F$, 记 $\nu_k := \mathbb{E}[|X|^k]$ 。那么 $\nu_{2k+1}^2 \leq \nu_{2k}\nu_{2k+2}$, 因此不难证明

$$\overline{\lim}_{n \rightarrow \infty} \frac{\nu_n^{1/n}}{n} = r < \infty.$$

进而对任意 $\varepsilon > 0$, $\nu_n \leq [n(r + \varepsilon)]^n$ 对充分大 n 成立。

对任意 $n \geq 0$ 及实数 θ , 我们有下面的估计式:

$$|e^{i\theta} - \sum_{k=0}^{n-1} \frac{(i\theta)^k}{k!}| \leq \min\left(\frac{2|\theta|^{n-1}}{(n-1)!}, \frac{|\theta|^n}{n!}\right). \quad (\text{C.6})$$

那么对任意 $\theta, t \in \mathbb{R}$, $|e^{i\theta X}[e^{itX} - \sum_{k=0}^{n-1} \frac{(itX)^k}{k!}]| \leq \frac{|tX|^n}{n!}$ 。这意味着: 如果记 φ 为 F 的特征函数, 那么对充分大的 n

$$|\varphi(\theta + t) - \sum_{k=0}^{n-1} \frac{t^k}{k!} \cdot \varphi^{(k)}(\theta)| \leq \frac{|t|^n}{n!} \cdot \nu_n \leq \frac{|nt(r + \varepsilon)|^n}{n!}.$$

由 Stirling 公式 $n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\theta_n}$ (其中 $\theta_n = \frac{1+o(1)}{12n}$), 当 $|t| < \frac{1}{e(r+\varepsilon)}$ 时

$$\frac{|nt(r + \varepsilon)|^n}{n!} = \frac{|et(r + \varepsilon)|^n}{\sqrt{2\pi n}} \cdot [1 + o(1)] \rightarrow 0.$$

也就是说

$$\varphi(\theta + t) = \varphi(\theta) + \sum_{k=1}^{\infty} \frac{t^k}{k!} \cdot \varphi^{(k)}(\theta) \quad (\text{C.7})$$

对所有 $\theta \in \mathbb{R}$ 、所有 $|t| < \frac{1}{er} =: \delta$ 成立。易知 $\varphi(0) = 1$, $\varphi^{(k)}(0) = i^k \cdot m_k$, $k = 1, 2, \dots$ 。根据(C.7), φ 在 $(-\delta, \delta)$ 上由各阶矩 $\{m_k\}_{k=1}^{\infty}$ 唯一决定; 进而,

φ 在 \mathbb{R} 上的取值由各阶矩 $\{m_k\}_{k=1}^\infty$ 唯一决定。由特征函数的唯一性定理，本定理结论成立。 \square

例 C.1. 对于标准正态分布， $m_{2n+1} = 0, m_{2n} = (2n-1)!! = \frac{(2n)!}{2^n \cdot n!}$ ，由 *Stirling* 公式（其中 $\theta_n \rightarrow 0$ ）

$$(m_{2n})^{1/(2n)} = \left[\frac{\sqrt{4\pi n} \cdot \left(\frac{2n}{e}\right)^{2n} \cdot e^{\theta_{2n}}}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n \cdot e^{\theta_n} \cdot 2^n} \right]^{1/(2n)} \rightarrow \sqrt{\frac{2}{e}}.$$

由 *Carleman* 准则或上面的定理，它的矩问题具有唯一解。

下面的 *Krein* 条件则能给出各阶矩不能唯一决定分布的众多例子。

定理 C.1.5.（矩问题的 *Krein* 条件）设分布 F 具有概率密度 ρ ，它满足

$$\int \frac{-\ln \rho(x)}{1+x^2} dx < \infty,$$

并且具有有限的各阶矩 $\{m_n : n \geq 0\}$ 。此时，矩问题的解不唯一。即存在分布 $G \neq F$ ，它具有与 F 相同的各阶矩。

证明. 这是 Mark G. Krein（1907/4/3–1989/10/17；苏联数学家，犹太裔，他是当时苏联泛函分析学派的重要人物，以算子理论等方向的贡献知名，1982 年获得 Wolf 奖）在 1940 年左右的工作，证明参见 [1] 或 [22]。 \square

例 C.2. 对于对数标准正态分布（即 $X = e^Z$ 的分布，其中 $Z \sim N(0, 1)$ ），请读者验证其各阶矩存在且有限，并且 *Krein* 条件成立（而 *Carleman* 准则不成立）。因此这个分布的矩问题的解不唯一。在 [13, 5th Edition, §3.3.5] 中给出了一族连续型分布和一族离散型分布，它们都具有与对数标准正态同样的各阶矩。

讨论矩问题的一个目的是研究用矩方法来论证依分布收敛的可行性。现在我们可以给出如下的一个答案。

定理 C.1.6. 设随机变量 X 的各阶矩都存在，满足

$$\lim_{n \rightarrow \infty} \frac{(\mathbb{E}[X^{2n}])^{1/(2n)}}{2n} = r < \infty$$

或 *Carleman* 准则中条件。如果有随机变量序列 $\{X_n\}_{n=1}^\infty$ ，使得

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n^k] = \mathbb{E}[X^k], \quad k = 1, 2, \dots,$$

那么 $X_n \xrightarrow{d} X$ 。

C.2 Laplace 变换

在本节，我们考虑的分布或测度都是 $\mathbb{R}_+ := [0, \infty)$ 上的。这时，使用 *Laplace* 变换通常会比特征函数（或 *Fourier* 变换）更简单。由于应用的需要，我们有必要提供这方面的有关理论论述；但由于 *Laplace* 变换的很多性质和证明与特征函数部分的理论类似，在那样的场景我们将略去有关证明。本节大部分材料来源于 [18]；对更深入的理论感兴趣的读者可以参考文献 [39] 等专门讨论 *Laplace* 变换的文献。

C.2.1 Laplace 变换的定义与基本性质

通常，如果 $f: \mathbb{R}_+ := [0, \infty) \rightarrow \mathbb{R}$ 是 $L^1(\mathbb{R}_+)$ 中函数，对任意 $\lambda \geq 0$ ，我们可以定义

$$\mathcal{L}f(\lambda) := \int_0^\infty f(x)e^{-\lambda x} dx. \quad (\text{C.8})$$

我们称函数 $\mathcal{L}f: \mathbb{R}_+ \rightarrow \mathbb{R}$ 为 f 的 Laplace 变换。但在此处，我们可以对非负随机变量的分布函数 F 定义更广的 Laplace 变换（为了区分，我们记作 \mathcal{L}_F 或 $\mathcal{L}\mu_F$ ；个别场合我们也允许突破仅对非负随机变量的分布谈 Laplace 变换的限制）

$$\mathcal{L}_F(\lambda) := \int_0^\infty e^{-\lambda x} dF(x). \quad (\text{C.9})$$

容易知道，当 $X \sim F$ 是非负随机变量时（我们也记 $\mathcal{L}_X = \mathcal{L}_F$ ），

$$\mathcal{L}_F(\lambda) = \mathbb{E}[e^{-\lambda X}]. \quad (\text{C.10})$$

显然，当 F 具有密度 f 时， $\mathcal{L}_F = \mathcal{L}f$ 。

更一般的，我们可以对 \mathbb{R}_+ 上非负有限测度（甚至符号测度） μ 定义 Laplace 变换 $\mathcal{L}\mu: \mathbb{R}_+ \rightarrow \mathbb{R}$ 如下

$$\mathcal{L}\mu(\lambda) := \int_0^\infty e^{-\lambda x} d\mu(x), \quad (\text{C.11})$$

如果上述积分有意义。这些推广的 Laplace 变换有时也称为 Laplace-Steljes 变换。在本节，我们重点关注 \mathbb{R}_+ 上概率测度或概率分布的 Laplace 变换。

我们把 \mathbb{R}_+ 上概率测度的全体记作 $\mathcal{M}^1(\mathbb{R}_+)$ ， \mathbb{R}_+ 上有限（符号）测度的全体记作 $\mathcal{M}^f(\mathbb{R}_+)$ ， \mathbb{R}_+ 上有限正测度的全体记作 $\mathcal{M}_+^f(\mathbb{R}_+)$ 。一般的，我们可以把 \mathbb{R} 上有限（符号）测度的全体记作 $\mathcal{M}^f(\mathbb{R})$ 。对任意 $\nu_1, \nu_2 \in \mathcal{M}^f(\mathbb{R})$ ，我们可以定义卷积运算 $\nu_1 * \nu_2$ 如下：

$$\nu_1 * \nu_2(A) := \int \nu_1(A - x) d\nu_2(x) = \int \nu_2(A - x) d\nu_1(x), \quad \forall A \in \mathcal{B}.$$

容易知道此时 $\nu_1 * \nu_2 \in \mathcal{M}^f(\mathbb{R})$ 。

有时我们需要考虑 \mathbb{R}_+ 上 Radon 测度，其全体记作 $\mathcal{M}_+^R(\mathbb{R}_+)$ ；对应地， $\mathcal{M}^R(\mathbb{R}_+)$ 表示 Radon 符号测度的全体。容易知道，如果 \mathbb{R}_+ 上非负测度 μ 满足：存在 $a \in \mathbb{R}_+$ 使得

$$\mathcal{L}\mu(\lambda) := \int_0^\infty e^{-\lambda x} d\mu(x)$$

满足 $\mathcal{L}\mu(a) < \infty$ ，那么

$$\mu([0, x]) \leq e^{ax} \mathcal{L}\mu(a), \quad \forall x \in \mathbb{R}_+.$$

反之，如果 μ 满足：存在 $a \geq 0, C > 0$ 使得

$$\mu([0, x]) \leq Ce^{ax}, \quad \forall x \in \mathbb{R}_+, \quad (\text{C.12})$$

那么

$$\mathcal{L}\mu(\lambda) < \infty, \quad \forall \lambda > a;$$

这类测度我们可以称之为指数控制的，其全体记作 $\mathcal{M}_+^{EC,R}(\mathbb{R}_+)$ 。如果 μ 是 \mathbb{R}_+ 上符号测度，且其全变差测度 $|\mu| \in \mathcal{M}_+^{EC,R}(\mathbb{R}_+)$ ，则仍然称 μ 为指数控

制的，记作 $\mu \in \mathcal{M}^{EC,R}(\mathbb{R}_+)$ 。

关于 Laplace 变换 \mathcal{L} ，我们有下面一些基本性质（其中部分性质可以推广到更一般的形态，留给读者自行思考）。

性质（1）（线性性质）在 $\mathcal{M}^f(\mathbb{R}_+)$ （或 $\mathcal{M}^{EC,R}(\mathbb{R}_+)$ ）上， \mathcal{L} 是线性算子；

性质（2）（正算子/单调性）在 $\mathcal{M}^f(\mathbb{R}_+)$ （或 $\mathcal{M}^{EC,R}(\mathbb{R}_+)$ ）上， \mathcal{L} 是正算子。

特别的，对任意 $\mu \in \mathcal{M}_+^f(\mathbb{R}_+)$ ， $\mathcal{L}\mu$ 是单调非增函数，并且

$$0 \leq \mathcal{L}\mu(\lambda) \leq \mathcal{L}\mu(0) = \mu(\mathbb{R}_+) < \infty, \forall \lambda \geq 0;$$

性质（3）（卷积性质）如果 $\nu_1, \nu_2 \in \mathcal{M}^f(\mathbb{R}_+)$ ，那么：

$$\mathcal{L}(\nu_1 * \nu_2)(\lambda) = \mathcal{L}\nu_1(\lambda)\mathcal{L}\nu_2(\lambda);$$

性质（4）（连续可微性与矩）如果非负随机变量 $X \sim F$ ，则 $\mathcal{L}\mu_F$ 是一个连续函数；对于正整数 $n \in \mathbb{N}$ ，如果 $\mathbb{E}[X^n] < \infty$ ，那么 $\mathcal{L}\mu_F$ 是 n 阶连续可微函数，并且对任意 $\lambda \geq 0$ 及 $0 \leq k \leq n$

$$(-1)^k D^k \mathcal{L}\mu_F(\lambda) = \mathbb{E}[e^{-\lambda X} \cdot X^k].$$

特别的，对于“好的函数” f 以及任意 $k \in \mathbb{N}$

$$(-1)^k D^k \mathcal{L}f(\lambda) = \mathcal{L}[x^k f(x)](\lambda);$$

性质（5）（分部积分）设非负随机变量 $X \sim F$ ，那么

$$\int_0^\infty e^{-\lambda x} F(x) dx = \frac{\mathcal{L}\mu_F(\lambda)}{\lambda}, \quad \forall \lambda > 0.$$

上式也可以改写为

$$\int_0^\infty e^{-\lambda x} [1 - F(x)] dx = \frac{1 - \mathcal{L}\mu_F(\lambda)}{\lambda}, \quad \forall \lambda > 0.$$

性质（6）（尺度变换与平移变换）如果 X 是非负随机变量，那么

$$\mathcal{L}_X(a\lambda) = \mathcal{L}_{aX}(\lambda), \forall a > 0$$

以及

$$\mathcal{L}_{X+a}(\lambda) = e^{-a\lambda} \mathcal{L}_X(\lambda), \forall a \in \mathbb{R}.$$

另外，对任意 $\mu \in \mathcal{M}_+^{EC,R}(\mathbb{R}_+)$ ，如果 $\mathcal{L}\mu(b) < \infty$ ，则对任意 $a \in \mathbb{R}$

$$\mathcal{L}\mu(\lambda + a) = \mathcal{L}\mu_a^\#(\lambda), \forall \lambda \geq b - a,$$

其中 $\mu_a^\#(A) := \int_A e^{-ax} d\mu(x), \forall A \in \mathcal{B}(\mathbb{R}_+)$ 。

我们对上面性质（6）的后半部分加一点注记。事实上，如果非负 Radon 测度 μ 满足 $\mathcal{L}\mu(b) < \infty$ ，那么容易知道

$$\mu_b^\#(\mathbb{R}_+) = \int_{\mathbb{R}_+} e^{-bx} d\mu(x) = \mathcal{L}\mu(b) < \infty,$$

从而 $\mu_b^\# \in \mathcal{M}^f(\mathbb{R}_+)$ ，并且 $\mathcal{L}\mu_b^\#(\lambda) = \mathcal{L}\mu(\lambda + b), \forall \lambda \geq 0$ 。此时，我们还有

$$\frac{d\mu_b^\#}{d\mu}(x) = e^{-bx}, \quad \frac{d\mu}{d\mu_b^\#}(x) = e^{bx}.$$

因此 μ 与 $\mu_b^\#$ 之间是一一对应的关系。由此, 在后文的有关 Laplace 变换的 (连续性、唯一性与刻画等的) 理论讨论中, 仅限于有限测度、特别是概率测度来进行讨论, 只会使得有关探讨更为简单、表达更为简洁。

C.2.2 唯一性定理与 Laplace 变换的反演公式

与特征函数类似, \mathbb{R}_+ 上概率分布的 Laplace 变换能唯一确定对应的分布律。这也被称为 Laplace 变换的唯一性定理; 它也有所谓的反演公式。此处我们给出如下两个反演公式: 一者是有关分布的 Laplace 变换的反演公式; 一者是有关函数 (相当于密度) 的 Laplace 变换的反演公式 (这被称为 Widder 逆转公式)。

定理 C.2.1. (1) \mathbb{R}_+ 上概率测度 μ 由它的 Laplace-Steljes 变换 $\mathcal{L}\mu$ 在任一给定区间 $[a, \infty)$ 上的取值唯一决定。事实上, 如果 $F(x) = \mu([0, x]), x \geq 0$, 那么在 F 的连续点 $x > 0$ 上有

$$F(x) = \lim_{\lambda \rightarrow \infty} \sum_{0 \leq n \leq \lambda x} \frac{(-\lambda)^n}{n!} D^n \mathcal{L}\mu(\lambda). \quad (\text{C.13})$$

(2) 设 \mathbb{R}_+ 上连续函数 f 存在 Laplace 变换。则 f 由它在某区间 $0 \leq a < \lambda < \infty$ 上的普通 Laplace 变换 $\mathcal{L}f$ 的取值唯一决定。特别的, 如果 f 是有界一致连续函数, 则 $\mathcal{L}f(\lambda)$ 对 $\lambda > 0$ 有意义, 并且

$$f(x) = \lim_{n \rightarrow \infty} \frac{(-1)^{n-1}}{(n-1)!} \left(\frac{n}{x}\right)^n (D^{n-1} \mathcal{L}f)\left(\frac{n}{x}\right) \quad (\text{C.14})$$

证明. 我们只证明 (1), (2) 的证明请参见习题 C.1。

这里, 如果 $X \sim \mu$, 那么对任意 $\lambda > 0$, $\mathcal{L}\mu(\lambda) = \mathbb{E}[e^{-\lambda X}]$, 且

$$(-1)^n D^n \mathcal{L}\mu(\lambda) = \mathbb{E}[e^{-\lambda X} X^n].$$

此外, 对任意给定的 $\theta > 0, \varepsilon > 0$, 注意到如果 $Y_{\lambda, \theta} \sim \text{Poisson}(\lambda\theta)$, 那么当 $\lambda \rightarrow \infty$ 时

$$\mathbb{P}(|Y_{\lambda, \theta} - \lambda\theta| > \lambda\varepsilon) \leq \frac{\text{Var}(Y_{\lambda, \theta})}{\lambda^2 \varepsilon^2} = \frac{\theta}{\lambda \varepsilon^2} \rightarrow 0.$$

更细致的分析 (第 11 章的特征函数方法) 表明, 当 $\lambda \rightarrow \infty$ 时

$$\frac{Y_{\lambda, \theta} - \lambda\theta}{\sqrt{\lambda\theta}} \xrightarrow{d} N(0, 1).$$

这意味着

$$\mathbb{P}(Y_{\lambda, \theta} \leq \lambda x) = e^{-\lambda\theta} \sum_{0 \leq n \leq \lambda x} \frac{(\lambda\theta)^n}{n!} \xrightarrow{\lambda \rightarrow \infty} \begin{cases} 0, & \text{如果 } \theta > x \\ 1, & \text{如果 } \theta < x \\ \frac{1}{2}, & \text{如果 } \theta = x. \end{cases}$$

因此,

$$\lim_{\lambda \rightarrow \infty} e^{-\lambda X} \sum_{0 \leq n \leq \lambda x} \frac{(\lambda X)^n}{n!} = 1_{\{X < x\}} + \frac{1}{2} 1_{\{X = x\}}.$$

根据控制收敛定理, 对于 F 的连续点 $x > 0$, 我们有 (C.13) 成立。□

C.2.3 Laplace 变换的连续性定理

在本小节，为了后续应用的需要，我们考虑 \mathbb{R}_+ 上比概率测度概念略广泛一点的 Radon 测度的 Laplace 变换。我们列出如下关于 Laplace 变换的连续性定理而不给出证明。此处，对于 Radon 测度列 $\{\mu_n\}_{1 \leq n \leq \infty}$ ，定义 $F_n(x) := \mu_n([0, x])$ ，如果

$$F_n(x) \rightarrow F_\infty(x)$$

在 F_∞ 的所有连续点上成立，则认为 $\mu_n \xrightarrow{w} \mu_\infty$ 。

定理 C.2.2. (1) 设 \mathbb{R}_+ 上 Radon 测度列 $\{\mu_n\}_{1 \leq n \leq \infty}$ 满足 $\mu_n \xrightarrow{w} \mu_\infty$ ，且存在 $a \geq 0$ 使得 $\{\mathcal{L}\mu_n(a)\}_{n=0}^\infty$ 有界。那么 $\mathcal{L}\mu_n(\lambda) \rightarrow \mathcal{L}\mu_\infty(\lambda), \forall \lambda > a$ 。

(2) 设 \mathbb{R}_+ 上 Radon 测度列 $\{\mu_n\}_{1 \leq n \leq \infty}$ 满足：存在 $a \geq 0$ ， $\mathcal{L}\mu_n(\lambda)$ 对 $\lambda > a$ 有定义，并且 $\mathcal{L}\mu_n(\lambda) \rightarrow \varphi(\lambda), \forall \lambda > a$ 。那么 φ 是某个 Radon 测度 μ_∞ 的 Laplace 变换，并且 $\mu_n \xrightarrow{w} \mu_\infty$ 。

上述定理的证明请参见 [18, Chapter 13]。在那里，Feller 先讨论了概率测度版本的 Laplace 变换的连续性定理，之后再证明了上面的连续性定理；在他的著作中，这个定理被称为广义连续性定理。

读者也可以仿照之前特征函数的连续性定理来完成上述定理（概率测度版本）的证明：定理中（1）的证明仍然可以基于 Skorodhod 嵌入定理；定理中（2）的部分则可以基于 Helly 定理来完成。一般 Radon 测度情形下的结论可以基于概率测度情形的结果而给出证明，也留给读者自行尝试。

C.2.4 Laplace 变换的刻画定理

什么样的函数可以表示为 \mathbb{R}_+ 上某概率分布的 Laplace 变换？下面的 Bernstein 定理回答了这个问题。

定理 C.2.3. (Bernstein 定理, 1928) 函数 $\varphi: \mathbb{R}_+ \rightarrow \mathbb{R}$ 是 \mathbb{R}_+ 上某概率分布的 Laplace 变换的充分必要条件是：

(1) $\varphi(0) = 1$;

(2) φ 是完全单调的，即： φ 的各阶导数均存在，且

$$(-1)^n D^n \varphi(\lambda) \geq 0, \quad \forall \lambda > 0. \quad (\text{C.15})$$

证明. 必要性部分很简单，以下只证明充分性部分。为方便，我们不妨假定

$$\varphi(\infty) := \lim_{\lambda \uparrow \infty} \varphi(\lambda) = 0.$$

上述条件是用于保证 0 不是对应的分布的原子；否则，我们只需考虑

$$\tilde{\varphi}(\lambda) := \frac{\varphi(\lambda) - \varphi(\infty)}{1 - \varphi(\infty)}$$

即可。此时，本定理可以视作 Hausdorff 矩问题的一个应用。

事实上，不难知道，在所给条件下，对任意固定的整数 $m \geq 1$ ， $\{\varphi(\frac{n}{m})\}_{n=0}^\infty$ 是完全单调序列，于是 $[0, 1]$ 上存在唯一的概率分布函数 F_m 使得

$$\varphi(\frac{n}{m}) = \int_0^1 t^n dF_m(t), \quad n \geq 0.$$

显然 $\tilde{F}_m(x) := F_m(x^{1/m})$ 也是一个分布函数, 并且

$$\varphi(n) = \int_0^1 t^{mn} dF_m(t) = \int_0^1 x^n d\tilde{F}_m(x), \quad n \geq 0.$$

唯一性说明了 $\tilde{F}_m(t) = F_1(t)$ 。现在令 $F(t) = 1 - F_1(e^{-t}), t \geq 0$, 则

$$\varphi\left(\frac{n}{m}\right) = \int_0^1 t^{n/m} dF_1(t) = \int_0^\infty e^{-nt/m} dF(t), \quad n \geq 0, m \geq 1.$$

由连续性, 立即得到

$$\varphi(\lambda) = \int_0^\infty e^{-\lambda t} dF(t). \quad \square$$

C.2.5 Laplace 变换的 Tauber 定理

对数列 $\{a_n\}_{n=0}^\infty$, 令 $s_n := \sum_{k=0}^n a_k$ 。当存在 $A \in \mathbb{R}$ 使得 $s_n \rightarrow A$ 时, 我们称数列 $\{a_n\}_{n=0}^\infty$ (普通) 可和; 而当

$$\sigma_n := \frac{1}{n} \sum_{k=0}^{n-1} s_k \rightarrow A$$

时, 则称数列 $\{a_n\}_{n=0}^\infty$ Cesàro 可和。当实数列 $\{a_n\}_{n=0}^\infty$ 的母函数

$$f(x) := \sum_{n=0}^\infty a_n x^n \quad (\text{C.16})$$

对 $|x| < 1$ 收敛, 且存在 A 使得 $f(1-) = A$, 则称数列 $\{a_n\}_{n=0}^\infty$ Abel 可和。

分析学中常见的一个 Abel 定理如下:

定理 C.2.4. (Abel 定理) (1) 如果数列 $\{a_n\}_{n=0}^\infty$ 普通可和, 那么它也是 Abel 可和的, 并且

$$\lim_{x \rightarrow 1-} \sum_{n=0}^\infty a_n x^n = \sum_{n=0}^\infty a_n. \quad (\text{C.17})$$

(2) 如果数列 $\{a_n\}_{n=0}^\infty$ Cesàro 可和, 那么它也是 Abel 可和的, 并且

$$\lim_{x \rightarrow 1-} \sum_{n=0}^\infty a_n x^n = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} s_k. \quad (\text{C.18})$$

对于上述 Abel 定理: (a) 对数列 $\{a_n := (-1)^n\}_{n=0}^\infty$, 它不是普通可和的, 但 $f(x) = \frac{1}{1+x}$, 此时数列 $\{a_n := (-1)^n\}_{n=0}^\infty$ 对应的 Abel 和为 $f(1-) = \frac{1}{2}$, $s_{2n} = 1, s_{2n+1} = 0$, 从而数列 $\{a_n := (-1)^n\}_{n=0}^\infty$ 的 Cesàro 和恰为 $\frac{1}{2}$; (b) 对数列 $\{a_n := (-1)^n(n+1)\}_{n=0}^\infty$, 有 $s_{2n-1} = -n, s_{2n} = n+1$, 它不是 Cesàro 可和的, 但它是 Abel 可和的, 对应的 Abel 和为 $\frac{1}{4}$ 。

下面陈述的分析中常见的 Tauber 第一定理讨论的就是上述 Abel 定理中(C.17)左端极限存在时, 何时该式仍然成立; 所附加的条件通常称为 Tauber 条件。

☞ **定理 C.2.5.** (Tauber, 1897) 若数列 $\{a_n\}_{n=0}^{\infty}$ Abel 可和且 $a_n = o(\frac{1}{n})$, * 那么数列 $\{a_n\}_{n=0}^{\infty}$ 普通可和, 且(C.17)成立。

上述定理的一个证明思路如下: 令 $r = 1 - \frac{1}{N}$, $\epsilon := \sup\{n|a_n| : n \geq N\}$, 作估计

$$\left| \sum_{n=0}^N a_n - \sum_{n=0}^N a_n r^n \right| \leq \sum_{n=1}^N |a_n| [1 - (1 - \frac{1}{N})^n] \leq \sum_{n=1}^N |a_n| \frac{n}{N}$$

以及

$$\begin{aligned} \left| \sum_{n=1}^{\infty} a_n r^n - \sum_{n=1}^N a_n r^n \right| &\leq \sum_{n>N} |a_n| (1 - \frac{1}{N})^n \\ &\leq \sum_{n>N} \frac{n|a_n|}{N} (1 - \frac{1}{N})^n \\ &\leq \sum_{n>N} \frac{\epsilon}{N} (1 - \frac{1}{N})^n = \epsilon (1 - \frac{1}{N})^{N+1}. \end{aligned}$$

由 $na_n \rightarrow 0$ 不难得知其前 N 项和的算术平均也趋于零, 而 $(1 - \frac{1}{N})^{N+1} \rightarrow e^{-1}$. 综合这两个估计, 就证明了上述定理。

Tauber 进一步在他自己工作的基础上建立了如下的 Tauber 第二定理。

☞ **定理 C.2.6.** (Tauber 第二定理) $\sum_{n=0}^{\infty} a_n = A \in \mathbb{R}$ 成立的充分必要条件是: 数列 $\{a_n\}_{n=0}^{\infty}$ Abel 可和, 且

$$\lim_{x \rightarrow 1^-} \sum_{n=0}^{\infty} a_n x^n = A,$$

并且 Tauber 条件 $\sum_{k=0}^n k a_k = o(n)$ 成立。

上述 Abel 定理和 Tauber 定理的精神是通过构建数列 $\{a_n\}_0^{\infty}$ 的母函数来分析有关临界点处的极限行为。Hardy 和 Littlewood 进一步证明了如下形态的 Tauber 定理: †

☞ **定理 C.2.7.** 设 $f(x) := \sum_{n=0}^{\infty} a_n x^n$ 对 $|x| < 1$ 收敛, 且存在 $\alpha \geq 0$ 使得

$$(1-x)^{\alpha} f(x) \rightarrow A, \quad x \nearrow 1$$

且 $na_n \geq O(n^{\alpha})$, 那么

$$\frac{s_n}{n^{\alpha}} \rightarrow \frac{A}{\Gamma(1+\alpha)}, \quad n \rightarrow \infty.$$

*Littlewood 在 1910 年证明了 Tauber 第一定理中的条件 $a_n = o(\frac{1}{n})$ 可以减弱为 $a_n = O(\frac{1}{n})$ 。

†有关 Tauber 理论的更多、更深入讨论, 感兴趣的读者可以阅读 [21] 等文献。

类似的，我们有下面关于 Laplace 变换的 Tauber 定理。只不过我们需要引入规则变化函数的概念，以方便论述。设 $U : [0, \infty) \rightarrow (0, \infty)$ 为可测函数，如果存在 $\rho \in \mathbb{R}$ 使得：对任意 $x > 0$

$$\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^\rho, \quad (\text{C.19})$$

则称 U 是以 ρ 为指标的规则变化函数 (Regularly Varying Function/Function of Regular Variation)，记作 $U \in R_\rho$ 或 $U \in R_\rho(\infty)^*$ ； $\rho = 0$ 时又称 U 为缓慢变化函数 (Slowly Varying Function/Function of Slow Variation)。规则变化函数的概念是 J. Karamata 于 1930 年引入，这一概念已经被证明在许多方面富有成效，且在概率论中的应用越来越广泛；关于规则变化函数的系统论述，读者可以参见 [3]。

☞ 定理 C.2.8. 设 μ 是 \mathbb{R}_+ 上 Radon 测度，记 $F(x) = \mu([0, x]), x > 0$ 。设

$$\mathcal{L}\mu(\lambda) := \int_0^\infty e^{-\lambda x} dF(x)$$

对任意 $\lambda > 0$ 存在。设 $\tau t = 1, \gamma \geq 0$ 。那么以下两个关系式

$$\frac{\mathcal{L}\mu(\lambda\tau)}{\mathcal{L}\mu(\tau)} \rightarrow \lambda^{-\gamma}, \quad \tau \searrow 0, \quad (\text{C.20})$$

$$\frac{F(tx)}{F(t)} \rightarrow x^\gamma, \quad t \nearrow \infty \quad (\text{C.21})$$

之间是等价的，且都蕴含了

$$\mathcal{L}\mu(\tau) \sim F(t) \cdot \Gamma(1 + \gamma), \quad \tau \searrow 0. \quad (\text{C.22})$$

上述结论在 $\tau \nearrow \infty, t \searrow 0$ 情形也成立。用规则变化函数的术语，可以把上面的结论表述为：

$$\mathcal{L}\mu \in R_{-\gamma}(0) \Leftrightarrow F \in R_\gamma(\infty), \quad \mathcal{L}\mu \in R_{-\gamma}(\infty) \Leftrightarrow F \in R_\gamma(0).$$

证明. 以下证明思路来自 [18, Chapter 13]；读者也可参考 [39, Chapter V] 中的证明方法。

(a) 先讨论 $\gamma > 0$ 的情形。假定 (C.20) 成立，它左边对应的是 $G_\tau(x) := \frac{F(tx)}{\mathcal{L}\mu(\tau)}$ （这对应于 \mathbb{R}_+ 上一个 Radon 测度）的 Laplace 变换。由连续性定理，有

$$G_\tau(x) = \frac{F(tx)}{\mathcal{L}\mu(\tau)} \rightarrow \frac{x^\gamma}{\Gamma(1 + \gamma)}.$$

对上式，取 $x = 1$ ，就得到了 (C.22)。把它代回上面表达式，就得到了 (C.21)。

当 $\gamma = 0$ 时，容易知道上面讨论仍然成立，只不过 $G_\tau(x)$ 对应的 Radon 测度收敛到 Dirac 测度 δ_0 。

(b) 假定 (C.21) 成立，即 $F_t(x) := \frac{F(tx)}{F(t)} \rightarrow x^\gamma$ ，设 F_t 对应的测度为 μ_t 。我们验证 $\frac{\mathcal{L}\mu(\tau)}{F(t)} = \mathcal{L}\mu_t(1)$ 有界。事实上，

$$\mathcal{L}\mu(\tau) = \int_0^\infty e^{-\tau x} dF(x) \leq \sum_{n=0}^\infty e^{-2^{n-1}} F(2^n t).$$

* 记号 $R_\rho(\infty)$ 是用以强调极限 $t \rightarrow \infty$ ；我们也可以谈论 $t \rightarrow 0$ 下的规则变化函数，此时以 ρ 为指标的规则变化函数的全体对应记作 $R_\rho(0)$ 。

由 F 是单调递增的规则函数, 我们知道 (参见 [3]): 存在 $C \geq 1$ 使得

$$\frac{F(xt)}{F(t)} \leq Cx^{1+\gamma}, \forall x \geq 1, t > 0.$$

于是

$$\frac{\mathcal{L}\mu(\tau)}{F(t)} \leq C \sum_{n=0}^{\infty} e^{-2^{n-1}} 2^{n(1+\gamma)} < \infty.$$

由连续性定理, 对 $F_t(x) := \frac{F(tx)}{F(t)} \rightarrow x^\gamma$ 取 Laplace 变换, 就得到

$$\frac{\mathcal{L}\mu(\tau\lambda)}{F(t)} \rightarrow \frac{\Gamma(1+\gamma)}{\lambda^\gamma}.$$

由此, 同 (a) 部分推理就得到 (C.22) 和 (C.20)。 \square

但更多时候, 我们关心概率分布函数 F 的尾部 $\bar{F}(x) := 1 - F(x)$ 的极限行为。设 F 诱导分布测度 μ 。定义

$$\bar{\mu}(A) := \int_A \bar{F}(x) dx.$$

则回顾 Laplace 变换的分部积分性质 (5)

$$\mathcal{L}\bar{\mu}(\lambda) = [1 - \mathcal{L}\mu(\lambda)]/\lambda, \forall \lambda > 0.$$

由上面的定理立即得到 (其中 $R_0(\infty)$ 表示缓慢变化函数):

定理 C.2.9. 设 μ 是 \mathbb{R}_+ 上概率分布测度, 记 $F(x) = \mu([0, x]), x > 0$ 。对任意 $\alpha \in [0, 1], \ell \in R_0(\infty)$, 我们有下面的结论 (其中 $\tau \searrow 0, t \nearrow \infty$):

(a) 当 $\alpha \in [0, 1)$ 时,

$$1 - \mathcal{L}\mu(\tau) \sim \tau^\alpha \ell(1/\tau) \Leftrightarrow 1 - F(t) \sim \frac{\ell(t)}{t^\alpha \Gamma(1-\alpha)};$$

(b) 当 $\alpha = 1$ 时,

$$\int_{[0,t]} x dF(x) \sim \ell(t) \Leftrightarrow \int_0^t [1 - F(x)] dx \sim \ell(t).$$

上面定理来自 [3, Corollary 8.1.7]。更多相关结果请参见 [3, §8.3]。

注 C.2. David Vernon Widder (威德, 1898/03/25–1990/07/08) 是美国数学家。他在 G. Birkhoff 指导下于 1924 年毕业于哈佛大学, 并留在哈佛任教。他是 *Duke Mathematical Journal* 杂志的联合创始人之一, 著有以下有影响力的教材或专著: (1) 《Advanced Calculus》; (2) 《The Laplace transform》(在此书中他给出了关于 Dirichlet zeta 函数的 Landau 问题的第一个解答); (3) 《An introduction to transform theory》; (4) 《The convolution transform》(与 I. I. Hirschman 合著)。

Alfred Tauber (陶伯, 1866/11/05–1942/07/26) 是出生在匈牙利的奥地利数学家, 以在数学分析和单复变函数方向的贡献而知名; 以他的名字命名的 Tauber 定理有多个, 应用于数学分析、调和分析、数论等多个方向。他 1884 年就读 Vienna 大学, 1889 年获得博士学位。1892–1908 年作为首席数学家供职于 Phönix 保险公司, 之后正式在 Vienna 大学任教 (之前从 1901 年起在 TU Wien 和 Vienna 大学有兼职)。他 1942 年被杀害于 Theresienstadt 集中营。

习 题 C

习题 C.1. 本题来自 [61]。设 $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ 是连续函数, 且 $f \in L^r(\mathbb{R}_+)$, 其中 $r > 1$ 。设 f 的 *Laplace* 变换为

$$\mathcal{L}f(\lambda) := \int_0^\infty e^{-\lambda t} f(t) dt,$$

则

$$f(x) = \lim_{n \rightarrow \infty} \frac{(-1)^{n-1}}{(n-1)!} \left(\frac{n}{x}\right)^n (D^{n-1} \mathcal{L}f)\left(\frac{n}{x}\right)$$

在每个有限区间中一致成立。此处 D 表示求导函数的算子。【提示: 对每个 $\lambda > 0$, 取 $\{X_k\}_{k=1}^\infty \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(1)$, $S_n := X_1 + \cdots + X_n$, 则

$$\mathbb{E}[f(S_n/\lambda)] = \frac{(-1)^{n-1}}{(n-1)!} \lambda^n (D^{n-1} \mathcal{L}f)(\lambda),$$

且当 $\lambda = n/x$ 时, $S_n/\lambda \rightarrow x$ 依概率 (几乎处处) 成立。这是 *Laplace* 变换的 *Widder* 逆转公式的稍容易一些的改述。】

习题 C.2. 证明正态分布 $N(\mu, \sigma^2)$ 的各阶矩唯一确定其分布。

习题 C.3. 证明 *Poisson* 分布的各阶矩唯一确定其分布。

习题 C.4. 证明指数分布的各阶矩唯一确定其分布。

习题 C.5. 证明对数标准正态分布的各阶矩不能唯一确定其分布。

习题 C.6. 设 ξ 的矩问题具有唯一解, 它的特征函数为 f 。 ξ_n 的特征函数是 f_n , 且 $f_n(t) \rightarrow f(t)$ 在某区间 $[-\varepsilon, \varepsilon]$ 上成立, 那么 $\xi_n \xrightarrow{d} \xi$ 。

习题 C.7. 任给 $\kappa \in [0, 1]$, 定义

$$M_\kappa(t) := \sum_{n=0}^{\infty} \frac{t^n}{\Gamma(1+n\kappa)}, \quad t \in (-1, 1).$$

求证: 存在唯一分布函数 F_κ (它是集中在 $[0, \infty)$ 上的分布) 使得

$$M_\kappa(t) = \mathbb{E}[e^{tX}], \quad \text{其中 } X \sim F_\kappa.$$

容易知道, F_0 是标准指数分布函数, F_1 是常数 1 对应的分布函数。分布函数 $\{F_\kappa: \kappa \in [0, 1]\}$ 称为 **Mittag-Leffler** 分布族 [32], 在 *Darling-Kac* 理论中它们是占位时在适当尺度变换下的极限分布律 (见 [3, Theorem 8.11.2])。

习题 C.8. 本习题的目的是引导读者给出 *Hausdorff* 定理的一个证明。必要性部分已经由方程 (C.3) 给出; 以下只证明充分性部分。这部分只需注意到泛函分析中的 *Gelfand* 的表示定理: 设 M 为紧空间, 那么 $C(M)$ 上的连续、线性、正泛函与 M 上有限测度一一对应。显然,

$$t^n \mapsto \mu_n$$

确实能诱导 $C[0, 1]$ 上的一个线性泛函 α 。具体实现可以按照如下步骤。对任意实系数 n 阶多项式 $P_n(t) := \sum_{k=0}^n a_k t^k$, 我们可以定义

$$\alpha(P_n) := \sum_{k=0}^n a_k \mu_k.$$

下面我们说明 α 可以延拓为 $C[0, 1]$ 中线性泛函:

(1) 令 $\lambda_{n,\ell}(x) := C_n^\ell x^\ell (1-x)^{n-\ell}$, $n, \ell \geq 0$ (此处, 按照通常的约定, 对整数对 (n, ℓ) , 当 $\ell > n$ 时 $C_n^\ell := 0$), 则

$$\alpha_{n,\ell} := \alpha(\lambda_{n,\ell}) = C_n^\ell \cdot (-1)^{n-\ell} \Delta^{n-\ell} \mu_\ell \geq 0.$$

容易算出 $\sum_{\ell=0}^n \alpha_{n,\ell} = \mu_0 = 1$ 。

(2) 对任意 $f \in C[0, 1]$, 令 $B_n f(x) := \sum_{\ell=0}^n f(\frac{\ell}{n}) C_n^\ell x^\ell (1-x)^{n-\ell}$; 这是 f 对应的 *Bernstein* 多项式, 且当 $n \rightarrow \infty$ 时, $\|f - B_n f\| \rightarrow 0$, 其中 $\|f\| := \sup\{|f(x)| : x \in [0, 1]\}$ 。

(3) 不难知道, 对任意正整数 $k \geq 1$, 下式对 $x \in [0, 1]$ 一致收敛:

$$\lim_{n \rightarrow \infty} \prod_{i=0}^{k-1} \frac{x - \frac{i}{n}}{1 - \frac{i}{n}} = x^k.$$

由此不难验证: 如果 $f_k(x) = x^k, \forall k \geq 0$, 那么

$$\mu_k = \lim_{n \rightarrow \infty} \alpha(B_n f_k).$$

进而可以对任意 $f \in C[0, 1]$, 定义

$$\alpha(f) := \lim_{n \rightarrow \infty} \alpha(B_n f).$$

上面极限的定义方式说明了 α 实际上是一个连续的线性泛函; 注意到

$$B_n f(x) = \sum_{\ell=0}^n f(\frac{\ell}{n}) \lambda_{n,\ell}(x), \alpha(B_n f) = \sum_{\ell=0}^n f(\frac{\ell}{n}) \alpha_{n,\ell},$$

α 是正算子。于是存在有限测度 α , 使得

$$\alpha(f) = \int_0^1 f(x) d\alpha(x).$$

进而由 $1 = \mu_0 = \alpha(1)$ 知道 α 是概率测度。

参考文献

- [1] Akhizer, N. I.: *The Classical Moment Problem: and some related questions in analysis*, (Translated by N. Kemmer), University mathematical monographs, The University Press, Glasgow.
- [2] Berkes, I.: *A remark to the law of the iterated logarithm*, Studia Sci. Math. Hungar. **7** (1972), No. 1-2, 189–197.
- [3] Bingham, N. H.; Goldie, C. M.; Teugels, J. L.: *Regular Variation*, Cambridge University Press, Cambridge, 1987.
- [4] Birkel, T.: *A note on the strong law of large numbers for positively dependent random variables*, Statistics & Probability Letters **7** (1989), 17–20.
- [5] Birkel, T.: *Law of large numbers under dependence assumptions*, Statistics & Probability Letters **14** (1992), 355–362.
- [6] Breiman, L.: *Probability*, Addison-Wesley, Cambridge, Mass.
- [7] Chen, Louis H. Y.; Goldstein L., Shao, Q.-M.: *Normal Approximation by Stein's Method*, Springer.
- [8] Chen, X.-X.; Xie, J.-S.; Ying, J.-G.: *Range-Renewal Processes: SLLN, Power Law and Beyonds*, arXiv:1305.1829.
- [9] Chung, K.-L.: *Note on Some Strong Laws of Large Numbers*, Amer. J. Math., Vol. **69**, No. 1 (Jan., 1947), pp. 189–192.
- [10] Chung, K.-L.: *Markov Chains with Stationary Transition Probabilities*, 2nd edition, 1967. Springer, Berlin. 301pp.
- [11] Chung, K.-L.: *A Course in Probability Theory*, 2nd edition, 1974. Acad. Press, New York.
- [12] Csörgő, S.; Tandori, K.; Totik, V.: *On the strong law of large numbers for pairwise independent random variables*, Acta Math. Hungar. **42** (1983), no. 3-4, 319–330. MR0722846

- [13] Durrett, R.: *Probability: Theory and Examples*, Fifth edition. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019. xii+419 pp.
- [14] Dvoretzky, A.; Erdős, P.: *Some problems on random walk in space*. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950. pp. 353–367. University of California Press, Berkeley and Los Angeles, 1951. MR0047272
- [15] Erdős, P.: *On the strong law of large numbers*, Trans. Amer. Math. Soc. **67** (1949), 51–56.
- [16] Etemadi, N.: *An Elementary Proof of the Strong Law of Large Numbers*, Z. Wahrscheinlichkeitstheorie verw. Gebiete **55**, 119–122 (1981).
- [17] Feller, W.: *A limit theorem for random variables with infinite moments*, Amer. J. Math., vol. **68** (1946), pp. 257–262.
- [18] Feller, W.: *An Introduction to Probability Theory and Its Applications*, Volume 2, 2nd Edition.
- [19] Grafakos, L.: *Classical Fourier Analysis* (Third Edition), GTM **249**, Springer.
- [20] Klenke, Achim: *Probability Theory: A Comprehensive Course*, Springer-Verlag, Berlin Heidelberg 2006. ISBN: 978-1-84800-047-6.
- [21] Korevaar, Jacob: *Tauberian Theory: A Century of Developments*, 科学出版社, 国外数学名著系列（影印版）36.
- [22] Kreĭn, M. G.; Nudelman, A. A.: *The Markov moment problem and extremal problems. Ideas and problems of P. L. čebyšev and A. A. Markov and their further development*. Translated from the Russian by D. Louvish. Translations of Mathematical Monographs, Vol. 50. American Mathematical Society, Providence, R.I., 1977. v+417 pp. ISBN: 0-8218-4500-4
- [23] Lehman, E. L.: *Some concepts of dependence*, Ann. Math. Statist. **37** (1966), 1137–1153.
- [24] Littlewood, J.: *Littlewood's Miscellany*. Edited by Bèla Bollobás. First published in 1953 as 'A Mathematician's Miscellany'. Cambridge University Press, Cambridge, 1986. p. 26. ISBN:0-521-33058-0(hard covers);0-521-33702-X (paperback).
- [25] Loève, M.: *Probability Theory*, 3rd edition, 1963. Van Nostrand, Princeton, N. J. 685pp.
- [26] Marcinkiewicz, J.; Zygmund, A.: *Sur les fonctions indépendantes*, Fundamenta Mathematicae, vol. **29** (1937), pp. 60–90.

- [27] Matuła, P.: *A note on the almost sure convergence of sums of negatively dependent random variables*, Statistics & Probability Letters **15** (1992), 209–213.
- [28] Newman, C. M.: *Asymptotic independence and limit theorems for positively and negatively dependent random variables*, in: Y. L. Tong, ed., *Inequalities in Statistics and Probability* (Institute of Mathematical Statistics, Hayward, CA) pp. 127–140.
- [29] Owen, D. B.: *Handbook of Statistical Tables*, Addison-Wesley, Reading, MA (1962).
- [30] Plackett, R. L.: *Principles of Regression Analysis*, Oxford, 1960.
- [31] Petrov, V.V.: *On a relation between an estimate of the remainder in the central limit theorem and the law of the iterated logarithm*, Teor. Veroyatn. Primen. **11**, No. 3, 514–518. Engl. transl.: Theory Probab. Appl. **11** (1966), No. 3, pp. 454–458.
- [32] Pollard, H.: *The completely monotonic character of the Mittag-Leffler law $E_a(-x)$* , Bulletin Amer. Math. Soc. **54** (1948), 1115–1116.
- [33] Ross, Sheldon M.: *Introduction to Probability Models* (11th Edition), ISBN 978-0124079489, 2014.
- [34] Ross, Sheldon M.: *A First Course in Probability* (9th Edition).
- [35] Ross, Sheldon M.; Peköz, Erol A.: *A Second Course in Probability*, ISBN 0-9795704-0-9, 2007.
- [36] Schilling, R. L.: *Measures, Integrals and Martingales*, Cambridge University Press, 2005.
- [37] Shohat, J. A.; Tamarkin, J. D.: *The Problem of Moments*, Mathematical Surveys and Monographs Volume 1, American Mathematical Society, (Revised edition issued in 1950) Fourth printing of revised edition, 1970.
- [38] Simon, B.: *Functional Integration and Quantum Physics*, 2nd Edition. AMS Chelsea Publishing, American Mathematical Society, Providence, Rhode Island.
- [39] Widder, D. V.: *The Laplace Transform*, Princeton University Press, 1946.
- [40] 陈家鼎, 郑忠国: 概率与统计, 北京大学出版社, 2006。
- [41] 陈希孺: 数理统计引论, 科学出版社, 1981 年 11 月。
- [42] 程士宏: 高等概率论, 北京大学出版社, 1996。
- [43] 程士宏: 测度论与概率论基础, 北京大学出版社, 2004。

- [44] 格罗斯 (B. Gross), 哈里斯 (J. Harris), 里尔 (E. Riehl): 哈佛概率论公开课 (Fat Chance: Probability from 0 to 1), 薄立军、李本崇译, 机械工业出版社, 2020。
- [45] 何书元: 概率论, 北京大学出版社, 2006。
- [46] 李心灿: 当代数学大师—沃尔夫数学奖得主及其建树与见解 (第四版), 高等教育出版社, 2013。
- [47] 梁宗巨 (主编): 数学家传略辞典, 山东教育出版社, 1989; 杜瑞芝、王青建、陈一心等参与合编。
- [48] 齐民友 (主编): 概率论与数理统计 (第二版), 刘禄勤、王文祥、龚小庆编, 高等教育出版社, 2011 年 8 月。
- [49] A. H. 施利亚耶夫: 概率 (第一卷), 周概容译, 高等教育出版社, 2007。
- [50] A. H. 施利亚耶夫: 概率 (第二卷), 周概容译, 高等教育出版社, 2008。
- [51] (美) 塔巴克著; 杨静译: 概率论和统计学: 不确定的科学, 商务印书馆, 2007。ISBN: 7-100-05158-4。
- [52] 汪嘉冈: 现代概率论基础, 复旦大学出版社, 1988。
- [53] 王丽霞: 概率论与随机过程: 理论、历史及应用, 清华大学出版社, 2012。
- [54] 吴文俊 (主编): 世界著名科学家传记 I, 科学出版社, 1990。
- [55] 谢践生: 现代概率论基础 (电子讲义, 未出版)。
- [56] 徐传胜: 从博弈问题到方法论学科: 概率论发展史研究, 科学出版社, 2009。
- [57] 严加安: 测度论讲义, 科学出版社, 1998。
- [58] 杨振海: 拟合优度检验, 合肥, 安徽教育出版社, 1994。
- [59] 应坚刚, 何萍: 概率论, 复旦大学出版社, 2006。
- [60] 张奠宙: 20 世纪数学经纬, 华东师范大学出版社, 2002。
- [61] 钟开莱: 概率论教程, 刘文、吴让泉译, 上海科学技术出版社, 1989.6。

索引

Fourier 变换, 185

Laplace 变换, 224

不等式

Cauchy 不等式, 110

Chebyshev 不等式, 157

Hölder 不等式, 106

Jensen 不等式, 106

Kolmogorov, 172

Minkowski 不等式, 106

矩不等式, 106

事件, 2

De Morgan 律, 19

事件域, 31

互斥/不交, 19

极限, 20

上限集, 20

下限集, 20

保测, 106

不变测度, 106

保测变换, 106

保测映射, 106

保测系统, 106

公式

Bayes 公式, 8, 48

Jordan 公式, 33

Poicare 恒等式, 33

乘法公式, 42

全期望公式, 115

全概率公式, 43, 44

积分变换公式, 107, 108

逆转公式, 186

函数

Cantor 函数, 147

Lipschitz 函数, 81

全变差, 81

分布函数, 92

离散型, 98

单调函数, 81, 94

可测函数, 84

右连续函数, 94

奇异 (连续), 147

密度函数, 128

有界变差函数, 80, 81

特征函数, 184

示性函数, 20

经验分布函数, 204

绝对连续函数, 80, 82

缓慢变化函数, 231

联合分布函数, 92

规则变化函数, 231

边缘分布函数, 92

非负简单函数, 84

分布, 91

先验分布, 48

分布函数

广义逆, 149

分布律, 91, 94

分布函数, 92, 107

分布测度, 93, 107

密度函数, 127, 128

概率分布列, 97

特征函数, 184, 185

联合分布函数, 92

联合密度函数, 128

- 联合概率分布列, 97
- 同分布, 95
- 后验分布, 48
- 奇异 (连续) 型, 147
- 独立同分布, 95
 - 样本量, 95
 - 简单样本, 95
- 边缘分布函数, 92
- 边缘分布测度, 93
- 连续型, 127
- 分布律
 - Mittag-Leffler 分布, 233
 - 离散型, 98
 - $0-1$ 两点分布, 98
 - ζ -分布/Zipf 分布, 59
 - Bernoulli 二项分布, 7, 10, 23, 98
 - Pascal 分布, 99
 - Poisson 分布, 10, 100
 - 两点分布, 98
 - 几何分布, 99
 - 单点分布/退化分布, 98
 - 均匀分布, 98
 - 负二项分布, 99
 - 超几何分布, 99
 - 连续型, 127
 - F -分布, 143
 - $U(a, b)$, 128
 - t -分布, 142
 - 反正弦律, 136
 - Beta 分布, 136
 - Cauchy 分布, 144
 - Erlang 分布, 134
 - Gamma 分布, 135
 - Gauss 分布/广义正态分布, 140
 - Kolmogorov 分布, 205
 - Rayleigh 分布, 144
 - 一维标准正态分布, 136
 - 一维正态分布, 136
 - 卡方分布, 141
 - 指数分布, 134
 - 标准均匀分布, 127
 - 标准指数分布, 134
 - 正态分布/Gauss 分布, 8
 - 高维标准正态分布, 138
 - 高维正态分布, 138
- 单调类定理, 213
- 原理
 - 不充分理由原理, 10, 29
 - 乘法原理, 22
 - 加法原理, 22
 - 反射原理, 26
 - 小概率事件原理, 56
 - 极大似然原理, 56
- 可数生成, 213
- 可测
 - Lebesgue 可测, 67
 - 可测函数, 60, 61
 - 可测矩形, 53
 - 可测集, 27, 60
- 定律
 - 小数定律, 10
 - 弱大数律
 - Bernoulli, 5, 169
 - Chebyshev, 170
 - Khintchine, 171, 194
 - Markov, 170
 - Poisson, 10
 - 强大数律
 - Borel, 11, 170
 - Kolmogorov, 171
 - Rachman, 170
 - 重对数律, 11
- 定理
 - Bochner-Khintchine, 193
 - Carathéodory 扩张, 69
 - Fubini 定理, 82, 85
 - Helly-Bray, 190
 - Helly 选择定理, 191
 - Jordan 分解定理, 81
 - Kochen-Stone 定理, 168
 - Kolmogorov 定理, 205
 - Lévy 连续性定理, 191
 - Lebesgue 控制收敛定理, 76, 85
 - Pearson-Fisher 定理, 208
 - Pearson 定理, 206
 - Radon-Nikodym 定理, 86

- Skorokhod 嵌入定理, 190
- 中心极限定理, 7
 - De Moivre-Laplace, 9
 - Lindeberg-Feller, 195
 - Lindeberg-Lévy, 193
 - Lyapunov, 195
- 单调收敛定理/Levi 定理, 74, 85
- 单调类定理, 213
- 微积分基本定理, 80, 85
- 特征函数
 - 刻画定理, 193
 - 唯一性定理, 79, 186
 - 连续性定理, 191
- 引理
 - Borel-Cantelli 引理, 165
 - Borel-Cantelli 第二引理, 166
 - Fatou 引理, 75, 85
 - Kronecker, 172
 - Riemann-Lebesgue 引理, 78
- 悖论
 - Bertrand 悖论, 10, 17, 29
 - Littlewood “无穷” 悖论, 43
 - 圣彼得堡悖论, 7, 8, 15
 - 辛普森悖论, 15
- 扩张
 - Carathéodory 扩张, 69
- 收敛
 - L^p -收敛, 157
 - 以概率 1 收敛, 85
 - 依分布收敛, 157
 - 依概率收敛, 85, 157
 - 依测度收敛, 85
 - 几乎处处收敛, 85, 157
 - 几乎必然收敛, 85
- 方法
 - Bayes 方法, 8
 - Carathéodory 扩张, 67
 - Monte-Carlo 方法, 8, 29
 - 典型方法, 107
 - 截断手术, 11
 - 极大似然估计, 57, 101
 - 概率微元法, 130
 - 特征函数方法, 11, 165
 - 矩方法, 111
- 条件密度, 124
- 概率, 3, 5
 - 先验概率, 48
 - 后验概率, 48
 - 条件概率, 8, 39, 40
- 概率模型
 - Polyá 坛子模型, 23
 - 推广模型 A, 47
 - 推广模型 B, 47
 - 无放回, 23
 - 有放回, 23
 - 几何概率模型, 8, 17, 27
 - 古典概率模型, 17, 21, 53
 - 赌徒破产模型, 58
- 概率理论
 - 狄氏型理论, 12
 - 倒向随机微分方程理论, 12
 - 大偏差理论, 12
 - 狄氏型理论, 12
 - 随机分析理论, 12
 - 鞅论, 12
- 概率空间
 - 乘积概率空间, 52, 53
- 测度
 - Borel 测度, 93
 - 原子, 147
 - Dirac 测度, 98
 - Lebesgue 测度, 67
 - 体积测度, 63, 67
 - 外测度, 67
 - 奇异测度, 63
 - 概率测度
 - 乘积概率测度, 53
 - 相互奇异, 85
 - 相互等价, 85
 - 符号测度, 64
 - Hahn 分解, 64
 - Jordan 分解, 64
 - 全变差, 64
 - 全变差测度, 64
 - 正部, 64
 - 正集, 64

- 负部, 64
- 负集, 64
- 绝对连续, 82, 85
- 计数测度, 63
- 边缘测度, 93
- 非负测度, 62
- σ -有限测度, 62
- Borel 测度, 62
- Radon 测度, 62
- 有限测度, 62
- 测度的扩张, 69
- 预测度, 66
- 现象, 2
- 矩问题, 220
 - Carleman 准则, 223
 - Hamburger 定理, 221
 - Hamburger 矩问题, 221
 - Hankel 矩阵, 221
 - Hausdorff 定理, 221
 - Hausdorff 矩问题, 221
 - Krein 条件, 224
 - Stieltjes 矩问题, 223
 - 截断矩问题, 222
- 积分
 - Lebesgue 积分, 72, 77, 83
 - Riemann 积分, 77
- 空间
 - Banach 空间, 90
 - Lebesgue 空间, 150
 - 乘积空间, 53
 - 可测空间, 31
 - 乘积可测空间, 53
 - 样本空间, 21, 27
 - 样本点, 21, 27
 - 概率空间, 31
 - 乘积概率空间, 54
 - 子概率空间, 40
 - 条件概率空间, 39, 40
 - 测度空间, 32, 62
 - 赋范空间, 90
- 简单样本, 95
- 似然函数, 141
- 最大次序统计量, 150
- 最小次序统计量, 150
- 样本均值, 141
- 样本方差, 141
- 样本标准方差, 142
- 样本量, 95
- 次序统计量, 150
- 结构
 - 可测结构, 31
 - 乘积可测结构, 53, 54
 - 概率结构, 31
 - 乘积概率结构, 54
 - 测度结构, 32, 62
- 过程
 - Brown 运动, 12
 - Lévy 过程, 12
 - 平稳过程, 12
 - 随机游动, 12
 - 鞅, 12
 - 马氏过程, 12
 - 马氏链, 11
- 问题
 - Buffon 投针问题, 8, 28
 - Littlewood “无穷” 悖论, 42
 - Monty Hall 问题, 14, 41
 - Ramsey 问题, 26
 - “三枚银币” 骗局, 14
 - 三门问题, 41
 - 伯努利-欧拉装错信封问题, 33
 - 估算池塘里的鱼, 101
 - 医学诊断问题, 49
 - 同卵双生问题, 45
 - 同生日问题, 24
 - 唱票问题, 25
 - 常规赛表现与季后赛, 48
 - 抓阄的公平性, 24, 46
 - 敏感性调查问题, 45
 - 是否作弊问题, 51
 - 约会问题, 28
 - 赌金分配问题, 4, 5, 46, 103
 - 选举得票率问题, 44
 - 选举得票率问题后续, 48
 - 频率代替概率, 101

随机

- 事件, 1, 55
 - 不可能事件, 18
 - 必然事件, 18
 - 相互独立, 55
- 现象, 1, 55
- 随机变量, 91
 - 生成的 σ -代数, 92
- 随机向量, 92
- 随机实验
 - 条件实验, 55
 - 重复实验, 55
- 随机变量, 93
 - n 阶矩, 109
 - 一阶矩, 109
 - 分布函数, 92
 - 分布测度, 93
 - 协方差, 110
 - 原子, 96
 - 可积, 105
 - 数学期望, 103, 105
 - 方差, 109
 - 相互独立, 95
 - 相关系数, 110
 - 离散型, 5, 96
 - 概率分布列, 97
 - 退化随机变量, 98
 - 线性不相关, 110
 - 补丁程序, 93
 - 连续型, 9, 128
 - 均匀分布, 127
 - 均匀分布 $U(0, 1)$, 127
 - 密度函数, 128
- 随机向量
 - 协方差矩阵, 137
 - 原子, 96
 - 条件分布, 121

- 条件分布函数, 121
- 条件密度, 132
- 生成的 σ -代数, 92
- 相互独立, 95
- 离散型, 96
 - 联合概率分布列, 97
- 联合分布函数, 92
- 联合分布测度, 93
- 边缘分布函数, 92
- 边缘测度, 93
- 连续型
 - 均匀分布, 127
 - 条件密度, 129
 - 联合密度函数, 128
 - 边缘密度, 129

随机小, 163

随机控制, 163

随机数发生器

Inverse Transformation Method,
149

Rejection Method, 153

集合类, 210

- λ -系、 D -系, 211
- π -系, 68, 211
- σ -代数, 31
 - Borel, 27
 - 乘积 σ -代数, 53
 - σ -代数、 σ -域, 210
 - Borel σ -代数, 213
 - 代数/域, 68
 - 代数、域, 210
 - 半环, 68, 211
 - 单调系, 211
 - 环, 68, 211

频率, 3

极限频率, 5