窥探天机

--科学地认识随机现象–

应坚刚

我思故我在

René Déscartes

序

这里我们说三个问题, 课程讲什么? 为什么要开这个课? 怎么用这本讲义?

自然界万物的演变是否都遵循一个确定的规律? 这一直是一个充满争议的科学哲学问题, 实际上, 不管对上述问题持何种立场, 世界上的许多现象在人类看来是不确定的, 或者是不可知的. 引用 17 世纪后期数学家 Jakob Bernoulli 在其著作猜度术中的一段话: **普通人谁能确定, 例如, 疾病的数目, 以及在哪个年龄, 哪个疾病会侵入人的无数器官中的哪一个而导致人的死亡, 以至于我们可以猜度生死的未来状态? 谁能够数清楚每天空气所经历的不可数的变化然后猜度它在一个月后的状态, 更不要说一年后? 谁能够足够清楚地知道人类心灵或者身体的构成以至于敢于说可以确定在一个游戏中的参与者最终的胜负呢?** 也许是因为规律本质上不存在, 也许是规律存在但人类能力所限而不可知, 总之, 我们可以说随机现象无处不在, 且与我们的生活息息相关. 随着时间推移, 人们渐渐地发现, 尽管随机现象看似没有规律可言, 但它发生的机会, 或者说可能性是有一定规律的, 这个规律可以理解为概率, 概率论是研究概率的数学理论, 而所谓的猜度, 其实就是窥探天机. 本课程介绍概率论的背景与历史, 阐述概率论的思想以及它怎么用于解释和解决真实世界的问题. 我们的目的是想让学生们发现概率论对于生活是真实的. 正如概率学家 W. Feller 曾经说: *如果概率论对于生活是真实的, 那么每一个经验应该对应一个定理.*

21 世纪是 AI 的世纪, 其中概率论变得越来越重要, 进入了中学数学. 2018 年我荣幸地参与由李大潜先生担任主编的上海高中数学教材的编写, 并写了其中的概率部分. 但遗憾的是, 由于篇幅限制, 这些内容对于理解随机现象和概率是远远不够的, 而多数学生在大学里可能不再有机会继续学习概率论. 为此我设计了这门通识课程, 并为此写了这本讲义, 作为中学内容的延续和补充, 目的是引导这些学生如何正确认识随机现象中的规律, 帮助学生了解概率的思想与意义, 提升认知能力, 更好地应对来自不确定性的挑战.

因为随机现象的普遍性, 所以人们对与随机现象相关的名词与术语并不陌生, 受过适当教育的人不难知道怎么来估算一些事件发生的可能性的大小, 这些估算的方法差不多是源于直觉. 需要强调的是, 在学习概率论的时候, 直觉是非常重要的, 甚至比理论本身还要重要. 著名物理学家杨振宁曾经说: “人生来就有直觉, 学习的根本是获得更正确的直觉.”讲义基本上是问题组成的, 很多问题来源于生活或者游戏, 希望学生在看到问题时独立思考, 寻找问题的数学表达, 寻找答案, 再用生活的语言来解释答案. 尽管这不是一本数学书, 但每个推理都是认真严谨的, 学习严谨性也是课程的目的之一. 总的来说, 讲义中用到的数学知识是简单的, 以复旦学生的基础, 只要有兴趣, 肯思考, 花点时间, 无论文科理科, 都可以愉快地学习.

概率论的教材很多, 但通识课的教材还很少见, 我们为此设想了一堆的目标也付出很多的时间精力, 希望在教学实践中能实现一二, 但因个人学识有限, 讲义不免有许多瑕疵, 欢迎读者指正. 最后, 我的研究领域是概率论, 但属于 Kolmogorov 所说的忘记了现实世界概率的那一类人, 感谢我的同学和朋友罗震, 他关于概率论应用的评论让我受益匪浅.

应坚刚
复旦大学

目录

第一章　引言

概率论是数学的分支之一, 它在数学中的正式身份应该是 1933 年建立公理体系之后才确认的, 但这并不是说概率是 1933 年凭空出现的, 实际上在这之前已经有快 300 年的有文字记载的历史, 概率有自己的背景, 有自己的问题, 有对它感兴趣的天才数学家, 是人们感兴趣的学问, 所以概率其实一直在发展. 自几何原本以来, 很多数学的分支也是这样发展起来的, 例如微积分, 拓扑, 代数等等. 本章是引言, 目的是聊聊随机现象与概率论的历史背景, 在我看来, 这要比写概率论的数学部分难得多.

1.1　认识随机现象

简单地说, 随机现象是现实世界中的真实存在, 概率论是认识和研究随机现象的一种方式. 关于随机现象与概率, 需要关注下面几个基本问题.

问题: 什么是随机现象?

问题: 人类怎么认识随机现象?

问题: 什么是概率?

问题: 什么是概率的本质意义?

前两个问题是哲学和历史问题, 我们将在本章中浮光掠影地聊聊, 但因为作者文学功底差, 所以可能词不达意. 后两个问题才是本课程的主要内容.

随机现象无处不在

首先要说的是随机现象是无法预测结果的现象, 具有不确定性, 在生活中无处不在. 人类生活在一个称为地球的星球上, 与宇宙相比如同海洋里的一滴水, 但人类认识的世界远远超过地球, 也称为自然, 如头顶上太阳东升西落, 星空璀璨闪烁, 大地上植物春绿秋黄, 动物生老病死, 等等. 如果我们关注这些自然现象的因果, 那么会发现它们有的是确定的, 即什么原因产生什么结果是可以预见的, 有的是不确定的, 即在

同一个原因下无法预见会出现什么结果, 且确定与不确定可能互相交织纠缠.

人类对于确定的自然规则的探索有很长的历史, 发展出丰富的成果, 经典的科学领域正是研究确定的自然规则的. 对于不确定性的研究相对短暂, 也更为困难. 不确定的现象也通常称为随机现象. 面对确定的现象, 人没有什么选择; 面对不确定的现象, 人会有各种选择, 也称为机会或者可能性. 例如, 人类社会有各种各样的赌场, 小的像家里的麻将桌, 大的如澳门 Las Vegas 那里的赌城, 香港的赛马场, 还有像彩票, 像充斥全球的股票市场, 等等, 这些地方有机会让人一夜致富, 当然最大的赌场是人本身的命运, 命即生死, 运即机会, 人在命运的赌场上, 希望能够比别人多窥得一丝天机, 就足以受用终生. 提起随机, 大多数人可能首先想到一些常见的游戏, 例如掷硬币骰子, 打牌打麻将, 还有抽签买彩票等. 其实随机现象充斥生活的每个角落, 无时不刻会遇到, 只不过我们对此已经习以为常. 例如你是一个银行前台职员, 在一个普通工作日的早上, 你知道 6 点多天差不多亮了, 当然夏天亮得早一点, 冬天亮得晚一点, 但不会有太大不同, 如果天晴, 太阳肯定会从东方升起, 这是确定的, 但是你可能对气温, 风力, 还有是否下雨会感到不确定, 要看看天气预报, 但天气预报也不一定每次都能预报准确, 这就是不确定性. 然后你吃好早饭, 在计划好的时间出门, 这是你可以控制的, 也是确定的, 接着去坐公交车上班, 车上有没有座位? 会不会堵车? 什么时间到办公室? 这些都是不确定的. 上班之后, 有多少人来办事? 办什么业务? 中间有没有空闲? 这些也是不确定的. 再比如一位小学一年级的教师站在讲台后, 看着面貌不同性格迥异的学生, 遥想他们的未来, 从机会来看, 这些学生应该是差不多的, 但是最终的结果肯定大相径庭, 有的顺风顺水, 有的逆风奋斗, 有的英年早逝, 不一而足, 令人不由得感慨命运无常. 也就是说, 机会平等, 而结果却迥异.

啰啰嗦嗦地写了那么多, 可能仍然没有表达清楚. 让我们用文字下一个粗略的定义, 不止一个结果且无法预测哪个会出现的自然现象称为随机现象. 这肯定不是一个毫无争议的定义, 但它包含了大多数我们感兴趣的现象. 要注意的是, 定义中排除了人的意志, 因为在我们看来, 有人为意志起主要作用的现象不能称为自然现象, 这不是我们的课程所关心的. 例如, 有一个人随便给你写正反两个字, 这当然也无法预测, 但这不应该认为是随机现象. 定义也许会引起争议, 但我们不要将时间浪费在争议上.

世界是可知的吗

上面所谈及的是普通人对于随机现象的认识, 但概率论的诞生最终还是要归功于人类中的智者, 是他们汲取人类思想的精华总结出探索随机现象的理论基础. 那么他们是怎么理解随机现象并抽象出这样一个数学分支的呢? 人类内心里普遍认为所有的现象背后总有规律, 认为现象呈现出随机性不是因为它不能被了解, 而是因为人类的认识水平所限而无法理解. 例如前面所提及的, 人类不懈的努力使得我们知道原本以为是随机出现的日全食以及彗星其实是确定的, 天气预报在慢慢地越来越准确, 可以治愈的病也越来越多. 早期的人类信仰神, 认为即使是掷骰子的结果也是由神来决定的. 实际上, 大多数著名学者也认为世界本质上是受规律支配的. 17 世纪伟大的科学家与数学家 Newton 用漂亮的公式揭示了物体运动规则, 他说给我一个支点, 我能撬动地球. 这话实际上是宣告没有什么是人类不能解决的. 同时代著名数学家, 哲学家, Leibniz, 19 世纪著名数学家 Laplace, 20 世纪伟大的科学家 A. Einstein 等都是决定论者, 他们和 Newton 有同样的信念, 认为宇宙完全是由因果定律支配, 不确定是我们人类对于世界认识不完全而导致的, 完全否定随机现象的存在性, 即使是掷硬币这样普通人看来完全随机的现象, 也被认为是一个服从 Newton 力学规则的确定性过程, 只是因为过于复杂而无法确定结果.

应该说, 一直以来, 对于世界的认识无疑是有争议的. 20 世纪初量子力学的诞生对于物理学来说是一次革命, 一系列实验发现, 在量子尺度上, 任何的观测对被观测的对象会产生不可知的干扰. 例如一个光子被射入电气石晶体 (观测) 时以确定的概率可能被吸收可能被射出, 无法预测. 物理学家 P. Dirac 说: “因果律仅仅适用于那些免于扰动的系统. 如果系统是小的, 不可能观测它而不干扰它. 因为科学仅仅关心可观测的事物, 所以我们不能期望得到观测结果之间的任何因果关联.” 这些现象的解释引起了 Bohr 和 Einstein 为代表的两个阵营关于世界观的世纪论战, 使得越来越多的科学家认为自然现象的部分是本质地不确定的. 有意思的是, 即使一些学者信仰

使然, 也不表示他的研究与信仰一致, 例如 Laplace, 他是决定论者, 但他也是概率和统计推断理论的创始人之一. Einstein 是决定论者, 但也是量子力学的创始人之一. 19 世纪英国生物学家托马斯 • 亨利 • 赫胥黎是不可知论者, 但作为科学家, 他相信自然界受客观规律支配.

科学地认识随机现象

尽管人类对随机现象的探索要远远晚于且热情也远远低于对确定性规律的探索, 但对它的认识和利用一直在进行, 这在历史的长河里时有记载. 随机现象的特征是它无法预测, 古时人们就把这种人无法掌控的东西当作是天的意志, 当对某事迟疑不决的时候, 就运用占卜, 求卦等手段来进行决策. 从考古结果看, 世界上各个文明都有类似卦卜的手段. 即使现在, 人们在生活中遇到无法决策的事情时, 也常求助抛掷硬币或者抽签. 随机是传达上天意志或者体现公平的一种方法. 这里有个真实的故事, 发生在 19 世纪, 一艘从英国开往美国的轮船在途中遭遇风暴, 船要沉没, 船长首先命令把所有的行李扔掉以减轻重量, 但是还不够, 为了大多数人的安全, 船长指定了一些乘客把他们放入救生艇自生自灭, 最后这些乘客很多都遇难了, 但船安全抵达美国. 这时一些乘客起诉船长谋杀, 法官审理后认为船长有罪, 认为他为了大多数乘客而放弃一些乘客并没有错, 但是船长错在不应该私自决定放弃哪些乘客, 而应该交给上帝来决定, 也就是用抽签或者其他随机手段来决定.

人类对确定的现象有一种天生的亲切感, 相反地, 对于不确定或者随机现象却会有一种天生的厌恶和畏惧. 人们的基因里对于无规则时期人类的生活状态一定是非常排斥的, 正如俗语所说: “宁为太平狗, 不做乱世人.” 为了秩序, 人们愿意让渡自己的部分权利, 建立起一个有法的社会, 这里的秩序实际上就是确定性. 对于自然现象, 看着四季依次交替, 太阳东升西落, 可以坦然面对, 甚至欣赏. 而对于突如其来的灾难, 例如地震, 台风, 洪水等, 内心充满畏惧和不安, 俗话说: “天有不测风云, 人有旦夕祸福,” 正是这种情感的写照. 然而, 随着时间推移, 人类对随机现象会形成有用的直觉, 即学会通过持续观察并总结经验来进行推断, 从而应对随机性. 例如上班的人需要估计早上该几点出门使得迟到的可能性小到可接受的程度. 关于天气预报, 民间有很多谚语, 例如: “一雾三晴, 重雾三日必大风.”“满天乱飞云, 雨雪下不停.”“喜鹊枝头叫, 出门晴天报.”“风大夜无露, 阴天夜无霜.” 这些都是古人总结出来的, 不一定准确, 但有一定的参考价值. 还有些现象, 例如日全食, 古时人们对此很害怕, 称为天狗吞日, 以为它会伴随灾难降临, 殊不知后来人类认识到这是某种自然规律, 是可以预测的, 因此不再神秘和可怕.

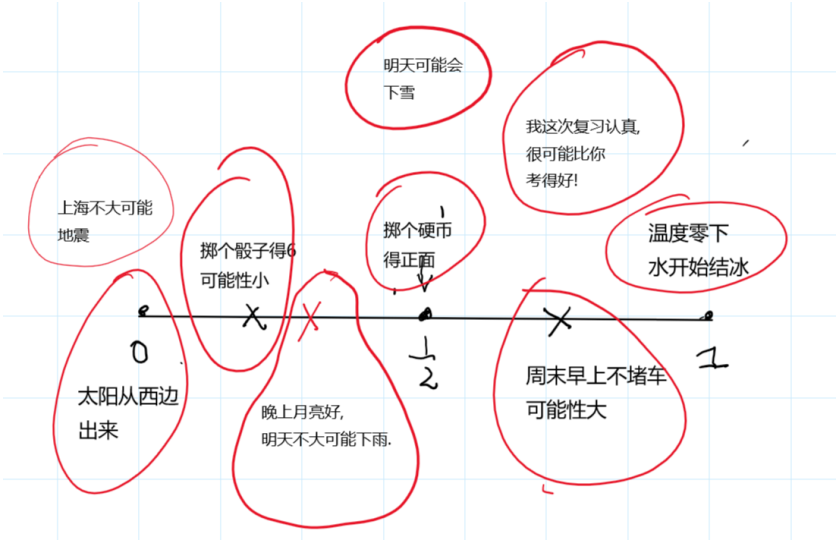
不可否认, 人类的进步是巨大的, 特别是最近的几百年来, 我们见证并享受了人类在科学技术方面取得了巨大的进展. 尽管如此, 仍然有许多现象及其问题让人束手无策. 在这种情况下, 常规的科学研究手段 (称之为推断) 可能失效, 但概率论及以其为理论依据的统计推断方法, 通过研究现象中事件出现机会的规律, 渐渐地成为一种新的研究手段, 提供了新的思路. 很多科学家认为, 统计推断是科学研究的辅助手段, 是一种权宜之计, 最终会被确定性推断取代. 但也有很多科学家相信, 在某些领域, 统计推断即使不是唯一有效的手段, 至少也是不可或缺的手段. 为什么怎么说呢? 让我们以掷硬币为例, 设想你站在一个大桌子前, 用大拇指顶住硬币一边用力向上弹出使得硬币旋转向上然后因重力落下, 这时你能判断片刻后硬币平躺在桌面上时是哪一面朝上吗? 物理学家认为, 在无外力作业时, 硬币的运动服从牛顿定律, 其最终状态是由初始状态决定的. 但可以想象, 初始状态的测量, 硬币与桌子材质的测量, 空气阻力的测量, 运动方程的建立, 以及中间状态的计算之中的任何一点误差都可能导致预测结果与真实结果的截然不同. 这实际上是一个对初始状态极其敏感的复杂系统, 谁相信人类有一天能够确定地预测硬币的状态吗? 回顾 Jakob Bernoulli 的那一段话: **普通人谁能确定, 例如, 疾病的数目, 以及在哪个年龄, 哪个疾病会侵入人的无数器官中的哪一个而导致人的死亡, 以至于我们可以猜度生死的未来状态? 谁能够数清楚每天空气所经历的不可数的变化然后猜度它在一个月后的状态, 更不要说一年后? 谁能够足够清楚地知道人类心灵或者身体的构成以至于敢于说可以确定在一个游戏中的参与者最终的胜负呢?** 上面提到的几种情况的任何一种都不会比掷硬币更简单, 整个系统的复杂程度让人绝望地怀疑是否真的可以有搞清楚的那一天. 即使有, 人类也不能袖手等待, 科学地认识随机现象是必由之路. 正如数学家 Hilbert 曾经说: 与其考虑这种愚蠢的不可知论, 不如让我们铭记我们的箴言: 我们必须知道, 我们必将知道.

1.2 概率: 不确定性的度量

人类认识世界是一个漫长的过程, 从定性到定量往往经历几百上千年. 例如几个牛几只鸡开始有了数字, 从定性地认识到很远的地方和很大的田地开始到提问多远的距离和多大的田地, 再到定量地用数字来表示两地的距离与表示一块地的面积等等, 这样可以更加方便地进行比较和交易. 类似这样, 人们慢慢地给生活中的很多事物进行度量, 如重量, 体积, 温度, 时辰, 随之也发展出相应的数学. 关于度量的数学是

数学的一个重要组成部分.

人类对于随机现象中随机事件的度量的认识是相对比较晚的. 人们肯定很早就认识到随机现象中的事件有可能发生有可能不发生, 事先无法预测, 也许有人认为是神决定的, 但至少承认人类是没有能力预知的. 当然, 我也毫不怀疑即使在远古, 人们对很多的随机现象中的事件发生的可能性大小是有感觉的. 很多人在很小的年龄就知道怎么正确地使用抽签, 抛硬币和掷骰子这样的随机物件, 说明人类很早就明白什么是机会平等. 另外, 有记载说很早人类就懂得怎么选择相对安全的居住地建造房屋, 这说明人类对某些自然灾害发生的可能性有比较大小的意识. 下面的图简略说明人在日常生活中是怎么直观地度量可能性的.



因此, 思考怎么对随机事件发生的可能性大小进行度量并且最终产生这个度量是一个历史的必然, 尽管在哪个时间点产生这个度量是偶然的. 现在没有资料显示人类什么时候认识到机会是可以度量的这件事情的. 我相信这一定远远早于下一节所介绍的有记载的概率历史. 可以想象, 怎么度量机会这个问题并不是一开始就非常清晰, 在 16 世纪 Cardano 的书上, 他还是用有利场合数与不利场合数的比来表达机会的, 一直到 17 世纪 Fermat 与 Pascal 通信之后才统一为有利场合数在总场合数总所占比例来度量机会的大小, 现在通常叫做概率.

不幸的是, 在实际问题中, 获取随机现象中事件的概率并不总是一件容易的事情. 在类似硬币或者骰子的情况下, 随机现象的结果是显然的, 量化也很容易, 因为所有这

些结果直观上是等可能发生的. 但是在一些其他问题上却几乎不可能实现, 参见上面引用的 Bernoulli 著作中的那段话. 因此, 我们知道, 除了类似硬币和骰子这样被先验地认定的等可能随机现象, 我们很难获取随机现象中的具体事件的概率. 但是, 我们并不是束手无策, J. Bernoulli 在他的著作猜度术中也谈到了这种情况, 他的著作的主要贡献是建议用事件发生的频率来估计概率, 也就是说用统计方法来对随机现象进行有依据的量化, 从而获得概率的合理估计值, 这就是著名的大数定律, 会在第五章中详细解释.

对于一个随机现象中的每个给定事件发生可能性大小进行度量是一个好想法, 也许不一定总是能够做到. 但即使我们获得了事件的概率, 它的意义是什么, 它能够被证伪吗? 例如, 两个人争议上海在未来十年内是否会发生地震, 一个人说概率是 20%, 另一个人说概率是 80%, 我们有没有手段来判断谁的说法更准确呢? J. Bernoulli 所说的用频率来猜度概率实际上可能是判断一个概率是否准确的唯一途径, 也许也说明了概率的意义. 遗憾的是, 不是所有随机现象中事件的概率都能用频率来估计. 显然, 频率这个词本身就要求事件所在的那个随机试验可以用某种方式进行重复. 如果随机试验可以随心所欲地重复, 那么事件的度量是否合适是可以验证的, 这也是概率的本质意义所在. 从上面那个图中所说的事件来看, 像 '掷一个硬币的正面', '掷一个骰子得 6' 这样的事件显然是可以通过重复随机现象来进行验证的, 这也是 Bernoulli 的书中所说的例子, 但有一些是不可能进行重复的, 例如 '我这次复习认真, 很可能比你考得好', '明天可能会下雪', 对这样的事件赋予任何概率作为度量都不具有科学上的意义, 因为无法证伪. 这时候如果我们仍然谈论概率, 那么这样的概率只能称为主观概率. 也有一些不能主动重复但可以观察, 例如 '上海不大可能地震', 我们也许从过去的记录中获取一些数据, 但一是数据的量不足, 二是没有什么科学依据说明过去的数据对现在或者未来有任何的指导意义, 因此这样得到的概率的意义也是值得怀疑的.

总之, 应用数学来研究随机现象问题是极其困难的, 一般地, 我们不仅无法判断哪个结果会发生, 也很难获得其发生的概率. 其中只有极少数问题可以通过数学方法来计算或者数学思想来解释, 例如随机分析在金融定价领域的应用是数学应用于解释真实世界的一个典型案例. 也有少数可以通过重复或者 '近似地重复' 随机试验获取统计数据的方法来进行估计, 这实际上就是统计学家在从事的工作, 但由于上面所说的原因, 统计学家根据数据推断出的结论不一定值得信任. 最后, 对于极大多数实际问题中的随机现象而言, 人类仍然很无知, 唯有不懈地探索.

1.3　概率论简史

概率论无疑是人类智慧的结晶, 但是, 与其他学科一样, 智者在其中扮演了关键的角色. 在本节中, 我们只是尽可能简单地介绍概率论 (有记载的) 历史以及为概率论的诞生作出最重要贡献的几位学者: Cardano, Fermat, Pascal, Bernoulli, DeMoivre, Bayes, Laplace, Kolmogorov.

早期 (17 世纪前)

关于概率论的历史, 要分两个时期讲. 如果只是谈人类对随机现象的兴趣和认识, 那么应该说是很早很早不可考的早期. 据考古发现, 远古时候人类就有占卜, 算卦等活动了, 所谓占卜, 就是用龟壳、蓍草、铜钱、竹签、纸牌或占星等手段和征兆来推断未来的吉凶祸福, 除此之外, 民间也很早就有抽签掷骰子打骨牌等赌博游戏了. 这些都属于人类对随机现象的利用. 现在生活在地球上的人类把有确切且连续记载的历史称为人类文明史, 大概是 6000 年. 可以想象, 自从有智慧的人类诞生之后, 对于不确定的未来的担忧是一直存在的, 即使到今天依然如此. 据世界各地文献记载, 骰子出现在几千年前的印度埃及罗马还有中国, 人们使用骰子玩游戏以及算命. 骰子也逐渐地演变为现在常见的正方体模样. 历史文献上也偶尔出现关于几个骰子有多少种可能结果的讨论, 中国历史上古老的易经总结讨论 64 种卦象对于未来的预测. 所谓: 易有太极, 太极生二仪, 二仪生四象, 四象生八卦。易经的本身的历史已不可考, 大概是在公元前一千年. 在这个时期, 人类对随机现象中可能性的大小并没有明确的记载, 但是显然不能排除人类已经对可能性的大小有感觉也有兴趣知道. 在日常言语中, 可能性与机会差不多是等义的, 可能性的大小在现在也称为概率或者几率. 据记载, 真正对概率这个数字进行描述和计算, 要归功于 16 世纪中期意大利数学家 G. Cardano, 卡尔达诺, 生于 1501 年, 死于 1575 年, 他最著名的工作是关于三次方程的解. 他喜欢赌博, 对于赌博的热情驱使他在晚年写了题为“机会的游戏”(Game of Chance) 一书, 其中阐述了很多概率的问题与思想. 例如, 他在此书中给出的概率是以“甲乙各自的有利场合数之比 $\mathrm{r}:\mathrm{s}$”这种形式给出, 而不是现在通常定义的概率: $\frac{\mathrm{r}}{\mathrm{r}+\mathrm{s}}$. 他还讨论了独立重复, 大数定律等等的问题, 但只是有叙述, 没有严格的证明. 他的观点还比较粗糙, 处于探索概率的初期, 他的书中也还有很多的错误. 但是因为与教会的恩怨, 这本书实际上在他死后近百年才出版, 没有在合适的时候呈现, 否则的话, 他应该至少是概率的开创者之一.

Cardano, Fermat, Pascal

法国数学家 P. de Fermat (1601-1665) 与 B. Pascal (1623-1662) 在 1654 年关于分奖金问题以及其他一些概率问题讨论的通信, 被认为是概率理论的萌芽. 这些问题是赌博游戏中的一些实际问题, 是法国著名的贵族赌徒 de Méré 写信问当时法国著名物理学家和数学家 Pascal, 所以可以说是游戏驱动了概率.Fermat 与 Pascal 的探讨富含思想, 推理相当严格, 为后续概率的数学理论大厦放下了第一块基石.

差不多同时, 荷兰著名科学家 C. Huygens, [1] 得知 Fermat 和 Pascal 的通信, 对这些问题也很有兴趣, 在 1657 年左右出版了《关于机会游戏的推理》一书, 对当时关注的概率问题有系统的研究, 例如赌徒输光问题, 超几何分布等. 显然, 他把概率论理解为机会的游戏. 如果说 Fermat 和 Pascal 只是把通信中的问题当作游戏讨论, 那么 Huygens 被认为是第一个把机会问题当作学术问题来研究的人物.

中期 (18-19 **世纪**)

Jakob Bernoulli 是 Bernoulli 兄弟中的兄长, 生于 1654 年, 正是 Fermat 与 Pascal 通信讨论分赌注问题那一年. Jakob 和弟弟 Johann 都是那个时代伟大的数学家与科学家, 是 Leibniz 微积分的主要传道者. 例如 Johann 提出并解决了最速降线问题, 解决了 Jakob 提出的悬链线问题, 但是他们两个关系不好. 学界对兄弟的普遍的评论是: Johann 思路更快, 但 Jakob 更深刻.

[1] 他是著名数学家和哲学家 Leibniz 的一位老师

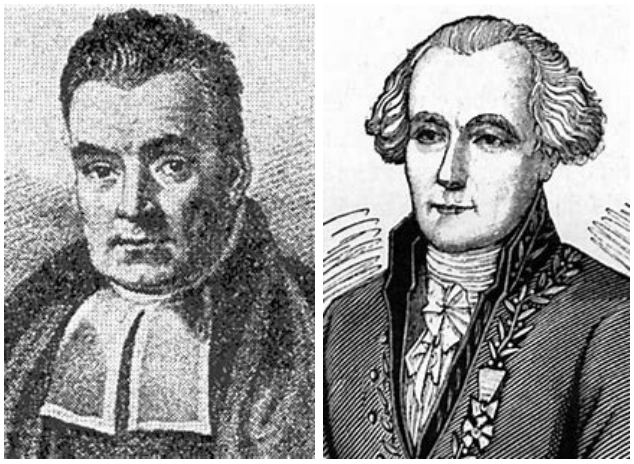
J. Bernoulli 和 A. deMoivre

在 1713 年, 去世后八年, Bernoulli 的 Ars conjectandi (The art of conjecturing) 出版, 中文直译为《猜度术》, 书中不仅讨论了很多复杂概率问题的计算, 例如二项分布, 而且讲了很多概率在实际生活中怎么应用的问题, 最重要的是给出并严格证明了大数定律 (law of large numbers), 即在独立地重复一个随机试验 n 次, 当 n 很大时, 成功的频率趋向于成功的概率. [2] 大数定律很直观, 也是一个非常深刻的定理, 它影响了整个概率统计学科. 很多人都或多或少地对此有感觉, Cardano 曾经叙述过这个思想, 另外 Bernoulli 在还没有极限定义的时候严格地证明了一个关于极限的定理, 因此对微积分也有巨大的影响, 这种影响也许被数学史严重地低估了. 他在书中还完整地叙述了为什么要证明以及怎么使用该定理, 是探讨概率思想的宝库. 因此 Jakob Bernoulli 被称之为第一个伟大的概率学家也是当之无愧的.

同时期因为避难而旅居英国的法国数学家 A. de Moivre (1667-1754) 也是一个在概率论著名学者, 他在 1718 年出版《机会理论》(也译为: 计算游戏中事件概率的方法) 一书, 然后在 1733 年发表的论文中提出并粗略地证明了最初形式的中心极限定理: 二项分布标准化之后的极限是正态分布. 因为后来 Laplace 完善了该结果的证明, 现在此定理通常称为 de Moivre-Laplace 中心极限定理, 它可以被看作是大数定律的精细化. 尽管没有大数定律那么直观, 但它的重要性并不下于大数定律, 证明难度更是远远超过大数定律. 实际上, 后来数学家发现, 大数定律与中心极限定理具有非常广泛的普适性.

[2] Bernoulli 自己在他的书中把这个定理称为黄金定理, 大数定律是后来由 Poisson 命名的.

另外我们还应该提到在 18 世纪的中叶, 1763 年发表在哲学刊物上的论文, 题为《解决机会理论中一个问题的论文》, 作者是 Thomas Bayes (1702-1761), 这篇论文是他去世后两年由他的朋友 R. Price 整理出版的, 值得注意的是, Bayes 是个职业教士, 生前并没有发表过真正的数学论文. 论文是从概率论角度来回答哲学家 David Hume (休谟) 关于归纳推断的基础问题. 现在, 文中的公式称为 Bayes 公式, 它是统计推断理论中的重要观点.

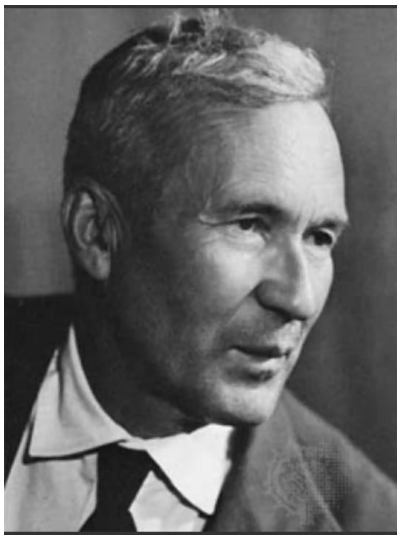


Bayes 和 Laplace

在整个 18 世纪, 概率的应用逐渐从机会的游戏转向科学与社会科学问题. 到 19 世纪初, 概率的应用发展到达高潮, 法国数学家 P.-S. Laplace 对概率论的发展与应用也做出了重要贡献, 他分别在 19 世纪初期的 1812 年和 1814 年出版了两本著作《概率的分析理论》与《概率的哲学探讨》, 给出古典概率的定义及概率应该服从的一些规则, 讨论概率在哲学, 科学, 社会科学, 法律审判中的应用, 也就是用于统计推断.

现代 (20 世纪初)

20 世纪是个伟大的世纪, 尤其对于概率论. 由于概率本身的发展以及概率论在自然科学和社会科学中的广泛应用, 人们清楚地知道, 概率论自身正在孕育着一个伟大的学科, 但是概率论还不是数学. 什么是数学? 古希腊数学家 Euclid 的标题为几何原本的不朽著作为数学建立了一个范本, 简单地说, 数学从具体问题中抽象而来, 是建立在公理体系上的学问. 而概率还没有自己的公理体系. 到 1930 年, 在许多数学家的工作中, 概率论的公理化已经呼之欲出了.



A.N. Kolmogorov

俄罗斯数学家 A.N. Kolmogorov 是给与临门一脚的人物, 他于 1933 年 (前苏联) 发表了一篇长论文 'The foundation of Probability Theory', 不仅清晰地列出了概率论所需的公理, 还在公理基础上比较完整地搭建了概率论的基本框架.[3] 这个体系迅速地被学界接受. 同样, Kolmogorov 的工作不是凭空产生的, 他一样站在巨人的肩膀上. 在公理化方面的早期贡献者有 S. N. Berstein (1880—1968), R. von Mises (1883—1953), E. Borel (1871—1956) 与 P. Lévy (1886-1971) 等, 其中最重要的是法国数学家 Lebesgue, 他于 1900 年发表了历史上最著名的博士论文, 构造出 Lebesgue 测度, 建立了 Lebesgue 积分. Lebesgue 测度的关键是可列可加性, 它进而成为一般测度的关键, Kolmogorov 的公理化概率恰是一个总量为 1 的测度. 自此, 概率论才真正成为数学的一个分支, 数学家们可以根据逻辑来推导概率的性质, 用于解释并研究生活以及其他学科中的随机现象问题. 概率论的概率是否是随机现象的概率呢? 这应该由读者自己来回答.

最后我从 [5] 上引用一句话, 以怀念所有为概率论诞生做出贡献的学者. Today, what Fréchet and his contemporaries knew is no longer known. We know Kolmogorov and what came after; we have mostly forgotten what came before. This is the nature of intellectual progress.

[3]Kolmogorov 出生于 1903 年沙皇时期的俄罗斯, 在他成年的时候已经变成苏联.

第二章　概率初步

在这一章, 我们主要介绍概率论语言与古典概率中的一些问题, 这些问题是概率论公理化之前几百年所积累的, 非常适合用来诠释概率论的思想.

2.1　概率论的语言

伽利略曾经说过：“一个人要理解自然, 必须要先学会自然向我们沟通时所用的语言和符号, 然而这份语言就是数学, 它的符号就是数学符号.” 康德也曾表态：“我的观点是, 一门自然科学中包含了多少科学真理, 取决于它包含了多少数学真理.” 著名数学家李大潜先生在主编高中教材概率论部分时说: 学习概率论, “尽管所涉及的数学实际上是相当初等的, 但第一次接触随机现象, 真正理解起来可能并不容易, 和学习以往的内容不同, 不仅要换一种思路, 还要换一套语言.” 这里我们就来谈谈概率论的语言, 当然在历史长河中, 它也是在演变的, 相比数学其他领域是非常独特的.

2.1.1　集合

概率论是一门有现实背景或者说是来自于现实问题的数学学科, 一开始的时候, 它的问题和回答都是用普通语言来表达的, 其实其他的数学也差不多. 语言是内容的载体, 我们这个课程是关于概率的通识课程, 概率论直观, 与生活密切相关, 且为了通俗易懂, 我们将刻意使用不很严谨的日常语言描述概念和问题, 但数学是严谨的学科, 使用严谨的语言, 我们总归需要把日常语言翻译成数学语言. 因此, 实际上我们要学会两套语言并且熟练地切换, 避免因为对文字理解不同而产生歧义.

19 世纪后期所发展起来的集合与映射是现代数学的通用基础语言, 使得数学各学科之间的交流更加方便. 本节的目的是学习怎么用集合语言表达概率论的基本概念. 要注意的是, 语言只是用来表达, 不是最本质的, 但好的语言的确更简洁准确.

假设读者已经在高中初期阶段学过集合的概念, 所以只是在下面简单复习一下. 一些明确指定的东西放在一起就组成一个集合, 集合中的东西称为集合的元素. 集合中元素是不能重复的, 也不排序. 符号 $a \in A$ 表示 a 在集合 A 中, 或者 a 是 A 的元素. 例如所有自然数的集合 $\mathbf{N}$, 所有整数的集合 $\mathbf{Z}$, 所有实数的集合 $\mathbf{R}$, 所有中国人的集合, 所有女性的集合.

没有任何元素的集合称为空集, 记为 $\emptyset$. 设有两个集合 A, B, 如果 A 中的任何元素也出现在 B 中, 那么我们说 A 是 B 的子集, 或者 A 包含于 B, 或者 B 包含 A, 记为 $A \subset B$, 也可以写 $B \supset A$. 例如 $\mathbf{N} \subset \mathbf{Z} \subset \mathbf{R}$. 中国人的集合和女性的集合是互不包含的. 空集是任何集合的子集, 当然 A 自己也是 A 的子集, 它们称为是 A 的平凡子集, 其他子集是非平凡子集. 如果 A 是 B 的子集, B 也是 A 的子集, 那么我们说两个集合是一样的, 记为 $A = B$.

集合也可以运算得到新的集合. 集合有三种基本的运算, 一个是交, 两个集合 A, B 的交是它们公共元素的全体, 记为 $A \cap B$. 如果它们没有公共元素, 即 $A \cap B = \emptyset$, 那么说它们不交, 或者互斥. 再一个是并, 两个集合 A, B 的元素放在一起 (重复的只放一个) 组成的集合, 称为两个集合的并, 记为 $A \cup B$. 交与并运算可以对多个集合进行. 还有一个运算是减, 把在集合 A 中不属于集合 B 的元素组成的集合, 称为 A 减去 B 的差, 记为 $A \setminus B$, 即

$$A \setminus B = \{a \in A : a \notin B\}.$$

如果我们谈论的一切集合都是 Ω 的子集, 那么 $\Omega \setminus A$ 是不属于 A 的元素全体, 称为 A 的余集, 记为 A^c.

集合 X 到集合 Y 的映射 $f : X \to Y$ 是指按照此规则, X 中任何一个元素 x 对应 Y 中唯一的一个元素 y, 记为 $y = f(x)$, 称 y 是 x 对应的像. 当 Y 都是数集时, 这样的映射也称为集合 X 上的函数. 例如在给定时间下, 身高是人类这个集合上的函数, 体重也是. 中学所说的函数更特殊一点, 要求 X, Y 都是数的集合. 映射是函数概念的推广, 因此它尽管抽象, 但还是有很多具体的例子.

2.1.2　样本空间与事件

实际生活中的随机现象是多种多样的, 大多数随机现象是观察型的, 人类只能被动地观察. 也有一些随机现象是人类可以实践的, 像掷硬币, 掷骰子, 抽签, 摸球, 抽卡片等, 这样的随机现象通常称为随机试验, 因为它们像试验那样可以随意地做. 相比于

观察型, 随机试验更简单, 更容易研究. 我们对于概率论的直觉, 通常也来自随机试验.

从哪里开始呢? 实际上, 对于一个随机现象, 我们关心的是它的随机性. 什么是随机性? 随机性是指每个结果发生的可能性. 例如掷一个硬币的随机性和从一个有一黑一白两个球的袋子中摸球的随机性是一样的. 为了描述随机性, 首先需要记录随机现象的结果. 一个随机现象要可以用数学来描述, 它的所有结果必须是可以明确表达的.

抛一枚硬币, 可能出现正面与反面; 掷一个骰子, 可能出现 1,2,3,4,5,6 之中的一个数字. 明天上午到某个餐厅吃饭的可能人数, 是某个范围内的正整数; 从远处往一面墙上投掷飞镖, 可能落到整面墙上的所有点. 一般的随机试验也是这样, 我们可以描述其所有可能的结果. 这样, 我们把一个随机现象的所有可能出现的结果放在一起组成一个集合, 称为该随机试验的样本空间, 通常用希腊字母 Ω 表示, 样本空间的元素称为结果或者样本点, 一般用 ω 表示.

一个随机试验“所有可能出现的结果”这句话并不是完全客观的, 不同的人对此可能有不同的观点. 例如摸一张牌, 有人关心的是花色, 那么只有四个结果, 有人关心的是数字, 那么有 13 个数字; 例如掷一个骰子, 一般地说所有可能结果是

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

也有人说只有两个结果: 6 与非 6, 另外只关心是不是掷出一个偶数的人眼中只有两个结果: 偶数与奇数. 因此所有可能出现的结果由观察者决定.

样本空间的选取不唯一, 但也不是随意取的, 它需满足以下条件: 观察随机现象, 样本空间中的所有可能结果必须**有且只有**一个出现. 例如掷骰子, 1, 偶数, 素数三个结果不能组成样本空间, 尽管在掷一个骰子之后, 三个结果至少有一个会出现, 但是当 2 出现时, 它是偶数也是素数, 因此它违反“有且仅有”的标准.

接着, 人们关心什么问题呢? 掷三个硬币, 至少有一个正面朝上的可能性; 抛一个骰子, 点数是偶数的可能性; 抛两个骰子, 点数和等于 7 的可能性; 摸一张扑克牌, 摸到黑桃的可能性; 抽签, 抽到幸运签的可能性, 等等. 人们在看到一个随机现象时又关心什么? 一个城市在十年内发生七级地震的可能性; 下一个台风在某地登录的可能性; 明天下雨的可能性; 高速上发生拥堵的可能性, 等等.

概率是随机现象中某一件可能发生也可能不发生的事情发生的可能性大小. 这个事情在概率论中的名称是事件, 所以概率是随机现象中一个事件发生的可能性大小.

问题: 在样本空间已经数学化之后, 事件究竟是什么呢?

大多数学生在中学时已经学过什么是事件, 但如果你恰好不知道, 那么这个问题值得你合上书本好好地思考. 其实样本空间是所有可能的结果全体, 而事件通常是样本空间中的某些结果组成的集合, 即事件对应于若干个结果, 该事件发生相当于其中一个结果发生. 该若干个结果是样本空间的一个子集, 所以我们用子集表示事件. 这样问题就清楚了, 样本空间是所有结果的集合, 一个事件是某些结果组成的子集. 例如掷骰子出现偶数对应于子集 $\{2,4,6\}$, 出现素数对应于 $\{2,3,5\}$.

因为至少有一个结果出现, 所以集合 Ω 作为事件是肯定发生的, 称为必然事件; 没有结果发生是不可能的, 即空集 $\emptyset$ 也称为不可能事件. 不可能事件与必然事件是特别命名的两个特殊的事件. 必然与不可能其实是确定的, 我们一般用大写字母 A, B 表示事件及对应的子集.

事件之间可能是有大小关系的, 如果事件 A 发生蕴含着事件 B 发生, 那么事件 A 中的结果必然都在事件 B 内, 这意味着 A 是 B 的子集, 即 $A \subset B$.

事件是可以复合的, 例如两个事件同时发生相当于一个新的事件, 两个事件至少有一个发生也是一个新的事件. 有意思的是, 事件的复合也对应于集合的运算, 事件 A 不发生也是个事件, 对应补集 A^c, 可以读作 A 的否定, 也称为 A 的对立事件; 事件 A 与 B 同时发生, 那就是事件 $A \cap B$; 事件 A 或者 B 发生, 即至少一个发生, 那就是事件 $A \cup B$. 同时发生与至少一个可以应用于多个事件. 如果事件 A, B 不同时发生, 那么它们就不会有共同的结果, 即 $A \cap B = \emptyset$, 这时我们说两个事件互斥.

这样就可以用集合的语言来叙述随机现象中我们所关注的问题, 与生活中的语言不同, 集合的语言非常严格与清晰, 不会引起歧义. 我们可以用生活中的语言来解释事件, 但是语言的精确度和数学是无法相比的, 如果有争议的话, 最终要以数学表达式为准. 因此从事件到子集, 实际上就是从生活到数学.

练习 2.1 设有 A, B, C 三个事件, 请用集合运算表示下列事件:

1. 仅有 A 发生;
2. A, B 都发生, C 没发生;
3. 三个事件都发生;
4. 至少一个事件发生;
5. 至少两个事件发生;
6. 仅有一个事件发生;

7. 仅有两个事件发生;

8. 不多于两个事件发生.

练习 2.2 设 $A_1, A_2, \cdots, A_n$ 是事件. 写出下面事件的表示 (1) 它们都不发生; (2) 它们不都发生.

2.1.3 直觉

前面说的只是语言, 用来表达每个人在多年的思考和经验中形成的关于随机现象的直觉, 直觉对于这个课程来说是最重要的部分. 当然, 理论结果不一定总是符合直觉, 有时候理论结果“出乎意料”, 甚至和直觉相反, 这时候应该好好思考一下是哪里出了问题, 然后“重建直觉”, 让自己的认知得以提升.

现在对于一个给定的随机现象, 有了样本空间, 也知道了什么是事件, 下一步就是问事件发生的可能性. 我们相信, 每个人都有关于事件发生的可能性及其大小的直觉, 早上出门的时候会估计迟到的可能性有多大, 某次考试之后会估计成绩 90 分以上的可能性多大, 等等. 除了纯粹的可能性之外, 我们也会考察和估计事件发生的可能性之大小, 正如对于物体之重的直觉感受产生重量的概念一样, 对可能性大小的直觉感受也自然地产生了概率的概念, . 但要注意的是, 概率只是表达事件发生的可能性大小, 并不能明确预言事件会不会发生.

问题: 可能性大小是不是可以度量?

其实, 度量无非是赋予一个数字, 因此无所谓可不可以度量, 这个问题的本质是可能性度量是不是有意义. 在某些情况下, 例如随机试验, 可能性大小是可以度量, 而且度量的意义是可以被检验的. 对于其他很多随机现象, 可能性大小是不是可以度量是个很难回答的问题, 因为我们很难解释可能性度量的价值. 例如某地在未来十年发生七级以上地震的可能性可否度量, 某个人未来十年会不会死亡的可能性可否度量, 这些问题依赖于你怎么看待和检验这个度量, 也就是说, 这问题的答案是主观的.

问题: 为什么要度量可能性大小?

我们时常需要在面对随机现象时作决策, 尽管我们不能知道事件会不会发生, 但是事件发生的可能性大小仍然是一个非常重要的指标, 有很大的参考价值.

最后一个问题.

问题: 怎么度量一个事件的可能性大小?

这其实要分两步: 首先是要已知随机性, 这通常是通过分析问题得到的, 其次才是计算一个事件的可能性. 例如, 测量一根木头的长度, 首先我要一把度量尺, 知道 1 尺是多长, 然后我才能测量这个木头有多长. 第一步实际上是假设或者题意, 第二步才是学生发挥聪明才智的地方, 当然大家可以看得出来, 重要的是第一步.

概率的背景就在生活中, 也就是说, 大多数人在生活中自然地形成对于可能性的认识和感觉, 把其中的一般性抽象出来, 就是我们所要的理论.

问题: 人们对生活中的概率有什么样先验的认识和感觉呢?

这问题是很主观的, 难以完全说清楚. 在这里列出关键的几点, 大家可以思考一下是否认可:

1. 不可能事件的概率是 0, 必然事件的概率是 1, 其他事件的概率是处于 0,1 之间;
2. 概率是可累加的, 简单地称为可加性或者加法法则: 两个 (不同时发生的) 事件至少有一个发生的概率等于两个事件概率之和. 例如不输的概率等于胜的概率加平局的概率;
3. 如果随机试验重复, 那么事件发生的频率与概率有密切关系, 概率大的事件发生得更经常.

第一点只是约定俗成, 这是 '率' 的通常意义; 第二点称为概率的可加性, 即可能性的大小是可以累加的, 它是非常本质的. 要注意的是, 互斥的条件是必须的. 从 Venn 图可以看出, 集合的并 $A \cup B$ 只是简单地把两个集合的元素放在一起组成一个集合, 但其中可能有重复的地方, 所以一般情况下, 并的概率与概率的和不同; 第三点是直觉, 也是可能性大小的意义所在.

可加性在实际生活中不仅不陌生, 而且很常见, 物体的重量有可加性, 这很容易验证, 只要拿一块物体来称一下重量, 然后切成两块再分别秤一下重量, 就可以验证了. 线段的长度有可加性, 两根线段连接组成新的线段的长度等于原先两个线段的长度的和. 测量一根很长木头, 我们可以把尺子复制之后交给几个人分段测量, 量好之后加起来. 物体的体积也有可加性. 但概率的可加性不像秤重量那样显然, 是人们的直觉, 仅在一些特殊场合可以证明是成立的.

现在我们用符号来表述, 这是为了方便和精确. 用 $\mathsf{P}(A)$ 来表示某个随机现象中事件 A 发生的概率, 按照定义, 它是事件集合上的一个函数, 满足下面的条件

(1) 非负: 对任何事件 A 有 $\mathsf{P}(A) \geqslant 0$; 且总量为 1: $\mathsf{P}(\Omega) = 1$;

(2) 可加性: 如果事件 A, B 互斥, 那么 $P(A \cup B) = P(A) + P(B)$.

(3) 稳定性: 重复随机试验, μ_n 是事件 A 发生的频率, 即发生次数与总次数之比, 则 $\mu_n \sim P(A)$.

用数学归纳法可以证明, 可加性只要对两个对就对任意有限个对, 即对任何有限个互斥的事件 $A_1, \cdots, A_n$ 有

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n).$$

练习 2.3 用严谨的逻辑证明: 从上面的 (1)(2) 推出 (a) $P(\emptyset) = 0$, (b) 如果 $A \subset B$ 则 $P(A) \leqslant P(B)$, (c) 设 A, B 是两个事件, 则 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. 因此有不等式 $P(A \cup B) \leqslant P(A) + P(B)$.

从历史的角度看, 概率论是从游戏中一些具体问题开始的, 让我们沿着前人的脚步开启概率之旅. 像掷硬币或者骰子这样的随机试验, 结果是明确的, 掷硬币的结果是正面反面, 掷骰子的结果是 123456, 除了这个, 人们非常自然地认为这些结果发生的可能性相同, 简称等可能, 这样每个结果发生的概率自然地是总结果数的倒数, 硬币正面的概率是 1/2, 骰子得 1 点的概率是 1/6. 事件是结果集的一个子集, 它的概率应该是其中结果数与总结果数的比值, 这样的概率恰好满足上面所说的条件. 早期考虑的概率问题基本上都是这种类型的: 随机现象有有限个等可能发生的结果. 等可能性是最简单的一种随机性. 这样的随机现象统称为古典概率模型.

2.2　古典概率模型

有记录的概率问题探讨是从 1654 年 Fermat 与 Pascal 的通信开始的, 大家自然认为那个时间所讨论的概率算是古典概率, 但实际上不完全如此, 古典概率是一个专有名词, 指的是一个所有结果都是等可能发生的随机现象. 早期的大多数概率问题与赌博中的游戏有关.

2.2.1　有限结果与等可能

中学学的概率通常是用排列组合来计算的, 使得很多学生认为概率就是排列组合. 那为什么中学时期算概率经常是使用排列组合呢? 这是因为古典概率的本质是等可能性, 这时算事件概率的本质就是数数, 样本空间小的时候就直接数, 否则就得利用

排列组合. 现在我们从简单的随机试验开始. 首先是掷硬币, 硬币有两面, 就称为正反面吧. 落地的时候究竟哪一面朝上是随机事件, 随便问一个人, 他都会说两面出现的可能性一样, 所以概率都是 $1/2$, 其中 2 是样本空间的元素个数.

同样, 掷一个骰子, 掷出任何一个数的可能性是一样的, 所以概率都是 $1/6$, 其中 6 是样本空间的元素个数.

仔细琢磨这两个例子, 这里有两个关键点, 一是任何结果的可能性一样, 所谓等可能性; 二是一个约定, 也就是说, 必然事件概率约定为 1. 类似地, 只要样本空间中的基本事件是等可能的, 那么每个事件发生的可能性就是 $1/|\Omega|$, 其中 $|\cdot|$ 表示集合中的元素个数.

问题: 等可能性能被证明吗?

也许有人要问, 为什么硬币两面是等可能的? 这个问题也不是一个数学问题, 因为这在数学上无法证明对或者不对, 只能说是合乎自然且理想的假设. 合乎自然的意思是, 硬币的两面物理上几乎没有差别, 我们没有理由作其他假设.

当然这只是一种假设, 如若考虑问题的角度不同, 完全可以作其他假设. 这个问题与我们熟知的平行公理有点类似. 回忆平面几何中的的平行公理: 过直线外一点能且只能作一条直线与已知直线平行. 这个假设直观上容易接受, 但是它无法被证明对或者不对. 假设它对就是 Euclid 几何, 假设它不对, 也是一种几何, 但不再是 Euclid 几何. 不同的假设适用于不同的世界.

现在来看概率的可加性: 两个不会同时发生的事件 A, B 至少有一个发生的概率是它们各自概率之和, 用数学语言说: 如果 $\mathsf{A} \cap \mathsf{B} = \emptyset$, 那么

$$\mathsf{P}(\mathsf{A} \cup \mathsf{B}) = \mathsf{P}(\mathsf{A}) + \mathsf{P}(\mathsf{B}).$$

例如掷三枚硬币至少有两个正面朝上的概率等于恰好有两个正面朝上的概率与有三个正面朝上的概率之和.

还是要回到简单的掷骰子问题, 我们已经知道, 在等可能的假设下, 掷得每个点的概率都是 $1/6$, 然后掷得 1 或者 2 的概率自然是 $2/6$, 因为掷得 1 或者 2 这个事件对应子集 $\{1, 2\}$, 其中有两个元素; 掷得偶数的概率是 $3/6$, 因为掷得偶数这个事件对应子集 $\{2, 4, 6\}$, 其中有三个元素. 因此, 如果 A 是样本空间为 Ω 的古典概率模型中的一个事件所对应的子集, 那么概率理所当然应该是

$$\mathsf{P}(\mathsf{A}) = \frac{|\mathsf{A}|}{|\Omega|}. \tag{2.2.1}$$

这从数学上非常完美, 与之前所说的完全相容, 例如每个基本事件的概率都是

$$\frac{1}{|\Omega|},$$

不可能事件的概率是 0, 必然事件的概率是 1. 现在看起来, 这样来定义概率是很自然的, 但这实际上是 Fermat 和 Pascal 的通信之后才达成共识的, 例如在 G. Cardano 留下的著作中概率并不是这样定义的. 另外, 真正把(2.2.1)作为 (古典) 概率的定义写出来的是 18 世纪末的 S. Laplace.

从定义可以看出, 计算概率的关键是计数, 或者说数元素个数, 不需要知道具体有些什么元素. 因此在中学阶段计算概率经常是使用排列组合.

显然, 这样定义的概率有可加性, 因为当 A,B 互斥时, 并集 $\mathsf{A}\cup\mathsf{B}$ 中元素个数等于他们各自元素个数之和,

$$|\mathsf{A}\cup\mathsf{B}| = |\mathsf{A}| + |\mathsf{B}|,$$

这是元素个数的可加性, 推出概率的可加性

$$\mathsf{P}(\mathsf{A}\cup\mathsf{B}) = \mathsf{P}(\mathsf{A}) + \mathsf{P}(\mathsf{B}).$$

前面说过, 对于一个随机试验来说, 样本空间是由结果决定的, 但是什么是随机试验的结果? 这不是个数学概念, 是主观决定的, 因此样本空间可以有不同的取法.

例 2.2.1　掷一枚硬币三次. 如果依次记录其正反面, 则

$$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\};$$

如果只记录正面出现的次数, 则

$$\Omega = \{0, 1, 2, 3\},$$

其中数字表示正面次数.

第一个样本空间是等可能的, 每个概率都是 1/8; 第二个样本空间不是等可能的, 例如正面次数等于 0,1,2,3 分别对应于第一个样本空间的子集

$$\{000\}, \{001, 010, 001\}, \{011, 101, 110\}, \{111\},$$

所以概率分别是 1/8,3/8,3/8,1/8.

练习 2.4　掷 8 个硬币, 求正面比反面多的概率.

当我们说第一个样本空间是等可能的时, 实际上是从 '每个硬币的两面是等可能的且任何两次掷硬币的结果之间是独立的' 这个前提而来的. 这个前提的后半段 '任何两次掷硬币的结果之间是独立的' 和前半段那样都是属于数学上无法证明的 '合理' 假设.

例 2.2.2 掷两枚骰子. 每个骰子 6 个结果, 所以分别看两个骰子的数字, 有 36 个等可能的结果

$$\Omega = \{(\mathrm{i}, \mathrm{j}) : 1 \leqslant \mathrm{i} \leqslant 6, 1 \leqslant \mathrm{j} \leqslant 6\}.$$

若只看两个骰子的点数之和, 那么样本空间是

$$\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\},$$

其中的概率依次为 1/36,2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36. 算事件 A: "一个数是另一个数的倍数" 的概率, 从样本空间中找, A 共 22 个结果, 因此 $\mathsf{P}(\mathsf{A}) = 22/36$. 再算事件 B: "两个数互素" 的概率, B 共 23 个结果, 因此 $\mathsf{P}(\mathsf{B}) = 23/36$.

练习 2.5 列出上例事件 A, B 的所有元素.

样本空间可以按照观察者的关注角度选取, 只要一个随机试验存在一个等可能的样本空间, 我们就认为它是古典概率模型, 而且一般总是就取等可能的样本空间.
注意: 对于一个随机现象来说, 样本空间会可能因为视角不同而有差异, 但是, 事件的概率只依赖于随机性的确定, 与样本空间选取无关. 因此我们总是愿意选取一个容易计算概率的样本空间.

练习 2.6 一排 12 个车位停了 8 辆车, 求四个空位在一起的概率.

练习 2.7 从标准的 52 张扑克牌中取 5 张牌, 求下列情况各有多少不同选择:

1. 至少有一个对子;
2. full house(三张数字一样再加一个对子);
3. 正好是个顺子;
4. 正好是个同花顺子;
5. 四种花色都出现.

2.2.2　独立性

在前面几个简单的例子中, 读者可能已经看到我们实际上潜在地使用了独立性. 独立是古典概率模型中另外一个重要的概念.

计算古典概率时, 一个经常使用的方法是把一个随机试验分解为若干个独立的随机试验, 或者发过来把几个随机试验独立地复合在一起. 例如, 掷 10 枚银币可以分解为独立地掷 10 次银币. 独立性是古典概率的另外一个重要概念, 可以理解为一个随机试验的结果对另一个随机试验的结果没有影响. 例如掷一个硬币和一个骰子, 两者的结果是没关系的, 或者说是独立的. 分别属于独立的两个随机试验之下的事件同时发生的概率等于它们各自概率的积. 例如, 硬币正面朝上且骰子掷出 6 点的概率等于 1/2 与 1/6 的乘积, 是 1/12.

这通常被认为是非常显然的, 但也应该问为什么, 把其中的道理理解透彻. 硬币有两个结果, 骰子有 6 个结果, 总共有 12 个结果, 独立本质上是说这 12 个结果等可能, 反之亦然. 一般地, 如果 A 是一个随机试验中的事件, B 是另外一个随机试验中的事件. 如果这两个随机试验独立, 那么就有

$$P(A \cap B) = P(A)P(B).$$

这是直觉, 但也是因为由乘法原理, 两个随机试验合在一起, 其样本空间总数是先前的样本空间总数之乘积, 且结果是等可能的. 现在我们说, 如果上式成立, 那么称 A, B 独立.

练习 2.8　设 A, B 独立. 证明: A 与 B^c 也独立. 这是说事件 A 与事件 B 发生不发生独立.

上面所说的是独立性的本质. 读者可能注意我们在使用独立性时没有解释或者定义, 感觉完全依赖直觉, 不严谨. 实际上, 现在这个场合下独立性是个幌子, 不是必须的, 引入它的目的是直观和方便. 本质还是等可能性, 掷三个硬币 '结果是独立的' 与有 '八个等可能结果' 等价, 不过独立听起来更直观一些. 也就是说, 在古典概率模型中, 独立性假设本质上就是等可能性假设, 因此用不用独立性无关紧要,

当然在一般概率模型中, 随机试验是没有数学定义的, 独立性有, 且具有不可替代的意义, 后面将反复地解释独立性. 事先提醒大家, 独立的概念不是表面所见的那么简单.

2.3　古典概率典型问题

下面我们来讨论一些简单经典的概率问题, 解决这些问题基本上只需要直觉, 不需要太多的数学工具, 因此大家在思考这些问题的时候, 首先尽量发挥想象力, 使用直觉-加法法则与乘法法则; 其次在使用直觉解决问题时, 要能够判别使用了什么隐藏的假设或者工具.

解古典概率问题的目的是加深大家对于随机性及其概率的理解. 但很多古典概率问题有难度, 更像是数学中的智力游戏, 通过解决有挑战性的问题获得愉悦感, 对理解概率不一定很有帮助.

2.3.1　分苹果

在一个社会中, 资源分配是否公平是一个永恒的话题, 永远会有争议. 为了公平, 经常引入随机性. 例如, 我小的时候, 是人民公社年代, 计划经济, 很多分配都是村干部说了算, 但年底分鱼塘里的鱼, 春天分山上的桃子, 是抽签决定的. 因为鱼或者桃子有大有小, 按照户数分成若干堆, 不可能完全一样, 在不同的人眼里价值不一, 所以强行指定的话群众会有怨言, 而通过抽签决定能让公社社员们觉得比较公平, 这是很重要的心理安慰剂.

现在我们看看怎么把两个苹果公平地分给两个人? 这看起来是个简单问题, 但实际上在物理学及经济学中有深刻的背景. 平均主义分配, 一人一个. 这算是表面看来最公平的确定的方法. 但两个人未必都会赞同, 因为苹果可能有大小和品质的区别. 谁来决定分配方案? 如果由人来分配, 就不能排除旁人的怨言. 这时候, 还可以引入竞争机制, 比如回答一个问题, 完成一个任务等等. 如果一个社会都以这种方式分配, 那就可能会形成能力强的人总是获得资源, 最后的结果可能是贫富悬殊, 贫富悬殊的结果是社会动荡, 谁也没有安全感. 还有一个办法就是机会均等, 通过随机的方法分, 这并不是多么高明, 只是相当于把决定权从有形的手交给无形的手, 使得资源分配不再那么集中, 也使得众人即使不满也无处抱怨. 这实际上也是一种社会治理模式.

问题: 怎么公平地把两个苹果分给甲乙丙三个人?

如果不切开分, 那么随机的方法就是必须的, 没拿到苹果的人肯定抱怨. 随机分配的方案有很多, 下面是一些例子. 有趣的是, 苹果可以看作粒子, 人可以看作位置, 分苹果可以看作粒子占据位置, 这样分苹果问题就与统计物理的粒子占位问题有关了. 注意下面说的等可能决定是指产生等可能概率的方法, 可以掷骰子也可以抽签, 假设

大家可以设计出这样的方法.

1. 两个苹果一起, 抽签决定归谁.
2. 依次拿一个苹果, 然后抽签决定归谁. 这个方案在物理上称为 Maxwell-Boltzmann 统计, 实际上是个假设.
3. 两个苹果分成三份 (可以是 0）有六种方法: $002, 020, 200, 110, 101, 011$, 其中第 1,2,3 个位置的数字是分给甲乙丙的苹果数. 然后抽签决定哪一种. 这在物理上称为 Bose-Einstein 统计. 注意如果苹果好坏不一, 那么这个分法是不公平的.
4. 限制每人最多一个苹果, 那么有三种分配方法: $110, 101, 011$, 然后抽签决定. 这在物理上称为 Fermi-Dirac 统计. 这相当于先用 1/3 概率分一个苹果, 然后再用 1/2 概率分剩下的苹果给其它两人.

注意, 两个苹果无法区分好坏时 (例如在统计物理中), 3,4 两种分法是公平的, 否则这两种分法不公平, 可能还需要掷硬币来决定谁先拿.

练习 2.9 就上面罗列的四种情况, 求 (1) 甲恰好分到一个苹果的概率; (2) 甲没有分到苹果的概率.

练习 2.10 请列出第 2 种分配的样本空间. 请解释 2,3 有何不同.

一个随机的分配方案是不是公平, 一般人都能够判断. 但是其内蕴的依据是什么呢? 后面我们会再讨论这个问题.

2.3.2 分赌注: Fermat 的想法

第一个问题自然是分赌注问题, 分赌注问题是法国贵族赌徒 de Méré 请教数学家 B. Pascal 的, Pascal 写信给 Fermat 讨论这个问题. 这个问题催生出概率这个学科, 可以说是概率论的开山之题.

问题: 两个赌徒甲乙各出 32 块钱作为赌注, 每赌一局胜者得一分先得 3 分者赢得全部赌注. 现在甲得 2 分, 乙得 1 分时, 甲乙同意终止赌局, 问应该怎么分赌注才是公平的?

当时有很多不同的答案, 很多人倾向于按比分分配, 但主要的问题不是答案本身, 而是让大家相信你的答案. Fermat 和 Pascal 都是天才数学家, 下面是 Fermat 的思路.

现在我们来看, 如果继续赌局的话, 甲乙最终赢得赌注的概率比是多少? 甲已经有 2 分, 乙有 1 分, 继续赌局. 甲先得三分有两个途径: 1. 接着的第一局甲胜, 概率为 1/2; 2. 接着的第一局乙胜, 但第二局甲胜, 概率为 1/4, 即 1/2 乘以 1/2, 其中假设了两次赌局是独立的. 因此甲最终赢的概率是 3/4, 那么乙最终胜的概率是 1/4, 他们两者最终赢得赌注的概率比就是 3:1. 也就是说, 赌注按照赢的概率之比例来分配就是公平的.

2.3.3　分赌注: Pascal 的想法

问题: 为什么上面这样分是公平的呢?
答案是因为公平就是满足预期, 而上述按概率的分法恰好代表了预期. 这也是分赌注问题中 Pascal 的思想. Pascal 同意 Fermat 的想法, 但是他有自己的思路.
我们假设第一个人得了两分, 另外一个人得了一分, 他们现在再玩一次, 如果第一个人赢, 那么他就拿走全部赌注, 如果另一个人赢了, 那么他们二比二平. 接着, 如果他们愿意终止, 那么他们每人拿一半赌注.
那么, 先生, 考虑到如果第一个人赢, 他取 64 元, 如果输, 他取 32. 这时, 如果他们不想玩这一局, 就此终止, 那么第一个人会说: “对我来说, 32 元是保证的, 对于另外的 32 元, 也许可能是你的, 也许可能是我的, 机会是相等的. 因此我们应该平分这 32 元, 且给我另外的 32 元.”那么他就拿 48 元, 另一个人拿 16 元. 这实际上是以“若必赢拿全部, 若对等则均分”这一大家认可的规则来分配.
最后我们看到, 按照概率比的分配和按照期望的分配是一致的, 或者说概率符合人的心理预期. 这也某种程度上解释了期望的直观意义.

练习 2.11　5 局 3 胜. 甲得 1 分而乙得 0 分时终止赌局, 他们应该怎么分赌注?

2.3.4　掷硬币

问题: 甲乙各掷 $\mathfrak{n}$, $\mathfrak{n}+1$ 枚硬币. 问乙得到的正面数比甲多的概率是多少?
当 $\mathfrak{n}=1$ 时, 总共三枚硬币, 8 个等可能结果,

$$\Omega=\{000,001,010,100,011,101,110,111\},$$

其中第一个符号表示甲的硬币, 后两个符号表示乙的两个硬币. “乙得到的正面数比甲多”这个事件对应于子集 $\{001,010,011,111\}$, 其中 4 个元素, 所以所求概率是 4/8=1/2.

当 $n=2$ 时, 总共 5 枚硬币, 32 个等可能结果, 每个概率都是 $1/32$,

$$\begin{aligned}\Omega=\{&00000,00001,00010,00011,00100,00101,00110,00111,\\&01000,01001,01010,01011,01100,01101,01110,01111,\\&10000,10001,10010,10011,10100,10101,10110,10111,\\&11000,11001,11010,11011,11100,11101,11110,11111\},\end{aligned}$$

其中前两个是甲的硬币, 后三个是乙的硬币. 怎么列出所有可能结果? 这里是利用二进制表示的方法从小到大列出所有结果. 这个事件对应于子集

$$\begin{aligned}\{&00001,00010,00011,00100,00101,00110,00111,\\&01011,01101,01110,01111,\\&10011,10101,10110,10111,\\&11111\},\end{aligned}$$

其中共有 8 个元素, 所以所求概率是 $8/16=1/2$.

因此我们猜测不论 n 多大, 概率总是 $1/2$. 但是当 $n=3$ 时, 总共 7 个硬币, 样本空间有 $2^7=128$ 个元素, 上面那种数个数的方法就很麻烦了. 因此我们需要换个聪明点的方法.

用字母 A 表示“乙得到的正面数比甲的正面数多”这个事件, 用 X,Y 分别表示甲乙的正面个数. 那么

$$A=\{X<Y\}=\bigcup_{i<j}\{X=i,Y=j\}.$$

因此

$$P(A)=\sum_{i=0}^{n}\sum_{j=i+1}^{n+1}P(X=i)P(Y=j),$$

其中用到甲乙两人硬币正面数之间是独立的这个假设.

概率 $P(X=i)$ 是指掷 n 个硬币得到 i 个正面数的概率. 掷 n 个硬币这个随机试验的样本空间有 2^n 个等可能的元素, 如上所示, 每个是 n 位置放上 0 或者 1. $X=i$ 这个事件相当于 n 位置中恰有 i 个 1, 因此有 $\binom{n}{i}$ 个可能结果, 因此

$$P(X=i)=\binom{n}{i}/2^n.$$

因此

$$P(A)=\frac{1}{2^{2n+1}}\sum_{i=0}^{n}\sum_{j=i+1}^{n+1}\binom{n}{i}\binom{n+1}{j}.$$

由二项式定理,

$$\frac{1}{2^{2n+1}}\sum_{i=0}^{n}\sum_{j=0}^{n+1}\binom{n}{i}\binom{n+1}{j}=1.$$

现在我们要说其中组成 $P(A)$ 的与剩下的和 $P(A^c)$ 一样

$$\sum_{i=0}^{n}\sum_{j=i+1}^{n+1}\binom{n}{i}\binom{n+1}{j}=\sum_{i=0}^{n}\sum_{j=0}^{i}\binom{n}{i}\binom{n+1}{j}.$$

这由组合的对称性 $\binom{n}{i}=\binom{n}{n-i}$ 推出. 为什么呢? 看下面的矩阵

$$\begin{array}{ccccccc}
\binom{n}{0}\binom{n+1}{0} & \diagdown & \binom{n}{0}\binom{n+1}{1} & \binom{n}{0}\binom{n+1}{2} & \cdots & \cdots & \binom{n}{0}\binom{n+1}{n+1}\\
\binom{n}{1}\binom{n+1}{0} & \binom{n}{1}\binom{n+1}{1} & \diagdown & \binom{n}{1}\binom{n+1}{2} & \cdots & \cdots & \binom{n}{1}\binom{n+1}{n+1}\\
\binom{n}{2}\binom{n+1}{0} & \binom{n}{2}\binom{n+1}{1} & \binom{n}{2}\binom{n+1}{2} & \diagdown & \cdots & \cdots & \binom{n}{2}\binom{n+1}{n+1}\\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\
\binom{n}{n}\binom{n+1}{0} & \binom{n}{n}\binom{n+1}{1} & \binom{n}{n}\binom{n+1}{2} & \cdots & \cdots & \diagdown & \binom{n}{n}\binom{n+1}{n+1}
\end{array}$$

$P(A)$ 中的组合数是矩阵中斜线右部的和, 即 $i<j$ 的所有 $\binom{n}{i}\binom{n+1}{j}$, 由组合公式

$$\binom{n}{i}\binom{n+1}{j}=\binom{n}{n-i}\binom{n+1}{n+1-j}.$$

因此 $\{(i,j):0\leqslant i\leqslant n,i+1\leqslant j\leqslant n+1\}$ 与 $\{(i,j):0\leqslant i\leqslant n,0\leqslant j\leqslant i\}$ 通过映射 $(i,j)\mapsto(n-i,n+1-j)$ 是一一对应的, 因此 $P(A)$ 与 $P(A^c)$ 一样, 推出 $P(A)=1/2$.
这个方法虽然可以解决问题, 但看起来很繁琐, 有没有更聪明的方法呢? 既然答案是 $1/2$, 是不是有什么诀窍在其中呢? 实际上关键在于两人的硬币总数是奇数, 正因如此, 如果两人得到的正面数一样, 那么反面数肯定不同. 用字母 B 表示“乙得到的反面数比甲的反面数多”这个事件. 因为正反面对称 (其实也就是等可能性), 所以 $P(A)=P(B)$. 下面我们证明 A,B 是对立事件, 即至少有一个发生, 但不会同时发生.
如果它们都不发生, 即乙得到的正面数不比甲的正面数多, 且乙得到的反面数也不比甲的反面数多, 那么乙的硬币总数不比甲的总数多, 矛盾.
如果它们同时发生, 即乙得到的正面比甲的正面数多, 且乙得到的反面数也比甲的多, 那么乙的硬币数至少比甲的硬币数要多两个, 矛盾.

所以 $\mathsf{P}(A)+\mathsf{P}(B)=1$, 推出 $\mathsf{P}(A)=1/2$.

解答问题的三个方法: 第一个是数数, 第二个还是数数, 只不过用排列组合来数, 第三个是利用概率性质. 数数是解决古典概率的基本方法, 可能繁琐一点, 第三个方法更本质一点但不容易想到. 也许很多人欣赏后一种方法. 但是从学习的角度讲, 前一种方法更为自然, 更容易入手, 适用性也更广. 学习数学的初心是学习怎么从概念出发进行有逻辑的思考, 刻意追求技巧是不可取的. 只要不断地思考, 也许有一天, 这样有深度和奇妙的想象力的方法会自己来找你.

练习 2.12　一个盒子里有红球与黑球, 现在任取两个球都是红球的概率是 1/2. 问:

1. 盒子中至少有几个球?
2. 已知有偶数个黑球, 盒子中至少该有几个球?

练习 2.13　投三颗骰子, 求下列事件的概率 (1) 其中恰有两颗点数一样; (2) 其中至少有两颗点数一样.

练习 2.14　掷 5 个骰子, 点 1 可以替代任何点, 求 5 个骰子点数一致的概率.

2.3.5　生日问题

生日对于一个人是个奇妙的日子, 一年 365 天, 两个人恰好生日相同的可能性是 1/365, 大概是 0.3%, 所以碰到同生日的人通常是一种惊喜.

问题: 设有 $\mathfrak{n}$ 个人, 问至少有两个人生日相同的概率是多少?

把 $\mathfrak{n}$ 个球随机地放入 N 个盒子中, 每个球等可能地放入任何一个盒子. 那么每个球有 N 种放法, 因此样本空间 Ω 共有 $\mathsf{N}^{\mathfrak{n}}$ 种可能的放法, 每一种放法是等可能的. 这相当于粒子占位中所说的 Maxwell-Boltzmann 统计. 现在设 $\mathfrak{n}\leqslant\mathsf{N}$, 考虑事件: 每个盒子至多只有一个球. 我们只需计算这个事件有多少基本事件就可以了. 由乘法原理知道, 它应有

$$\mathsf{N}(\mathsf{N}-1)\cdots(\mathsf{N}-\mathfrak{n}+1)$$

个, 即 N 个数中取 $\mathfrak{n}$ 个的排列 $(\mathsf{N})_{\mathfrak{n}}$. 因此每个盒子至多只有一个球的概率为 $(\mathsf{N})_{\mathfrak{n}}/\mathsf{N}^{\mathfrak{n}}$.

我们可以近似地把生日问题看作粒子占位问题. 一年是 365 日, 一个人的生日假设是从 365 日中随机选一日, 不同人的生日也可以假设是独立的, 这样 $\mathfrak{n}$ 个人生日各

不相同的概率为 $(365)_n/365^n$, 因此至少有两人生日相同的概率为

$$p_n = 1 - \frac{(365)_n}{365^n}.$$

计算这个概率, 直觉不是那么好用, 但是读者可以想一想你直觉认为多少人可以保证有 99% 的概率有两个人生日相同?

好, 现在让我们利用计算器甚至计算机来计算. 比如

$$\begin{array}{rccccccccc} n = & 15, & 20, & 25, & 30, & 35, & 40, & 45, & 50, & 55, \\ p_n = & 0.25, & 0.41, & 0.57, & 0.71, & 0.81, & 0.89, & 0.94, & 0.97, & 0.99, \end{array}$$

例如 55 个人就有 99% 的概率有两人生日一样, 有一本书题目叫做“计算出乎意料”. 这个数字是否出乎你的意料之外呢?

你知道什么时候可以说这件事一定发生呢? 一定是指概率为 1. 只要 n 不超过 365, 有两个人生日相同的概率必小于 1, 也许非常非常接近 1, 但不是一定的. 所以只有当 n 大于 365, 我们才可以说一定会有两个人生日相同, 但只要有 55 个人就可以说有 99% 的概率会有两个人生日相同. 也就是说, 需要多 6 倍的人数才能弥补最后 1% 的概率, 这对很多人来说是出乎意料的.

2.3.6　抽签与顺序

抽签是生活中经常用的一个工具, 例如小时候生产队分桃子, 一堆一堆的桃子, 个数重量差不多, 但是各人对哪堆好意见不同, 那通常就用抽签方法解决, 这样公平. 一般来说, 抽签是依次一个个抽, 那么我们自然会问抽到好签的机会是否是顺序有关, 这是大家都好奇的问题. 有人觉得可能有关系, 因为他觉得如果他在以后抽的话, 可能前面的人把好签抽走了, 机会没了. 有的人会比较镇定, 因为他觉得前面的人不一定抽到, 如果抽不到, 那么留给自己的机会就大了. 因此需要理性地分析一下.

问题: 假设是 4 个签, 2 个写字母 A, 2 个写字母 B. 4 个人顺序不放回抽签, 问抽到字母 A 的机会是否与顺序有关?

把 4 根签编号为 1,2,3,4, 其中 1,2 写字母 A, 3,4 写字母 B. 按照 4 个人顺序不放回抽签得到的数字写样本空间, 应该是 1,2,3,4 的所有排列

$$\begin{aligned} \Omega = \{&1234, 2134, 3214, 3124, 1324, 2314, \\ &1243, 2143, 3241, 3142, 1342, 2341, \end{aligned}$$

$$1423, 2413, 3421, 3412, 1432, 2431,$$
$$4123, 4213, 4321, 4312, 4132, 4231\}.$$

A_i 表示第 i 个抽签者抽到 A 这个事件. 那么 A_i 就是第 i 个位置是 1 或者 2 的那些基本事件, 所以

$$\begin{aligned}
A_1 =& \{1243, 2134, 1324, 2314, 1243, 2143,\\
& 1342, 2341, 1423, 2413, 1432, 2431\}\\
A_2 =& \{1234, 2134, 3214, 3124, 1243, 2143,\\
& 3241, 3142, 4123, 4213, 4123, 4231\}\\
A_3 =& \{3214, 3124, 1324, 2314, 1423, 2413,\\
& 3421, 3412, 4123, 4213, 4321, 4312\}\\
A_4 =& \{3241, 3142, 1342, 2341, 3421, 3412,\\
& 1432, 2431, 4321, 4312, 4232, 4231\}.
\end{aligned}$$

每个事件的元素都是 12 个, 所以

$$P(A_1) = P(A_2) = P(A_3) = P(A_4) = \frac{1}{2}.$$

因此抽签得中的机会与顺序无关, 这其实也是经验或者直觉.
一般地, 把 n 个球编号, 其中 k 个白, $n-k$ 个黑. 然后 n 个人顺序不放回取球等价于把球随机排列. 那么对称性说明 $n!$ 种全排列是等可能的. 无论哪个位置, $1, 2, 3, \cdots, n$ 上是白球有 k 种可能, 该位置放好之后, 其他 $n-1$ 个位置 $n-1$ 个球有 $(n-1)!$ 种放法, 所以概率是

$$\frac{k \cdot (n-1)!}{n!} = \frac{k}{n}.$$

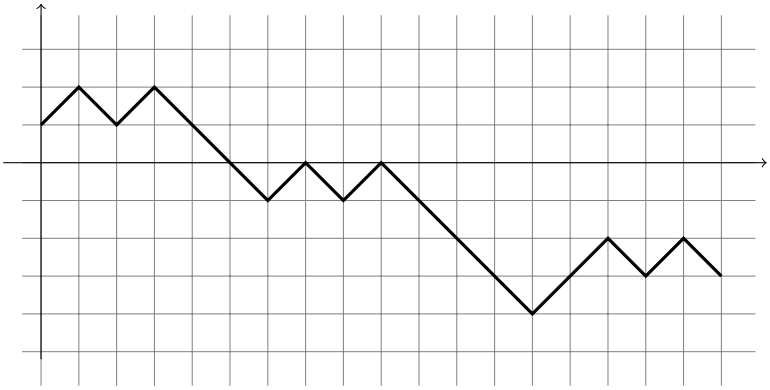
因此与顺序无关的性质其实就是排列的对称性.

2.3.7　格点轨道: 反射原理

前面所说的问题比较简单, 下面我们说两个仍然是初等但数学上有难度且有意思的问题. 格点轨道问题实际上是简单随机游动问题, 而随机游动是概率论的核心问题,

其极限是著名的 Brown 运动. 下面我们介绍一种利用对称性计算格点轨道数的方法, 称为反射原理.

平面上的格点是指坐标都是整数的点, 把横轴看成时间轴, 纵轴看成状态轴. 一个格点轨道是指在格点上向右方移动, 每次向右上或者右下移动一格, 的轨道, 即在 (i,j) 处的点等可能地移到 $(i+1,j+1)$ 或者 $(i+1,j-1)$.



格点轨道对应于简单对称的随机游动, 设甲乙两个人掷一枚硬币进行赌博, 正面甲赢一元, 反面乙赢一元. 用 $x(n)$ 表示甲在时刻 n 的钱数, 它是简单对称随机游动, 它的图像 $\{(n,x(n)) : n \geqslant 0\}$ 是个格点轨道. 上图是掷硬币产生的一个轨道. 研究这样的随机游动与研究格点轨道是一样的, 前者通常是概率方法, 后者通常是组合方法. 在本节中, 我们使用组合方法, 后面我们会用概率方法再次回到随机游动.

固定起点, 时间长度为 n 的格点轨道一共有 2^n 条, 它们是等可能的.

问题: 从点 $(0,j)$ 出发, 是否可以到达点 (m,n)? 如果可以到达, 总共有多少条?

从点 $(0,j)$ 到 (m,n), 时间跨度为 m, 假设在这段时间掷了 r 个正面, s 个反面, 那么 $(0,j)$ 可达 (m,n) 当且仅当存在非负整数 r,s 使得 $r+s=m$, 且 $r-s=n-j$, 即

$$\begin{cases} 2r = m + (n-j); \\ 2s = m - (n-j). \end{cases}$$

因此方程有解的充分必要条件是 m 与 $(n-j)$ 有相同的奇偶性且

$$|n-j| \leqslant m.$$

当上述条件满足时, $(0,j)$ 到达 (m,n) 相当于掷 m 次硬币得

$$[m-(n-j)]/2$$

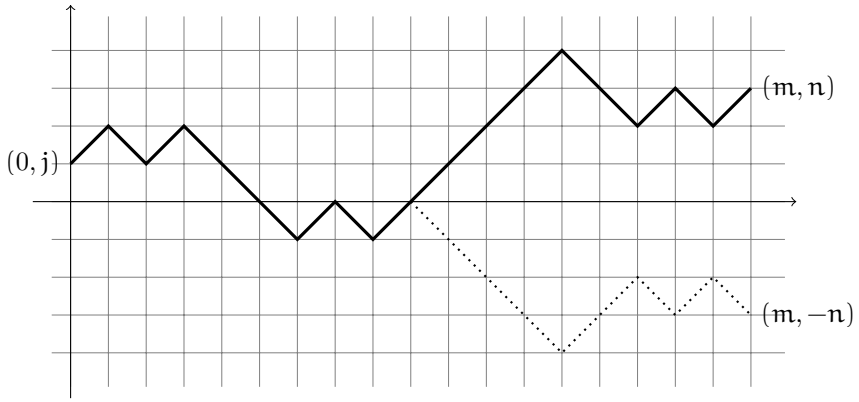
次正面, 用组合原理, 共有

$$\binom{m}{[m-(n-j)]/2}$$

种方法, 因此到达的轨道也是这么多.

问题: 设 $j>0, n>0$, 问从 $(0,j)$ 到 (m,n) 的轨道有多少是碰到时间轴的? 即轨道上有一个点的纵坐标是零.

用 A 表示这样碰到时间轴的轨道全体, B 表示从 $(0,j)$ 到 $(m,-n)$ (它是 (m,n) 关于时间轴的对称点) 的轨道全体. 为了比较两个集合中元素的多少, 建立 A 到 B 的一个变换: 任取 A 的一条轨道, 把其上最后一个纵坐标等于 0 的点之后的轨道关于时间轴反射下来, 之前保持不动, 得到新的轨道, 这条轨道的终点是 $(m,-n)$, 所以属于 B. 见下图



这个变换是 A 到 B 的一一对应 (请读者证明之), 即单 (不同的轨道的像不同) 且满 (B 中轨道必定是 A 中轨道反射来的). 因此 A 中的轨道总数为

$$|A| = |B| = \binom{m}{[m+n+j]/2}.$$

这个结论称为反射原理. 它的主要原因是轨道的对称性. 通过构建一一对应来数数是一个非常重要的思想.

问题: (投票问题) 总数为 $m+n$ 的投票人依次投票给甲乙两个候选人, 已知甲乙的票数分别为 m, n 且 $n>m$, 问在投票过程中乙一直领先的概率是多少? 这里一直领先是指乙票数一直多于甲.

这是个经典的问题, 相当于求 $(0,0)$ 到达 $(m+n, n-m)$ 的格点轨道中不碰到时间轴 (起点除外) 的格点轨道比例.

首先 $(0,0$ 到 $(m+n,n-m)$ 的格点轨道总数是 $\binom{n+m}{m}$.
其次, $(0,0)$ 到 $(m+n,n-m)$ 的格点轨道中不碰到时间轴 (起点除外) 的格点轨道数, 等于 $(1,1)$ 到 $(m+n,n-m)$ 的格点轨道中不碰到时间轴的格点轨道数, 等于 $(1,1)$ 到 $(m+n,n-m)$ 的格点轨道总数, $\binom{m+n-1}{n-1}$, 减去其中碰到时间轴的轨道数. 后者可以用反射原理计算.
由反射原理, $(1,1)$ 到 $(m+n,n-m)$ 碰到时间轴的格点轨道数, 等于 $(1,1)$ 到 $(m+n,m-n)$ 的格点轨道总数, 即等于 $\binom{m+n-1}{m-1}$.
因此所求概率为

$$\frac{\binom{m+n-1}{n-1}-\binom{m+n-1}{m-1}}{\binom{n+m}{m}}=\frac{n-m}{n+m}.$$

练习 2.15　设有 n 个持 50 元钱的人和 m 个持 100 元钱的人在一个窗口排队买票, $n\geqslant m$, 票价是 50 元且窗口开始没有零钱. 求所有买票的人都不需要等待找钱的概率. 等同于上面的投票过程中乙一直不落后的概率.

练习 2.16　假设有 $2n$ 个人投票, 每个人等可能地投甲或者乙. 证明: 投票过程中甲一直不落后的概率与最终甲乙得票一样的概率.

2.3.8　配对问题: 容斥原理

问题: n 对夫妇参加舞会, 舞会将 n 位男士与 n 女士随机地配成 n 对舞伴, 问没有一对夫妇配成舞伴的概率是多少?
用 A 来表示问题中的这个事件, 当 $n=2,3$ 时, 样本空间总数是 2,6, 很简单可以计算 $P(A)$ 分别是 1/2, 1/3. 我们认真地看一下 $n=4$ 的情况, 1,2,3,4 的所有排列

$$\begin{aligned}\Omega=\{&1234,2134,3214,3124,1324,2314,\\&1243,2143,3241,3142,1342,2341,\\&1423,2413,3421,3412,1432,2431,\\&4123,4213,4321,4312,4132,4231\}.\end{aligned}$$

去掉其中数字与位置有相同的元素, 例如 1243, 3241 等, 得 $A=\{2143, 3142, 2341, 2413, 3421, 3412, 4123, 4321, 4312\}$, $P(A)=9/24=3/8$. 对一般的 n, 这样的方法就没用了.
下面我们介绍容斥原理, 它是一个复杂但有用的公式. 当然读者可以尝试其他方法, 数学的想象空间是无限的, 我们永远不能断言什么是解决这个问题的唯一方法. 容

斥定理最早讨论怎么数多个集合并的元素个数. 例如两个集合的并的元素个数, 先把各自集合元素个数加起来 (容), 那么相交部分其实数了两次, 所以要减去相交部分的元素个数 (斥). 如果数三个集合的并的元素个数, 先数各自集合元素个数加起来, 那么两个集合相交部分多数了一次, 三个集合相交部分多数了两次, 先减去两个结合相交部分个数, 那么三个几何相交部分被减了三次, 因此需要再加回来一次. 现在我们来看看它的一般形式是什么样的, 怎么证明.

用 A_i 表示第 i 对夫妇配对了这个事件. 那么

$$A = A_1^c \cap A_2^c \cap \cdots \cap A_n^c = \left(\bigcup_{i=1}^{n} A_i\right)^c.$$

因此我们只需要算

$$P\left(\bigcup_{i=1}^{n} A_i\right).$$

这些事件可能会同时发生, 所以不能直接用概率的可加性, 这是集合 '并' 和数字 '加' 的区别. 当 $n = 2$ 时有

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2),$$

事实上, $A_1 \cup A_2 = A_1 \cup (A_2 \setminus A_1)$, 可加性推出

$$P(A_1 \cup A_2) = P(A_1) + P(A_2 \setminus A_1).$$

然而, $A_2 = (A_2 \setminus A_1) \cup (A_2 \cap A_1)$, 所以

$$P(A_2 \setminus A_1) = P(A_2) - P(A_1 \cap A_2).$$

因此结论成立. 该公式被称为容斥定理或者原理, 它的一般形式如下, 实际上也是可以想象的,

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) = &\sum_{i=1}^{n} P(A_i) \\ &- \sum_{i<j} P(A_i \cap A_j) \\ &+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \\ &- \cdots \end{aligned}$$

$$+(-1)^{n-1}P(A_1\cap A_2\cap\cdots\cap A_n).$$

理解容斥定理的难点不在于证明, 而在于理解右边的那个复杂表达式. 首先它是一正一负的 n 个和, 第一个和是单个事件的概率和, 共有 n 项, 第二个和是两个事件交的概率和, 共有 C_n^2 项, 第三个和是三个事件交的概率和, 共有 C_n^3 项, 等等, 要注意的是任何几个事件交在右边的和中出现且只能出现一次. 奇妙的是, 容斥原理与二项展开式有类似的地方, 可以帮助我们记忆. 将容斥定理的两边保留其项数与符号, 那么就有

$$1=n-C_n^2+C_n^3-\cdots+C_n^n(-1)^{n-1},$$

而这恰好就是二项展开公式,

$$1-n+C_n^2-C_n^3+\cdots+C_n^n(-1)^n=(1-1)^n.$$

也就是说, 左边简单的 1 可以表达为右边的复杂式子, 数字正好相同于容斥定理的项数, 这是巧合吗? 有时, 数学表达式本身的神奇是无法用文字描述的.

一般的情况的证明可以应用数学归纳法, 例如 $n=3$,

$$P(A_1\cup A_2\cup A_3)=P(A_1\cup A_2)+P(A_3)-P((A_1\cup A_2)\cap A_3),$$

交对并有分配律 $(A_1\cup A_2)\cap A_3=(A_1\cap A_3)\cup(A_2\cap A_3)$, 然后应用 $n=2$ 时的公式即可得证.

练习 2.17　就 $n=4$ 证明容斥定理.

容斥定理是将事件或发生的概率用事件同时发生的概率来计算, 这似乎是典型的化简为繁, 与数学通常做的化繁为简背道而驰, 说明数学思想的不拘一格. 现在我们来展示容斥定理的力量, 用它解决配对问题. 配舞伴可以看成以下过程: 男士依数字顺序站好, 女士随机排列. 这个模型有许多等价的表述: 1. 标号为 $1,2,\cdots,n$ 的 n 个数卡随机地排放在标号为 $1,2,\cdots,n$ 的 n 个位置上, 每个位置只能放一个数, 2. 标号为 n 个数字的人抽取编号为 n 个数字的签, 等等. 样本空间是 n 个数字的全排列, 所以 $|\Omega|=n!$. A_i 表示事件“第 i 对夫妇配对”, 或者第 i 个位置上恰好是编号为 i 的数卡, 这样 $|A_1|=(n-1)!$. 因此 $P(A_i)=1/n$. 同理

$$P(A_i\cap A_j)=\frac{1}{n(n-1)}.$$

当 $1<i_1<\cdots<i_m\leqslant n$ 时,

$$P(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_m})=\frac{1}{n(n-1)\cdots(n-m+1)}.$$

应用容斥定理,

$$\begin{aligned}P\left(\bigcup_{i=1}^{n} A_i\right) &= C_n^1 \cdot \frac{1}{n} - C_n^2 \frac{1}{n(n-1)} + \cdots + (-1)^{n-1}\frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!}.\end{aligned}$$

推出

$$P(A) = 1 - P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{m=0}^{n} \frac{(-1)^m}{m!}.$$

当 $n = 2, 3, 4$ 时, 答案分别是 1/2,1/3,3/8, 与最初的数数吻合. 学过 Taylor 展开的同学知道, 当 n 趋于无穷时, 上述概率的极限是 e^{-1}, 其中 e 是欧拉数, 位于 2.7182 与 2.7183 之间的一个无理数.

练习 2.18　5 个球 n 个盒子, 每个球等可能地放入 n 个盒子. 就 $n = 2, 3, 4, 5$ 四种情况下求没有空盒的概率. 是否可以总结出 m 个球, n 个盒子的情况下怎么做?

2.4　几何概率

前面所说的古典概率模型是指仅有有限多个等可能发生的结果的随机试验, 还有一种直观的古典概率模型称为几何概率模型. 例如从 $[0, 1]$ 区间随机地取一个点, 记为 X, 那么 X 有无限的可能, 但是 X 落在 $[1/2, 2/3]$ 中的概率应该是此区间的长度 1/6. 这是因为当我们说在区间内随机地取一个点时, 也就意味着“均匀”或者“等可能”地取得每一个点. 可是点有无穷多, 取得具体每个点的可能性必须是零, 所以描述这种等可能的方法应该是“点落在区间 I 中的概率与区间的长度成比例”, 而落在整个区间上的概率是 1, 所以比例系数必须是整个区间长度的倒数.

设 Ω 是空间 (可以是 1 维, 2 维, 也可以是高维) 的一个有界区域, 从 Ω 中随机取一个点, 记为 X, 那么 X 落在 Ω 中的一个区域 A 的概率为

$$P(X \in A) = \frac{|A|}{|\Omega|},$$

其中 $|\cdot|$ 表示区域的体积, 注意在 3 维及以上的空间时, 称为体积, 在 2 维空间时, 称为面积, 在 1 维空间时, 称为长度. 在解决前面的古典概率问题时, 方法是数元素个数, 在解决几何概率问题时, 方法是计算长度, 面积, 或者体积, 这些概念和计算是中学数学知识. 从定义看, 几何概率模型类似于古典概率.

2.4.1　约会问题

问题： 两人约在 8:00-9:00 这个时间段在某咖啡店见面, 并约定先到之人最多等 15 分钟, 问他们能够见面的概率是多少?

首先要把问题化成在某个区域中随机地取一个点. 用 x, y 分别表示甲乙两人抵达的时间, 由于随机性, 两人抵达的时间相当于在平面区域

$$\Omega = \{(x, y) : 8 \leqslant x, y \leqslant 9\}$$

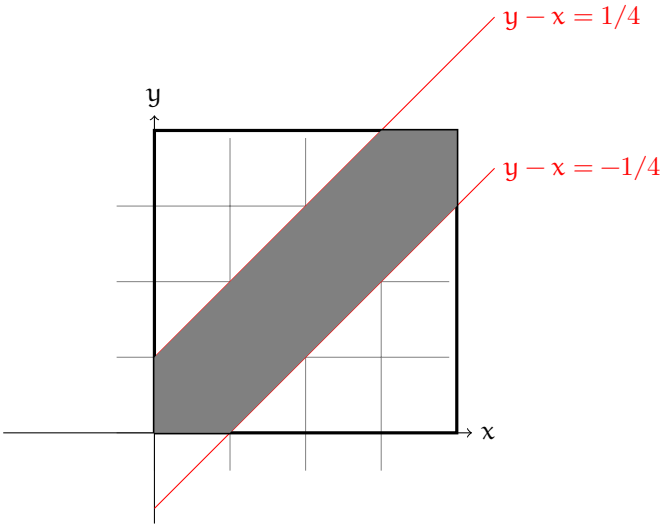
中随机地取一个点, 这等价于假设,

1. 他们两人在约定时间段的任何时间抵达的可能性都是一样的, 也就是等可能性;
2. 他们到达的事件是互相独立的.

而两人能见面这个事件相当于这个点落在区域

$$A = \{(x, y) \in \Omega : |x - y| \leqslant \frac{1}{4}\}$$

内. 它的概率就等于这个集合的面积与整个样本空间的面积的比, 即 7/16.

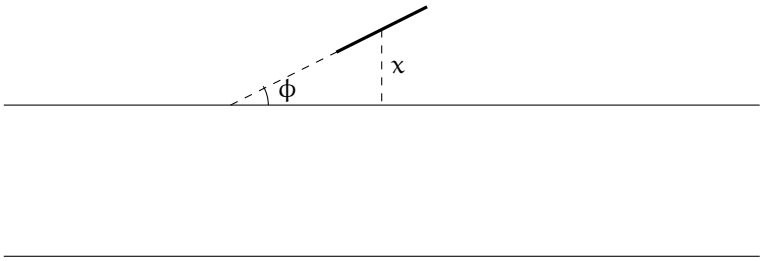


练习 2.19　一个人试图扔一个一元硬币 (直径 3cm) 到 2 米外的一个边长为 4cm 的正方形台面上, 现在他的硬币已经落在台面上, 问硬币完全在台面里的概率是多少?

2.4.2 Buffon 问题

下面的问题称为 Buffon 投针问题. Buffon 是生于 1707 年的法国数学家, 他在 1777 年提出下面的投针问题.

问题: 向一个画着等距离平行线的平面上投针, 平行线间的距离为 l, 针的长度为 a, $l > a$, 问此针与平行线相交的概率是多少?



蒲丰问题及其解答几乎可以在所有概率论教材中发现, 解答方法的思想就是下面所介绍的, 读者可参考王梓坤先生的概率论教材 [7] 的 §1.1 的例 8.

设针的中点与最近的平行线的距离是 x, 针 (或其延长线) 与平行线的夹角为 ϕ, 我们假设 (x, ϕ) 是在区域

$$[0, \frac{l}{2}] \times [0, \frac{\pi}{2}]$$

上随机取的一个点, 这等于假设

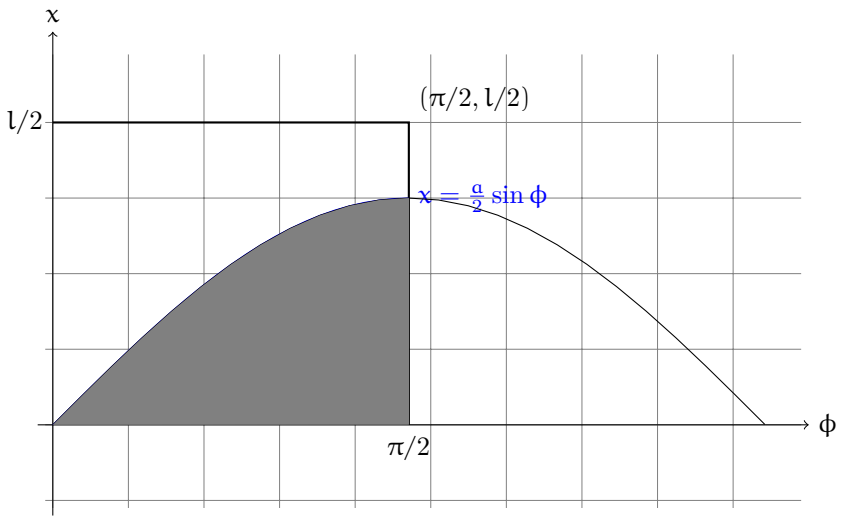
1. x, ϕ 在各自范围内是等可能的;
2. x 与 ϕ 是独立的.

事件 A: 针与平行线相交发生当且仅当 (x, ϕ) 落在区域

$$\{(x, \phi) : x \leqslant \frac{a}{2} \sin \phi\}.$$

因此概率是面积的比

$$\mathbb{P}(A) = \frac{\int_0^{\frac{\pi}{2}} \frac{a}{2} \sin \phi d\phi}{\frac{l}{2} \cdot \frac{\pi}{2}} = \frac{2a}{\pi l}.$$



Buffon 惊奇地发现这个概率是一个包含 π 的数, Monte Carlo 算法告诉我们通过投针可以估计 π 的值, 下面列出历史上记录的一些尝试.

实验者	时间	投掷次数	相交次数	π 的估计值
Wolf	1850	5000	2532	3.1596
Smith	1855	3204	1218.5	3.1554
DeMorgan	1860	600	382.5	3.137
Fox	1884	1030	489	3.1595

在解决实际问题中, 判断是否可以合理假设两个随机试验独立通常会容易一些, 而判断同一个随机试验中的两个事件或者两个随机变量是否独立需要通过定义完成.

掷两个骰子, '点数和为 7' 这个事件 A 与 '其中一个点数是 3' 这个事件 B 是不是独立呢? 那需要通过等式

$$\mathsf{P}(\mathsf{A}\cap\mathsf{B})=\mathsf{P}(\mathsf{A})\mathsf{P}(\mathsf{B})$$

是否成立来判断. 容易看出 $\mathsf{P}(\mathsf{A})=1/6$, $\mathsf{P}(\mathsf{B})=11/36$, $\mathsf{A}\cap\mathsf{B}=\{(3,4),(4,3)\}$, 所以 $\mathsf{P}(\mathsf{A}\cap\mathsf{B})=1/18$, 不满足等式, 因此 A 与 B 不独立.

换一个问题, 掷黑白两个骰子, '点数和为 7' 这个事件 A 与 '白色骰子点数是 3' 这个事件 B 是不是独立呢? 这时 $\mathsf{P}(\mathsf{A})=1/6$, $\mathsf{P}(\mathsf{B})=1/6$, $\mathsf{P}(\mathsf{A}\cap\mathsf{B})=1/36$, 所以等式成立, 故两个事件是独立的.

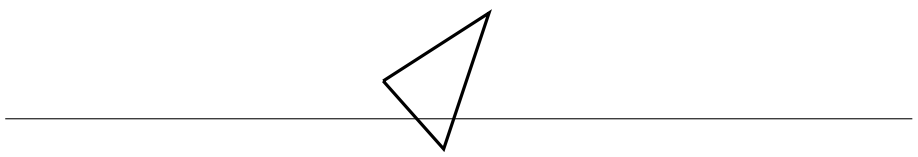
Buffon 问题的解答需要两个假设, 第一个假设说 x 与 ϕ 是等可能的, 这个也许没有什么争议, 可以认为是合理的. 但是第二个假设说角度 ϕ 和距离 x 是独立的, 这其

实不是那么容易让人信服. 一般地, 我们可以对两个看上去没有什么关系的随机试验假设独立, 但这里, 这两个随机量来自同一个随机试验的同一根针, 它们的独立性不应该通过假设获得, 而且该假设也不是那么理所当然. 直观上也许无法说它不对, 但也没足够理由说它对. 能不能避开独立性假设来解答这个问题呢? 答案是肯定的. 在讨论这个 Buffon 问题时, 我们自然会假设针与平行线相交的概率只与针的长度有关. 这个假设直观上容易接受. 设 A_x 是 '针上长为 x 的一段与平行线相交' 这个事件. 那么上面的假设可以推出等可能性假设: 针上的任意一段碰到平行线的概率只与该段的长度有关, 与位置无关, 即 $P(A_x)$ 只与 x 有关.

现在设 x, y 是针上不相交的两段, 则 $A_{x+y} = A_x \cup A_y$ 且 A_x 与 A_y 互斥, 因为这两段不可能都与平行线相交. 令 $f(x) = P(A_x)$. 利用概率的可加性推出: 对 $x, y > 0$, $x + y \leqslant a$, 有

$$f(x + y) = f(x) + f(y).$$

因此由实变函数的一个结论, 存在常数 k 使得对 $0 \leqslant x \leqslant a$ 有 $f(x) = kx$, 下面我们只需要计算 k.



进一步看一个三角形 abc 扔在平面上与平行线相交的概率, 其中 a, b, c 是三角形三条边及其长度. 用 A, B, C 分别表示 a, b, c 与平行线相交的这个事件, 那么 $A \cup B \cup C$ 就表示三角形与平行线相交这个事件. 由容斥定理,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$
\begin{aligned}
&- P(A \cap B) - P(B \cap C) - P(C \cap A) \\
&+ P(A \cap B \cap C),
\end{aligned}
$$

事实上, 一个简单的观察: 三角形与平行线相交必定是其中的两条边和平行线相交, 所以 $P(A \cap B \cap C) = 0$, 而

$$
\begin{aligned}
P(A) &= P(A \cap B) + P(A \cap C) \\
P(B) &= P(B \cap A) + P(B \cap C) \\
P(C) &= P(C \cap A) + P(C \cap B).
\end{aligned}
$$

由此推出三角形与平行线相交的概率为

$$
\begin{aligned}
P(A \cup B \cup C) &= \frac{1}{2}(P(A) + P(B) + P(C)) \\
&= \frac{1}{2}(f(a) + f(b) + f(c)) \\
&= \frac{1}{2}k(a + b + c),
\end{aligned}
$$

因此这个概率是周长的 $k/2$ 倍.

同样的方法证明, 任何一个直径不超过 l 的凸多边形随机扔在平面上与平行线相交的概率也是周长的 $k/2$ 倍. 而凸图形是凸多边形的极限, 由概率的连续性推出, 一个直径不超过 l 的凸图形与平行线相交的概率同样是其周长的 $k/2$. 注意一个图形的直径是指其中最远两个点的距离.

最后, 一个非常直观的事实是, 直径为 l 的圆扔在平面上以概率 1 与平行线相交, 即有

$$
l\pi k/2 = 1, \text{ 即 } k/2 = \frac{1}{l\pi}.
$$

因此, 长度为 a 的针与平行线相交的概率是 $\dfrac{2a}{l\pi}$. 答案与前面方法一致.

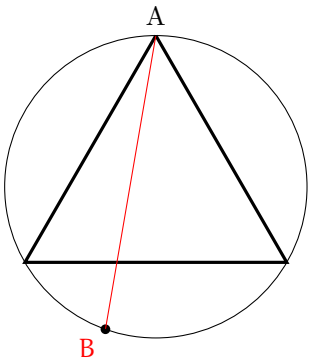
练习 2.20　向一个画着等距离平行线的平面上任意地投一个长为 2 的针, 平行线间的距离为 1, 求针与平行线相交的概率.

2.4.3　Bertrand 悖论

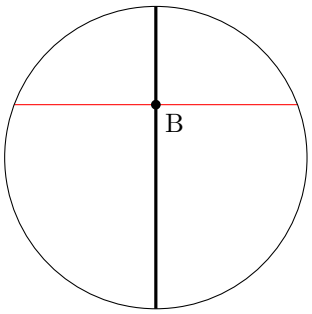
问题: 在一个圆周上随机地任取一根弦, 问其长度大于圆的内接等边三角形边长的概率是多少?

在这个问题里, 随机性至少有三种理解:

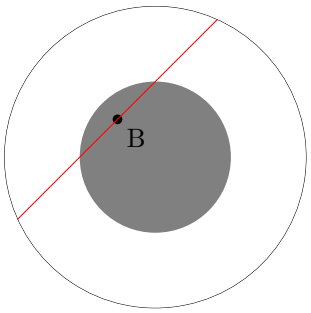
(1) 先在圆周上取定一点 A 然后再在圆周上随机地取一个点 B, 连接 A 与 B 成为弦;



(2) 先取定一条直径, 然后在直径上随机地取一个点 B 作一条过此点与直径垂直的弦;



(3) 以圆内的任何点作为中点的弦是唯一决定的, 因此以这个对应, 随机地取一条弦就等同于随机地在圆内取一个点 B.



说这三种取法的随机性不同是说如果按 (1) 随机地取弦 AB, 那么它与圆心的距离就

不可能是在半径上等可能的, 它的中点也不会等可能地出现在圆内. 关于此, 在后面会有进一步解释.

一条弦长大于圆的内接等边三角形边长当且仅当弦的中点与圆心的距离小于 1/2 的半径长, 那么在 (1) 的情况下, 点 B 必须落在点 A 对面的 1/3 圆弧上, 因此概率为 1/3, 在 (2) 的情况下, 点 B 必须与圆心距离不小于 1/2 半径长, 因此概率为 1/2, 在 (3) 的情况下, 点 B 必须落在半径为原来圆半径长的 1/2 的圆内, 因此概率为两圆面积的比 1/4.

Bertrand 在 1889 年出版了一本概率论讲义, 里面列出三个悖论, 这是其中之一, 称为 Bertrand 悖论.

悖论的悖的意思是矛盾. 数学是从公理出发且用逻辑进行推理, 原则上, 只要公理不矛盾, 数学体系就不应该出现矛盾. 确切地说, 悖论的英文是 paradox, 意思是貌似矛盾或者似非而是, 也就是说看上去有矛盾实际没有. 一个问题看上去是悖论, 是因为我们对问题没有理解清楚, 只要能够把问题理清楚, 矛盾就不存在了. 但后人通常还把它叫悖论, 是提醒大家这里有一个需要搞清楚的问题. 例如前面介绍的问题在刚提出的时候让人非常困惑, 被称为悖论, 但后来人们认识到造成困惑的原因是我们对“随机地取一根弦”这句话没有理解清楚.

练习 2.21　在圆周上固定一点 A, 在园内随机取一个点 B, (1) 问线段 AB 长大于圆半径的概率是多少? (2) 问 AB 延长线决定的弦长大于 $\sqrt{3}$ 的概率?

2.4.4　Bertrand 大圆悖论

上面的 Bertrand 悖论说明几何概率的等可能性是个很容易导致误解的问题, 下面又是这样的一个例子. 在单位球面上任取两点, 它们的球面距离不超过 $x \in (0, \pi/2)$ 的概率是多少. 一般的想法是这样的, 先取一个点 A 固定, 再在球面上任取一个点 B, B 在球面上均匀分布, 所求概率是以 A 点为圆心, x 为半径的球面圆的面积与球面面积的比值, 即

$$p = \frac{2\pi(1-\cos x)}{4\pi} = \frac{1-\cos x}{2}.$$

另外一个想法是这样的, 因为 B 在过 A 点的唯一大圆上, 所以 AB 距离小于 1 的概率是 x/π. 不一样, 它一定是错的.

问题: 问题出在哪里?

几何概率模型的样本空间通常是清楚的, 问题是当你假设等可能性的时候, 你可能会无意识地改变样本空间. 例如从单位圆内取一个点, 样本空间就是单位圆, 清清楚

楚, 明明白白, 但如果按照极坐标那样先等可能地取方位角, 再等可能地取与圆心的距离, 那就实际上改变了等可能性, Bertrand 这两个悖论实际上都犯了这样的错误. 另外, 几何概率模型的样本空间从元素个数上讲是无限的, 但是从测度 (重量, 长度, 体积等) 上来说是有限的且等可能的. 正因为如此, 所以几何概率模型也被认为是古典概率模型. 另外, 对于几何概率模型来说, 因为一个点的测度是零, 所以样本空间上每个结果发生的可能性都是零. 可能会有读者奇怪, 既然每个结果发生的可能性是零, 而必然事件是所有可能的结果, 那么问题来了, 必然事件的概率是所有结果的概率之和, 一大堆的 0 加起来怎么就变成了 1? 这个问题对很多学生来说是难以理解的, 让我们在下一章解释这个问题.

第三章　从有限到无限

本章的目的是解释概率的可列可加性及其数学意义和实际意义. 可列可加性的重要性相当于极限在微积分中的重要性.

3.1　无穷和简介

为了下一步的学习, 我们在数列极限的基础上补充一点数学知识, 简单介绍一下无穷和的概念. 一个无穷项数列的和称为无穷和或者级数:

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + \cdots + a_n + \cdots,$$

其中 a_n 称为通项. 对任何 n, $S_n = a_1 + \cdots + a_n$, 那么 S_n 被称为部分和序列. 无穷和等于部分和序列的极限

$$\sum_{n=1}^{\infty} a_n = \lim_n S_n.$$

极限存在称级数收敛, 否则发散. 解释一下和号 $\sum$, 这个符号是对一些数求和的意思, 下标指示对什么范围的数求和, $\sum_{n=1}^{\infty} = \sum_{1\leqslant n<\infty}$, 在上下文清楚的时候简写为 $\sum_n$, 或者 $\sum$. 通项非负的级数称为正项级数. 正项级数的部分和数列 S_n 递增, 所以最终只有两种可能: 或者有限或者无限. 有限时级数收敛, 否则发散. 因为 $a_n = S_n - S_{n-1}$, 所以级数收敛蕴含着通项趋于零.

注意级数求和可以从任意项开始, 这会影响级数的和, 但是对级数是否收敛没有影响. 一个熟悉的级数是等比数列求和

$$\sum_{n\geqslant 0} x^n = 1 + x + x^2 + \cdots + x^n + \cdots.$$

它的部分和

$$S_n = 1 + x + x^2 + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x},$$

显然当 $0 < x < 1$ 时, 级数收敛, 因此我们可以写

$$\sum_{n \geqslant 0} a x^n = \frac{a}{1 - x},\ x \in [0, 1);$$

当 $x \geqslant 1$ 时, 级数发散.

显然当级数收敛的时候, 通项的极限一定是零. 反过来, 通项的极限是零不能保证级数收敛, 例如调和级数 $\sum_{n=1}^{\infty} \frac{1}{n}$ 是发散的, 尽管通项 $1/n$ 的极限是零.

因为部分和

$$S_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$$

不能简单地用一个数列表示, 所以我们只能采用不等式来进行估算. 显然对任何 $n \geqslant 1$ 有

$$\frac{1}{n+1} + \cdots + \frac{1}{2n} > \frac{1}{2}.$$

因此

$$\begin{aligned} 1 + \frac{1}{2} + \cdots = 1 + \frac{1}{2} + (\frac{1}{3} + \frac{1}{4}) \\ + (\frac{1}{5} + \cdots + \frac{1}{8}) + (\frac{1}{9} + \cdots \frac{1}{16}) + \cdots \\ > \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots, \end{aligned}$$

右边是无限多个 $1/2$ 求和, 肯定是无穷大.

再考察级数

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \frac{1}{1 \cdot 2} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{n(n+1)} + \cdots,$$

它的部分和是

$$S_n = \sum_{k=1}^{n} \frac{1}{k(k+1)} = \sum_{k=1}^{n} \left(\frac{1}{n} - \frac{1}{n+1} \right) = 1 - \frac{1}{n+1},$$

其极限是 1.

级数 $\sum_{n=1}^{\infty} 1/n^2$ 可以和上面级数比较,

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \leqslant 1 + \sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 2.$$

练习 3.1　(比较判别法) 设 $0 \leqslant a_n \leqslant b_n$, 证明: $\sum_n b_n$ 收敛蕴含 $\sum_n a_n$ 收敛.

练习 3.2　设 α 是个实数. 证明级数 $\sum_{n\geqslant 1} n^{-\alpha}$ 当 $\alpha \leqslant 1$ 时发散, 当 $\alpha \geqslant 2$ 时收敛. 问当 $\alpha \in (1,2)$ 时会怎么样?

最后, 我们介绍三个极限公式. 第一个极限是

$$\lim_n \left(1+\frac{r}{n}\right)^n = e^r,$$

其中 e 是一个实数, 称为 Euler 数, 和 π 一起是数学里最重要的两个常数, 它们都是无理数. 左边的数列可以看成年利率 $r>0$ 分为 n 期的复合利率, 它是关于 n 递增且有界的数列, 因此极限存在, 当 $r=1$ 把极限记为 e. 由此可以证明

$$e = \sum_{n\geqslant 0} \frac{1}{n!}.$$

第二个极限是 Stirling 公式

$$\lim_n \frac{n!}{\sqrt{2\pi n}(n/e)^n} = 1,$$

这个公式用指数来估计 $n!$. 第三个极限是 Taylor 展开, 如果 $y = f(x)$ 是 0 点附近定义的无穷次可导函数, 那么在一定的条件下, 函数在 0 点附近有无穷级数的表示

$$f(x) = \sum_{n\geqslant 0} \frac{f^{(n)}(0)}{n!} x^n,$$

其中 $f^{(n)}$ 表示 f 的 n 阶导数. 特别地, 设 $f(x) = (1-x)^{-1/2}$, 则

$$f^{(n)}(0) = \frac{1}{2}\frac{3}{2}\cdots\frac{2n-1}{2} = \frac{(2n-1)!!}{2^n},\ n\geqslant 1,$$

因此

$$\frac{1}{\sqrt{1-x}} = \sum_{n\geqslant 0} \frac{(2n-1)!!}{n!2^n} x^n = \sum_{n\geqslant 0} C_{2n}^n (x/4)^n.$$

无穷级数是微积分的一大组成部分, 有很多内容, 对于我们的课程来说这点就够了.

3.2　概率的无穷和

人们对概率的可加性无疑是认同的, 但下面我们会看到, 要计算一些更有意思的概率, 需要有无限的可加性, 也称为可列可加性, 或者说无穷和, 即如果有一列互斥的事件, 那么他们至少一个发生的概率等于各自发生的概率之和. 先来看两个简单问题, 从中获得一点启发.

3.2.1　等待正面出现

在古典概率模型中, 因为样本空间有限, 所以不会生出无限. 掷一次硬币是古典概率模型, 掷有限次硬币也是古典概率模型, 但是往前再走一步, 重复不断地掷硬币, 就会出现无限多个结果, 这将不再是古典概率模型. 这时会有更多更有趣的问题. 让我们从一个简单的问题开始.

问题: 重复不断地掷一枚硬币, 是不是一定可以会得到正面?

问题似乎很简单, 大家都相信答案是肯定的. 根据我们的经验, 掷一枚硬币, 一般来说, 不需要几次, 就可以看到正面的, 如果打赌, 很少有人会赌 5 次都不是正面. 但要理性地回答这个问题, 并不能依靠经验. 认真地思考一会儿, 就会发现这里有几个问题需要澄清.

首先, 什么叫一定? 一定等同于肯定和必然, 但是这里的一定是指概率等于 1, 实际上是几乎肯定, 后面再详细解释. 概率小于 1 只能叫可能, 不管可能性多大或者多小.

其次, 按照经验, 是否能说 10 次或者 100 次或者 10000 次之内一定可以得到正面? 用 X 表示首次掷得正面时所掷的次数. 例如 $X = n$ 表示第 n 次掷得正面, 但前面都是反面. 显然

$$P(X = n) = \frac{1}{2^n}.$$

再例如 $X \leqslant n$ 表示前 n 次出现正面, 而相反地, $X > n$ 表示前 n 次掷得的都是反面, 没看到正面, 因此 $P(X > n) = 1/2^n$. 例如掷 10 次都不是正面的概率是 $1/2^{10} \sim 0.1\%$, 掷 100 次都不是正面的概率是 $1/2^{100}$, 这个可能性已经小到无法想象了, 在实际生活中, 它肯定被当作零看待了, 但在数学中, 它不是零. 因此, 对于上面的问题, 答案是不论 n 多大, 都不一定能在 n 次内掷得正面, 因为 $P(X \leqslant n) < 1$.

最后, '会得到' 是什么意思? 实际上, '会得到' 是一个口语, 真正的意思应该是 '有限次之内会得到'. 有限次没有也不能限定次数, 正整数每一个都是有限的, 但总数是无限的, 有限多次是指某个正整数, 但具体不确定是多少. 例如愚公移山的精神: 只要子子孙孙努力搬山, 总有一天会成功的. 这是指有限时间内会成功. 还有古语说: 一尺之棰日取其半万世不竭. 这是指有限多次是取不完的. 注意, 有限与无限有本质不同, 如非零与零有本质不同一样. 我们可以用归纳法证明有限多个有理数的和是有理数, 但是上一节的一个关于 e 的公式告诉我们无限个有理数的和可能是无理数.

有限次内看到正面是事件 $X < \infty$. X 是有限的, 即 X 等于某个正整数, 具体地说

$\mathsf{X}=1$, 或者 $\mathsf{X}=2$, 或者 $\mathsf{X}=3$, 一直下去, 没有尽头. 数学上

$$\{\mathsf{X}<\infty\}=\{\mathsf{X}=1\}\cup\{\mathsf{X}=2\}\cup\{\mathsf{X}=3\}\cup\cdots\cdots,$$

右边是无穷个互斥的事件至少有一个发生. 上面的可加性可能会自然地驱使很多读者在计算左边的概率时把右边各事件的概率加起来. 这样得

$$\begin{aligned}\mathsf{P}(\mathsf{X}<\infty)&=\mathsf{P}(\mathsf{X}=1)+\mathsf{P}(\mathsf{X}=2)+\mathsf{P}(\mathsf{X}=3)+\cdots\\&=\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\cdots=1,\end{aligned}$$

理论上证明了我们 100%(概率 1) 肯定或者说一定在有限次内可以看到正面出现, 反过来, $\mathsf{P}(\mathsf{X}=\infty)=0$, 也就是说有限次内看不到正面的概率是零.

因此, 无限的可加性对于概率是非常重要的, 如果没有, 那就永远无法计算 $\{\mathsf{X}<\infty\}$, “有限次出现正面”, 这个事件的概率. 可以说, 没有无限可加的概率, 就像没有极限的微积分.

3.2.2　三人比赛

问题: A, B, C 三人下棋, 规则如下: 两人下, 赢者与第三人下, 一直到其中一人连赢两局为胜. 三人的水平相当, A 与 B 先下. 问各人最终胜的概率是多少?

用 ACBACBB 表示这样的一次比赛: A 赢, C 赢, B 赢, A 赢, C 赢, B 连赢. 最终的结果是 B 胜, 共 7 局, 这样的结果出现的概率应该是 $\dfrac{1}{2^7}$. 那么样本空间 Ω 是这样的排列全体, 是一个无限集.

A 最终胜出这个事件包含样本空间里如下的序列:

(1) 首局 A 赢: AA, ACBAA, ACBACBAA, ACBACBACBAA, $\cdots$, 它们发生的概率依次为 $\dfrac{1}{2^2},\dfrac{1}{2^5},\dfrac{1}{2^8},\cdots$;

(2) 首局 A 输: BCAA, BCABCAA, BCABCABCAA, $\cdots$, 它们发生的概率依次为 $\dfrac{1}{2^4},\dfrac{1}{2^7},\dfrac{1}{2^{10}},\cdots$.

因此 A 最终胜出的概率为

$$\frac{\frac{1}{2^2}}{1-\frac{1}{2^3}}+\frac{\frac{1}{2^4}}{1-\frac{1}{2^3}}=\frac{5}{14}.$$

由对称性知 B 胜出的概率与 A 胜出的概率是一样的, 而类似的计算得 C 胜出的概率为 $\dfrac{4}{14}$, 三者概率和是 1.

这也回答了另外一个问题: 游戏是不是在有限时间内结束? 因为三个概率之和是 1, 所以比赛终有一个人会赢, 也就是说会在有限时间结束的概率是 1. 但游戏会无限进行下去并不是不可能事件, 例如 ABABABABAB....., 但概率为零.

练习 3.3　A,B 两人连续下棋, 每一局, A 胜的概率 p. 分别在下面两种规则下列出 A 最终赢得比赛的所有可能情况并计算概率.

1. 先连胜两局者赢;
2. 先净胜两局者赢.

上述两规则下的概率哪个更大?

3.3　概率论-数学的分支

随机现象是客观存在, 概率是其中可能性的度量, 可以说人们对什么是概率理解不完全统一. 但是概率论是数学理论, 建立在一个公理体系之上, 它毫无歧义, 但说概率就是概率论中的概率就太武断了, 可能也不是历史上研究概率论那些伟大学者的本意, 研究概率论和研究实际概率问题完全不同. 在这一节中, 我们来聊聊作为数学分支的概率论.

3.3.1　无限的可加性

这是个符合预期的结果, 但是从数学上说, 这是有问题的, 它超越了前面所说的可加性了, 可加性是说, 如果事件 A, B 互斥, 则 $\mathsf{P}(\mathsf{A}\cup\mathsf{B})=\mathsf{P}(\mathsf{A})+\mathsf{P}(\mathsf{B})$. 然后从中学生熟悉的数学归纳法推出对任何有限个互斥的事件 $\mathsf{A}_1,\cdots,\mathsf{A}_n$ 有

$$\mathsf{P}(\mathsf{A}_1\cup\cdots\cup\mathsf{A}_n)=\mathsf{P}(\mathsf{A}_1)+\cdots+\mathsf{P}(\mathsf{A}_n).$$

那么无限个互斥的事件呢? 直观上似乎是自然的, 但逻辑上不能从有限推到无限, 从有限到无限在数学上通常是一条鸿沟, 在认识论上也是一条鸿沟.

问题: 无限的可加性是否成立?

上面的例子告诉我们无限的可加性符合预期, 实际上, 对于这个特别的例子, 即独立地不断地重复一个随机试验, 可以证明 (非常困难的数学问题) 这样的可加性是成立的. 但也有例子说明无限的可加性不成立. 考虑 Bertrand 问题, 从单位圆上任取一

条弦, 按照第一种方式, 用 X 表示所取弦的长度, 那么 $P(0 \leqslant X \leqslant 2) = 1$, 但是对任何 $0, 2$ 之间的一个数, 使得弦长恰好是 a 的 B 最多只有两个点, 所以概率是 0. 显然

$$\{0 \leqslant X \leqslant 2\} = \bigcup_{a \in [0,2]} \{X = a\},$$

右边的无限个事件是互斥的且概率为零, 如果可加性可以推广到无限, 那么左边的概率是 1, 右边的概率是 0, 导致矛盾. 这表明这样的可加性不成立. 问题出在哪里呢? 问题出在无限上. 在大多数人看来, 有限有多少或者大小的分别, 而无限只有一种. 在早期, 数学家们也不能区分无限, 但到了 19 世纪, 数学家 Cantor 发现无限也是可以区分的. 怎么比较两个无限集合的多少呢? 采用一一对应的方式, 这是人类学会数数之前比较多少的方式, 学会数数之后, 这种方式就逐渐被遗忘了. 如果两个集合的元素之间可以建立一个一一对应, 那么就说它们一样多. 例如正整数有无限个, 但它们可列, 即可以排成一个数列. 这样可以排成一个数列的集合和正整数一样多, 称为是可列的. 可以证明 (不是很容易) 有理数 (集合) 可以排成数列, 是可数的, 而实数 (集合) 无法排成一个数列, 即不能和正整数建立一一对应. 自然地, 不能排成一个数列的集合称为不可列的.

现在, 可列可加性是指对任何可列个互斥的事件 $A_1, A_2, \cdots, A_n, \cdots$ 有

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

3.3.2　零概率事件

古典概率模型与几何概率模型看起来类似, 但某些方面差别很大. 古典概率模型只有有限个结果, 每个结果的概率都是正的. 几何概率模型有无限个结果, 每个结果的概率是零. 在我们谈论一个事件可能或不可能发生的时候, 零概率事件是比较特别的, 本节我们谈谈零概率事件及相关的小概率事件.

这个问题以及上面讨论的几个问题中出现一个新的让人疑惑的事件, 零概率事件, 应该不可能会发生但它又不是不可能事件. 那么零概率事件到底是什么东西?

让我们整理一下. 不可能事件, 对应的是空集. 不可能发生的事件发生的概率一定是零. 再来看' 可能' 这个词. 这个词不是个数学概念, 平常我们说可能发生的事件应该指概率大于零的事件, 至少在本讲义中是如此认定的. 在人们熟悉的古典概率模型中, 零概率事件只能是不可能事件, 两者是一致的, 因此除了不可能发生的事件就

是可能发生的事件. 但上面的例子告诉我们, 在一般情况下, '可能' 与 '不可能' 之间居然还有东西?

好在 '可能' 这个词是生活语言, 所以这听起来矛盾的事情只是语文问题, 不是数学问题. 零概率与概率 1 事件被赋予一个新的名称, 零概率事件也叫做几乎不可能事件, 对应地, 概率 1 的事件也叫做几乎必然事件. 现在我们有不可能事件, 几乎不可能事件以及 '可能' 发生的事件. 但如果学生思考这些事情, 很可能问一个概率课老师最害怕的问题.

问题: 几乎不可能发生的事件是不是可能发生?

19 世纪的数学家古诺是这么说的 (1843 年) It may be mathematically possible for a heavy cone to stand in equilibrium on its vertex, but it is physically impossible. 我通常会这么回答: 几乎不可能事件不是不可能的, 但几乎是不可能的. 这时, 我能想到学生的反应: 老师, 这绕口令等于什么也没说啊. 是啊, 在不可能发生与可能发生之间还有几乎不可能发生这样看得见摸不到的影子, 太不可理喻了. 但是我也没办法说更多了, 这时语言是无力的. 为什么难以回答呢? 作为思考的主体, 人生活在一个有限的世界里, 周围的一切尽管数不清, 但却是有限的, 而零概率的非不可能的事件与数学的诸多概念一样触及到无限 (或者无穷), 在生活实践中不可能产生, 所以, 数学中有很多与人直觉相悖的例子, 例如我们平常看到碰到学到的几乎都是有理数, 但在整个数轴上有理数集合的长度为零, 微不足道. 用概率论的语言说, 在区间上随便取个数, 取得有理数是几乎不可能的.

古诺那句话的意思是不可能事件应该称为数学不可能, 零概率事件应该是物理不可能. 如果换我说的话, 那么不可能事件也许应该称为绝对不可能事件. 前面我们说过, 样本空间上有不同的概率法则, 但是, 不管换什么概率法则, 不可能事件还是不可能的; 相对地, 几乎不可能事件应该称为相对不可能事件, 因为它的不可能是相对于某个概率法则而言的, 换个概率法则它也许就变得可能了. 举一个不太准确的例子, 由于种种不可预知的疾病与意外, 人的寿命是随机的, 但这概率法则一直在变化, '人到七十古来稀' 说明在古代一个人活到七十是几乎不可能的, 但现在, 很多人七十岁才退休. 为了方便, 我们通常把事件分为正概率事件与零概率事件, 或者可能的事件与几乎不可能事件. 不那么严格的话, 把几乎不可能也当作不可能事件.

3.3.3　公理化的概率论 (*)

问题: 什么是公理化? 为什么要公理化?

中国古代有许多惊才绝艳的学者, 研究并解决过很多有趣的数学问题, 但是为什么中国的数学没有走到古希腊的那一步, 原因之一是没有把数学的问题抽象化概念化. 所谓抽象, 就是找到问题的本质和普遍性. 数学没有抽象, 就如同有车而没有路. 有了抽象, 就可能循路找到起点, 那就是公理化.

在现实生活中, 公理是一个经常听到的词. 我们要说的公理是指大家都遵循的规则. 大到一个国家, 宪法是全体公民认可的基本规则, 小到平时玩游戏, 我们也先需要预定规则,

人类在进行推理的时候使用一样的逻辑, 正确的运用逻辑是科学推断的基础. 数学与科学相同的是都要讲道理. 讲道理就是把公认的基本事实与结论用逻辑的链条连起来. 在讲道理的时候, 公理是基础, 逻辑是语言. 科学的基本事实是自然, 科学家通过实验来读取自然的规则. 数学的基本事实叫做公理.

数学和其他科学-例如物理或化学-是有区别的. 什么是科学? 简单地说, 存在可行的方法来验证对错的学问, 称为证伪性. 同学们在中学阶段应该已经知道, 在物理中回答一个质疑的最好方法是实验, 而在数学中回答一个质疑的最好方法是证明. 物理学家可能对导致某个现象的解释会有争议, 但数学家几乎没有争议, 一个从公理出发的证明之对错是一个完全基于逻辑的过程. 科学是以自然为标准来判断真伪的, 而数学是以人为设立的公理为标准来判断真伪的. 例如, Euclid 的几何原本中制定了几何公理, [1] 它定义了点线面等基本几何元素, 然后不加证明地叙述了几个性质, 称为公理或者共设. 尽管公理是人创造的, 但不是所有人创造的公理都有意义, 所以我们也许应该说有意义的公理是自然借人的手写出来的.、要注意的是, 自然只有一个, 而公理可以有很多. 例如, 相矛盾的欧几里得几何和黎曼几何都是对的, 因为它们基于不同的公理.

在本节中, 我们将简单介绍概率论的公理. 公理在理论上很重要, 而且非数学专业的学生有机会完整地体验什么是公理化肯定是很有意义的.

概率来源于生活, 许多没有受过高等教育的人也可以直观地思考概率问题, 但是对于一个正在接受高等教育的学生来说, 只是直观思考是不够的, 因为每个人都有自己的直觉, 所以凭直觉是无法区分真假的. 要区分真假, 就要讲道理讲逻辑, 或者说我们要学习理性地思考, 为了这个目的, 概率论需要公理. 另外, 如同有坚实的地基才能建造高楼, 也只有有公理才能让理论走得更远.

所有我们学过的数学都是建立在公理体系上的, 但它们可能不是一开始就建立在公

[1] 注意读者所学习的教材上使用的术语不一定是公理.

理体系上的, 我们可能也不是按照公理体系来学习的, 而是按照人的认知能力学习的. 最开始学习数学的时候, 学习自然数及其加法. 什么是自然数? 这是非常抽象的概念, 但老师举起一个手指头说这是 1, 两个手指头是 2, 三个手指头是 3, 十个手指头用 10 表示了, 然后 1 个手指头再加两个手指头得到 3 个手指头, 所以

$$1+2=3.$$

从具体的手指头到抽象的数字, 我们慢慢地学会了自然数的表示和加法运算. 但等我们长大了, 学到更多, 理解更深刻, 那时候有一些同学会明白自然数实际上是个体系, 我们只需要了解其中的规则就可以了, 例如它对加法封闭, 加法满足交换律结合律. 至于 1,2,3 等等, 是数的一种表示, 我们完全可以用其他方式表示. 体系的本质是规则而不是表示, 这个规则其实就是数的公理体系. 这个公理被大家接受无疑是因为它实际上看上去是最自然的规则. 当然, 我很难用一个例子来说清楚公理化, 你也很难用一个例子理解公理化. 但碰到的例子多了, 思考得多了, 也许你慢慢地就理解了.

概率的公理化是描述事件的概率应该是什么. 设 Ω 是一个集合, 称为样本空间, 它通常表示随机试验的所有可能的结果. 注意, 在每次随机试验中, 有且仅有其中一个结果会出现.

从简单的例子中我们看到, 概率是一个随机事件 (简称为事件) 发生的可能性大小, 事件通常是某些结果组成的, 所以在数学上, 我们说事件是结果组成的集合, 或者说是样本空间的一个子集, 事件的全体用符号 $\mathcal{F}$ 表示, 数学上, 它需要满足以下条件:

1. $\emptyset, \Omega \in \mathcal{F}$;
2. 若 $A \in \mathcal{F}$, 则 $A^c \in \mathcal{F}$;
3. 若 $A_1, \cdots, A_n, \cdots \in \mathcal{F}$, 则 $\bigcup_n A_n \in \mathcal{F}$.

什么是概率呢? 概率是事件发生的可能性大小, 是一个数值, 通常说是百分之多少, 所以是 0, 1 之间的数. 但概率不是随便什么数都可以, 而是一个满足可加性的体系.

定义 3.3.1 用数学的语言说, 对于事件 $A \in \mathcal{F}$, $P(A)$ 是 A 发生的概率, 它需满足下列条件:

P1. 概率 $P(A)$ 在 0,1 之间且 $P(\Omega) = 1$;

P2. 可列可加性成立.

或者说, 满足上面条件的 P 称为是一个概率测度.

可以证明, 前面所有讨论的概率问题都是一个满足定义条件的概率框架中的问题, 但是要证明这一点, 有时容易, 有时不容易, 我们在这里不详细展开. 公理是理论的基础, 所有的定理结论要从公理出发用严格逻辑证明, 即从公理出发重建直觉. 例如, 从公理出发可以推出下面的结论,

1. $P(\emptyset) = 0$;
2. 可加性成立;
3. 如果 A 发生蕴含 B 发生, 那么 $P(A) \leqslant P(B)$.

练习 3.4 从定义出发证明 1,2.

公理来自于人们对于概率的直观认识, 对于概率的直观认识将会是本课程的主要内容. 公理体系建立之后, 根据逻辑会展开成为一个理论体系, 根据 Euclid 的几何公理展开的理论体系是欧几里得几何, 根据概率公理所展开的理论体系就是概率论, 这正是大学概率论课程要学习的内容.

公理的目的是使得该理论最大程度地与我们直观中的概率论吻合, 但是公理一旦确立之后, 从公理出发得到的结论才是正确的. 直觉依然重要, 但是在发生矛盾的时候, 代表感性的直觉需让位于代表理性的公理. 因此在数学领域, 理性是真理的最终守门员.

最后, 有的读者可能会问, 上面的公理是概率论的唯一选择吗? 这是个好问题, 是个哲学问题, 答案是很可能是. 到现在为止, 世界上肯定还有人不接受上面的公理作为概率论基础, 但是绝大部分或者说几乎所有数学家都愉快地接受了.

练习 3.5 从定义出发证明在一个概率空间中不可能有无穷多个互斥的正概率的等可能事件.

可列可加性是对于互斥的事件列 (A_n) 有 $P(\bigcup_n A_n) = \sum_n P(A_n)$. 令 $B_n = \bigcup_{k=1}^n A_k$. 显然 $\lim_n B_n = \bigcup_{n=1}^\infty A_n = \bigcup_{n=1}^\infty B_n$. 下面的概率连续性在下一章将会用到.

练习 3.6 证明: 概率连续性对 (B_n) 成立, 即 $\lim_n P(B_n) = P(\lim_n B_n)$, 如果 X 是正面出现时间, 则 $P(X < \infty) = \lim_n P(X \leqslant n)$.

为什么叫连续性? 对于一个函数 f 来说, 如果极限和它可交换 $f(\lim x_n) = \lim f(x_n)$, 我们就说 f 连续. 上式类似于此. 实际上, 任何递增事件列都可以表示成互斥事件列的并, 因此概率连续性对任何递增事件列成立.

练习 3.7　证明: 概率连续性对递增事件列成立, 也对递减事件列成立.

这节的几个习题很数学化, 对于锻炼自己的数学能力很有帮助.

3.4　独立与重复

独立性在概率论计算中是最重要的两个性质之一, 其实在前面的计算中, 我们实际上已经在使用独立性了. 例如分赌注问题中假设每局赌博是独立的, 在约会问题中假设两人到达时间是独立的, 在蒲丰问题中假设针到平行线的距离和角度是独立的, 等等. 其中有些独立性假设是自然的, 有些似乎不那么自然, 但无论如何, 我们不能这样肆无忌惮地使用独立性, 至少我们需要知道什么是独立性, 什么情况下我们可以使用独立性, 这些使用是不是合理. 否则的话, 很可能会导致错误. 在公理体系中, 独立性首先是对两个事件定义的, 仿照古典的独立随机试验.

定义 3.4.1　在概率空间 $(\Omega, \mathcal{F}, \mathsf{P})$ 中, 两个事件 A, B 独立, 是指下式成立

$$\mathsf{P}(\mathsf{A} \cap \mathsf{B}) = \mathsf{P}(\mathsf{A})\mathsf{P}(\mathsf{B}).$$

事件 A, B 独立等价于 $\mathsf{P}(\mathsf{A}|\mathsf{B}) = \mathsf{P}(\mathsf{A})$. 两个随机试验的独立通常是合理假设, 不可以证明, 就如同假设一个硬币是圆形一样. 例如我们直观地知道, 掷一个硬币, 再掷一个硬币, 因为它们不会互相影响, 所以它们可以被合理地认为是独立的, 因此我们假设它们是独立的; 即使是同时掷多个硬币, 也可以假设是独立的, 因为相互的影响可以忽略; 摸一个球, 放回去, 再摸一个球, 可以说是独立的; 摸一个球, 不放回, 再摸一个球, 应该不独立, 因为互相有影响, 第一次摸到的球不会被第二次摸到了.

在一个不可分割的随机试验中的事件是否独立是很难直观判断的, 例如从一副牌 (52 张, 没有大小王) 中摸一张牌, A 是摸得红桃, B 是摸得 10, 它们是否独立? 这个需要用定义来判断. $\mathsf{A} \cap \mathsf{B}$ 是摸得红桃 10, 所以

$$\mathsf{P}(\mathsf{A} \cap \mathsf{B}) = \frac{1}{52} = \frac{1}{4} \cdot \frac{1}{13} = \mathsf{P}(\mathsf{A})\mathsf{P}(\mathsf{B}),$$

因此这两个事件是独立的. 同样, 独立随机试验各自的随机变量是独立的. 例如甲乙两人各掷 $\mathfrak{n}$ 枚硬币, 他们得到的正面个数是互相独立的. 但同一个随机试验中的随机变量之间的独立性并不直观, 回忆前面的 Buffon 问题, 假设针的角度与针中点到平行线的距离之间独立也不是非常自然.

上面我们理解了两个事件的独立性意义. 可能我们觉得三个或者更多事件的独立性是类似的.

问题: 设有三个事件 A, B, C, 说它们独立是什么意义?

从语言上理解, A, B, C 独立应该是其中任何一个发生与否与其他两个是否发生是独立的. 什么叫其他两个是否发生呢? 严格地说, A 与 B, C 是否发生独立应该是指 A 与 $B \cap C$, $B^c \cap C$, $B \cap C^c$, $B^c \cap C^c$ 中的任何一个独立. 类似地可以表达其他两种情况. 这看起来很复杂, 但是我们可以证明这等价于下面的叙述: A, B, C 两两独立且

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

因此我们说 A,B,C 独立是指它们满足上面两个条件. 这里有个问题是 A,B,C 两两独立是否可以推出它们三个是独立的. 这听起来似乎是对的, 但实际上严格来说是不对的, 下例为证.

现在有 1,2,3,4 四个人, 1,2,3 各说一种语言, 分别是中英法, 而 4 会说所有三种语言. 任选一个人, 用 A, B, C 分别表示选出的人分别会说中英法三国语言这三个事件. 那么 $A = \{1, 4\}$, $B = \{2, 4\}$, $C = \{3, 4\}$, 故而 $P(A) = P(B) = P(C) = 1/2$. 显然, 会其中两国语言或者三国语言的都只能是选到 4, 即

$$A \cap B = B \cap C = C \cap A = A \cap B \cap C = \{4\},$$

推出

$$P(A \cap B) = P(B \cap C) = P(C \cap A) = 1/4,$$
$$P(A \cap B \cap C) = 1/4.$$

因此看出, A, B, C 是两两独立但不是独立的. 直观地看, A, B, C 中任何两个发生导致剩下一个一定发生, 这就失去了独立性.

这就是逻辑战胜直觉的一个例子, 它提醒我们要小心数学概念之间的微小区别, 要遵循定义, 不能想当然.

练习 3.8　设有事件 A, B, C, A 与 C 独立且 B 与 C 独立. 问 $A \cap B$ 与 C 独立吗?

练习 3.9　设某个人每次投球命中的概率是 p, 且两次投球命中是独立的. 求投两次至少命中一次的概率.

练习 3.10　A,B 两人下棋比赛, 每一局, A 胜的概率 $p < 1/2$. 对于 A 来说, 一局定胜负与三局两胜这两种规则哪个获胜概率更大?

练习 3.11　父亲为了鼓励儿子打网球, 宣称如果他能赢得三场与父亲和教练的比赛中连续的两场, 他将获得一笔奖金. 他可以选择比赛的顺序为

1. 父亲 – 教练 – 父亲;

2. 教练 – 父亲 – 教练.

教练比父亲打得好. 问为了增加获得奖金的机会, 他应该选择哪个顺序?

3.4.1　等待成功

另外一个有意思的随机变量是等待成功的时间, 用 T 表示首次成功的时间, 则 $\mathsf{X}=\mathsf{T}$ 表示前 $\mathrm{k}-1$ 次试验是失败, 而第 k 次试验是成功. 由独立性

$$\mathsf{P}(\mathsf{T}=\mathrm{k})=\mathrm{q}^{\mathrm{k}-1}\mathrm{p},\ \mathrm{k}=1,2,3,\cdots.$$

有限次之内成功的概率等于 1, 实际上,

$$\mathsf{P}(\mathsf{T}<\infty)=\sum_{\mathrm{k}=1}^{\infty}\mathrm{q}^{\mathrm{k}-1}\mathrm{p}=1.$$

这个结果看上去似乎没什么惊人的, 但如果换一个说, 就有点不同了. 我们知道世界上的所有信息可以编成一个 01 的有限序列, 小到人名, 大到国家图书馆的所有图书可以编成一个长度为 r 的 01 序列 A.

问题: 你能通过掷硬币掷出国家图书馆吗?

不仅可能, 而且是肯定的. 掷 r 次硬币得到的序列恰好是 A 的可能性很低, 但是一个正数

$$\mathrm{p}=\mathsf{P}(\mathsf{A})=\frac{1}{2^{\mathrm{r}}}.$$

我们把掷 r 次硬币作为一个随机试验, 如果正好是 A 就称为成功, 然后我们重复这个随机试验, 那么上面的结论是说 A 肯定会成功地出现的, 也就是说, 你可以通过不懈地掷硬币掷出你的名字, 掷出国家图书馆. 惊人吗?

3.4.2　无穷多次成功

上面说重复掷硬币肯定可以掷出国家图书馆, 下面我们说实际上肯定可以掷出无穷多次. 先让我们学会表达“无穷多次”. 设

$$\mathsf{A}_1,\mathsf{A}_2,\cdots,\mathsf{A}_{\mathrm{n}},\cdots$$

是一个事件列. 则至少有一个发生是 $\bigcup_{n=1}^{\infty} A_n$, 同时发生是 $\bigcap_{n=1}^{\infty} A_n$. 怎么表示其中有无限多个发生以及最多只有有限多个发生这样的事件呢? 用 A 表示事件 '(A_n) 中有无限多个发生'. 那么对任何 k, $A_k, A_{k+1}, \cdots$ 必然有一个发生, 即 $\bigcap_{n=k}^{\infty} A_n$ 发生, 因此

$$A = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n.$$

用 B 表示事件 '(A_n) 中仅有有限个发生'. 那么

$$B = A^c = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c.$$

由此可以推出事件 '(A_n) 中最多仅有有限多个不发生' (也称为 '几乎都发生') 这个事件是

$$\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n.$$

下面的性质称为次可加性.

定理 3.4.1　设 A_n 是事件列. 那么

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

因此一列零概率事件 '至少有一个发生' 还是零概率事件.

设 $A_0 = \emptyset$, 那么

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} (A_n \setminus (A_0 \cup \cdots \cup A_{n-1})),$$

即事件的并总是可以写成为右边的不交并, 这在有限个事件时是对的, 对一列事件也是对的. 那么由可加性推出

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n \geqslant 1} P(A_n \setminus (A_0 \cup \cdots \cup A_{n-1})) \leqslant \sum_{n \geqslant 1} P(A_n).$$

独立地重复一个成功概率为 $p \in (0,1)$ 的 Bernoulli 试验. 用 A_n 表示事件 '第 n 试验是成功', 那么 (A_n) 是一个事件列. 前面证明了, 至少有一次成功的概率是 1. 还有一个问题.

问题: 成功是不是会无限多次地发生?

现在计算事件 $A =$‘(A_n) 仅发生有限多次’ 的概率. 由次可加性得

$$P(A) = P\left(\bigcup_{k\geqslant 1}\bigcap_{n\geqslant k} A_n^c\right) \leqslant \sum_{k\geqslant 1} P\left(\bigcap_{n\geqslant k} A_n^c\right) = 0,$$

其中因为 (A_n) 独立且 $P(A_n^c) = 1-p$, 所以 $P\left(\bigcap_{n\geqslant k} A_n^c\right)$ 是无穷个 $1-p$ 的乘积, 显然是零. 因此, 几乎必然地有无限多次成功会发生.

3.4.3　输了还会赢回吗

现在我们考虑随机游动. 一个醉汉站在街道上漫无目标地来回走动俗称为随机游动. 数学上说, 随机游动如下描述, 一个粒子在数轴的整数点上移动, 从某个点出发, 粒子以相同的概率在每个单位时间向左或者向右移动一个位置. 把时间与位置分别作为横轴和数轴在平面上画出粒子的运动, 就是一条格点轨道. 格点轨道可以看成随机游动, 随机游动也可以看成格点轨道. 可以这么说, 时长为 n 的任何一条格点轨道出现的概率都是 2^{-n},

回到格点轨道那个问题, 我们可以再问一个问题.

问题: 从零点出发的格点轨道是否一定返回零?

如果把这个过程看成是一个赌博, 那么这也就是问输了还能赢回来吗? 或者赢了还能输回去吗? 用 T 表示格点轨道首次返回零的时间, 称为首次返回时. 显然 $\{T<\infty\}$ 是 ‘格点轨道会回到 0 点’ 这个事件, $\{T=\infty\}$ 是 ‘格点轨道不再返回 0 点’ 这个事件. 首先, 我们可以画出无穷多条不碰到 0 点的格点轨道, 也就是说 $\{T=\infty\}$ 包含有无穷多个元素样本点. 也就是说, 格点轨道不是必然返回零. 那么能不能计算它返回零的概率呢?

应用反射原理 来计算 $P(T=2n)$.

首先从 $(0,0)$ 出发到 $(2n,0)$ 途中一直在横轴之上的格点轨道数等于 $(1,1)$ 到 $(2n-1,1)$ 不碰到横轴的格点轨道数, 这等于格点轨道的总数减去碰到横轴的轨道数, 前者等于 C_{2n-2}^{n-1}, 而由反射原理, 后者等于 $(1,1)$ 到 $(2n-1,-1)$ 的格点轨道总数: C_{2n-2}^{n}. 因此

$$P(T=2n) = \frac{2(C_{2n-2}^{n-1} - C_{2n-2}^{n})}{2^{2n}} = \frac{C_{2n}^{n}}{(2n-1)2^{2n}}.$$

那么

$$P(T<\infty) = \sum_{n\geqslant 1} \frac{C_{2n}^{n}}{(2n-1)2^{2n}}.$$

算这个级数不是那么容易, 需要一点技巧. 记 $u_n = C_{2n}^n \cdot 2^{-2n}$, 它是格点轨道从 $(0,0)$ 到 $(2n,0)$ 的概率. 利用等式

$$\frac{1}{2n-1} = \frac{2n}{2n-1} - 1 = \frac{2n \cdot 2n}{2n(2n-1)} - 1,$$

很容易计算出

$$P(T < \infty) = \sum_{n=1}^{\infty} (u_{n-1} - u_n) = u_0 = 1.$$

因此格点轨道以概率 1 返回 0, 或者说不返回 0 点的概率是零.

但上面这个方法太巧合, 或者说太诡异, 不能推广. 下面我们介绍一个更平凡的方法考虑更一般的情况.

应用马氏性

设甲持续地玩一个游戏, 每一局成功时得一块钱, 概率为 p, 失败时失一块钱, 概率是 $q = 1-p$. 用 X_n 表示第 n 局游戏甲的所得, 那么

$$P(X_n = 1) = p, P(X_n = -1) = 1 - p.$$

且从开始到第 n 次结束, 甲赢的的财富为

$$S_0 = 0,\ S_n = X_1 + \cdots + X_n,$$

其中 $X_1, \cdots, X_n, \cdots$ 是独立同分布的. (S_n) 被称为简单随机游动, 当 $p = 1/2$ 时称为对称的.

定义首次返回零的时间

$$T = \inf\{n > 0 : S_n = 0\}, \tag{3.4.1}$$

称为首次回归时. 上面的问题等价于计算概率 $P(T < \infty)$. 注意 $P(T < \infty) = \sum_{n=1}^{\infty} P(T = 2n)$.

令

$$f_n = P(T = n),$$

因为 T 的取值一定是偶数, 所以该数列的奇数项为零. 而

$$P(T < \infty) = \sum_{n=1}^{\infty} P(T = n).$$

令 $u_n = P(S_n = 0)$, $n \geqslant 1$. 显然 $S_n = 0$ 只有当 n 是偶数时才有可能, 而事件 $S_{2n} = 0$ 相当于说, 在 $2n$ 局游戏中, 成功次数与失败次数一样多, 所以

$$u_{2n} = P(S_{2n} = 0) = \binom{2n}{n} p^n q^n.$$

现在我们来推导概率 $P(T = 2n)$. 如果 $S_{2n} = 0$, 则它一定在 $2n$ 前的某个时间 $2k$ 处首次回归零, 然后从 $2k$ 时的零出发, 到 $2n$ 时再回到零, 即 $S_{2n} - S_{2k} = 0$. 不难看出, 这两个事件是独立的, 并且 $S_{2n} - S_{2k}$ 与 S_{2n-2k} 分布相同. 由此推出对 $n \geqslant 1$ 有

$$\begin{aligned} u_{2n} =& P(S_{2n} = 0) = \sum_{k=1}^{n} P(T = 2k, S_{2n} - S_{2k} = 0) \\ &= \sum_{k=1}^{n} P(T = 2k) P(S_{2n} - S_{2k} = 0) \\ &= \sum_{k=1}^{n} f_{2k} u_{2n-2k} \\ &= f_2 u_{2n-2} + f_4 u_{2n-4} + \cdots + f_{2n-2} u_2 + f_{2n}. \end{aligned}$$

原则上, 当 $n = 1$ 时, $f_2 = u_2$. 当 $n = 2$ 时, 从 $u_4 = f_2 u_2 + f_4$ 解出 $f_4 = u_4 - (u_2)^2$, 然后原则上可以依次解出 f_6, f_8 以及所有的 f_{2n}. 但实际上很难, 问题是即使写出 f_{2n} 表达式, 还要算级数和, 也很难. 怎么办呢? 所幸我们的目标是计算级数和

$$\sum_{n=1}^{\infty} f_{2n},$$

而不是数列 f_{2n}.

让我们把所有的方程列出来, u 的下标从小到大, f 的下标从大到小.

$$\begin{aligned} u_2 &= f_2; \\ u_4 &= f_4 + u_2 f_2; \\ u_6 &= f_6 + u_2 f_4 + u_4 f_2; \\ u_8 &= f_8 + u_2 f_6 + u_4 f_4 + u_6 f_2; \\ u_{10} &= f_{10} + u_2 f_8 + u_4 f_6 + u_6 f_4 + u_8 f_2; \\ &\cdots\cdots \end{aligned}$$

全部加起来, 右边每一列相加, 看出

$$\begin{aligned}\sum_{n=1}^{\infty} u_{2n} &= \sum_{n=1}^{\infty} f_{2n} + u_2 \sum_{n=1}^{\infty} f_{2n} + u_4 \sum_{n=1}^{\infty} f_{2n} + \cdots \\ &= (1 + u_2 + u_4 + \cdots) \sum_{n=1}^{\infty} f_{2n}.\end{aligned}$$

由此推出

$$P(T < \infty) = \sum_{n=1}^{\infty} f_{2n} = \frac{\sum_{n\geqslant 1} u_{2n}}{1 + \sum_{n\geqslant 1} u_{2n}}.$$

这说明什么呢? 当级数 $\sum_{n\geqslant 1} u_{2n}$ 收敛时, $P(T < \infty) < 1$; 否则 $P(T < \infty) = 1$. 因此问题归结为判断级数 $\sum_{n\geqslant 1} u_{2n}$ 是否收敛.

回答这个问题不难, 利用 Stirling 公式来估计 u_{2n},

$$u_{2n} = \frac{(2n)!}{(n!)^2}(pq)^n \sim \frac{\sqrt{2\pi 2n}(2ne^{-1})^{2n}}{(\sqrt{2\pi n}(ne^{-1})^n)^2} \cdot (pq)^n = \frac{(4pq)^n}{\sqrt{\pi n}}.$$

因此级数 $\sum_{n\geqslant 1} u_{2n} = +\infty$ 当且仅当 $p = 1/2$. 这回答了一开始的问题, 在对称的情况下, 输的人一定会赢回来

估计的方法无法在不对称时计算 $P(T < \infty)$ 的值. 这时, 利用 Taylor 展开,

$$\sum_{n\geqslant 1} u_{2n} = \sum_{n\geqslant 1} C_{2n}^n (pq)^n = \frac{1}{\sqrt{1-4pq}} - 1.$$

因此

$$P(T < \infty) = \frac{\frac{1}{\sqrt{1-4pq}} - 1}{\frac{1}{\sqrt{1-4pq}}} = 1 - \sqrt{1-4pq} = 1 - |p-q|.$$

对于一开始的问题, 答案是, 在对称时, 只要游戏继续, 输了一定会赢回来, 赢了也一定会输回去. 在不对称时, 就不一定了. 如果 $p > 1/2$, 也就是一个游戏者有优势, 那么他输了应该可以赢回来, 赢了不一定会再输掉.

第四章　概率初步 (续)

平均寿命是人类文明的一个重要指标. 在谈论寿命的时候, 我们可以谈论一个国家全体人民的平均寿命, 也可以谈论一个省或地区的平均寿命. 同样, 我们可以讨论整体的概率, 也可以谈论某些条件下的概率, 也就是条件概率.

4.1　条件概率与全概率公式

条件概率是一个非常重要的概念, 直观但难以叙述. 什么是条件概率呢? 在一个给定的概率模型中, 假设某个事件发生, 将得到一个新的概率模型. 或者说在给定的条件下, 概率模型的随机性会发生改变. 例如掷两个骰子是一个概率模型, 如果假设两个骰子的和是 4 实际上就是说已知这个事件发生了, 那么这是一个新的概率模型, 样本空间改变了, 是 $\{(1,3),(2,2),(3,1)\}$. 如果我们问至少有一个 1 的概率, 那么在第一个模型下, 答案是 11/36, 在新的模型下, 答案明显变大, 是 2/3. 这样假设某个事件发生所在的模型下得到的概率即条件概率. 条件概率在现实世界中也有体现, 期权或者期货这样的随机商品的价值会随着时间推移而逐渐清晰, 这实际上就是条件在改变, 使得随机性在变化, 概率在变化. 下面的定义适用于一般情形.

定义 4.1.1　设有两个事件 A,B, 且 $\mathsf{P}(\mathsf{A})>0$, 在 A 发生的条件下考虑 B 发生的概率称为条件概率, 记为 $\mathsf{P}(\mathsf{B}|\mathsf{A})$.

一个概率问题是否是一个条件概率问题依赖于我们怎么建立概率空间. 如果把其中的某些条件看成是一个更大概率空间中的一个事件, 那么它是一个条件概率问题, 而如果直接按照条件建立概率空间, 那它就是一个普通概率问题.

条件概率与原概率有什么关系? 这要从分布的原意说起, 分布粗略地理解为是将某个总量 (不一定是 1) 分布到若干点上. 因为我们只是关心分布的相对性, 所以各点所分布的量除以一个共同常数不改变分布态势. 特别地, 各点分布的量除以总量后

得到的分布是概率分布. 在考虑条件概率时, 事件 A 发生后, 样本空间改变了, Ω 上的分布留在 A 上的部分成为 A 上的分布. 如果原样本空间是等可能的, 新样本空间 A 也是等可能的, 这时事件 B 的条件概率是

$$\frac{|A \cap B|}{|A|} = \frac{P(A \cap B)}{P(A)}.$$

一般地, 条件概率与原概率有以下的关系

$$P(B|A) = \frac{P(B \cap A)}{P(A)},$$

称为条件概率公式. 概率 $P(B)$ 与条件概率 $P(B|A)$ 一般不一样, 当且仅当两者独立时一样.

设一个袋子里有两个球, 一白一黑, 甲摸一个球, 不放回, 乙再摸剩下的球, A, B 分别表示甲乙摸到白球. 显然

$$P(A) = P(B) = 1/2.$$

如果考虑条件概率, 那么甲若摸走白球, 乙就不可能摸到白球, 所以 $P(B|A) = 0$; 甲若摸走黑球, 乙就肯定摸到白球, 所以 $P(B|A^c) = 1$. 注意区别概率与条件概率, 事件的概率是不变的, 条件改变导致的概率变化是条件概率.

从条件概率公式马上得到乘法公式: 事件同时发生的概率表示为条件概率的乘积,

$$\begin{aligned}
&P(A_2 \cap A_1) = P(A_2|A_1)P(A_1);\\
&P(A_3 \cap A_2 \cap A_1) = P(A_3|A_2 \cap A_1)P(A_2|A_1)P(A_1);\\
&\cdots\cdots .
\end{aligned}$$

设随机试验的结果可以分成 n 种情况, 即设样本空间 Ω 可分成 n 个两两不同时发生 (两两不相交) 的事件 $\Omega_1, \cdots, \Omega_n$:

$$\Omega = \Omega_1 \cup \cdots \cup \Omega_n.$$

做一个随机试验, 必定是这 n 种情况之一发生. 一个事件发生自然也可以看成是在不同情况下分别发生, 即

$$A = (A \cap \Omega_1) \cup \cdots \cup (A \cap \Omega_n).$$

因此, 由概率的可加性和乘法公式得,

$$P(A) = P(A \cap \Omega_1) + \cdots + P(A \cap \Omega_n)$$

$$= P(A|\Omega_1)P(\Omega_1) + \cdots + P(A|\Omega_n)P(\Omega_n).$$

因此有下面的

全概率公式：

$$P(A) = \sum_{k=1}^{n} P(A|\Omega_k)P(\Omega_k).$$

因为

$$\sum_{k=1}^{n} P(\Omega_k) = 1,$$

所以全概率公式实际上是说概率 $P(A)$ 是条件概率 $P(A|\Omega_k)$, $1 \leqslant k \leqslant n$, 的加权平均, 而条件概率 $P(A|\Omega_k)$ 的权重为 $P(\Omega_k)$, $k = 1, \cdots, n$, 简单直白地说, 是不同情况下事件发生的条件概率的加权平均.

在古典概率之下, 和独立性一样, 引入条件概率不是必须的, 但它可以让概率计算更加直观. 例如配对问题中的概率, 那里我们是直接使用排列组合来计算概率, 但也可以用条件概率. A_i 表示事件“第 i 对夫妻恰好配对”. 条件概率 $P(A_2|A_1)$ 是在第一对夫妻配对成功时第二对也配对了的概率. 实际上, 第一对配对后, 问题就转化成剩下的 $n-1$ 对的配对问题, 这时, 第二对配对的概率是 $1/(n-1)$. 因此

$$P(A_1 \cap A_2) = P(A_2|A_1)P(A_1) = \frac{1}{n(n-1)}.$$

练习 4.1　设一个随机试验 E 有两个事件 $A \subset B$, 重复此随机试验一直到 B 出现为止, 问这时 A 出现的概率多大?

练习 4.2　有标号为 $1, 2, \cdots, m$ 的 m 个卡片, 随机一张一张不放回抽取, 已知第 k 张是前 k 张中最大的, 求它是 m 的概率.

4.1.1　抽签与顺序无关

问题： n 个人依次抽 n 个签, 不放回, 其中有一个大奖签, 问每个人抽到大奖签的概率是多少?

用 A_k 表示第 k 个人抽到大奖签. 如果第 k 个人抽到大奖签了, 前面的人肯定都没有抽到, 即

$$A_k = A_1^c \cap \cdots \cap A_{k-1}^c \cap A_k.$$

依次地计算 $P(A_1^c) = (n-1)/n$, $P(A_2^c|A_1^c) = (n-2)/(n-1)$, $\cdots$,

$$P(A_k|A_1^c \cap \cdots \cap A_{k-1}^c) = \frac{1}{n-k+1}.$$

因此由乘法公式得

$$\begin{aligned} P(A_k) &= \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdots \frac{n-k+1}{n-k+2} \cdot \frac{1}{n-k+1} \\ &= \frac{1}{n}. \end{aligned}$$

由此可以看出每个人, 不管什么时候抽, 抽到大奖的机会是一样的. 简单地说抽签与顺序无关. 这是说抽到什么签的可能性与顺序无关. 上面的方法当只有一个大奖时是有用的, 当有两个大奖签时, 这个方法不管用.

一般情况下, 这可以用对称性来证明, 也可以用全概率公式来论证. 抽签有两种：放回与不放回. 放回的情况很简单, 因为每次的结果是相互独立的, 所以当然与顺序无关. 对不放回的情况, 我们用前面不放回摸球的例子来说明. 一个袋子中有 3 个白球 2 个黑球, 5 个人依次不放回地摸球, 我们来验证他们每个人摸到白球的概率都是 $3/5$.

A 表示第一个人摸到白球, B 表示第二个人摸到白球. 显然 $P(A) = 3/5$. 第二个人摸球的时候有两种情况: 一种是 A 发生, 即第一个人摸去白球, 这时袋子中剩有 2 白 2 黑; 另一种是 A 没发生, 第一个人摸去黑球, 这时袋子中剩有 3 白 1 黑. 在第一种情况下条件概率为 $P(B|A) = 2/4$, 而在第二种情况下条件概率为 $P(B|\overline{A}) = 3/4$. 由于上述两种情况发生的概率分别是 $3/5$ 与 $2/5$, 直观地可以看出 B 发生的概率应该是两个条件概率的下述加权平均:

$$P(B) = \frac{2}{4} \cdot \frac{3}{5} + \frac{3}{4} \cdot \frac{2}{5} = \frac{12}{20} = \frac{3}{5},$$

其值与 $P(A)$ 相等.

第三个人再摸, 记他摸到白球的事件为 C, 其概率是多少呢? 前面两个人摸球会产生四种情况:

$$\Omega_1 = A \cap B,\ \Omega_2 = A \cap \overline{B},\ \Omega_3 = \overline{A} \cap B,\ \Omega_4 = \overline{A} \cap \overline{B}.$$

用乘法公式, 这四种情况发生的概率分别是

$$\frac{3}{5} \cdot \frac{2}{4} = \frac{6}{20},\ \frac{3}{5} \cdot \frac{2}{4} = \frac{6}{20},\ \frac{2}{5} \cdot \frac{3}{4} = \frac{6}{20},\ \frac{2}{5} \cdot \frac{1}{4} = \frac{2}{20}.$$

例如

$$\mathsf{P}(\Omega_2) = \mathsf{P}(\mathsf{A} \cap \overline{\mathsf{B}}) = \mathsf{P}(\mathsf{A})\mathsf{P}(\overline{\mathsf{B}}|\mathsf{A}) = \frac{3}{5} \cdot \frac{2}{4} = \frac{6}{20},$$

其余同理. 在这四种情况下, 袋子中剩下的黑白球的个数分别是：1 白 2 黑, 2 白 1 黑, 2 白 1 黑, 3 白 0 黑, 因此事件 C 的条件概率分别是 1/3、2/3、2/3、1. 再应用全概率公式, 就推出

$$\begin{aligned}\mathsf{P}(\mathsf{C}) &= \sum_{\mathrm{k}=1}^{4} \mathsf{P}(\mathsf{C}|\Omega_{\mathrm{k}})\mathsf{P}(\Omega_{\mathrm{k}}) \\ &= \frac{1}{3} \cdot \frac{6}{20} + \frac{2}{3} \cdot \frac{6}{20} + \frac{2}{3} \cdot \frac{6}{20} + 1 \cdot \frac{2}{20} \\ &= \frac{36}{60} = \frac{3}{5},\end{aligned}$$

其值仍与 $\mathsf{P}(\mathsf{A})$ 相等.

类似地, 第四、第五个人摸到白球的概率仍然是 3/5, 也就是说, 摸到白球的可能性是与摸球的顺序无关的.

练习 4.3　连续掷 8 次硬币, 分别求其中没有连续出现 (1) 2 次 (2) 3 次 (3) 4 次正面的概率.

4.1.2　骰子游戏

问题: 美洲有一种骰子游戏, 赌徒掷两个骰子. 如果掷出的点数之和是 7 或 11, 他就赢了. 如果掷出 2, 3 或 12, 他就输了. 如果掷出其他点数和, 记下这个数, 再掷骰子一直到掷出这个数或者 7 为止, 如果是这个数, 则赢, 如果是 7, 则输, 问赌徒赢的概率是多少?

用 A 表示赌徒赢这个事件, 求 $\mathsf{P}(\mathsf{A})$. 这个问题比较综合性, 一步一步解释. 先用 X 表示骰子和, X 取值 2 到 12.

然后用全概率公式

$$\mathsf{P}(\mathsf{A}) = \sum_{\mathrm{n}=2}^{12} \mathsf{P}(\mathsf{A}|\mathsf{X} = \mathrm{n})\mathsf{P}(\mathsf{X} = \mathrm{n}).$$

由条件所说,

$$\begin{aligned}&\mathsf{P}(\mathsf{A}|\mathsf{X} = 7) = \mathsf{P}(\mathsf{A}|\mathsf{X} = 11) = 1, \\ &\mathsf{P}(\mathsf{A}|\mathsf{X} = 2) = \mathsf{P}(\mathsf{A}|\mathsf{X} = 3) = \mathsf{P}(\mathsf{A}|\mathsf{X} = 12) = 0.\end{aligned}$$

但是其他情况比较复杂, 需要动脑子.

掷两个骰子, 它们的点数是独立的. 样本空间 $\Omega = \{(\mathtt{i},\mathtt{j}) : 1 \leqslant \mathtt{i},\mathtt{j} \leqslant 6\}$, 36 个结果等可能出现. 通过数事件元素个数, 可以看出, X 分布为

$$\begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36} \end{pmatrix}$$

第一步很简单, 和 X 为 7 或 11 的概率是 $6/36 + 2/36 = 8/36$.

第二步, 当 $\mathsf{X} = 4$ 时, 按照规则, 我们需要计算赌徒在重复掷骰子时在得到 7 之前先得到 4 的概率.

令 B_4 为事件: 和是 4, B_7 为事件: 和是 7, $\mathsf{B} = \mathsf{B}_4^{\mathsf{c}} \cap \mathsf{B}_7^{\mathsf{c}} = (\mathsf{B}_4 \cup \mathsf{B}_7)^{\mathsf{c}}$ 是事件: 和不是 4 也不是 7.

赌徒在得到 7 之前先得到 4 这个事件含有下面的结果

$$\mathsf{B}_4, \mathsf{B}\mathsf{B}_4, \mathsf{B}\mathsf{B}\mathsf{B}_4, \mathsf{B}\mathsf{B}\mathsf{B}\mathsf{B}_4, \cdots.$$

其概率是

$$\begin{aligned} &\mathsf{P}(\mathsf{B}_4) + \mathsf{P}(\mathsf{B})\mathsf{P}(\mathsf{B}_4) + \mathsf{P}(\mathsf{B})^2\mathsf{P}(\mathsf{B}_4) + \mathsf{P}(\mathsf{B})^3\mathsf{P}(\mathsf{B}_4) + \cdot \\ &= \frac{\mathsf{P}(\mathsf{B}_4)}{1 - \mathsf{P}(\mathsf{B})} = \frac{\mathsf{P}(\mathsf{B}_4)}{\mathsf{P}(\mathsf{B}_4) + \mathsf{P}(\mathsf{B}_7)} = \frac{3}{9} = \frac{1}{3}. \end{aligned}$$

因此

$$\mathsf{P}(\mathsf{A}|\mathsf{X} = 4)\mathsf{P}(\mathsf{X} = 4) = \frac{1}{3} \cdot \frac{3}{36} = \frac{1}{36}.$$

有意思的是, 这个例子告诉我们条件概率在实践中是怎么实现的. 赌徒重复不断地掷两颗骰子, 一直到骰子的点数和等于 4 或者 7 为止. 这时, 最后出现的数字是 4 的概率恰好是条件概率

$$\frac{\mathsf{P}(\mathsf{B}_4)}{\mathsf{P}(\mathsf{B}_4 \cup \mathsf{B}_7)} = \mathsf{P}(\mathsf{B}_4|\mathsf{B}_4 \cup \mathsf{B}_7).$$

其他的情形 $\mathsf{X} = 5, 6, 8, 9, 10$ 之下的条件概率计算是类似的.

1. $\mathsf{P}(\mathsf{A}|\mathsf{X} = 10)\mathsf{P}(\mathsf{X} = 10) = \mathsf{P}(\mathsf{A}|\mathsf{X} = 4)\mathsf{P}(\mathsf{X} = 4) = \dfrac{1}{36}$,
2. $\mathsf{P}(\mathsf{A}|\mathsf{X} = 9)\mathsf{P}(\mathsf{X} = 9) = \mathsf{P}(\mathsf{A}|\mathsf{X} = 5)\mathsf{P}(\mathsf{X} = 5)$
   $= \dfrac{4}{4+6} \cdot \dfrac{4}{36} = \dfrac{16}{360}$,

$$3.\ \mathsf{P}(A|X=8)\mathsf{P}(X=8)=\mathsf{P}(A|X=6)\mathsf{P}(X=6)$$
$$=\frac{5}{5+6}\cdot\frac{5}{36}=\frac{25}{396}.$$

现在应用全概率公式把所有概率加在一起

$$\mathsf{P}(A)=\frac{8}{36}+\frac{2}{36}+\frac{32}{360}+\frac{50}{396}=\frac{244}{495},$$

比 $1/2$ 稍微小一点.

这是一个很有意思的问题, 赌徒和赌场赌博, 一般来说, 要满足两个条件: 1. 游戏要复杂一点, 至少不像掷硬币那么简单; 2. 赌场的赢面要大一点, 但不能大很多. 上面这个游戏设计得非常好, 因为它足够复杂, 概率又稍微小于 $1/2$.

练习 4.4　设 A 盒有三个白球, B 盒有三个白球, 一个黑球, 每次从两盒子中任取一个球交换, 求 n 次后黑球仍然在 B 盒的概率.

练习 4.5　设有 A,B 两个盒子, 分别有三个白球和三个黑球, 每次从两盒子中任取一个球交换, 求 n 次后盒子中的球的颜色仍然是相同的概率.

4.1.3　赌徒输光问题

赌徒输光问题是 Huygens 提出的经典问题.

问题: 赌徒甲乙掷硬币或者其他方式赌博, 每局输赢的概率各半, 每次输赢都是一元钱, 赌徒甲乙分别带有赌注总数 a,b, 输光结束, 问赌徒甲先输光的概率是多少?

用 A 表示甲先输光这个事件, X 表示甲现在掌握的赌注, 我们要算 $\mathsf{P}(A|X=a)$, 我们应用全概率公式来推得 $\mathsf{P}(A|X=x)$, $0\leqslant x\leqslant a+b$. 显然 $X=0$, $X=a+b$ 这两种情况分别是甲输光和乙输光的情况, 比赛结束, 即有边界条件: $\mathsf{P}(A|X=0)=1$, $\mathsf{P}(A|X=a+b)=0$. 在 $0<x<a+b$ 时, 比赛要继续, 且这局的结果是赢一元或输一元, 那么

$$\mathsf{P}(A|X=x)=\mathsf{P}(A|X=x-1)/2+\mathsf{P}(A|X=x+1)/2.$$

用 $f(x)$ 表示 $\mathsf{P}(A|X=x)$, 得

$$f(x)=\frac{f(x-1)+f(x+1)}{2}.$$

$f(x)$ 是等差的, 即 $f(x)=cx+d$. 再利用边界条件, $f(0)=1$, $f(a+b)=0$, 推出 $d=1$, $c=-1/(a+b)$, 因此

$$\mathsf{P}(A|X=x)=\frac{a+b-x}{a+b}.$$

特别地

$$P(A|X=a)=\frac{b}{a+b}=\frac{1}{a/b+1}.$$

当 b 相比于 a, 即赌徒乙带的赌注远远超过甲时, 甲输光的概率几乎是 1.

如果甲乙水平不同, 甲赢的概率是 p, 输的概率是 $q=1-p$, 那么上面的递推方是

$$f(x)=qf(x-1)+pf(x+1).$$

因为 $p+q=1$, 所以有以下递推公式

$$f(x+1)-f(x)=\frac{q}{p}(f(x)-f(x-1)).$$

它在边界处 $f(0)=1$, $f(a+b)=0$. 怎么解这个带边界条件的递推问题?

练习 4.6　算赌徒输光问题中当 $p\neq 1/2$ 时最终 A 输光的概率.

4.1.4　抢座位

问题: 设有编号为 1 到 n 的 n 个座位和 n 个人. 第一个人随机选个座位坐下, 第二个人如果看到同号的座位空, 那么他坐该座位, 否则任选一个座位坐下; 第三个人如果看到同号的座位空, 那么他坐该座位, 否则任选一个座位坐下; 这样一直下去, 问最后一个人的同号座位被占的概率是多少?

设 $n=2$ 时, 很容易看出概率是 $1/2$. 当 $n=3$ 时, 第一个人坐自己位置的概率是 $1/3$, 坐 2 号位的概率是 $1/3$, 在第一种情况下, 第 3 人肯定坐自己位置. 第二种情况下, 2 号人坐 1 号位的概率是 $1/2$, 只有这时, 3 号坐在自己位置, 所以 3 号坐自己位置的概率是 $1/3+1/3\cdot 1/2=1/2$.

问题: 一般情况下是不是 $1/2$?

练习 4.7　A,B 两人连续下棋, 每一局, A 胜的概率 p. 使用全概率公式求在下面两种规则下 A 最终赢得比赛的概率,

1. 先连胜两局者赢;
2. 先净胜两局者赢.

4.2　Bayes 公式

归纳推断是人类获取知识的途径之一, 但哲学家 David Hume 认为归纳推断缺乏严谨的依据. Bayes 公式. 这个公式出现在在 Bayes 去世后朋友帮助整理发表的论文上, 论文致力于回答 David Hume 提出的疑问.

用 $\mathsf{P}(A)$ 表示 A 发生的概率, 如果事件 B 发生了, 就是有了新的信息, 在这个信息再看 A 发生的概率

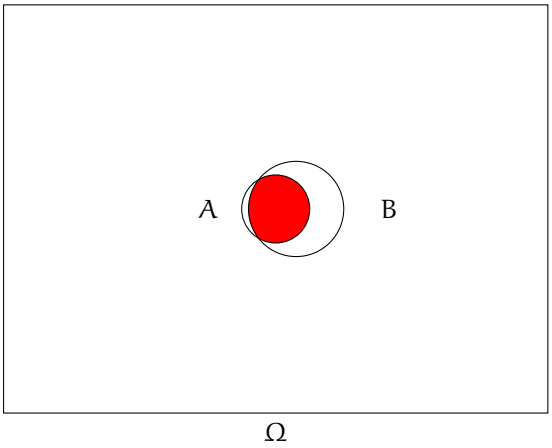
$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(B|A)\mathsf{P}(A)}{\mathsf{P}(B)}.$$

这个公式给出条件概率 $\mathsf{P}(A|B)$ 与 $\mathsf{P}(B|A)$ 的关系, 称为 Bayes 公式.

4.2.1　检测方法的有效性

问题: 对于某种致命疾病而言, 现有的检测方法昂贵而复杂, 某公司开发了一种成本低廉且快速的医学检测方法, 对有这种疾病的人检测时, 95% 是阳性. 对健康的人检测时, 有 95% 是阴性. 已知该疾病的患病率大约是 5%, 问该检测方法准确率高吗?

上面两个指标听起来会感觉这个方法不错, 但这两个指标并不反映真正的准确率. 检测方法的准确率应该是指当一个人检测是阳性时, 他的确有这种疾病的概率的大小. 用 A 表示一个人患有此疾病的事件, A 的概率是患病率, B 表示他的检测有阳性反应的事件. 那么我们要计算条件概率 $\mathsf{P}(A|B)$. 由条件, $\mathsf{P}(A) = 0.05$, $\mathsf{P}(B|A) = 0.95$, $\mathsf{P}(B|A^c) = 0.05$, 如图



由 Bayes 公式,

$$\begin{aligned}\mathsf{P}(A|\mathsf{B}) &= \frac{\mathsf{P}(\mathsf{B}|A)\mathsf{P}(A)}{\mathsf{P}(\mathsf{B}|A)\mathsf{P}(A)+\mathsf{P}(\mathsf{B}|A^c)\mathsf{P}(A^c)}\\ &= \frac{0.95\times 0.05}{0.95\times 0.05+0.05\times 0.95} = 50\%.\end{aligned}$$

比较 $\mathsf{P}(\mathsf{B}|A)=95\%$, $\mathsf{P}(A|\mathsf{B})=50\%$, 不是一个很让人放心的数据. 其实这个概率与患病率大小有很大关系. 例如, 当 $\mathsf{P}(A)=50\%$ 时, $\mathsf{P}(A|\mathsf{B})=95\%$.
为什么问题中 $\mathsf{P}(\mathsf{B}|A)$ 很大, 但关键的指标 $\mathsf{P}(A|\mathsf{B})$ 不大, 其实这容易理解, $\mathsf{P}(\mathsf{B}|A)$ 是 $A\cap\mathsf{B}$ 在 A 中的比例, 而 $\mathsf{P}(A|\mathsf{B})$ 是 $A\cap\mathsf{B}$ 在 B 中的比例.

4.2.2　家里有几个女孩

古典概率问题差不多都是用文字表述, 文字表述的问题就可能引起误解或者理解不够精确. 这一节中, 读者需要真正领会文字表达与数学严密性的不同.
首先问个简单的问题. 一个家庭有两个孩子, 假设每个孩子性别是等可能且互相间是独立的.
问题: 这家有个女孩子的概率是多少?
这个很简单, 假设两个孩子四种等可能的结果, 所以概率是 3/4. 下一个问题. 假设每天晚上两个孩子中会有一个会去阳台浇花, 远远地看到可以辨别性别却无法辨别具体的外表. 先问个简单的问题.
问题: 某一天, 看到阳台上浇花的是个女孩 (事件 A), 问两孩子都是女孩的概率是多少?
我们该怎么解决这个问题呢? Bayes 公式. 考虑事件 C: 两个孩子都是女孩. 则 $\mathsf{P}(\mathsf{C})=1/4$, $\mathsf{P}(A)=1/2$ 且

$$\mathsf{P}(\mathsf{C}|A)=\frac{\mathsf{P}(A|\mathsf{C})\mathsf{P}(\mathsf{C})}{\mathsf{P}(A)}=\frac{1}{2}.$$

现在再问一个问题.
问题: 如果昨天和今天看到阳台上浇花的都是女孩, 问两个孩子都是女孩的概率是多少?
用 A_1 表示 '昨天看到阳台上浇花的是个女孩', 用 A_2 表示 '今天看到阳台上浇花的是个女孩', 你可能一开始马上觉得这两个事件是独立的. 那么它们两个真的独立的吗? 如果是独立的, 那么 $\mathsf{P}(A_1\cap A_2)=1/4$, 这会导致一个明显错误的事情. 事实上,

$\mathsf{P}(A_1 \cap A_2|C) = 1$, 推出

$$\mathsf{P}(C|A_1 \cap A_2) = \frac{\mathsf{P}(A_1 \cap A_2|C)\mathsf{P}(C)}{\mathsf{P}(A_1 \cap A_2)} = \frac{1 \cdot 1/4}{1/4} = 1.$$

这是说接连两天看到阳台浇花的是女孩的条件下, 这家的两个孩子一定都是女孩. 这显然是错误的. 那么问题在哪里呢? 分子的计算没有问题, 问题在分母上, A_1 与 A_2 不可能是独立的. 具体怎么解释?

现在我们一起回答上面所说的两个问题. 通常考虑的样本空间是两个孩子性别的四种可能性: { 男男, 男女, 女男, 女女 }. 例如至少有个女孩的额概率是 3/4, 然后已知有一个是女孩的条件下, 两个都是女孩的概率是 1/3. 但是遗憾的是, 上面的事件 '阳台上浇花的是个女孩' 是不能在这个样本空间中表达的 (想想为什么). 这不妨碍我们计算其概率. 这相当于求不知具体性别的两个孩子中任取一个是女孩的概率. 因为两个孩子有 0,1,2 个女孩所对应的概率分别是 $1/4, 2/4, 1/4$, 所以用全概率公式

$$\mathsf{P}(A) = 0 + \frac{1}{2} \cdot \frac{2}{4} + 1 \cdot \frac{1}{4} = 1/2.$$

这个问题对应的样本空间是什么? 仔细分析, 这个问题中有两个随机性, 一个随机性是孩子性别的随机性, 还有一个随机性是选择一个孩子去浇花的随机性. 明白这个道理, 同学们应该可以把样本空间表达出来, 然后也可以通过数数来计算概率.

现在考虑第三个问题. 孩子的性别情况是随机的, 但是确定了就不会改变, 会变的是每天选一个孩子的随机性, 比如是用掷硬币的方式决定. 这个步骤是独立的. 因此事件 A_1 与 A_2 中有独立的元素, 但不是完全独立的.

为了计算 $A_1 \cap A_2$ 的概率, 需要分离两种随机性, 其中一种随机性在不同的两天是独立的, 另一种随机性在不同的两天是一样的. 先计算它在性别的四种不同情况下的条件概率: (1) 两个女孩; (2) 哥与妹; (3) 姐与弟; (4) 两个男孩, 然后

$$\mathsf{P}(A_1 \cap A_2) = 1 \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} = \frac{3}{8}.$$

我们也应该学会直接构建样本空间然后用计数的方法计算概率. 把前面说的两个随机性放在一起, 真正的样本空间是两个孩子性别的四种可能: 姐妹, 兄妹, 姐弟, 兄弟, 与选择的四种可能: 大大, 大小, 小大, 小小, 所组合的 16 种结果. 事件 $A_1 \cap A_2$ 包含下面几种情况: 如果孩子是姐妹两人, 那么所有 4 种选择都可以; 如果孩子是兄妹两人, 那么选择只能是 '小小' 一种; 如果孩子是姐弟两人, 那么选择只能是 '大大' 一种. 因此 $\mathsf{P}(A_1 \cap A_2) = 6/16 = 3/8$. 因此

$$\mathsf{P}(C|A_1 \cap A_2) = \frac{2}{3}.$$

这个例子很好地诠释了 Bayes 公式的意义, 这个家庭里孩子的性别是随机的, 我们通过观察来认识它. 一开始我们没有任何信息, 所以 '先验地' 认为性别是等可能的, 两个都是女孩的概率是 1/4. 如果观察到一个是女孩, 那么两个都是女孩的概率变大, 是 1/2; 如果两次观察到是女孩, 那么两个都是女孩的概率变得更大, 是 2/3.

练习 4.8 一家庭有两个孩子, 假设等可能性与独立性.

1. 求两个都是女孩的概率.
2. 已知大的一个是女孩, 求小的是女孩的概率.
3. 已知一个是女孩, 求另一个也是女孩的概率.
4. 看到一个孩子是个女孩, 求另一个也是女孩的概率.

练习 4.9 求在 §4.2.2'家里有几个女孩' 这个问题中, 连续 n 天都看到女孩子在浇花的条件下, 这家里两个都是女孩的概率.

4.2.3 Bertrand 3 首饰盒悖论

下面是 Bertrand 的三个悖论中的一个. 设有三个看上去一样的盒子 A,B,C, 分别有两个金手镯, 两个银手镯, 一个金手镯一个手镯, 这些球手镯无法由触摸来区分. 随机取一个盒子, 取的盒子是 C 的概率是 1/3. 现在想象我从中摸一个球出来, 摸出的球是金或者银的机会是一样的. 如果是金, 那么排除了我取的盒子是 B, 所以我取的盒子是 C 的概率是 1/2. 如果是银, 那么排除了我取的盒子是 A, 所以我取的盒子是 C 的概率也是 1/2. 因此由全概率公式, 取的盒子是 C 的概率是 1/2. 它肯定是错的.

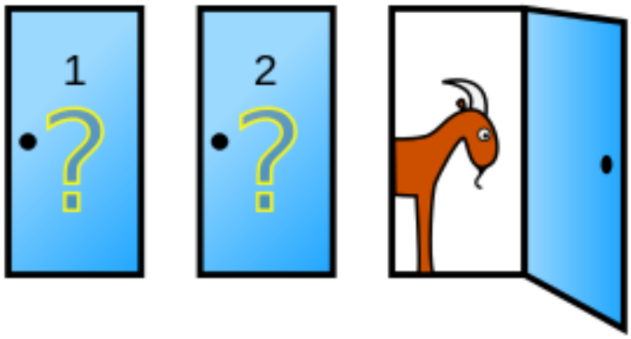
问题: 错在哪里?

练习 4.10 三枚硬币, 一枚两面涂白色, 一枚两面涂黑色, 一枚两面各涂白色与黑色. 现在任取一枚硬币放在桌面上, 呈现白色, 问另一面也是白色的概率是多少?

4.2.4 Monty Hall 问题

在美国的某刊物有一个栏目名为: ask Marilyn, 回答读者提出的各种问题, 其中有一个后来被称为是 Monty Hall 的问题. 有一个由 Monty 主持的电视游戏栏目是这样的: Monty 让参与人 Voila 在三个完全一样的大门 $1, 2, 3$ 中任选一个, 三门后分别

有两只羊与一部汽车, Monty 知道门后是什么. 当 Voila 选定, 比如, 1 后, Monty 打开一个放有羊的门, 比如, 3, 然后告诉 Voila 可以再选择.



问题: Voila 该选择换还是不换?

Marilyn 的答案是换, 理由是换比不换得车的概率更大. 但答案在当时引起了很大的争议. 因为很多人认为在 Monty 打开门之后, 剩下的两个门有汽车的概率是一样的, 所以换不换没有区别. 还有许多人编了电脑程序来模拟.

对于这个问题, 最直接的方法是写出完整的概率空间, 按照规则, 车在的门与 Voila 选的门是等可能且独立的, 然后如果这两个门一样, 那么 Monty 等可能地开其他两个门之一, 否则 Monty 只能开剩下的门. 三个门依次列出, $\Omega = \{112, 113, 221, 223, 331, 332, 123, 132, 213, 231, 312, 321\}$, 其中头两个字一样的结果概率为 1/18, 三个字都不同的结果概率为 1/9. 计算出换而得车的概率是 2/3, 不换得车的概率 1/3.

也可以用全概率公式. Voila 做了选择且 Monty 打开了一个羊门, 计算事件 H=‘换(一个关闭的门) 而得车’ 的概率. 用 (i, j) 表示事件 ‘车在门 i’ 与 ‘Voila 选择门 j’. 那么当 $i = j$ 时, $\mathsf{P}(\mathsf{H}|(i,j)) = 0$, 当 $i \neq j$ 时, 车在另外一个没打开的门内, $\mathsf{P}(\mathsf{H}|(i,j)) = 1$. 因此

$$\mathsf{P}(\mathsf{H}) = \sum_{i,j} \mathsf{P}(\mathsf{H}|(i,j))\mathsf{P}((i,j)) = 6 \cdot \frac{1}{9} = 2/3.$$

推出不换而得车的概率是 1/3, 小于换而得车的概率.

下面我们尝试用其他方法. 由对称性, 可以假设 Voila 选择了门 1, Monty 打开了一个门, 建立概率空间. 用 A_j 表示车在门 j 这个事件, 那么当 $j = 2, 3$ 时, $\mathsf{P}(\mathsf{H}|\mathsf{A}_j) = 1$,

用全概率公式得

$$\mathsf{P}(\mathsf{H}) = \sum_{\mathsf{j}=1}^{3} \mathsf{P}(\mathsf{H}|\mathsf{A}_\mathsf{j})\mathsf{P}(\mathsf{A}_\mathsf{j}) = 2/3.$$

在 Voila 看来, 主持人在可以选择开哪个门的时候总是等可能地选择的, 那么由对称性, 所求概率与 Voila 选择门 1, Monty 打开门 2 的条件下, 换而得车的概率相同. 还是假设 Voila 选择门 1, 建立概率空间. 用 M_2 表示事件'Monty 打开门 2', 这时计算条件概率 $\mathsf{P}(\mathsf{H}|\mathsf{M}_2)$. 由 Bayes 公式得

$$\mathsf{P}(\mathsf{H}|\mathsf{M}_2) = \frac{\mathsf{P}(\mathsf{M}_2|\mathsf{H})\mathsf{P}(\mathsf{H})}{\mathsf{P}(\mathsf{M}_2)} = 2/3,$$

其中 $\mathsf{P}(\mathsf{H}) = 1/3$, $\mathsf{P}(\mathsf{M}_2|\mathsf{H}) = 1$, 以及

$$\mathsf{P}(\mathsf{M}_2) = \sum_{\mathsf{j}} \mathsf{P}(\mathsf{M}_2|\mathsf{A}_\mathsf{j})\mathsf{P}(\mathsf{A}_\mathsf{j}) = 1/2 \cdot 1/3 + 0 + 1 \cdot 1/3 = 1/2.$$

还有一个简单思考. 对于 Voila 开始的选择, 两个结果: G='选到车', 概率是 1/3; G^c='没选到车', 概率是 2/3. Monty 打开一个羊门, 如果 Voila 选择换, 那么结果'选到车' 导致 '没选到车', 而 '没选到车' 导致 '选到车', 因此 Voila 换而得车的概率是 2/3. 一般情况下, 在 Voila 做出选择且 Monty 按规则打开一个门之后, 我们建立一个概率空间. 显然 $\mathsf{P}(\mathsf{H}|\mathsf{G}) = 0$, $\mathsf{P}(\mathsf{H}|\mathsf{G}^\mathsf{c}) = 1$, 因此

$$\mathsf{P}(\mathsf{H}) = \mathsf{P}(\mathsf{H}|\mathsf{G})\mathsf{P}(\mathsf{G}) + \mathsf{P}(\mathsf{H}|\mathsf{G}^\mathsf{c})\mathsf{P}(\mathsf{G}^\mathsf{c}) = 2/3.$$

严格地说, 需要说明 $\mathsf{P}(\mathsf{G}) = 1/3$.

4.3　笑话的背后

讲个笑话, 说是有个统计学家非常害怕坐飞机, 因为他害怕有人携带炸药之类的爆炸物. 他看到一个数据, 有一个人携带爆炸物上飞机的概率是百分之一, 有两个人以上携带爆炸物的概率就小于万分之一. 他灵机一动, 每次坐飞机时自己携带一个爆炸物, 他认为这样就保证还有别人携带爆炸物上飞机的概率小于万分之一.

问题: 笑话的笑点在哪里?

实际上这样的笑话还是经常会发生的, 比如在赌场里, 不管是什么赌博, 都会有输红了眼的赌徒, 大吼: "见鬼了，已经连开了 10 把大了，下一把肯定是小，都压上, 我就不信我还会输。" 这个话可能很多人不会觉得是笑话, 但实际上它和上面的笑话是

完全一样的可笑. 开大开小的概率一样, 连开 11 把都是大的概率小于 1/2000, 现在前面 10 把都是大, 你是不是觉得第 11 把还是小的概率太小了. 但问题不在于第 11 把, 问题出在前十把, 全部都是大真的太奇怪了. 但是不管怪不怪, 这个事件已经发生了, 你就得承认现实, 下一把结果怎么样与前面毫无关系, 开大开小的概率还是一样的. 因此你永远不要指望下一把会有更多的运气.

也许, 赌徒是把掷硬币和抽奖两件事情混淆了, 如果是 11 个签里有一个大奖, 那么抽一个签不是大奖就使得下一次抽到大奖的概率变大, 而且前 10 次都没抽到大奖的条件下赌徒第 11 次抽到的肯定是大奖. 但掷硬币不同, 每次掷得的结果是独立的, 不管前面怎么样, 下一次都是重新开始.

4.4　随机变量及其分布

一个随机试验, 有样本空间 Ω, 其中的元素一般用 ω 表示. 一个事件 A 是指 Ω 的一个子集, 事件发生的概率是一个 0,1 之间的数字, 用 $P(A)$ 表示. 用花写的字母 $\mathscr{F}$ 表示事件的全体.

在随机试验中, 为了方便和清晰, 我们经常用数值来表示结果, 即考虑样本空间上的函数, 这样的函数称为随机变量 (注意: 随机变量是个函数). 随机变量的概念是如此自然, 以至于我们可能没有意识到前面已经出现过很多次. 读者对函数是熟悉的 $y = f(x)$, 自变量 x 对应唯一的 y, 自变量通常有个取值范围, 称为定义域. 随机变量也是函数, $y = X(\omega)$, 定义域是 Ω. 在概率论中, 习惯只是写 X, 省略自变量 $\omega \in \Omega$ 与因变量 y. 例如掷硬币, 正面用 1 表示, 反面用 0 表示, 就是随机变量; 实际上, 给定事件 A, 我们总可以定义一个随机变量, 记为 1_A, 它像一个小旗, 在 A 发生的时候往上挥舞, 等于 1, 否则往下挥舞, 等于 0. 这样的随机变量称为 Bernoullli 随机变量, 也称为事件 A 的指标. 这样的随机变量是最基本的, 其他随机变量是这样的随机变量的线性组合之后取极限. 你可以把 Bernoulli 随机变量理解为砖头, 其他随机变量是用砖砌起来的墙. 掷 n 个硬币, 正面个数也是随机变量, 然后你可以想想样本空间是什么, 对应法则是什么.

取有限多个值, 例如掷 n 个硬币的正面个数, 或者可列多个值, 例如等待成功出现的时间, 称为简单或者离散型随机变量. 把取值的概率写出来就是随机变量的分布. 我们只需要把所有取值的概率写出来就可以了, 例如掷 n 个硬币得到正面的个数为 X,

则

$$P(X=i)=\binom{n}{i}/2^n,\ 0\leqslant i\leqslant n,$$

左边读成正面个数是 i 的概率. 分布的表示方法很多, 简单的情况可用图或者表来表示, 这比较直观易懂, 第一排表示随机变量的取值, 下面对应的数表示取此值的概率, 例如, 掷 3 个硬币, 正面次数的分布

$$\begin{pmatrix}0 & 1 & 2 & 3\\ 1/8 & 3/8 & 3/8 & 1/8\end{pmatrix}.$$

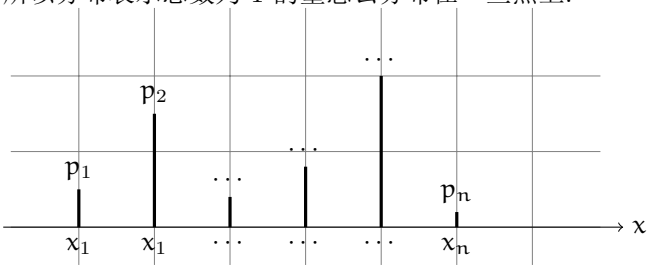
一般地, 一个如下形式的图表被称为是一个分布：

$$\begin{pmatrix}x_1 & x_2 & \cdots & \cdots & x_n\\ p_1 & p_2 & \cdots & \cdots & p_n\end{pmatrix},$$

其中 x_1、x_2、$\cdots$、x_n 是实数, p_1、p_2、$\cdots$、p_n 是非负数, 作为概率值, [1]其总和为 1, 即成立

$$p_1+p_2+\cdots+p_n=1.$$

所以分布表示总数为 1 的量怎么分布在一些点上.



前面所说的随机变量的取值一般是整数, 它的分布通常叫做分布律, 但是随机变量可以取值是实数, 例如身高体重, 还有水位, 寿命等, 这些量其实是连续变化的量, 它们的分布无法用分布律的方式表示, 而是采用分布函数来表示. 对于随机变量 X, 定义

$$F(x)=P(X\leqslant x),\ x\in\mathbf{R},$$

这是一个递增的取值在 0,1 之间的函数, 称为 X 的分布函数. 显然, 对于区间 (a,b),

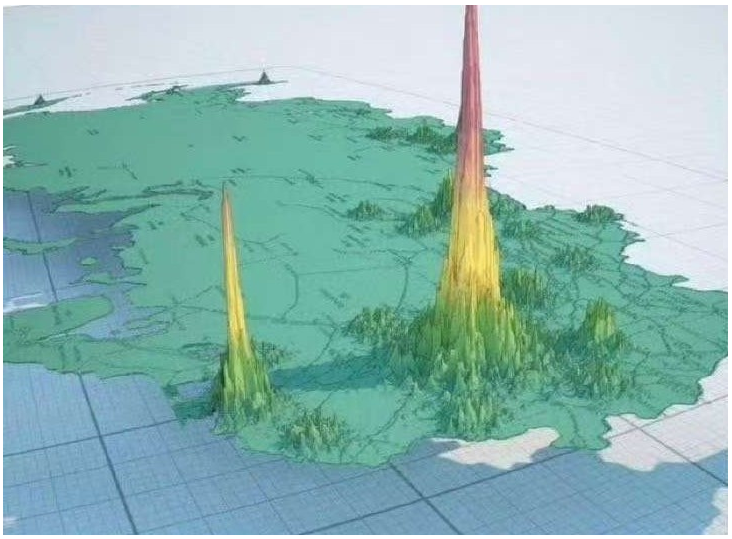
$$P(X\in(a,b))=F(b)-F(a),$$

[1]一般要求是正数, 因为 0 概率值在分布中所在的这一列总可以删去.

读作 X 落在区间 (a,b) 的概率等于 $F(b)-F(a)$. 通常, 对于许多随机试验来说, 尽管随机变量的结果是不可预知的, 但分布是清楚的. 例如买彩票, 中奖的概率是清晰的, 公开的, 但中奖号码是不可预知的. 这也是随机现象的本质.

练习 4.11　设 X 是掷骰子的点数. 写出 X 的分布函数并画出图像. 并讨论分布律与分布函数的关系.

分布也是一个生活中常用的词, 人口分布, 收入分布, 年龄分布, 商品品牌的市场占有率分布, 等等. 媒体上给大众看的分布表示也是形式多样, 有条状, 圆饼状, 立体状等等, 例如下图中的俄罗斯人口分布图, 清晰地显示出绝大部分人口分布在莫斯科和圣彼得堡两个地区.



当然概率其实就是一种分布, 体现总量为 1 的量怎么分布在一个集合的元素上, 而随机变量的分布体现它怎么分布在实数上.

独立与同分布这两个概念是非常重要的. 两个随机变量 X,Y 独立, 是指对任何实数 x,y, 事件 $\{X\leqslant x\}$ 与 $\{Y\leqslant y\}$ 独立. 两个随机变量 X,Y 同分布, 是指它们的分布函数一样. 同分布的随机变量不一定一样. 例如一个骰子的点数与另一个骰子的点数未必一样, 但是独立同分布的.

随机变量的分布主要分成离散型与连续型. 直观地说, 离散型是指以点方式分布的, 用和表示即可; 连续型是指以密度方式分布的, 需要用积分来表示. 在物理学中, 我们学过密度, 物体的质是连续地分布的, 每个点的质量是零, 点聚在一起构成区域,

质量是体积和密度 (常数时) 的乘积. 随机变量非常类似地分布在实数轴上, 可以用可变的密度来表示分布情况. 怎么表示呢? 画一个直角坐标系, 假设随机变量 X 是分布在横轴上, 然后在横轴上方画一个函数图像, 称为密度函数, 记为 $y = f(x)$. 形象地说, 随机变量分布在区间 (a, b) 上的概率等于区间上曲线下这个区域的面积, 假设读者学过积分, 那么分布函数与密度函数有下面的关系

$$F(b) - F(a) = P(X \in (a, b)) = \int_a^b f(x)dx.$$

这样这个曲线非常直观地表达了随机变量的分布情况. 分布函数是累积的, 所以用增量来表示分布多少, 相比之下, 密度函数的面积表示更直观, 类似于分布律. 前面提到的几何概率模型中的分布是均匀分布, 是连续型分布, 它的密度函数在一个区间上等于常数, 其他地方等于零. 中学介绍过的正态分布的密度函数是一个钟形的函数

$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

系数的选取为了保证图像下的面积等于 1. 本课程不假设学生学过微积分, 所以不涉及除均匀分布外的连续型分布.

4.5　分布的期望与方差

期望 (expectation) 是个很朴素的概念, 任何人在等待不确定的结果时都会有期望. 例如孕妇生孩子的日期称为预产期, 英文是 expected date, 不是一个确定的日子. 期望的字面意思大概是合理的预期. 比如考试的结果通常是不确定的, 付出得多, 期望就高; 再比如掷 100 个硬币, 正面的个数自然是随机的, 但你会期望正面的个数在 50 个左右; 在历史上, 期望的概念和概率一样自然, Pascal 与 Fermat 通信讨论的分赌注问题也可以认为是期望问题, 它是这样叙述的: 甲乙两个赌徒通过掷硬币三局两胜来分 64 块钱赌注, 正 (反) 面是赌徒甲 (乙) 胜. 在未掷硬币时, 双方对自己能拿到的钱的期望是一样的, 因为他们赢得概率是一样的. 现在掷一次硬币, 假若甲胜, 问如果游戏这时终止, 甲乙应该怎么分钱才合理? 这实际上是问他们各自对于应该获得的份额的期望是多少. 因为甲已经胜一局, 所以他最终赢得机会就大, 对钱的份额期望就会增加, 乙的期望相应就会减少, 那么具体数额呢? Pascal 是这样说的: '假设再掷一次, 如果是甲胜, 那么游戏结束, 甲拿走所有的钱; 如果是乙胜, 那么比分一样, 两人的期望又一样了, 所以无论何种情况, 甲至少获得 64 的一半 32 块, 剩下的

一半应该对半分, 最终甲应该得 48, 乙得 16.' 这个比例 3:1 恰好是两个人赢的概率比, 也就是说, 期望与概率成比例.

尽管期望这种直观的意思早已被人认识, 但文献中最早使用期望这个词的大概是荷兰数学家物理学家 C. Huygens, 他得知 Fermat 与 Pascal 的通信之后对机会的问题非常感兴趣, 于 1656 年完成了最早的一本关于概率的著作, 其中有这样一段话直观地解释了期望: ''如果一个人在一个手中放 3 块钱在另一个手中放 7 块钱并让我选择, 那么这个机会或者期望等同于直接给我 5 块钱." 因为等概率地选择 7 块与 3 块的平均是 5 块, 所以任何一个即使没有学过概率的人也会体会到自己的期望是 5.

随机变量有分布, 分布就有中心, 称为数学期望, 简称为期望或者平均. 设随机变量 X 的取值有限多个, 分布为

$$P(X = x_i) = p(x_i),\ 1 \leqslant i \leqslant n.$$

实际上, 随机变量的取值可列个 (是个数列) 也是可以的, 那样的话, 公式中的求和是可列和.

定义 4.5.1　X 的期望 (也叫做均值) 定义为

$$E[X] = \sum_{i=1}^{n} x_i p(x_i),$$

即 x_i 的概率权平均.

上面的定义简单直观地写成

$$EX = \sum_{x} xP(X = x),$$

右边解释为对所有可能的 x 求和, 可能的 x 是指 $P(X = x) > 0$. 实际上 X 可以取可列无穷多个值, 这是上面的和变成无穷级数. 随机变量分布的期望类似于物体的重心: 位置与位置的权重的乘积之和. 另外 X 的期望用 EX, $E(X)$, 或者 $E[X]$ 表示都可以, 没有区别.

数学期望是随机变量组成的集合到实数的一个映射, 它有下面的性质.

命题 4.5.1　期望的三个性质: X, Y 是随机变量, c 是常数, 则

1. $E[cX] = cE[X]$;
2. $E[X + Y] = E[X] + E[Y]$.

3. 如果 X, Y 独立, 那么

$$E[XY] = E[X]E[Y].$$

练习 4.12 第二个性质是需要证明, 并且很不容易, 请读者自己思考, 把它搞清楚.

从定义看出, 一个非负随机变量的期望也是非负的, 因此期望还有

命题 4.5.2 单调性: 如果 $X \geqslant Y$, 那么 $E[X] \geqslant E[Y]$.

这里的 $X \geqslant Y$ 是指在任何情况下 X 的值都大于等于 Y 的值, 或者说对任何 $\omega \in \Omega$ 有 $X(\omega) \geqslant Y(\omega)$.

定义 4.5.2 随机变量的方差定义为

$$D[X] = E[(X - EX)^2].$$

右边 $(X - EX)^2$ 是随机变量偏离中心的距离, 因此方差是随机变量偏离中心的程度, 从某种意义上, 它是衡量 X 的随机程度, 不确定性大小或者说波动程度的一个指标. 特别地, 当 $D[X] = 0$ 时, X 是个常数, 没有随机性.

练习 4.13 证明: 函数 $f(x) = E[(X - x)^2]$ 在 $x = E[X]$ 时最小.

将定义的右边展开,

$$D[X] = E[X^2 - 2X \cdot EX + (EX)^2] = EX^2 - (EX)^2,$$

这个计算公式可能更加方便, 其中

$$E[X^2] = \sum_{i=1}^{n} x_i^2 p(x_i).$$

一般地, 设 U 是个函数, 那么随机变量 $U(X)$ 的期望为

$$E[U(X)] = \sum_{i=1}^{n} U(x_i) p(x_i).$$

命题 4.5.3 方差的三个性质:

1. $D[cX] = c^2 D[X]$;
2. 平移不变: $D[X + c] = D[X]$;
3. 如果 X, Y 独立, 那么

$$D[X + Y] = D[X] + D[Y].$$

事件 A 可以看成是随机变量 1_A, 它在 A 发生时取值 1 没发生时取值 0, 或者说在集合 A 上值为 1, 其他地方为 0, 称为 A 的示性函数或者指标. 因此期望与概率本质上是一个东西, 只不过使用的范围更广. 概率本质上是一个度量, 度量总有先用于基本的事物, 然后渐渐扩大. 例如当我们讨论面积的时候, 小学只能谈论长方形的面积, 中学可以理解多边形的面积, 大学生可以理解任意图形的面积.

尽管期望和方差是随机变量的期望和方差, 但实际上它们只依赖于分布, 同分布的随机变量有相同的期望方差. 另外, 期望和方差是分布的两个用数字表示的特征, 称为数字特征, 如同年龄身高体重是人的特征一样, 特征帮助我们认识分布, 但是不能确定分布.

如果一个随机变量是有度量单位的, 例如距离, 面积等等, 那么标准差 $\sqrt{D[X]}$ 可能是更好的选择, 因为它与随机变量的度量单位一致.

练习 4.14　设 $\lambda > 0$, 随机变量 X 的分布律为

$$P(X = n) = e^{-\lambda}\frac{\lambda^n}{n!},\ n \geqslant 0.$$

求 X 的期望和方差.

练习 4.15　某人用 n 把外形相似的钥匙去开门, 只有一把能打开. 今逐个任取一把试开, 直至打开门为止. 分别考虑每次试毕放回与不放回两情形, 求试开次数 X 的期望.

Bernoulli **随机变量**

取值 0,1 的随机变量称为 Bernoulli 随机变量, 它是最基本的. 对于一个事件 A, 它的指标 1_A 是 Bernoulli 随机变量, 反过来, Bernoulli 随机变量 X 是事件 $A = \{X = 1\}$ 的指标. 这时, $E[X] = p$, 其中 $p = P(X = 1)$. 另外 $E[X^2] = p$, 所以 $D[X] = E[X^2] - (E[X])^2 = p - p^2 = pq$, 其中 $q = 1 - p$. 可以看出 $D[X]$ 在 $p = 1/2$ 时最大, 即随机性最大.

回过头看分赌注问题, 在甲乙两个赌徒 2:1 时甲最终赢这个事件记为 A, 那么赌徒甲最终赢得赌注数为

$$X = 64 \cdot 1_A,$$

这其实是说, 或者赢得 64 元, 或者 0. 那么其期望为

$$E[X] = 64P(A),$$

即与赢的概率成比例. 因此赌注实际上恰好是赌徒未来所得的期望, 即赌注按照期望划分. 实际上, Fermat 与 Pascal 当年的通信里正表达了期望这个概念. 但是, 期望 (expectation) 这个词是 C. Huygens 在 1657 年出版的著作 (可能是第一本概率论著作) 中首先使用的.

分苹果

再回到 2.3.1 中的苹果的随机分配方案. 我们现在可以算分配方案的分布和期望了. 三个人分二个苹果, 每个人对分得的期望是多少呢? 不用算就知道: 三分之二. 用 X 表示各分配方案中一个人 (例如甲) 得到的苹果个数.

1. $\mathsf{P}(\mathsf{X}=0)=2/3$, $\mathsf{P}(\mathsf{X}=2)=1/3$. $\mathsf{E}[\mathsf{X}]=2/3$.
2. 直接数 $\mathsf{P}(\mathsf{X}=0)=6/9, \mathsf{P}(\mathsf{X}=1)=4/9, \mathsf{P}(\mathsf{X}=2)=1/9$. 期望为
$$\mathsf{E}[\mathsf{X}]=0\cdot 6/9+1\cdot 4/9+2\cdot 1/9=2/3.$$
3. $\mathsf{P}(\mathsf{X}=0)=3/6, \mathsf{P}(\mathsf{X}=1)=2/6, \mathsf{P}(\mathsf{X}=2)=1/6$. 期望为 $\mathsf{E}[\mathsf{X}]=0\cdot 3/6+1\cdot 2/6+2\cdot 1/6=2/3$.
4. $\mathsf{P}(\mathsf{X}=0)=1/3, \mathsf{P}(\mathsf{X}=1)=2/3, \mathsf{P}(\mathsf{X}=2)=0$. 期望为
$$\mathsf{E}[\mathsf{X}]=0\cdot 1/3+1\cdot 2/3+2\cdot 0=2/3.$$

尽管它们的分布不同, 但是期望的确都是 $2/3$. 公平分配的本质是什么? 本质是同分布, 只要同分布, 就可以推出同期望, 且期望为 $2/3$. 前面我们说方差在某种意义上描述随机变量的随机性. 为什么说方差只是在某种意义上表达了随机性大小呢? 因为随机性不是一个有确定定义的词, 而是一个主观的有争议的词. 方差描述随机变量取值的分散程度, 分散程度越大说明随机变量越难以估计或者猜测, 因此分散程度只是随机性或者不确定性的一个部分.

这里我们也用上面的分苹果方案来来解释.

1. 方案 1: 方差是 $8/9$.
2. 方案 2: 方差是 $4/9$.
3. 方案 3: 方差是 $5/9$.
4. 方案 4: 方差是 $2/9$.

练习 4.16　具体计算上述四种方案的方差.

分配方案首先要看期望, 期望相同就是机会均等. 但是社会在进步, 只看期望是不够的, 因为这可能导致分配很不均匀, 有的人多得吃不了, 而有的人吃不饱. 因此我们就自然希望分配方案的方差不要太大. 这也就是为什么现代社会分配不能仅依赖自然的竞争, 需要用政府的手来调节, 对于富人的税收越来越高, 对于穷人的福利越来越全面. 上述方案方差从小到大的顺序是: 方案 4,2,3,1. 事实上, 如果 X 在 0,1,2 上的分布分别是 p_0, p_1, p_2, 它们是正且和为 1, 还是一个约束是期望为 2/3: $p_1 + 2p_2 = 2/3$ 那么方差为

$$D[X] = p_1 + 4p_2 - 4/9 = 8/9 - p_1.$$

因此 p_1 越大方差越小.

考虑 0,1,2 上的两个分布, 一个是分布为 1/2,0,1/2, 一个是均匀分布, 期望都是 1, 方差分别是 1 与 2/3, 有的人直觉上会认为均匀分布的随机性更大, 因为它更难猜测, 尽管它方差小. 很多人认为均匀分布的随机性大一些, 为什么呢? 是因为随机性的另一个方面: 混乱. 这里要谈到熵的概念. 熵是一个统计物理术语, 也是一个信息论概念, 是指一个系统的混乱程度的度量, 定义为

$$S = -c \sum_{i=1}^{n} p_i \log p_i,$$

其中 $(p_i : 1 \leqslant i \leqslant n)$ 是个分布, c 是正常数. 注意熵仅与概率值有关和随机变量取值无关. 这个系统什么时候熵最大, 或者说最混乱? 当分布是均匀的时候: $p_1 = \cdots = p_n = 1/n$. 什么时候方差最大, 当分布是两端分布时. 直观来看, 一个随机现象的随机性与混乱程度类似, 但从这个例子看, 方差与熵的意义显然不同, 它们各自从不同角度表达随机性, 这说明随机性是多元的.

成功的次数: 二项分布

只关心两个结果的随机试验通常称为 Bernoulli 试验, 一个结果称为成功, 概率是 $0 < p < 1$, 另一个结果称为失败, 概率是 q, 当然 $q = 1 - p$. 掷硬币是 Bernoulli 试验; 掷骰子, 如果掷出 6 算是成功, 否则算失败, 也是 Bernoulli 试验, 成功概率是 1/6. 通常用一个随机变量 ξ 来表示成功的指标, 即成功的话, $\xi = 1$, 失败的话, $\xi = 0$. 只取 0 与 1 两个值的随机变量称为 Bernoulli 随机变量.

独立地重复 Bernoulli 试验是我们经常碰到的随机试验, 例如重复掷硬币或者骰子等, 生活也仿佛是日复一日地在重复, 每天都在经历成功或者失败.

用 X 表示重复 n 次 Bernoulli 试验中成功的次数, 那么 X 的取值是 0 到 n 间的整数. 成功次数为 k 这个事件 $\{X = k\}$ 的概率怎么计算?

把 n 次试验看成具有 n 个标号的位置, 其中每个位置都有两种可能: 成功或者失败, 分别标记为 1 及 0. “成功次数为 k”的事件 $X = k$ 可以看成是从 n 个位置里拿出 k 个位置标记为 1, 而其他标记为 0, 这样的选择共有 $\binom{n}{k}$ 种. 因为独立性, 每种标记发生的概率是 $p^k q^{n-k}$. 由概率的可加性, 成功次数为 k 的概率为

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k},$$

其中 $0 \leqslant k \leqslant n$. 因为 $\binom{n}{0} = \binom{n}{n} = 1$, 可用下图表表示 X 的分布：

$$\begin{pmatrix} 0 & 1 & 2 & \cdots & k & \cdots & n \\ q^n & \binom{n}{1} pq^{n-1} & \binom{n}{2} p^2 q^{n-2} & \cdots & \binom{n}{k} p^k q^{n-k} & \cdots & p^n \end{pmatrix}$$

从而可以从这个角度证明二项式定理

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} \mathbb{P}(X = k) = 1.$$

这是这个分布称为二项分布的理由. 对任何的 n, n 之内至少有一次成功的概率

$$P(X > 0) = 1 - P(X = 0) = 1 - (1 - p)^n.$$

我们在上一章说过, 一般来说, 人是无法体会到概率零事件与小概率事件的差异的. 但上式告诉我们, 当重复该试验且次数 n 很大时, 成功概率为零与成功概率非零表现出本质的区别. 如果成功概率为零, 那么无论重复多少次, $P(X > 0) = 0$, 会有一次成功的可能性还是零. 但只要成功概率不是零, 那么无论多小, $\lim_n P(X > 0) = 1$, 至少成功一次的概率趋近 1.

把上面的结论翻译成生活语言. 做一件事情, 不管成功概率多小, 只要执着地努力, 重复的次数足够多, 终究有成功的一天. 如同俗话所说：失败是成功之母. 这是好的一面, 反过来说, 如果不断地重复, 小概率的坏事也终有可能发生. 例如在高速公路上开车, 只要发生一次事故就可能导致严重的后果, 所以, 为了使发生事故的可能性尽可能小, 不仅每次开车都要格外小心, 减小事故发生率 p, 而且要尽可能地减少开车次数 n.

问题：求二项分布的随机变量 X 的期望与方差.

设独立地重复一个成功概率为 p 的 Bernoulli 随机试验 n 次, 成功的次数为 X, 那么

$$P(X=k)=\binom{n}{k}p^k q^{n-k},\ k=0,1,2,\cdots,n.$$

$$\begin{aligned}E[X]&=\sum_{k=0}^{n}k\cdot\binom{n}{k}p^k q^{n-k}\\&=\sum_{k=0}^{n}\frac{kn!}{k!(n-k)!}p^k q^{n-k}\\&=\sum_{k=1}^{n}p\frac{n(n-1)!}{(k-1)!(n-k)!}p^{k-1}q^{n-k}\\&=np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{n-k}\\&=np(p+q)^{n-1}=np.\end{aligned}$$

成功次数的期望是 np, 试验次数乘以成功概率, 这与直观相符.
计算方差比较困难, 因为先要计算二阶矩

$$E[X^2]=\sum_{k=0}^{n}k^2\cdot\binom{n}{k}p^k q^{n-k}.$$

$$E[X^2]=\sum_{k=0}^{n}k(k-1)\cdot\binom{n}{k}p^k q^{n-k}+\sum_{k=0}^{n}k\cdot\binom{n}{k}p^k q^{n-k},$$

第二个和就是 $E[X]$. 那么我们就来算第一个和, 同上面一样的方法

$$\begin{aligned}\sum_{k=0}^{n}k(k-1)\cdot\binom{n}{k}p^k q^{n-k}&=\sum_{k=2}^{n}\frac{n!}{(k-2)!(n-k)!}p^k q^{n-k}\\&=\sum_{k=2}^{n}p^2\frac{n(n-1)(n-2)!}{(k-2)!(n-k)!}p^{k-2}q^{(n-2)-(k-2)}\\&=n(n-1)p^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}q^{n-k}\\&=n(n-1)p^2(p+q)^{n-2}\\&=n(n-1)p^2.\end{aligned}$$

因此

$$D[X]=E[X^2]-(E[X])^2=n(n-1)p^2+np-(np)^2=npq.$$

还有一个比较简单的方法, 利用期望方差的性质. 首先成功次数可以写成为

$$X = \xi_1 + \xi_2 + \cdots + \xi_n,$$

其中 ξ_i 是第 i 次试验成功的标志, 且 $\xi_1, \cdots, \xi_n$ 是互相独立的. 利用期望与方差的性质得

$$\begin{aligned} E[X] &= np, \\ D[X] &= D[\xi_1] + \cdots + D[\xi_n] = npq. \end{aligned}$$

期望等于试验次数与成功概率的乘积, 方差在 $p = 1/2$ 时最大. 这正是我们的直观感觉. 这个例子说明, 尽管期望和方差是通过方差定义的, 但它们的计算经常不需要知道具体分布. 这是因为期望方差仅仅是分布的特征, 二不是分布的全部, 例如你不需要知道一个人具体什么模样, 从背后看一眼就能知道他头发什么颜色, 大概有多高和多重.

练习 4.17　设 X 如上二项分布. 求 $E\frac{1}{1+X}$.

练习 4.18　设袋子里有 a 个白球与 b 个黑球, 随机地摸出 n 个球 (不放回), 用 X 表示其中的白球数, 求 X 的期望与方差.

练习 4.19　8 男与 7 女随机坐在一排 15 个座位上, X 表示其中相邻且异性的对数, 即集合 $\{1 \leqslant i < 15 : i$ 座与 $i+1$ 座为异性 $\}$ 的元素个数. 求 X 的期望.

等待成功: 几何分布

问题: 重复成功概率为 p 的 Bernoulli 试验, 首次成功的时刻 X, 称为等待首次成功出现的时间, 求其期望与方差.

其分布是

$$P(X = k) = q^{k-1}p, \ k \geqslant 1.$$

期望

$$E[X] = \sum_{k \geqslant 1} kq^{k-1}p = \frac{1}{p},$$

其中的和

$$\sum_{k \geqslant 1} kq^{k-1} = \left(\sum_{k \geqslant 0} q^k \right)' = \frac{1}{(1-q)^2} = \frac{1}{p^2}.$$

即平均等待成功的时间与成功概率成反比, 也是直观的.

要计算方差, 需要先计算二阶矩 $\mathsf{E}[X^2]$,

$$\begin{aligned}\mathsf{E}[X^2] &= \sum_{k\geqslant 1} k^2 q^{k-1} p \\ &= pq \sum_{k\geqslant 1} k(k-1) q^{k-2} + p \sum_{k\geqslant 1} k q^{k-1} \\ &= pq \frac{2}{(1-q)^3} + \frac{1}{p} \\ &= \frac{2pq}{p^3} + \frac{1}{p} = \frac{2-p}{p^2},\end{aligned}$$

其中, 读者需要验证

$$\sum_{k\geqslant 1} k(k-1) q^{k-2} = \frac{2}{(1-q)^3}.$$

因此

$$\mathsf{D}[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}.$$

练习 4.20　重复成功概率为 p 的 Bernoulli 试验, k 是正整数. 第 k 次成功的时刻记为 X_k, 求其分布, 期望与方差.

4.5.1　Coupon 问题

问题: 重复掷一个骰子, 等待所有数字都至少出现一次为止, 求所掷次数 X 的分布与期望.

这类问题统称为 Coupon 问题. 推销装有优惠券 (Coupon) 的卡片的某种方便面, 一套优惠券共 N 张. 在每包方便面中随机地放置一张优惠券卡片. 求买了 n 包方便面仍然没有收集到一整套优惠券卡片的概率, 与收集到一套优惠券卡片时所买方便面的包数的期望.

这等同于下面的放球问题: 往 N 个盒子里放球, 用 X 表示首次没有空盒时所放的球数. $X > n$ 表示投放 n 个球后仍然有空盒. 用 B_i 表示第 i 个盒子空. 那么由容斥定理,

$$\begin{aligned}\mathsf{P}(X > n) &= \mathbb{P}\left(\bigcup_{i=1}^{N} B_i\right) \\ &= \sum_{k=1}^{N} (-1)^{k-1} \binom{N}{k} \mathbb{P}\left(\bigcap_{i=1}^{k} B_i\right)\end{aligned}$$

$$= \sum_{k=1}^{N} (-1)^{k-1} \binom{N}{k} \cdot \frac{(N-k)^n}{N^n}.$$

那么

$$E[X] = \sum_{n \geqslant 0} \mathbb{P}(X > n) = \sum_{k=1}^{N} (-1)^{k-1} \binom{N}{k} \cdot \frac{N}{k}.$$

另外有个直接的方法可以算期望. 令 $\eta_1 = 1$, η_i 是从投放 η_{i-1} 个球之后算起, 等待第 i 个非空盒子出现时投放的球数. 那么 η_i 实际上是服从概率为 $(N-(i-1))/N$ 的几何分布. 而 $X = \sum_{i=1}^{N} \eta_i$, 因此

$$EX = \sum_{i=1}^{N} \frac{N}{N-i+1} = \sum_{k=1}^{N} \frac{N}{k}.$$

这样, 我们证明了一个恒等式

$$\sum_{k=1}^{N} (-1)^{k-1} \binom{N}{k} \cdot \frac{N}{k} = \sum_{k=1}^{N} \frac{N}{k}.$$

你可以试试用初等方法怎么证明这个等式.

练习 4.21　掷一个骰子, 平均要掷多少次才能让所有点数都至少出现一次?

4.5.2　配对数

在配对问题中, 用 X 表示配对数, ξ_i 表示第 i 对夫妇成功配对的指标, 那么 $E[\xi_i] = P(\xi = 1) = 1/n$ 且

$$X = \xi_1 + \cdots + \xi_n.$$

因此

$$E[X] = E[\xi_1] + \cdots + E[\xi_n] = 1.$$

即平均配对数总是 1.

为了计算配对数的方差, 需计算二阶矩

$$\mathbb{E}[X^2] = \sum_{i=1}^{n} E[\xi_i^2] + \sum_{i \neq j} E[\xi_i \xi_j].$$

因为 ξ_i 是 Bernoulli 的, 故 $\xi_i \xi_j$ 也是 Bernoulli 的. 因此 $E[\xi_i^2] = 1/n$, 而当 $i \neq j$ 时,

$$E[\xi_i \xi_j] = P(\xi_i = 1, \xi_j = 1)$$

$$= \mathsf{P}(\xi_j = 1|\xi_i = 1)\mathsf{P}(\xi_i = 1)$$
$$= \frac{1}{n-1} \cdot \frac{1}{n}.$$

这样推出

$$\mathsf{E}[X^2] = n \cdot \frac{1}{n} + n(n-1)\frac{1}{n(n-1)} = 2,$$
$$\mathsf{D}[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 = 1.$$

方差也是恒等于 1.

4.5.3　等待模式出现: 条件期望

重复地掷一个硬币, 依次记录掷出的结果, 1, 0 分别表示正反面, 那么得到一个 01 序列, 一个有限长度的 01 序列称为是一个模式, 例如 (1), (10), (101), (1111), (11101001) 等.

问题: 任意给定一个模式, 它是不是一定在有限时间内出现?

答案是肯定的. 对于一个给定的长度为 k 的模式, 把它看作成功, 成功的概率是 $p = 1/2^k > 0$. 然后把每 k 次掷硬币当作一个随机试验, 独立地重复, 前面已经说过, 有限时间内一定会成功, 而且期望等待时间是 $1/p = 2^k$.

你也许不明白这意味着什么. 一个模式可以很长, 例如一本书可以编译成为一个模式, 一个图书馆里所有的书也可以编译成为一个模式, 所以这结果告诉我们, 只要不断重复地掷硬币, 你肯定会看到这个图书馆的所有书出现.

问题: 求给定模式的平均等待时间.

用 ξ 表示给定模式首次出现的时间. 首先要注意, 如果把模式中的 01 互换得到的模式与原模式是对称的, 平均等待时间也是一样的. 因此我们只需考虑以 1 开始的模式; 其次, 上面这样分组考察模式出现方式不一定是模式首次出现的时间. 例如模式 (101), 在下面的亲测的掷硬币序列中

$$001\,110\,011\,010\,111\,001\,100\,110\,101\,111\,000\,001\,100\,110\,111\,100\cdots,$$

$\xi = 11$, 而若按 3 个一组观察, 那么模式出现的时间是第 9 组, 第 27 次. 所以我们需要有新的方法.

通过算 ξ 的分布来算期望是很复杂的, 这里我们引入一个有用的公式, 类似于全概率公式. 分不同的情况求条件下的期望, 然后平均

$$\mathsf{E}[\xi] = \mathsf{E}[\xi|\Omega_1]\mathsf{P}(\Omega_1) + \cdots + \mathsf{E}[\xi|\Omega_n]\mathsf{P}(\Omega_n),$$

其中符号 $\mathsf{E}[\xi|B]$ 表示事件 B 发生的条件下 ξ 的期望, 即

$$\mathsf{E}[\xi|B] = \frac{\mathsf{E}[\xi 1_B]}{\mathsf{P}(B)},$$

称为 (古典) 条件期望, 是条件概率的直接推广, 因为 $\mathsf{E}[1_A|B] = \mathsf{P}(A|B)$. 这个公式可以继续称为全概率公式, 也可以称为重期望公式.

怎么认识条件期望 $\mathsf{E}[\xi|B]$? 按照定义

$$\mathsf{E}[\xi|B] = \frac{\mathsf{E}[\xi 1_B]}{\mathsf{P}(B)} = \frac{\sum_i x_i \mathsf{P}(\{\xi = x_i\} \cap B)}{\mathsf{P}(B)} = \sum_i x_i \mathsf{P}(\xi = x_i|B),$$

其中 $\mathsf{P}(\xi = x_i|B)$ 就是 B 发生的条件下 ξ 的分布, 称为条件分布, 而条件期望 $\mathsf{E}[\xi|B]$ 是条件分布的期望.

用重期望公式来算期望时间特别直观且方便. 例如 ξ 是等待模式 (1), 那么, 在 1 出现时, 结束, $\xi = 1$, $\mathsf{E}[\xi|1] = 1$, 而在 0 出现时, 我们显然要重新等, $\mathsf{E}[\xi|0] = 1 + \mathsf{E}[\xi]$, 其中 $\mathsf{E}[\xi|1], \mathsf{E}[\xi|0]$ 分别表示 1, 0 出现时的条件期望, 后面的符号是类似的. 因此

$$\mathsf{E}[\xi] = \mathsf{E}[\xi|1]/2 + \mathsf{E}[\xi|0]/2 = 1/2 + 1/2(\mathsf{E}[\xi] + 1),$$

得 $\mathsf{E}[\xi] = 2$.

再 ξ 是等待模式 (10), 先掷一次, 如果是 1, 模式有望, 需继续等待; 如果是 0, 模式无望, 重新开始等待. $\mathsf{E}[\xi] = \mathsf{E}[\xi|1]/2 + \mathsf{E}[\xi|0]/2$, 因为 $\mathsf{E}[\xi|0] = 1 + \mathsf{E}[\xi]$, 所以 $\mathsf{E}[\xi] = 1 + \mathsf{E}[\xi|1]$. 再掷一次,

$$\begin{aligned} \mathsf{E}[\xi|1] &= \mathsf{E}[\xi|11]/2 + \mathsf{E}[\xi|10]/2; \\ \mathsf{E}[\xi|11] &= 1 + \mathsf{E}[\xi|1],\ \mathsf{E}[\xi|10] = 2; \\ \mathsf{E}[\xi|1] &= 3,\ \mathsf{E}[\xi] = 4. \end{aligned}$$

再设 ξ 是等待模式 (101), 反复利用上面的思想, 得方程

$$\mathsf{E}[\xi] = \frac{1}{2}(\mathsf{E}(\xi|1) + \mathsf{E}(\xi|0)) = \frac{1}{2}(\mathsf{E}(\xi|1) + 1 + \mathsf{E}[\xi]),$$

$$\mathsf{E}(\xi|1)=\frac{1}{2}(\mathsf{E}(\xi|10)+\mathsf{E}(\xi|11))=\frac{1}{2}(\mathsf{E}(\xi|10)+1+\mathsf{E}(\xi|1)),$$
$$\mathsf{E}(\xi|10)=\frac{1}{2}(\mathsf{E}(\xi|101)+\mathsf{E}(\xi|100))=\frac{1}{2}(3+3+\mathsf{E}[\xi]).$$

解方程得 $\mathsf{E}[\xi]=10$.

思考: 从这几个答案, 你能看出什么规律吗? 对一般的模式等待问题有没有猜测?

练习 4.22　独立地重复成功概率为 p 的随机试验. 用 X 表示连续 2 次成功发生时试验的次数. 求 $\mathsf{E}[X]$.

练习 4.23　求模式 (1111) 的平均等待时间? 求模式连续 n 个 1 的平均等待时间?

练习 4.24　n 个球, m 个盒子, 每个球等可能放入一个盒子, X 是空盒子个数, 求 $\mathsf{E}[X]$.

4.5.4　信封悖论

两个一样的信封放在桌子上, 里面各有卡片 A,B, 卡片上的数作为奖金数. 嘉宾只知道卡片 B 的数字是卡片 A 的两倍, 不知道具体是多少. 嘉宾甲有机会先选取一个信封, 剩下的那个给乙.

显然, 这是一个抽签问题, 其实两个人抽签, 不管先抽后抽, 他们抽到的结果是同分布的, 当然期望也是一样的, 因此先后并不重要.

但是当甲随机地抓了一个信封, 打开一看, 是 100, 在他打算离开时, 突然有了奇怪的想法. 他认为他拿到大数和小数机会均等, 所以推测另外一个信封有 50 块或者 200 块, 机会相等. 因此期望是 125, 大于 100. 按照这样的想法, 在他抓住信封时, 他感觉另一个信封中的数字的期望比自己拿住的数字的期望大, 这样他就应该选择换个信封, 然后在换了之后, 这个思路依然适合, 他应该选择再换回来. 也就是说, 无论他选择哪个信封, 他都应该选择换个信封. 这显然是个悖论, 那么问题出在哪里?

实际上, 甲看到自己的数字 a 后对剩下的信封中数字的期望是条件期望. 条件期望不一定等于 a, 但显然它不能总 (对所有可能的 a) 是大于 a, 因为那样就会产生悖论. 让我们解释一下为什么.

对于嘉宾来说, 信封问题有两个独立的不确定性: 小的数字 ξ 是不确定的, 以及甲拿的是哪个卡片是不确定的. 分别用 A,B 表示甲拿到卡片 A,B, 再设甲拿到的数字是 X, 剩下的是 Y. 那么认为随机变量 X, Y 真正的关系是, 当 X 等可能是 ξ 或者 2ξ 时, Y 等可能地是 2ξ 或者 ξ. 因此推出它们两个是同分布的, 期望也一样, 与直觉吻

合. 如果对任何 a, $\mathsf{E}(Y|X=a)>a$, 则由重期望公式推出矛盾

$$\mathsf{E}Y=\sum_a \mathsf{E}(Y|X=a)\mathsf{P}(X=a)>\sum_a a\mathsf{P}(X=a)=\mathsf{E}X.$$

甲是怎么想的? 设甲打开所选信封的数是 $a>0$, 即 $X=a$, 现在甲简单地认为 Y 等可能地是 $2a$ 或者 $a/2$, 这导致期望 $\mathsf{E}(Y|X=a)=5a/4>a$. 因此甲的想法肯定是错误的. 错在哪里呢? 错在甲认为条件分布是等可能的. 甲可能是这样想的: $\{X=a,Y=2a\}=\{X=a\}\cap A$, 因此

$$\mathsf{P}(Y=2a|X=a)=\mathsf{P}(A|X=a)=\frac{1}{2},$$

错误仅在最后一个等号, 它是说 A 与 X 独立. 实际情况是 A 与 X 不独立而与 ξ 独立且 $\{X=a,Y=2a\}=\{\xi=a\}\cap A$, 其中 ξ 被假设是个随机变量. 所以正确的条件分布为

$$\mathsf{P}(Y=2a|X=a)=\frac{\mathsf{P}(Y=2a,X=a)}{\mathsf{P}(X=a)}=\frac{\mathsf{P}(\xi=a)}{\mathsf{P}(\xi=a)+\mathsf{P}(\xi=a/2)};$$
$$\mathsf{P}(Y=a/2|X=a)=\frac{\mathsf{P}(Y=a/2,X=a)}{\mathsf{P}(X=a)}=\frac{\mathsf{P}(\xi=a/2)}{\mathsf{P}(\xi=a)+\mathsf{P}(\xi=a/2)}.$$

4.6　有界, 有限, 几乎肯定有限与期望有限

随机变量是以样本空间为定义域的一个函数

$$X:\Omega\longrightarrow\mathbf{R}.$$

这与中学学的函数稍有不同, 那里函数 $y=f(x)$ 的自变量 x 与因变量 y 都是实数, 而随机变量的自变量是样本空间的元素, 但取值是有限数. 因此随机变量是用数字来标识结果.

下面我们来谈谈有限和有界的区别. 前面几个例子中的随机变量取值都是非负整数, 但是它们有区别, 成功次数和配对数这两个随机变量是有界的, 也就是说取值是在一个有界的范围内. 但是等待成功时间 X 不一样, 按照定义我们无法马上判断它们的取值是有限的, 因为样本空间里有一个元素是一直失败, 它对应的 X 取值是 ∞. 如果 $\mathsf{P}(X=\infty)=0$, 那么 X 几乎肯定是有限的, 在概率论中, 我们不去区分肯定和几乎肯定了. 这样的 X 也当作随机变量了.

看前面对待成功时间 X, $P(X=n)=pq^{n-1}$, $n\geqslant 1$, 那么 $P(X<\infty)=\sum_n pq^{n-1}=1$, 且 $EX=1/p<\infty$, 所以它是有限的且期望也有限. 回顾第三章的格点轨道首次返回时 T, 它的分布是

$$P(T=2n)=\frac{C_{2n}^n}{(2n-1)2^{2n}},$$

已知 $P(T=\infty)=0$. 因此首次返回时是有限的. 它的期望是

$$E[T]=\sum_{n\geqslant 1}2nP(T=2n).$$

这个级数收敛吗? 通过 Stirling 公式 $n!\sim\sqrt{2\pi n}(n/e)^n$ 估计,

$$2nP(T=2n)=\frac{2n}{2n-1}\frac{(2n)!}{(n!)^2 2^{2n}}\sim\frac{\sqrt{2\pi 2n}(2n/e)^{2n}}{2\pi n(n/e)^{2n}2^{2n}}=\frac{1}{\sqrt{\pi n}},$$

最后得 $ET=\infty$. 因此 T 是有限的但期望无限.

4.6.1　两个简单公式

为了讨论下一个问题, 我们给出两个简单公式. 假设 X 是 Ω 上的取值为非负整数或者无穷的函数, 我们把它叫做广义的随机变量. 我们来看 X 的尾部概率 $P(X\geqslant n)$ 的极限行为.

命题 4.6.1

$$P(X=\infty)=\lim_n P(X\geqslant n). \tag{4.6.1}$$

前面第三章公理化概率论一节中证明过递增事件列的连续性, 注意到 $\{X<\infty\}=\bigcup_n\{X\leqslant n\}=\bigcup_n\{X<n\}$, 故而

$$P(X<\infty)=\lim_n P(X\leqslant n)=\lim_n P(X<n).$$

反过来必然有

$$\begin{aligned}P(X=\infty)&=1-P(X<\infty)=\lim_n(1-P(X<n))\\&=\lim_n P(X\geqslant n)=\lim_n P(X>n).\end{aligned}$$

对于等待成功的时间 X, 但是 $P(X\geqslant n)=q^{n-1}$, 所以 $P(X=\infty)=\lim_n P(X\geqslant n)=0$. 实际上对于一个广义的随机变量, 它的期望可以有如下表示.

命题 4.6.2

$$\mathsf{E}[X] = \sum_{n \geqslant 1} \mathsf{P}(X \geqslant n). \tag{4.6.2}$$

从期望的定义出发

$$\begin{aligned}\mathsf{E}[X] &= \sum_{n \geqslant 1} n\mathsf{P}(X = n)\\ &= \mathsf{P}(X = 1)\\ &+ \mathsf{P}(X = 2) + \mathsf{P}(X = 2)\\ &+ \mathsf{P}(X = 3) + \mathsf{P}(X = 3) + \mathsf{P}(X = 3)\\ &+ \cdots\cdots .\end{aligned}$$

很显然, 第一列是 $\mathsf{P}(X \geqslant 1)$, 第二列是 $\mathsf{P}(X \geqslant 2)$, 第三列是 $\mathsf{P}(X \geqslant 3)$, 推出结论. 把这个公式应用于等待成功时间

$$\mathsf{E}[X] = \sum_{n \geqslant 1} \mathsf{P}(X \geqslant n) = \sum_{n \geqslant 1} q^{n-1} = \frac{1}{p}.$$

这两个命题说 X 有限或者期望有限由尾部概率 $\mathsf{P}(X \geqslant n)$ 趋于零及其速度决定. 因为级数收敛蕴含通项趋于零, 所以期望有限蕴含 X 本身是 (几乎肯定) 有限的. 反之不然.

4.6.2　等待更高报价与对称性

问题: 某人在网上发布卖一个自行车, 记下第一个报价, 然后决定卖给下一个报价高于此报价的人, 问等待时间是否有限? 如果有限, 期望等待时间是多久?
为了用概率来回答问题. 我们假设报价依次为

$$X_0, X_1, X_2, \cdots, X_n, \cdots,$$

可以假设它们是独立同分布且分布在正整数上的随机变量. 定义

$$N := \inf\{n : \ X_n > X_0\},$$

即报价首次高于第一个报价的人的序号, 其中的高于等价于数学中的大于. 这时 N 是一个广义随机变量, 我们来看看它是否是随机变量, 如果是, 再看看期望是否有限.

我们非常严格地来分析一下, 首先, 当 $n > 1$ 时, 因为 $N \geqslant n$ 实际上说前 $n-1$ 个报价都不超过 X_0, 所以

$$\{N \geqslant n\} = \{X_0 \geqslant X_1\} \cap \cdots \cap \{X_0 \geqslant X_{n-1}\}.$$

因此,

$$\begin{aligned} P(N \geqslant n) &= P(\bigcap_{k=1}^{n-1} \{X_0 \geqslant X_k\}) \\ &= \sum_j P(\bigcap_{k=1}^{n-1} \{j \geqslant X_k\} \cap \{X_0 = j\}) \\ &= \sum_j P(X_0 \leqslant j)^{n-1} P(X_0 = j), \end{aligned}$$

最后一个等号成立是因为同分布.

练习 4.25 (*) 右边这个东西当 n 趋于无穷时的极限是不是零? 如果对任何 j, $P(X_0 \leqslant j) < 1$, 那么 $\lim_n P(N \geqslant n) = 0$. 否则, 存在正整数 J 使得 $P(X_0 \leqslant J) = 1$ 且 $P(X_0 = J) > 0$, 那么 $P(N = \infty) = \lim_n P(N \geqslant n) = P(X_0 = J) > 0$.

练习 4.26 (*) 在前一种情况下计算期望

$$\begin{aligned} E[N] &= \sum_{n \geqslant 1} P(N \geqslant n) \\ &= \sum_j \sum_{n \geqslant 1} P(X_0 \leqslant j)^{n-1} P(X_0 = j) \\ &= \sum_j \frac{P(X_0 = j)}{P(X_0 > j)}. \end{aligned}$$

这个级数是收敛还是发散?

实际上这两个问题对于大多数同学来说都是难题, 因此只是留给大家思考. 说第一个问题难是因为其中涉及到极限交换, 是数学中最难的问题.

换个思路, 由于 $N \geqslant n$ 相当于 $X_1, \cdots, X_{n-1}$ 都没超过 X_0, 换句话说 X_0 到 X_{n-1} 中 X_0 是最大的 (之一, 因为最大不一定唯一). 把这个事件记为 A_0. 类似地, 用 A_i 表示 X_0 到 X_{n-1} 中 X_i 是最大的, $i < n$. 那么这些事件概率一样 (因为报价是独立同分布的), 且至少有一个会发生的, 但可能会同时发生. 但次可加性总是成立的

$$\sum_{i=0}^{n-1} P(A_i) \geqslant P(\bigcup_i A_i) = 1.$$

推出 $\mathsf{P}(A_0) \geqslant 1/n$. 因此

$$\mathsf{E}[N] = \sum_{n \geqslant 1} \mathsf{P}(N \geqslant n) \geqslant \sum_{n \geqslant 1} \frac{1}{n} = \infty.$$

也就是说, 该问题中的期望等待时间总是无穷, 但不能回答等待时间是否是 (几乎肯定) 有限的.

奇妙的是当把问题中的高于 ($>$) 换成不低于 ($\geqslant$) 时, 问题的答案会有奇妙的变化. 用 N' 表示等待的时间, 即

$$N' = \inf\{n \geqslant 0 : X_n \geqslant X_0\}.$$

显然 N' 会早于 N, 所以如果 N 是有限的, 那么 N' 也是有限的. 但前面我们只证明了 N 是期望无限的, 这无法给我们关于 N' 的信息. 且 $N' \geqslant n$ 相当于 $X_0 > X_k$, $0 < k < n$. 仿照上面的方法

$$\mathsf{P}(N' \geqslant n) = \sum_{j} \mathsf{P}(X_0 < j)^{n-1} \mathsf{P}(X_0 = j).$$

因为当 $\mathsf{P}(X_0 = j) > 0$ 时, $\mathsf{P}(X_0 < j) < 1$, 故 $\lim_n \mathsf{P}(X_0 < j)^{n-1} = 0$. 这能推出 $\mathsf{P}(N' \geqslant n)$ 的极限是零吗? 严格地说不能.

换个方法, 用 B_k 表示事件 "$X_0, \cdots, X_{n-1}$ 中 X_k 比其他任何一个大", $0 \leqslant k < n$. 那么这些事件互斥, 由对称性它们的概率一样. 由可加性

$$\sum_{k=0}^{n-1} \mathsf{P}(B_k) = \mathsf{P}(\bigcup_{k} B_k) \leqslant 1,$$

其中最后为什么是 $\leqslant 1$? 因为这些事件的并不一定发生. 由此推出不等式 $\mathsf{P}(N' \geqslant n) = \mathsf{P}(B_0) \leqslant 1/n$, 即得 $\mathsf{P}(N' = \infty) = 0$, 但不能回答期望是否有限. 怎么办? 只能回到原始的公式

$$\begin{aligned}
\mathsf{E}[N'] &= \sum_{n} \mathsf{P}(N' \geqslant n) \\
&= \sum_{n} \sum_{j} \mathsf{P}(X_0 = j, X_1 < j, \cdots, X_{n-1} < j) \\
&= \sum_{j} \sum_{n} \mathsf{P}(X_0 < j)^{n-1} \mathsf{P}(X_0 = j) = \sum_{j} \frac{\mathsf{P}(X_0 = j)}{\mathsf{P}(X_0 \geqslant j)}.
\end{aligned}$$

这个级数实际上是发散的, 但需要较高的数学素养才能说清楚.

练习 4.27 (*) 这实际上等价于下面的级数问题, 如果 $a_n > 0$ 且 $\sum_n a_n < \infty$, 那么

$$\sum_n \frac{a_n}{\sum_{j \geqslant n} a_j} = \infty.$$

其中的分母 $\sum_{j \geqslant n}$ 改为 $\sum_{j > n}$ 结论也成立.

练习 4.28 依次记下报价, 决定卖给下一个报价高于 (不低于) 他/她之前报价的人, 问等待时间是否有限? 如果有限, 期望等待时间是多久?

第五章　概率论与现实世界

尽管我们说数学很有用, 但在实际生活中真正用到加减乘除四则运算之外的数学的地方并不多. 概率论在生活中是不是真的有用呢? 前面章节中的例子大部分是数学的, 离现实远了一点. 在本章中, 我们通过一些与现实更接近的例子来理解随机现象, 以及概率论方法怎么解释和解决实际问题.

5.1　大数定律与统计推断

Bernoulli 的大数定律是革命性的, 即使现在回过头看, 从 Fermat, Pascal 通信的那年, 即 Bernoulli 诞生的 1654 年到他去世的 1705 年, 概率还没有理论, 仅有一些零星讨论的问题, 想象一个被严格证明的概率论奠基性的定理产生在那个时代, 就像沙漠里开出了一朵鲜艳的花. 如果要问为什么, 那就是天才催生了它.

5.1.1　大数定律: Bernoulli 的黄金定理

对于一个随机现象来说, 结果是不可预测的, 结果发生的概率是可以知道的. 例如掷一次硬币得正面的概率比掷骰子得点 6 的概率要大不少, 但这不能告诉你正面或者 6 点会不会出现, 结果依然是随机的, 不可预知. 这样的话, 概率除了告诉我们可能性大一点或者小一点之外, 概率的数值究竟有什么意义呢? 公理的概率与实际的随机性问题会在哪里相遇呢?

这个数值会在多次试验时体现出直观的意义, 也就是说成功的频率, 成功次数与试验次数之比, 逼近成功概率. 这个规律是如此自然直观, 历史上肯定有很多人意识到. 16 世纪意大利著名数学家 G. Cardano (1501-1575), 也是一个赌徒, 对诸多的概率问题有所讨论并写在他的著作 The book of Games of Chance. 对于频率与概率的问题, 他大概是这么说的: when the probability for an event is p then by a large

number n of repetitions the number of times it will occur does not lie far from the value np. 如果一个事件的概率是 p, 那么重复次数 n 很大之后, 它发生的次数不会远离 np. 从这本书的内容看, Cardano 对概率论有重要的贡献, 也许不能说是决定性的. 但因为种种原因, Cardano 的这本书出版于 1663 年, 而且并没有引起很大的反响.

Jakob Bernoulli 在 1713 年出版的《猜度术》一书中精确地阐述并证明这个规律, 在书中, Bernoulli 说这个思想实际上出现在 20 年前, 算起来应该是 1685 年左右. 他自己显然非常看重这个结果, 称之为 Golden Theorem, 认为它在价值上超过了书中其他内容. 这个定理现在通常称为大数定律, 是由 Poisson 命名的, 大概是说它是可以由实验验证的规律. 奇妙的是, 在没有任何提示的时代, Bernoulli 是怎么找到定理的正确表达形式的?

尽管一些科学的发现源于纯粹的兴趣, 但大数定律却是一个本身有重要应用背景的问题, Bernoulli 在证明之前认真地解释了问题的实际意义. 我们从《猜度术》第四部分引用几句话, 而且把英文附在这里, 避免我们的翻译词不达意. 首先, Bernoulli 认为在生活中确定的现象是很少的, 很多情况是随机的, 我们需要计算某个事件发生的确定性程度或者某个证据的力度, 他称之为概率. 其次, 如在讲义序中所引用的一段话, 他认为在实际生活中, 只有在机会游戏 (骰子, 摸球, 抽签等) 中的随机现象中可以确切计算概率, 这样的现象是很少的. 大多数现象都无法像机会游戏那样用等可能假设来计算概率. 这段话就在讲义序的第一段中. 最后, Bernoulli 阐述了他的方法. Since this and the like depends on absolutely hidden causes, and, in addition, owing to the innumerable variety of their combinations always escapes our diligence, it would be an obvious folly to wish to find something out in this manner. Here, however, another way for attaining the desired is really opening for us. And, what we are not given to derive a priori, we at least can obtain a posteriori, that is, can extract it from a repeated observation of the results of similar examples. **大意: 因为这 (瘟疫产生和感染) 依赖于完全隐藏且变化繁杂的原因, 想要发现事件概率显然是愚蠢的. 但是, 得到所需概率的另一种方法就在这里, 也许我们不能得到先验的概率, 但我们可以得到后验的概率, 就是从重复试验中获取.**

It certainly remains to inquire whether, when the number of observations thus increases, the probability of attaining the real ratio between the number of cases, in which some event can occur or not, continually augments so that it finally exceeds

any given degree of certitude. 首先, 他说: ‘‘当然我们也要问当观察数增加时, 得到真实比例 (概率) 的概率是不是会持续地增加以至于超过任何给定的百分数. ” Bernoulli 还说, 如果这得不到保证, 那么这个问题 (重复观察估计概率) 就没有意义了. 在说明问题的意义之后, 他开始阐述这个问题的细节, 差不多就是定理的形式. 他说: 应该注意到我们想要用实验确定 (估计) 的概率不需要也做不到完全精确和严格, 但是可以要求一定的精度, 即让它包含在某个你所要求的范围内. 实际上, 在上面的小球例子中, 3:2 是袋子中黑白小球数的真实比例. 我们将假设两个比例, 301:200 与 299:200, 3001:2000 与 2999:2000, 等等, 一个很近但比 3:2 大一点, 另一个很近但比 3:2 小一点, 我们将证明, 可以增加观察数使得从观察得到的比例落在真实概率的这个范围内的可能性比任何给定的概率更大.”

最后一句话是关键, 观察数是 n, 观察得到的比例实际上就是频率, 他甚至说可以具体告诉你让 n 多大, 才能使得使得频率落在真实概率的某个范围内的可能性多么地接近 1. 通过这样的叙述, Bernoulli 在慢慢地接近问题的核心, 去掉一些具体的无关本质的东西, 把一个日常的经验概括为一个数学上可以验证的问题. 但是, Bernoulli 意识到, 尽管每个人都知道这个事实, 但是数学上的严格证明是必需的. 另外他有点自嘲地说, 他可能是做一件吃力不讨好的事情.

具体是怎么叙述的呢? 独立地重复一个成功概率为 p 的随机试验 n 次, 用 S_n 表示成功的次数, 那么 S_n/n 是成功的频率, 是一个随机变量. 它的期望为 $E[S_n/n] = np/n = p$, 方差为

$$D[S_n/n] = E[(S_n/n - p)^2] = \frac{npq}{n^2} = \frac{pq}{n} \leqslant \frac{1}{4n}.$$

现代概率论教材中的大数定律通常是这样叙述的.

Bernoulli 大数定律. 对任何的正数 ε, δ, 当 n 成分大时

$$P\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) < \delta.$$

这是说成功的频率 (经验概率) 与成功的概率之误差超过 ε 的可能性随着试验次数无限增大而趋向于零, 即概率的意义会在重复试验时体现, 也就是概率的科学价值.

看到这个定理, 至少有两点让人震惊, 第一是 Bernoulli 的非凡洞察力. 要知道频率的极限是概率这件事情是直观的, 那个时代及之前的许许多多智者可能都曾经想过, 但是想过这个问题和把问题精确地描述成一个可以验证的命题是完全不同的两件事情, 生活中有许多看似平凡的东西实际上却蕴含深刻的哲理, 能够把这些东西抽象出

来刻画成可读的命题绝对是天才级的洞察力. 概率历史上有不少这样的例子, 但大数定律是最早最重要的一个. 第二是数学方面. 要知道 17 世纪末期微积分刚刚诞生, 那时有导数和积分的概念, 但没有极限, 极限的概念成型是在 150 年之后的 19 世纪中叶. 再看 1685 年的概率论, 已经有差不多有半个世纪的历史, 但是研究的多半是一些具体的例子, 概率的处理技术基本是空白. 因此在那个时候把一个涉及在某个测度之下的函数列收敛的命题讲清楚并严格地证出来真的如同神迹.

5.1.2　大数定律的证明

Bernoulli 的原始证明主要是通过考察二项分布相邻两项的比例完成的, 较繁琐冗长, 且只适用于 p 是有理数的场合, 但方法初等且有原始证明的魅力. 这里介绍 19 世纪后半叶俄罗斯数学家 Chebyshev 的简单证明.

设 X 是一个随机变量, ε 是个正实数, 则有

Chebyshev **不等式**:

$$P(|X| \geqslant \varepsilon) \leqslant \frac{E[X^2]}{\varepsilon^2}.$$

证明的关键是应用期望的单调性. 因为

$$X^2 \geqslant X^2 1_{\{|X|\geqslant\varepsilon\}} \geqslant \varepsilon^2 1_{\{|X|\geqslant\varepsilon\}},$$

注意这是随机变量或者函数的比较, 所以

$$E[X^2] \geqslant E[\varepsilon^2 1_{\{|X|\geqslant\varepsilon\}}] = \varepsilon^2 P(|X| \geqslant \varepsilon),$$

因此推出所要证明的不等式.

应用 Chebyshev 不等式于频率与概率之差 $S_n/n - p$, 由二项分布的方差计算得

$$P\left(|S_n/n - p| > \varepsilon\right) \leqslant \frac{1}{\varepsilon^2} E\left[(S_n/n - p)^2\right] = \frac{pq}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}.$$

当 n 成分大时, 它小于任何正数. 有趣的是, Chebyshev 的证明与具体的分布无关, 而 Bernoulli 的原始证明只能用于二项分布.

前面我们说过, 可能性大小是被认为可以加的, 所以概率可加性是公理的一部分. 如果你继续追问概率为什么可加, 那可能很难得到令人信服的回答, 只能归因于人类的直觉, 但这个直觉不是虚无缥缈的, 它现在可以通过大数定律来解释. 频率的极限是概率, 频率有可加性: 在重复随机试验时, 两个互斥的事件至少有一个发生的频率等于各自发生的频率之和. 因此概率也应该有可加性.

Bernoulli 大数定律的推广形式如下: 一个随机试验中有个随机变量 X, 独立地重复随机试验, 观察此随机变量, 得到独立同分布的随机序列

$$X_1, X_2, X_3, \cdots, X_n, \cdots.$$

用 S_n 表示前 n 项的和, 那么 S_n/n 就是前 n 项的平均, 称为样本平均.
大数定律: 当 n 趋于无穷时, 样本平均趋于平均 (即期望):

$$\frac{S_n}{n} \longrightarrow E[X].$$

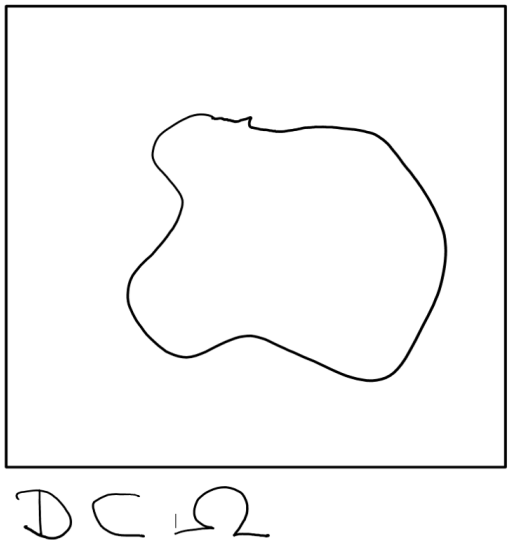
上面所说的收敛实际上是依概率收敛, 概率论中的随机变量序列有很多种不同的收敛, 另外一种重要的收敛是几乎肯定收敛, 也就是说, 不收敛这个事件的概率为零. 几乎肯定收敛比依概率收敛要强大, 在方差有限的条件下, 实际上有 S_n/n 几乎处处收敛于 $E[X]$, 称为强大数定律. 非概率专业的读者可能无法区分这两种收敛的不同, 可以忽略.
例如掷一个骰子, 骰子点数 X 的数学期望为

$$E[X] = \frac{1+2+3+4+5+6}{6} = 3.5,$$

那么掷 n 次骰子, 骰子点数的平均数差不多是 3.5.

5.1.3　蒙特卡洛算法

大数定律还给了我们一种新的算法, 称为蒙特卡洛算法. 假设我们要算平面上一个不规则区域 D (例如某个国家在地图上所占的区域) 的面积, 这样的图形用通常的面积计算方法是不容易求得的.
问题: 怎么应用蒙特卡洛算法呢?
看个简单的例子, 取一个正方形, 记为 Ω, 把所求的区域 D 放在此正方形中, 如下图所示



然后把这个包含区域 D 的正方形放置在墙上, 向它随意投掷 200 次飞镖, 记录到飞镖落在 D 中的次数有 82 次.

因为是随机投飞镖, 所以飞镖落在 D 中的概率等于面积比

$$\mathrm{p} = \frac{|\mathrm{D}|}{|\Omega|},$$

其中 |D| 与 |Ω| 分别表示两个区域的面积 (这是另一种古典概率：称为几何概率模型). 现在飞镖落在 D 中的频率为

$$\widehat{\mathrm{p}} = \frac{82}{200} = \frac{41}{100}.$$

按大数定律, 应有 $\widehat{\mathrm{p}} \approx \mathrm{p}$, 因此

$$|\mathrm{D}| \approx \frac{41}{100}|\Omega|,$$

其中正方形 Ω 的面积已知或者是容易计算的.

上面例子中所示的思想就是蒙特卡洛算法, 也称为随机化算法. 该算法的优点是实现简单, 但缺点是难以估计和控制误差, 因为大数定律告诉我们的是, 在试验次数很大时, 误差大于某个数的概率变得很小, 不能保证肯定地变得很小. 法国数学家蒲丰在 1777 年曾经设计过投针实验来计算圆周率 π 的近似值, 相关阅读留作习题.

通过频率来估计概率是统计的基本方法. 有时即使抽样不是独立同分布的, 我们也近似地使用蒙特卡洛算法来进行估计.

问题: 一个大水塘, 不把水抽干, 怎么来估计水塘里大概有多少鱼?

设水塘里共有 n 条鱼, 先从水塘里打捞出 1000 条鱼, 涂上一点不会掉落的红色, 然后放回水塘. 那么水塘里有红色的鱼的比例是 $1000/n$. 过几天再从水塘里打捞出 1000 条鱼, 查看有多少是红色的, 例如 150 条, 那么 $\frac{150}{1000}$ 就是水塘中有红色的鱼的比例的一个估计, 即

$$\frac{1000}{n} \approx \frac{150}{1000},$$

因此

$$n \approx \frac{1000^2}{150} \approx 6667.$$

5.1.4　统计推断

统计推断是根据一些观察来推断一个随机现象或者一个数据体中的规律. 统计推断属于归纳推理, 即从特殊归纳到一般. 统计推断的通常表现形式是推断分布或者分布中的参数. 例如通过掷硬币来推断硬币是不是真的两面等可能, 通过分析以往的数据来推断之后几天的天气情况, 通过以往的数据来推断股市的涨跌, 等等.

问题: 为什么要进行统计推断?

1713 年出版的 J. Bernoulli 的猜度术, The art of conjecture, 是概率统计领域最早的著作, 其中, 他证明了概率论中最基本的定理: 大数定律, 它也是统计的基础. 在这个著作中, Bernoulli 论述了为什么要进行统计推断.

在前一章证明了 -给定场合数, 其中有利于某事件的论点可能存在也可能不存在, 有证据也可能没有, 或者甚至证明了相反的结论- 它们所证明的力量和力量成比例的事件的概率能够通过计算进行推导和估计. 因此我们明白, 为了准确地猜度某件事情, 只需要精确计算场合数且求出其中某些场合怎么比另外的场合更容易发生. 这里, 我们会碰到困难, 因为这样的计算极少会成功, 除了机会游戏, 那里制作游戏的人会想方设法地设计输赢的场合数, 使得游戏是公平的.

可是, 对于大多数其他情况, 依赖于自然的产生以及人类的自由意志, 这样的计算不会成功. 普通人谁能确定, 例如, 疾病的数目, 或者在哪个年龄, 哪个疾病会侵入人的无数器官中的哪一个而导致人的死亡? 一个疾病 (例如瘟疫) 杀死人比另外一个疾病 (例如狂犬病, 或者狂犬病与发烧) 杀死人容易多少, 以至于我们可以猜度生死的未来状态? 谁能够数清楚每天空气经历的不可数的变化然后猜度它在一个月后的状

态, 更不要说一年后? 谁能够足够清楚地知道人类心灵或者身体的构成以至于敢于说可以确定在一个游戏中的参与者最终的胜负呢?

因为这样或那样的事情依赖于绝对隐藏的原因, 并且这些原因的无穷多组合变化总是能够逃过我们的努力, 所以在这样的事情上想要用数场合数的方式找出概率显然是愚蠢的. 但是, 还有另外一个得到结果的方法, 就是从类似例子的重复观察中获取, 由此, 虽然我们不能够得到先验的 (priori), 但至少可以得到后验的 (posteriori). 每个现象可能出现或者不出现, 观察的时候可能发生也可能不发生. 例如, 前面提到观察 300 个同 Titius 一样年龄和健康状态的男人, 其中 200 个在 10 年后去世了, 其他人仍然活着, 我们可以以足够的置信度断言 Titius 在十年之内去世的可能性是活着的可能性的两倍. 再者, 如果某人观察天气许多年, 注意到下雨或者晴天的天数, 或者经常出现在两个人进行的比赛注意到谁输谁赢, 那么他就能够发现将来, 在类似的情况下, 某个事件发生的比例会相似于过去.

由实验来确定场合数的经验方法不是新的. 每个人都会知道要做这样的推断, 基于一两次的观察是不够的, 需要很大数量的观察. 甚至最愚蠢的人, 按照其自然的直觉, 无需特别的指导, 就能够感觉到观察越多, 结论就越可靠.

要注意的是, 统计推断的结论依然是不确定的, 但这不是说推断是没有意义的. 关于这个问题, 统计学家 Fisher 是这么说的: We may at once admit that any inference from the particular to the general must be attended by some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of uncertainty may itself be capable of rigorous expression. 我们要承认任何从特别到一般的推断是有不确定性存在的, 但这不等于说这样的推断是不能绝对地严密的, 因为不确定的质与度本身是可以严密地表述的. 直到现在, 很多学者是不承认且极其反对归纳推理的, 认为归纳推理不严密, 可能因为滥用而导致严重的谬误, 但不可否认, 归纳推理推动了科学的发展, 是一种值得学习的推理方式, 同样不可否认, 像学者所担心的那样, 统计推断时常被严重滥用, 或者容易有意或者无意地被滥用. 甚至很多统计学者也这么认为. 因此在涉及写作和阅读一个统计推断时要特别小心, 作为一个统计学者, 你需要尽量准确地表述其中的不确定性的质与度, 分清楚其中的科学成分与统计成分. 而作为一个阅读者, 你需要尽量去理解结论所表述的不确定性的质与度, 理解其中的科学 (确定) 成分与统计 (不确定) 成分, 这样就可能避免一些现在所说的 ‘‘无意地被滥用’’ 现象: 只要有数据, 人们总是可以使用统计方法来得到一些貌似科学但其实荒谬的结论.

5.1.5　大样本随机双盲试验

下文是作者为上海高中教材《数学必修第三册》中第 12 章概率初步所写的一个课后阅读, 这是概率统计在生物医学领域的一个重要应用.

在医学上, 判断一种药物对治疗某种疾病有效, 是个严肃的事情, 不是可以随意下结论的.

问题: 怎么验证一个药是有效的呢?

从专业的角度说, 要先从病理和药理入手进行分析与研究, 说明该药物治疗该疾病理论上是成立的. 但即使理论上无懈可击, 药物在上市之前需进行严格的临床试验, 应用概率中的随机化思想检验其有效性.

首先, 需要找病人做试验, 看服用这个药的效果, 称为抽样. 大数定律告诉我们, 试验的人越多, 结论就越可靠. 这是对样本数量的要求, 称为大样本. 其次, 参加试验的病人须通过一个精心设计的程序来选定, 称为随机化, 其目的是保证药物的试验对象在病人群体中有代表性, 不偏向于某个特定人群 (例如年龄, 性别, 身体状况, 病情严重程度等等), 以减小对药效估计的偏差. 这是对样本质量的要求.

最后, 不能只看服用该药病人的情况, 例如服用某种药一周之后, 90% 以上的病人都康复了, 是不是可以说这个药有效呢? 答案是否定的, 这是因为动物先天具有的免疫能力使很多疾病能够自愈. 例如, 若我们观察到未服用该药的病人中 90% 多在一周之后也康复了, 则就不能证明该药是病人康复的原因. 因此我们不仅需要观察服用该药的病人, 也需要观察没有服用该药的病人. 但是为了防止因为服用该药以及没有服用该药所带来的心理差异对于疾病的影响 (医学界尚无法确定心理因素对疾病的影响), 临床上设计了称为双盲的方法: 让参加试验的一部分病人服用该药, 另外一部分病人服用外观及口感和真实药物一样的, 但不含任何药物成分的替代品, 医学上俗称为安慰剂的东西, 其中谁服用药物谁服用安慰剂的安排同样是由一个精心设计的随机化程序决定的, 其目的是使两组结果更可比, 减少混杂因素导致的偏差. 这个安排对病人和医生都是保密的, 故称为双盲. 为了排除其他药物干扰, 在整个试验过程中不使用任何其他药物.

现在, 对于这个试验群体, 用 A 表示该群体中服用药物的病人组, B 表示试验结束后该群体中达到某种治疗效果 (包括药物安全性) 的病人组. A, B 两组符合得好, 就说明该药物有效, 说明两组符合得好的关键指标是两个比值 (即频率或者经验概率): 1. 服用该药物的病人中达到疗效的比值; 2. 达到疗效的病人中服用该药物的比值. 在两个比值都充分大 (超过预先设定的标准) 时, A 与 B 两组符合得就比较好, 也就可

以断言该药物对治疗该疾病是有效果的, 而且具体效果也可以通过这两个比值来进行量化.

上述试验可称之为大样本随机双盲试验, 简称为双盲试验. 每一种药物在上市使用之前都需要通过双盲试验, 才能说明它在某种设定的科学标准下是安全有效的.

每个人都是在某种文化中成长的, 文化会潜移默化地影响人的思考, 这样所得到的结论多少会有主观意志的烙印. 双盲试验是一种科学方法, 它的要点是剔除分析问题过程中的主观因素, 不仅用于医学, 还可用于检验其他各种真假难辨的理论, 让许多有意无意的骗术现形, 是伪科学的克星. 因此, 双盲试验被认为是最能提高每个人认知能力的科学概念, 是基本科学素养.

5.1.6　平均与普遍

前面我们关注的是数学, 概率, 随机变量及其分布, 期望方差等, 它们怎么和现实世界联系起来, 怎么应用? 实际上, 概率的一个重要应用是统计, 也就是用概率的思想来分析数据. 先简单说说统计中的几个概念, 高中数学中有这些内容. 在实际的问题中, 我们假设具有某种共同模式的数据是从某个随机现象中抽取的具有给定随机性的量, 例如人的体重身高, 收入, 消费情况等. 这实际上是奇怪的假设, 也是必须的假设. 这时候, 这个量, 随机变量或者其分布, 统计中也称为总体, 统计的意义在于我们需要通过数据了解它的分布情况也就是随机性. 所谓数据就是从总体中抽取的样本, 用概率的语言说, 就是独立地重复一个随机试验得到的结果, 独立地重复意味着样本应该有代表性和不相关性. 抽取的样本个数称为样本容量. 简单地说, 数据是某个随机变量的独立拷贝, 这里重复和拷贝是指同分布. 所谓数据, 既可以认为是随机变量也可以代入具体的数值, 例如你可以说路人甲的身高是一个变量 X_1 或者是一个字母代表的数值 x_1 或者干脆是具体的数值 175 厘米.

期望在统计中对应于样本平均, 要注意的是, 这是两个不同的概念, 期望是指总体分布的期望, 而样本平均是样本数据的平均, 可以通过数据计算. 因为大数定律保证样本平均的极限是总体的期望, 所以我们用样本平均来估计期望, 在实际使用时, 可能不区分期望和样本平均. 类似的, 我们用数据获得总体的经验分布来估计 (代替) 总体的分布函数. 至于经验分布为什么可以代替总体分布, 这一节的最后小节会有解释. 举个例子. 在每个区域里统计居民收入, 当然最好的办法是看每一户居民的收入情况, 或者收入分布函数 $F(x) = P(X \leqslant x)$, 这里 X 是收入, $F(x)$ 是收入不超过 x 的概率. 对应的样本收入分布是 $\widehat{F}(x)$, 它是收入 (可以将收入规范在 0 与 1 之间) 不超

过 x 的人数比例. [1] 可以想象, 一个社会的低收入群体通常比例高, 所以这个函数曲线一开始是陡峭然后平缓, 且贫富悬殊的程度体现在陡峭程度.

实际情况体现在其分布函数上, 但大众能够理解的是简化的统计数字, 最简单的是平均收入, 大概能够说明这个区域居民的财富情况了. 通常来说这是对的, 但也很可能出问题, 例如这区域 99% 的户均财富在 10 万以下, 但有 1% 的居民很富有, 财富超过 10 亿, 那么平均就差不多是 1000 万了, 会严重地扭曲区域居民财富的真实情况. 也就是说财富差距很大的时候, 平均不一定是个好的指标, 预期不能反映真实结果, 平均不具有普遍性.

这时, 另外一个指标也很重要, 那就是中位数. 在概率中, 中位数是总体的分布函数等于 1/2 的地方, 在统计中, 样本中位数理解为把从总体抽取的数据按小到大排列之后那个位于中间位置的数据, 也就是样本分布函数等于 1/2 的那个地方. 可以证明样本中位数的极限是中位数. 中位数从一定程度上体现了普遍性, 它几乎不受极端数据的影响. 在很多情况下, 平均和中位数相差不大, 例如均匀分布或者正态分布的时候, 两者是一样的. 但是也有相差很大的时候, 那通常是分布极端不均匀或者极端不对称的时候.

在现在这个问题中, 那就是把居民财富按顺序排列之后位于中间的数字, 即中位数收入. 在以上特例中, 那就是 10 万元, 它差不多反映大部分人的真实情况, 但是它没有反映出富裕群体的情况. 如果要描述小区的收入水平, 可以说这个小区平均收入千万, 但普遍收入十万上下, 可以说这个区域普遍不富裕, 平均很富裕. 因此通过简单的统计数字不一定能看到真实情况, 我们应该千万小心.

实际上, 平均和中位数两个指标各有各的意义, 各有各的适用范围, 不能说哪个更好或者更不好. 例如在一个公司的劳资双方谈判时, 公司的老板说员工的平均工资已经达到 5 万年薪, 已经很高了, 但工会方说大部分工人的年薪才是 1 万多, 太低了. 这就是双方关注点不同, 公司的拥有者只关心他每年要支付的工资的总数, 所以关心平均工资, 而工人则关心自己拿到多少钱, 不关心老板总的付出多少钱, 因此劳资双方很难谈得拢. 以腾讯为例, 腾讯公司 2019 年上半年财报显示, 至 2019 年 6 月 30 日, 腾讯有 56310 名雇员, 上半年总薪酬成本达 242.59 亿, 人均半年 43 万元, 人均月薪 7.17 万. 实际情况呢? 少数高管把薪酬拉高了, 薪酬在 80-1500 万 1 人, 4000 万到 6500 万有 1 人, 6500 万到 1.15 亿有 4 人, 1.15 亿到 1.65 亿有 4 人, 2.15 亿到 8.15 亿有两人, 加上其他的中上级高管, 腾讯的中位数薪酬可能只有平均薪酬的

[1] 注意洛伦兹曲线是收入分布曲线的反函数.

一半甚至更少.

贫富差距是一个社会非常关注的问题, 但是怎么科学地度量贫富差距呢? 收入情况的原始数据是收入分布函数, 就是收入这个总体的分布函数或者经验分布函数. 学者想要从中得到能够更真实地反映贫富差距的数字特征. 例如平均收入与中位数收入之间的差可以一定程度上反映贫富差距. 平均和中位数差越大, 分布就越极端, 就当前的例子看, 是说贫富差距很大. 据国家统计局公布的数据, 2018 年中国的人均年可支配收入是 28228 元, 中位数是 24336 元. 为了可比性, 采用差与平均数的比, 这里 $(28226 - 24336)/28228 \sim 14\%$. 收入分布函数是一个从 0 递增到 1 的函数. 如果人的收入是一样的, 那么分布函数是均匀的或者说是直线. 贫富差距实际上体现在收入分布函数与均匀分布函数之间的差距. 因此衡量贫富差距的另一个常用指标是基尼系数, 它是收入分布函数与均匀分布函数之间区域的面积的规范化, 这里规范化的意思是除以这个面积的最大值, 使得基尼系数在 0 与 1 之间. 这些系数都是收入分布函数的一个侧面, 具体的真实情况还是需要看收入分布函数本身.

5.1.7　品茶女士

统计推断的另一个重要表现形式是假设检验, 就是根据数据以及某个置信度来判断一个断言是否是真的. 在日常的生活中, 我们经常看到很多听起来很玄妙的争议, 例如连接放大器和音箱的线材对声音有没有影响, 用煤气烧的饭和用柴烧的饭在味道上有没有差别, 等等. 这种争议通常是各说各的, 没有一个明确的结论, 结果总是不欢而散. 原因是说话的人不知道怎么证明自己和驳斥别人.

1920 年代后期夏天的某个下午在剑桥大学的校园里就有这样一个争议, 一个真实但带有点八卦的故事. 一些绅士, 大学教授, 以及他们的夫人们坐在一起喝下午茶, 就是红茶和牛奶的混合体: 英国奶茶. 一位女士提出了一个观点, 把茶倒在牛奶中得到的奶茶比把牛奶倒在茶中得到的奶茶味道更好. 这个观点立刻引来那些具有科学精神的男士们的反击, 他们嘲笑说, 这能有什么区别呢? 从化学角度看, 都是茶和奶的混合体而已. 学者的争议不同于普通人的争议, 马上一位戴着厚眼镜带尖髯胡的男士就提出一个实验来测试, 让那位认为有区别的女士来品尝一系列的茶, 其中有些是茶倒入奶中, 另一些是奶倒入茶中. 想到这样的实验不难, 普通人也许也能想到, 难的是怎么专业地解释实验得到的结果.

问题: 怎么设计一个实验并按实验的结果来理由充足地判断这位说自己可以区分奶茶味道的女士所言是真是假?

科学的发现往往来自于偶然的小问题. 女士能不能区分这两种茶的味道这个问题不是什么太有意义的问题, 有意思的是不是能找到一个有效的方法来判断这位女士说的是不是对的, 很快, 男士们就开始讨论怎么来做出判断. 但是, 稍微思考一下就知道, 科学研究中有同样的问题, 例如不同的肥料配方对于农业收成的影响, 什么样的植物成分对疾病有治疗效果, 等等. 因此, 小问题的讨论可以应用于大问题.

上面的品茶故事记载在于 2001 出版的《品茶女士》一书中的一开始，书中提到的这位戴厚眼镜尖髯胡的男士是英国著名统计学家 Sir Ranold Fisher. 有趣的是, 在 1935 年, Fisher 出版了一本题为《实验设计》的书, 在此书的第二章的第一句话是这样的: A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. 他一开始就讲述了上述女士品茶故事中所提到的实验, 讲述他怎么设计实验且以什么理由确定这位女士是不是真的可以区分两种方法做出来的奶茶的味道. Fisher 实际上是利用这个实验来讲解他关于假设检验的思想.

具体地说, 在书中, Fisher 设计了这么一个现在大学生熟知并称为假设检验的实验, 即提出一个假设, 然后用实验的数据来看是不是有足够的信心推翻它. 他设计的实验如下: 准备 8 杯茶, 4 杯是先倒奶, 然后茶倒入奶中, 另 4 杯是先倒茶, 然后奶倒入茶中. 8 杯茶依一个完全随机的顺序递给女士品茶, 把这些信息告知女士, 然后要求女士通过品尝把 4 杯茶倒入奶中的奶茶识别出来. 然后分析怎么按照这个女士可能给出的各种答案来推断女士是否真的可以区别茶的味道.

Fisher 设计的实验并不是要证明女士能够区分奶茶味道, 而是想要推翻下面的假设, 称为零假设 (或者原假设):

$$\mathrm{H}_0:\ \text{女士是随机地从 8 杯茶中选择 4 杯的.}$$

如果零假设成立, 那么从 8 杯茶随机地挑出 4 杯茶有 $\binom{8}{4} = 70$ 中不同的等可能的选择. 用 X 表示她选择正确的杯数, 那么 X 的分布是

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ \frac{1}{70} & \frac{16}{70} & \frac{36}{70} & \frac{16}{70} & \frac{1}{70} \end{pmatrix}.$$

即挑出的 4 杯茶完全正确的机会是 1/70. 在这里, Fisher 提出一个显著性的概念, 它认为只要在零假设下发生的概率小于 1/20 的事件就应该被认为是实际上不可能发生的. 因此, 如果它在某个实验中真的发生了, 那就说明假设不对, 所以应该可以安全地“拒绝”零假设了. 这也是显著性检验的意义.

在这个实验中, 也就是说, 如果最后的结果是品茶女士正确地挑出了 4 杯茶倒入奶中的茶, 那么 Fisher 的推理是这样的: 因为这个事件在零假设下发生的概率是 1/70, 远远小于显著性的指标 1/20, 所以就可以拒绝零假设, 也就是说拒绝女士是随机地选择的假设. 如果品茶女士准确地识别了 3 杯, 1 杯错, 在零假设下, 这个发生的可能性是 16/70, 远远超过 1/20, 没有显著性, 也就是说, 这个事件的发生概率虽然不大, 但也并不让人觉得不可接受. 因此我们不能拒绝零假设. 这时我们会问, 能不能接受零假设呢?

Fisher 本人对此是谨慎的, 他认为实验可以否定零假设但不能肯定, 他说:

It should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

“零假设永远无法证明或者确立，但是可以通过实验手段被否定. 可以这么说，每个实验存在的目的是为了给事实提供一个机会去推翻零假设. ”

与零假设相反的假设称为备择假设, 零假设与备择假设在数学上是没有区别的, 但在实际应用时, 零假设一般是想要推翻的假设. 如果上面的零假设被实验结果拒绝, 那么接受备择假设, 就是说女士并不是随机地分辨奶茶的, 或者说对两种做法的奶茶有一定的识别能力. 另外, 零假设需是确切的, 即在零假设下, 所考察的随机现象的分布是确定已知的. Fisher 还是以女士品茶为例来说明, 我们可以把女士可以正确地识别茶的味道作为零假设, 这时分布是确定的, 即 $\mathsf{P}(\mathsf{X} = 4) = 1$, 完全识别的概率为 1, 因此只要有一杯茶识别错误就可以推翻这个假设, 但是我们永远不能用有限次实验证明这个假设成立. 我们能不能把女士有一定的把握识别茶的味道当作零假设呢? 不能, 因为在这个假设下 X 的分布是不确定的. 另外, 假设检验是统计推断的一种, 无论是接受还是拒绝假设, 都不可能是绝对的, 而是基于某个给定的置信度而言的, 在这里也称为显著性.

Fisher 的这本书极大地推动了统计在科学研究中的应用, 在 Fisher 之前, 统计通常意味着只是收集数据, 统计很少会产生新的知识, Fisher 的书教给我们怎么读数据以及怎么从数据里获得新知识的一个规范性的方法. 到现在, 统计是科学研究的重要工具, 统计也是我们生活不可或缺的部分, 用数据说话慢慢成为我们的习惯.

最后提一句, Fisher 这本书中并没有提到实验的最终结果, 因为这对于他的书来说不重要. 但是据当时在场的其他人说, 这位女士果然胸有成竹, 准确地识别出所有的茶, 不管是奶倒入茶中还是茶倒入奶中. 不过这也许只是为了增加该故事的趣味吧.

5.1.8　小概率事件

仔细地阅读女士品茶这段故事, 你会感觉到推理中的缺陷. 人们日常使用的推理有演绎推理和归纳推理, 分别是从一般到具体和从具体到一般. 数学中使用的是演绎推理, 这是严密到没有缺陷的推理. 归纳推理是日常生活中更常使用的, 统计推断属于归纳推理, 但归纳推理从理论上讲是有缺陷的, 因为从具体个例永远不可能推出一般结论. 例如女士品茶, 不管女士准确地识别了多少杯茶, 我们都可以说她是碰巧了, 因为仅验证具体实例是无法排除这个碰巧的. 因此, 按照这样的逻辑, 我们永远无法解决这个争端, 或者说统计推断永远无法应用于解决实际问题.

幸运的是, 与一般的归纳推理不同, 统计推断利用概率进行推断. 如果我们相信当一个事件发生的概率充分小的时候, 它在一个试验中不会发生, 或者说如果假设:

“小概率事件在一次试验中不会发生,”

那么这个这个推断过程就可以继续下去, 就可以自圆其说, 而且可以知道我们是以多大的信念 (置信度) 进行这样的推理. 女士品茶这个问题的推断过程正是接受了这个假设.

为了理解这个假设, 我们需要聊聊小概率事件. 小概率和前面第三章提到的零概率有本质的不同, 这体现在重复试验. 无论重复多少次, 零概率事件发生的可能性还是零, 因此认为零概率事件不可能发生争议不大, 但小概率 (但非零) 事件在重复试验时发生的可能性会越来越大接近 1. 因此我们能有信心说小概率事件 (在一次试验中) 不会发生吗? 直观地说, 掷 5 个硬币都是正面这样的事情很少有人会相信, 掷 10 个硬币都是正面这样的事情几乎不会有人相信, 掷 20 个硬币都是正面这样的事情我敢说没有人会相信. 多小的概率是小概率事件呢? 统计学家有自己的标准, 理论上讲, 小概率事件当然有可能发生, 但实际上人们直觉认为不可能. Fisher 认为, 概率小于 5% 的事件大概被绝大多数人认为不可能会发生了, 或者说, 人的感知难以区别概率小于 5% 的那些事件,

在数学家的直觉中, 零概率事件以及不可能事件也许是明确的, 但小概率事件是模糊的. 多大的概率算是小概率? 小概率事件有什么特殊性? 18 世纪到 20 世纪初数学家就小概率事件有过很多看起来有点无聊的讨论, 这里我选择使用原文引用一些, 因为我恐怕我的翻译会曲解作者的意图. 在 1713 年, 他故去 8 年后, 出版的《猜度术》part 4 第一章中, 他首先说 probability is degree of certainty (概率是肯定的程度), 然后他精心地作了一些区别, 例如 probable 是概率可观地超过 1/2, doubtful

或者 undecided 是概率 1/2 左右, 他还说 something is possible if it has even a very small part of certainty, impossible if it has none or infinitely little. 他还具体说像 1/20, 1/30 是 possible, 看样子在他眼里, possible 是概率较小的意思了, 不是现在理解那样是中性的. 然后 Bernoulli 引入了 morally 这个词, 他说 something is morally certain if its probability comes so close to complete certainty that the difference cannot be perceived. 他解释 morally impossible 与 morally certain 是对立的, 即一个事件是大概率的当且仅当对立事件是小概率的. d'Alembert (大概 1760 年) 困惑一个概率很小的事件究竟会不会发生, 他用 physically 这个词, 觉得也许重复很多次时它 metaphysically possible, 但只做一次时它 physically impossible. Buffon 觉得这只是一个度的问题, 说 (大概 1777 年) An event with probability 9999/10000 is morally certain; an event with much greater probability, such as the rising of the sun, is physically certain. Boltzmannn 用了 vanishingly small 这个词, 说 (大概 1870 年) Dissipative processes are irreversible because the probability of a state with entropy far from the maximum is vanishingly small. Borel 曾经做过这样的划分, 说 (大概 1905 年) A probability of 10^{-6} is negligible at the human scale, a probability of 10^{-15} at the terrestrial scale, and a probability of 10^{-50} at the cosmic scale. 你觉得这有意义吗? 大概这些讨论的目的是为了说明上面这个假设可以接受. 实际上, 第一个表达这个思想的是 Bernoulli. 在如上讨论 moral certainty 之后, 在他的著作的第二章公理 9 中他说 Because it is rarely possible to obtain certainty that is complete in every respect, necessity and use ordain that what is only morally certain be taken as absolutely certain. 意思是完全肯定在实践中是做不到的, 所以必须把道义肯定当作绝对肯定. 这句话就是古诺假设. 他接着说要是由权威为 moral certainty 设立一个标准 (即置信度) 就更好了. 回顾大数定律, 假如我们需要重复随机试验来估计真概率, 实际上, 不管重复多少次都不能保证频率与真概率的差 (误差) 肯定不超过给定的正数 ε, 因此如果要追求绝对肯定, 我们永远无法停止随机试验. 因此 Bernoulli 建议找到充分大的 n 使上述误差超过 ε 的概率小于预先选定的数 δ (对立数 $1-\delta$ 达到 moral certainty 的标准), 或者说

$$P\left(\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right)\geqslant 1-\delta,$$

即误差不超过 ε 这个事件达到 moral certainty, 可以当作绝对肯定了, 这样随机试验就可以停止, 我们就可以认为频率是概率的估计了. 在发布统计数字时, 严肃的机

构通常会同时告知误差 ε 和置信度 $1-\delta$. 理论上讲, 精度和置信度都可以任意的提高, 越高越好, 但需要付出的代价是试验次数. Fisher 在权衡两者之后选了置信度为 $95\% = 1 - 5\%$.

练习 5.1　n 多大可以保证

$$P\left(\left|\frac{S_n}{n} - p\right| > 0.01\right) < 0.01?$$

上述假设称为古诺原理, 为什么把它命名给古诺呢? 尽管古诺不是第一个表达这个思想的人, 但他也许是第一个认识到唯有它能将概率论和现实世界联系起来. 他说 (1843 年) The physically impossible event is therefore the one that has in- finitely small probability, and only this remark gives substance- objective and phenomenal value-to the theory of mathematical probability.

20 世纪 30 年代之前, 概率主要是作为哲学登场的, 大家在做问题之前总是先讨论意义, 怎么理解概率, 怎么应用概率. 上面所谈的是那个年代关于对概率理解的片段, 是法国学术界的主流思想. 除此之外, 还有许多其它的观点, 英国的学者一开始就是以频率来理解概率, 而德国的主流学术界并不认同 Bernoulli 和古诺的观点, 有另外一套理论, 这里不作赘述. 在 1933 年 Kolmogorov 的公理概率论之后, 概率就作为数学登场了, 成为了主角. 其哲学成了配角, 很少有人再关心, 用 Kolmogorov 略带担忧的话来说: “数学家对公理化的概率似乎满意到甚至忘记了现实世界的概率.”

5.2　预期

对于一个随机变量来说, 结果是指它的实际取值, 期望被定义是结果的加权平均, 是个确定的数学概念. 我们通常相信或者希望期望真实地反映了预期, 期望是一个数学术语, 预期是日常用语, 是指心理或者直觉上的期望, 不是一个数学概念, 也许应该叫做心理预期. 我们将会看到, 尽管在多数情况下, 期望能够一定程度上反映预期, 但它们不一定总是一致的. 随机现象中的随机变量相当复杂, 期望只是随机变量分布的一个数字特征, 反映随机性的某个因素.

5.2.1　投资陷阱

在这一节中, 我们介绍一个投资的模型, 看看从这个模型能得到什么. 主要是关心下面的问题.

问题：怎么看待投资项目的期望收益与真实收益的关系?

假设我有一笔钱, 放在银行拿利息比较稳定, 设每期的利率为正常数 $r > 0$, 也就是说 a 元钱一期之后变成 $a(1+r)$ 元; 另一种方法是投资, 收益是随机的, 每期的投资收益率为随机变量 $X > -1$, 即 a 元钱一期之后变成 $a(1+X)$ 元钱.

由于投资收益是随机的, 也就是有风险, 那必然有风险溢价, 即期望收益率 $E[X]$ 应该高于利率, 这是说市场总会在期望上给予愿意冒险者一定的奖励, 否则这个投资就没有足够的吸引力. 例如现在, 银行年利率为 3%, 那投资的期望收益率应该在 3% 以上才会有吸引力, 否则大多数人愿意选择把钱存在银行, 市场就不可能活跃.

用 S_0 表示初始投资额, 这是常数. 用 S_n 表示 n 期投资之后的财富, 那么 S_n 可以如下表示

$$S_n = S_0(1+X_1)(1+X_2)\cdots(1+X_n),$$

其中 $X_1, X_2, \cdots, X_n$ 为第 $1, 2, \cdots, n$ 期的投资收益率. 假设投资项目稳定, 它们是独立的且与 X 同分布的. 特别地, 把钱存入银行时, 投资收益率就是利率 r, 投资收益为 $S_n = S_0(1+r)^n$.

利用期望的性质

$$E[S_n] = S_0(E[1+X])^n = S_0(1+E[X])^n.$$

因为 $E[X] > r > 0$, 所以期望收益 $E[S_n]$ 趋于无穷. 因此从预期看, 投资总是很乐观的. 事实是这样吗? 答案是不一定, 至少从数学上看有一个很深的陷阱. 什么是陷阱? 陷阱是指期望很乐观的投资却几乎总是失败的, 亏损的, 即对于几乎所有 $\omega \in \Omega$, $S_n(\omega)$ 趋于 0.

投资陷阱的存在性: 存在这样的随机收益率 X 使得

1. $E[S_n] \longrightarrow +\infty$;
2. $S_n \longrightarrow 0$.

因为 S_n 可以写成为

$$S_n = S_0 \exp\left(\sum_{i=1}^{n} \log(1+X_i)\right).$$

由强大数定律, 当

$$E[\log(1+X)] < 0$$

时, $\sum_{i=1}^{n} \log(1+X_i)$ 趋于负无穷, 这时 S_n 趋于零.

因此当

$$\mathsf{E}[1+\mathsf{X}]>1,\ (\text{等价地},\ \log\mathsf{E}[1+\mathsf{X}]>0),\ \text{且}$$
$$\mathsf{E}[\log(1+\mathsf{X})]<0$$

时, 投资陷阱呈现. 称之为陷阱条件.
这实际上很大程度上是因为对数函数的凹性. 因为对数的凹性, 由 Jensen 不等式知

$$\mathsf{E}[\log(1+\mathsf{X})]\leqslant\log(1+\mathsf{E}[\mathsf{X}]),$$

这样适当选取 X 的分布, 右边是正的, 但左边可能会负的, 见下例.
假设投资有亏损的可能, 即 X 可能大于 0 (盈利) 也可能小于 0 (亏损). 最简单地, 我们设 X 取两个值, 一个是 $\mathfrak{b}>0$ 一个是 $-1<\mathfrak{a}<0$, 它的分布为

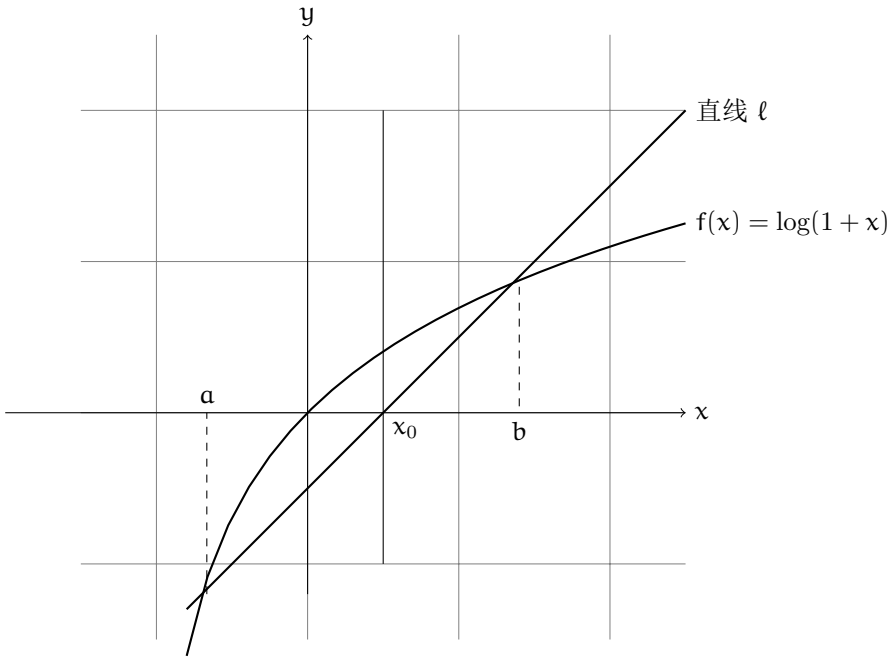
$$\begin{pmatrix}\mathfrak{a}&\mathfrak{b}\\1-\mathfrak{p}&\mathfrak{p}\end{pmatrix},$$

则 $\log(1+\mathsf{X})$ 的分布为

$$\begin{pmatrix}\log(1+\mathfrak{a})&\log(1+\mathfrak{b})\\1-\mathfrak{p}&\mathfrak{p}\end{pmatrix}.$$

那么陷阱条件等价于

$$\mathsf{E}[\mathsf{X}]=\mathfrak{p}\mathfrak{b}+(1-\mathfrak{p})\mathfrak{a}>0,$$
$$\mathsf{E}[\log(1+\mathsf{X})]=\mathfrak{p}\log(1+\mathfrak{b})+(1-\mathfrak{p})\log(1+\mathfrak{a})<0.$$



考察函数 $f(x)=\log(1+x)$, 通过两个点 $(a,\log(1+a))$ 与 $(b,\log(1+b))$ 作直线 ℓ, 方程为

$$\frac{y-\log(1+b)}{x-b}=\frac{\log(1+b)-\log(1+a)}{b-a}.$$

这条直线在区间 (a,b) 上位于在 $y=\log x$ 之下.

对于 $p\in(0,1)$, $pb+(1-p)a\in(a,b)$, 它在直线 ℓ 上对应的点恰是 $\mathsf{E}[\log(1+X)]$, 实际上,

$$\begin{aligned}y&=\log(1+b)+(pb+(1-p)a-b)\frac{\log(1+b)-\log(1+a)}{b-a}\\&=p\log(1+b)+(1-p)\log(1+a)\\&=\mathsf{E}[\log(1+X)].\end{aligned}$$

直观地可以看到, 在 $x>0$ 时, $f(x)>0$, 但直线上有一段却小于 0, 这就是我们要找的陷阱区. 现在我们就把这一段区间先找出来. 通过解方程得到直线与 $y=0$ 的交点的 x- 坐标

$$x_0=b-\frac{(b-a)\log(1+b)}{\log(1+b)-\log(1+a)}\in(0,b).$$

现在我们只需要找到 p 使得 $pb+(1-p)a \in (0, x_0)$ 就可以了. 实际上, 解方程

$$0 < pb + (1-p)a < x_0$$

得

$$0 < \frac{-a}{b-a} < p < \frac{x_0 - a}{b-a} < 1,$$

这时陷阱条件满足: $\mathsf{E}[1+X] > 1$ 但 $\mathsf{E}[\log(1+X)] < 0$, 投资陷阱现象出现.

投资陷阱存在的结果是, 投资普遍地说是失败的, 但是投资成功的人会非常赚钱, 财富向少数人聚集, 这样造成贫富差距会越来越大. 要注意的是, 真实的投资问题极其复杂, 无法用数学描述.

5.2.2　圣彼得堡悖论

概率论是研究与解释随机现象的数学理论, 从上面许多例子看, 概率论能够解释很多实际的概率问题, 其实是说期望与预期符合得好, 这里的预期是指心理预期, 或者说是直觉, 所以大多数学者对概率论是非常满意的, 认为它能够反映人对随机现象的感受.

但是, 还是有许多问题, 概率论的解释与心理感受不符, 特别是在概率很小的时候. 例如, 掷 10 个硬币, 10 个都是正面的概率约是千分之一. 如果压 10 个都是正面, 拿 1,000 元, 否则是零, 那么这个机会的期望是 1 块钱. 1 块钱是个小数, 在地上都不一定愿意弯腰去捡, 估计很多人愿意买这个机会. 那么放大一万倍, 你愿意花一万块钱赌 10 个硬币都是正面, 赢了拿一千万吗? 对于这个选择, 大多数人都不愿意, 因为大家心理感觉, 掷一次, 10 个硬币都是正面是几乎不可能的, 拿半年的生活费买这么小的机会不值得. 这大概就是数学期望和心理期望的差异吧, 概率论中的期望是线性的, 人的心理期望是非线性的.

圣彼得堡悖论是一个非常极端的例子, 能更清楚地说明期望与预期在某种情况下是相悖的. 它是 Bernoulli 家族的 Daniel Bernoulli(生于 1700 年) 于 1738 年的一篇论文中发表的, 他是 Jahanne Bernoulli 的儿子, 该问题的最初版本出现在他的堂兄 Nicolaus(I) Bernoulli 在 1713 年写给同行 (Pierre Rémond de Montmort) 的一封信中.

某赌场提供一种游戏: 掷一枚硬币到掷出正面为止. 如果这时掷的次数是 n, 那么玩者得到的奖金是 2^n 块钱, 因为首次正面出现在第 n 次投掷的概率是 2^{-n}, 所以所得

钱数 X 的分布是

$$\begin{pmatrix} 2 & 2^2 & 2^3 & 2^4 & \cdots & 2^n & \cdots \\ 2^{-1} & 2^{-2} & 2^{-3} & 2^{-4} & \cdots & 2^{-n} & \cdots \end{pmatrix}.$$

这个游戏的价值期望为

$$\mathsf{E}[X] = 2 \cdot 2^{-1} + 2^2 \cdot 2^{-2} + 2^3 \cdot 2^{-3} + \cdots + 2^n \cdot 2^{-n} + \cdots = +\infty,$$

因此按照期望来确定价值的话, 这个游戏的价值是无穷.

再来看, 如果玩 n 局, 每局所得记为 $X_1, X_2, \cdots, X_n$, 它们是独立同分布的, 它们的和用 S_n 表示, 那么应用大数定律可以证明, 平均所得 S_n/n 也趋于无穷.

概率论的知识告诉我们, 只要是花有限的钱, 100, 1,000, 10,000, $\cdots$, 这个游戏肯定是值得玩的, 但是我们的内心会这么认为吗? 按我们的通常想法, 平均 2 次正面就会出现 1 次, 5 次之内正面不出现的概率是 1/32, 就很小了, 10 次之内正面不出现的概率小于千分之一, 可能性几乎没有. 在几乎所有的问卷调查中, 很少有人愿意出多于 20 块钱玩这个游戏. 你可以试着做这样一个调查.

问题: 一个产品的数学期望与心理期望为什么会差别怎么大?

在 18 世纪 Buffon 做的 2048 次硬币抛掷实验 (注意一次游戏所进行的抛硬币次数是不定的) 中可以看到

- 这个游戏平均每次的所得是 4.91.

Hinners-Tobrägel (2003) 应用计算机模拟来探讨这个游戏, 他发现

- 玩这个游戏 10 次的平均所得大约是 2.1;
- 玩 250 次的平均所得是 4.5;
- 玩 1000 次的平均所得是 5.8;
- 玩 2500 次的平均所得是 13.7;

似乎游戏的平均所得随着次数的增加而增加. 理论上无法证明这个递增性, 但可以证明:

- S_n/n 以概率 1 趋于无穷;
- 但是速度是对数的, 即

$$\frac{S_n/n}{\log_2 n} \longrightarrow 1.$$

第一个结论的意思是如果你有钱, 不断地玩这个游戏, 最终的平均所得一定会超过付出的代价. 第二个结论的意思是平均所得的增长速度是对数速度, 很慢, 例如玩 1000

次, 平均所得也就是差不多 10 块钱, 这倒是差不多符合人的预期. 因此, 是不是可以说, 人们对这个游戏的心理预期很差也许是因为内心对其速度的感应? 如果是的话, 那么这可以说真的太神奇了.

为什么期望的数学理论和心理预期会相差这么大? 这是概率界一直难以解释的问题. 多年来, 学者们提出了很多不同的解释, D. Bernoulli 因此提出效用理论, 这些解释可以适用于某些场合, 可以说服某些人, 但都存在缺陷, 不能从根本上解释为什么我们通常认为与实际符合得很好的概率论在这里与心理预期如此地不一致.

注意, 再强调一句, 这个悖论不是数学上的悖论, 概率论本身在这里没有矛盾, 这里所说的悖论是指理论和直觉的矛盾, 如同 Euclid 几何的第五公设是符合直觉的, 否定第五公设的非 Euclid 几何是和直觉矛盾的, 但是非 Euclid 几何本身在数学上没有矛盾.

5.2.3　双臂老虎机问题 (*)

老虎机是赌场最简单的机器, 英文是 slot machine, 它有一个手柄, 叫做臂, 扳一下就会出来若干钱或者没有. 设有两台独立的 Bernoulli 老虎器, 扳臂以概率 $\mathrm{p}_1, \mathrm{p}_2$ 出一块钱, 或者没有. 这两个概率都是未知的.

问题: 现在你有机会玩 100 次, 问怎么玩才能使得获得的收益期望最大?

5.3　随机商品的定价与风险

现代社会是商品社会, 不仅有普通商品, 如食品, 消费品, 有服务品, 还有金融商品. 典型的金融商品是彩票, 银行存贷款, 股票, 保险, 期货, 衍生证券等等. 金融商品有个特点, 它的价值通常是随机的, 例如我们花 10 块钱买一张彩票, 彩票没中奖的话一钱不值, 中奖的话可能价值千万; 再例如买财产保险, 如果家里一直平安, 它似乎不值钱, 但是一旦家里遭遇火灾, 保险就会体现价值; 股票更是如此, 它的价值一直在波动, 如果你能低价时买入高价时卖出, 那就可以赚钱.

随机商品最早的雏形就是赌博, 现在它虽然改头换面, 称为投资, 或称为避险, 但是本质上还是赌博, 赌你对未来的预测是否准确.

随机商品为什么会出现? 这是一个经济学问题. 随机商品出现的原因主要是人对风险偏好的不同, 或者说人与人之间不仅需要交换通常的生活品, 还需要交换风险, 或者说交换不确定性. 不喜欢风险的人拿钱买保险, 降低风险, 把风险转移给那些喜欢

风险的人.

通常商品的价值通常是通过成本或者需求来计算, 那么随机商品的价值怎么计算呢?

5.3.1　随机商品的定价

先说随机商品期望的作用. 随机商品的价值在未知晓前是随机的, 可能很大也可能一文不值, 那么作为商品它应该卖多少钱, 也就是随机商品的定价. 你需要花钱买一个价值未定的商品.

问题: 怎么才算是随机商品的一个公平的定价呢?

实际上, 购买随机商品就是一个赌博, 运气好的话赚大钱, 运气不好的话损失惨重. 在一个典型的赌博中, 输赢是随机的, 输的人按照事先约定的规则付钱给赢的人, 因此对于赌博中一个固定的个体来说, 他的所得是随机变量, 记为 X, 正表示赢, 负表示输.

众所周知, 赌博的一个基本准则是公平. 什么是公平?

例如两个人赌博, 如果每次输赢的可能性一样, 那么每次的赌注是一样的. 如果每次的输赢可能性不同, 那么赌注就不同. 例如掷硬币, 每次输赢一块钱是公平的; 如果甲乙掷骰子, 甲掷出 6 才算赢, 其他算输, 那么每次输赢都是一块钱就不公平了, 怎么才算公平呢? 乙的赌注是甲的 5 倍时才是公平的, 即输赢数的比例应该等于赢输的概率比.

精确地说, 假如甲赢和输的概率分别是 p 和 $1-p$, 赌注分别为 x 与 y, 那么甲如果赢, 得 y, 如果输, 付 x, 这样的随机变量 X 的分布为

$$P(X=y)=p,\ P(X=-x)=1-p,$$

则当

$$\frac{y}{x}=\frac{1-p}{p}$$

时才是公平的, 实际上, 这就是说

$$E[X]=py-(1-p)x=0.$$

因此, 公平等价于输赢数的期望等于 0.

设一个随机商品未来的价值是 X, 假设每件商品的价值是独立的. 购买 n 件随机商品, 价值分别为 $X_1,\cdots,X_n$, 期望为 μ. 由大数定律, 未来的总价值

$$X_1+X_2+\cdots+X_n$$

差不多是 $n\mu$, 如果其定价 p 高于或者低于商品的期望值 μ, 对应第, 卖者或者买者将有期望为 $n|p-\mu|$ 的利益, 这都是不合理的套利, 称为统计套利, 因此, 从大数定律或者说重复试验的角度看, 随机商品的定价应该就是期望 μ. 但是, 实际问题是多种多样的, 例如前面所说的圣彼得堡悖论, 实际价值远远偏离期望, 也可能实际问题不容许重复试验, 等等, 总之在实际问题中, 期望不是唯一的标准, 我们需要考虑其他许多的因素, 其中最重要的是风险.

5.3.2　风险

随机商品和确定性商品不同, 确定性商品的价值是确定的, 买卖属于等价交换, 不存在损失, 而当我们购买随机商品之后, 因为其价值的不确定性, 所以我们说存在损失的风险. 那下一个问题就是

问题: 什么是风险? 怎么度量风险?

先来看同样期望的两个随机产品,

(1) 掷硬币, 掷出正面给 2 块钱, 掷出反面 0;

(2) 掷骰子, 掷出 6 点给 6 块钱, 其他点 0.

这两个游戏的期望价值是一样的, 都是 1, 所以按照期望, 玩一次的代价应该是 1 块钱, 但是有风险, 可能赚钱, 也可能损失掉一块钱而一无所得.

两个同样定价的商品, 你的感觉有什么不同? 主要的差别在哪里呢? 首先, 大多数人会感觉第二个游戏风险更大, 因为它赢的可能性比较小, 但若赢, 赢得的数目比较大. 这实际上是因为第二个游戏的随机性比较大, 而随机性的大小通常是用方差来刻画的, 第一个例子的方差是 1, 第二个方差是 5.

我们会认为风险是来自于不确定性, 不确定性大小与风险大小成比例, 而方差是刻画随机商品的随机性大小的, 所以我们通常把方差的大小理解为某种意义上风险的大小. 为什么说某种意义上呢? 因为随机性大小是没有方向的, 永远是个正数, 它既可能带来损失也可能带来好运, 而风险其实是有方向的, 造成好运的随机性不是风险, 造成损失的随机性才是风险, 因此方差具有风险的某些特征, 但远不是全部. 我们在后面会再详细解释.

5.3.3　风险厌恶与边际效用递减

在上两个小节我们说, 在大数定律看来, 用数学期望来给随机商品定价是一个理性的方案. 但是实际上, 从前一章的几个例子可以看到, 因为人的心理期望和数学期望可能差距很大, 所以用数学期望衡量随机商品的价值是有缺陷的.

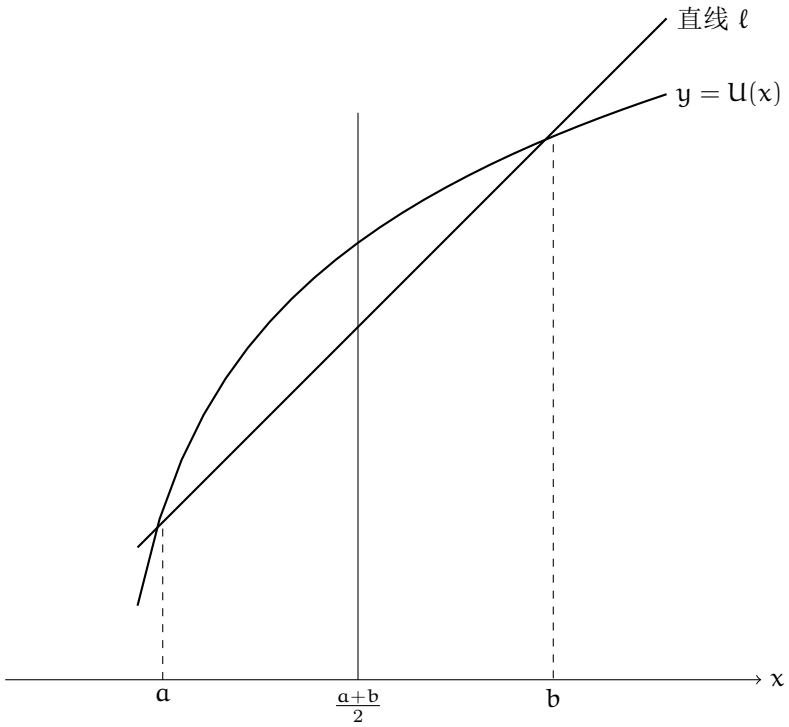
问题: 怎么解释财富的价值与它在人内心的价值的不一致性?

一个最受欢迎的办法是 Daniel Bernoulli 于 1738 年提出的效用理论, 他认为金钱的价值和它引起的心理感受 (心理价值) 不是成比例的. 直观上, 对于一个人来说, 财富肯定是越多越好, 但同样是发奖金 1 万元, 在总财富 10 万时候发 1 万与在总财富 1 万的时候发 1 万, 其数目一样, 但心理价值要降低, 也就是说效用更低, 或者说边际效用递减. 边际效用递减说明财富对人类心理的价值是非线性的, 而数学期望是线性的, 所以数学期望与心理期望是天然相悖的, 只是在一定的财富范围内, 心理期望才是近似于线性. 如同 Riemann 几何在小尺度内是 Euclid 几何.

也可以说, 金钱对于人的意义需通过效用来体现. 很多经济学家推崇这个解释. 不妨用 $U(x)$ 表示数量为 x 单位的财富 (钱) 产生的心理价值, 称为效用函数. 效用函数满足: (1) 财富越多越好, 这是指递增性; (2) 财富的边际效用递减, 即如果 $x_1 < x_2$ 且 $h > 0$, 那么

$$U(x_1 + h) - U(x_1) > U(x_2 + h) - U(x_2).$$

第二个性质通常称为上凸性 (或者下凹, 简称凹) 具有上凸性的函数称为上凸函数, 见图.



直观地看, 上凸性是指函数的增长越来越慢, 即一阶导数递减, 等价地, 二阶导数小于零. 上凸性还等价于函数图像上任何两点连接的线段在函数图像之下, 即对任何两点 a, b, 有

$$U(\frac{a+b}{2}) \geqslant \frac{U(a)+U(b)}{2},$$

这是说任何两点 a, b 的平均的函数值大于函数值 $U(a)$, $U(b)$ 的平均, 这等价于对任何两点 a, b 以及 $t \in (0,1)$ 有

$$U(ta + (1-t)b) \geqslant tU(a) + (1-t)U(b),$$

其中 $ta + (1-t)b$ 是 a, b 的凸组合, 它是 a, b 之间的一个点. 这也等价于函数图像之下的图形

$$\{(x, y) : y \leqslant f(x), x \in D\}$$

(其中 D 是函数定义域) 是凸图形. 凸图形很形象, 定义是其中任何两点的连线段在该图形内.

再可以证明上凸性等价于对任何 n 个点 $x_1, \cdots, x_n \in D$, 有

$$U\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \geqslant \frac{1}{n}\sum_{i=1}^{n} U(x_i).$$

还等价于看上去更一般的加权不等式: 对任何 n 个点 $x_1, \cdots, x_n \in D$, 及其凸组合

$$p_1 x_1 + \cdots + p_n x_n,$$

其中 p_i 都是正数且 $p_1 + \cdots + p_n = 1$, 有

$$U\left(\sum_{i=1}^{n} p_i x_i\right) \geqslant \sum_{i=1}^{n} p_i U(x_i).$$

最后一个不等式也就是概率论中的 Jensen 不等式: 对分布为

$$P(X = x_i) = p_i,\ i = 1, \cdots, n$$

的随机变量 X, 上式左边是 $U(E[X])$ 且右边是 $E[U(X)]$, 因此

$$U(E[X]) \geqslant E[U(X)].$$

实际上, Jensen 不等式可以推广至一般分布的随机变量.

问题: Jensen 不等式的直观意义是什么?

它也在说一个很直观的道理: 风险厌恶假设. 人的心理期望与数学期望的差距的另一个解释是人对风险的态度, 仅用期望来衡量风险商品的价值肯定是不够的. 我们通常认为在同样的期望收益的条件下, 人会选择风险小的一个. 例如, 有下面两个选择 (1) 直接拿 10000 元, (2) 1/2 的概率拿 20000 元, 1/2 的概率拿 0 元. 两个选择的期望是一样的, 但实验表明, 大多数人会选择 (1), 因为 (1) 没有风险, (2) 有风险. 这个现象通常称为人的风险厌恶. 实际调查表明大多数人甚至愿意为确定性付出额外的代价, 例如即使把 (2) 改为 1/2 的概率拿 30000 元, 1/2 的概率拿 0 元, 多数人还是会选择 (1), 尽管 (2) 的期望要超过 10000 元.

设随机变量 X 的期望是 μ, 那么 $U(\mu)$ 是财富 μ 对你的效用, 而 $U(X)$ 是期望同样是 μ 的随机商品对你的效用, 而 $E[U(X)]$ 是效用的期望值. 假设边际效用递减, 则有

$$U(\mu) \geqslant E[U(X)],$$

即确定性商品的效用大于同样期望的随机商品效用的期望, 也就是是说, 在同样期望水平之下, 会选择无风险的商品, 称为风险厌恶假设.

实际上, 风险厌恶假设与边际效用递减是等价的. 风险厌恶假设或者说边际效用递减似乎解释了为何数学期望和心理期望有时有巨大差异, 确实, 从定性的方面可以这么认为, 但该理论致命的缺陷是很难实用, 因为效用函数是主观的, 每个人的效用函数都不一样, 而且也无法从量上来确定每个人的效用函数. 因此效用理论也受到很多学者的批评.

D. Bernoulli 提出效用理论的最初动机是试图解释他自己提出的圣彼得堡悖论, 随机变量 X 的分布

$$P(X = 2^n) = 2^{-n},\ n \geqslant 1.$$

对 $a \in (0,1)$, $U(x) = x^a,\ x > 0$ 是一个上凸函数, 取它作为效用函数得到的效用期望为

$$E[U(X)] = \sum_{n\geqslant 1} (2^n)^a 2^{-n} = \sum_{n\geqslant 1} \frac{1}{2^{n(1-a)}} = \frac{1}{2^{1-a}-1}$$

还可以取 $U(x) = c\log_2 n$, 那么

$$E[U(X)] = \sum_{n\geqslant 1} \frac{cn}{2^n} = 2c.$$

这些值与 a, c 有关, 那么究竟应该取哪个呢? 不仅如此, 上凸函数无限多, 究竟应该用哪个呢? 另外, 如果效用理论解释成立的话, 财产数千亿的人和财产数万的人对于游戏的价值应该有极大的差异, 但是现实会是这样吗? 因此, 效用函数是一种解释, 但它并不能让所有的人都满意.

5.3.4　风险溢价

因为风险厌恶, 人们会去购买各种保险, 把生活中自己将会遇到的各种风险转给保险公司承担. 因为风险厌恶, 风险是有价值的, 例如保险公司因为承担风险可以收取额外的费用, 风险的价值也称为风险溢价. 因为风险溢价的存在, 人们才对投资有热情, 可以赚取超过银行利息的溢价.

问题: 怎么衡量风险溢价?

举个简单的例子, 设 w 是个人财富总额, X 是未来一年车损, 是随机的, U 是个人效用函数, 它是上凸的, 也可以看成是风险厌恶假设下的产物.

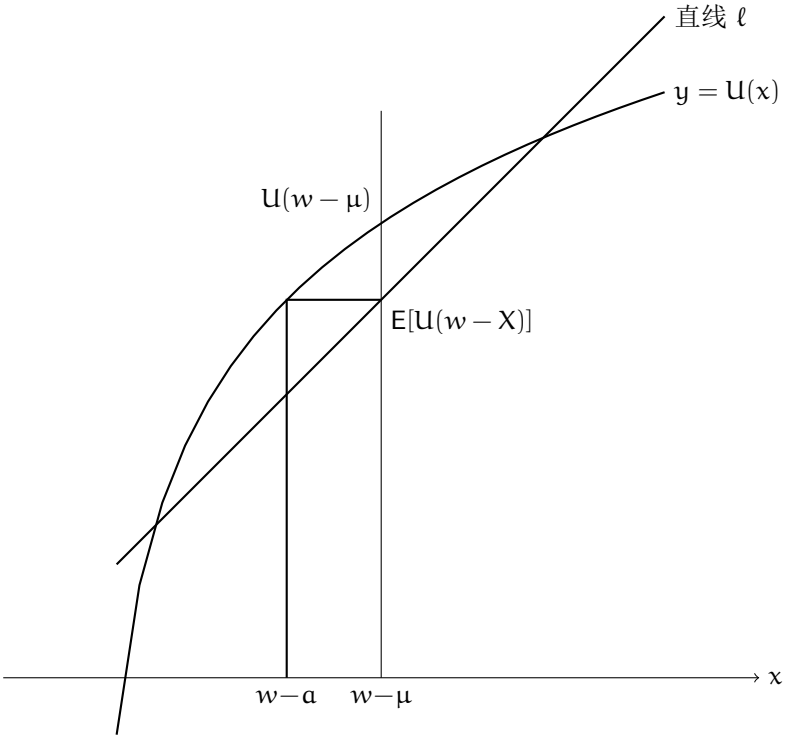
现在, μ 是期望车损 $\mu = E[X]$. 如果不买车险, 其效用衡量下的期望为 $E[U(w-X)]$, 因为 Jensen 不等式

$$U(w-\mu) \geqslant E[U(w-X)],$$

即确定地支付 μ 得到的效用高于不买保险的期望效用, 这是产生保险需求的原因. 但是保险公司该收取多少保费呢? 利润空间在哪里? 从大数定律看, 因为 μ 是客户的平均损失, 所以保费应该是期望值 μ, 超过该值就属于额外. 保险公司有没有理由收取额外费用呢? 从风险厌恶的角度看, 答案是肯定的. 因为效用函数的递增, 故存在 $a > \mu$ 使得

$$U(w-a) = E[U(w-X)],$$

即收取位于 $[\mu, a]$ 之间的保费 x 之后的效用仍然可以保证相对于无保险情况下效用的优势, 对客户来说是可以接受的. 这个 a 就是风险溢价. 但是可以看出 a 依赖于函数 U 也依赖于个人财富 w 以及车损的分布.



风险溢价只是描述因为普遍的风险厌恶心理而导致的市场对风险承担者所给予的额外奖赏, 在金融市场上是普遍存在的. 因为风险本身是个不确定的概念, 所以溢价也有多种不同的定义. 这里的风险溢价亦称为风险厌恶系数.

问题: 人真的厌恶风险吗? 除了随机性, 风险还有什么特征? 风险应该怎么度量?

人对风险的厌恶不是绝对的. 例如彩票, 赌博等游戏的盛行是因为人们迷恋风险带

来的机会而存在的. 从风险厌恶的角度看, 彩票完全没有存在的价值, 因为它的的售价不仅不低于彩票价值的期望, 反而是高于期望的, 为什么? 因为卖彩票的机构有运行费用, 专门的彩票, 例如福利彩票和体育彩票还要拿出一部分作为福利或者资助体育事业. 但即使价格高于期望, 还是有很多人会去买, 因为这是一个投入很低的发大财的机会, 尽管机会很小.

因此风险厌恶假设在这里不成立. 为什么? 这是因为一个彩票卖 10 块 20 块, 与一顿饭的代价差不多, 即使随机性很大, 也产生不了多么严重的风险. 尽管彩票有随机性, 有风险, 但是从损失的角度看, 风险其实不大. 因此风险不仅与随机性的大小有关, 也与随机商品本身的价格有关, 但如果一个月薪一万的个人一次买几张彩票无关痛痒, 但若一次性花一两万块钱买 1000 张彩票, 这就可能需要认真考虑了, 因为这个数量的损失对他会产生真正的风险. 也就是说, 风险与个人或者机构的承受能力有关.

假设一个公司一年的收入低于某个值 d 就会破产. 随机变量 X 表示公司未来一个年度的收入, 那么 $P(X < d)$ 大致反映了公司的风险, 它越大表示公司的风险越大. 这个风险称为 VaR: 在险价值.

拿保险公司的车险作为例子. 设有 n 个车损假设为独立同分布随机变量 $X_1, X_2, \cdots, X_n$ 的客户, 期望为 μ, 方差为 σ^2. 每人收取的保费为 $c > \mu$, 保险公司的准备金为 u, 那么保险公司的财富是

$$S = u + cn - \sum_{i=1}^{n} X_i,$$

这是最简单的保险模型. 财富的期望

$$E[S] = u + n(c - \mu),$$

趋于无穷. 但是方差

$$D[S] = n\sigma^2$$

也趋于无穷. 也就是说, 如果方差是风险的话, 那么保险公司的风险随着客户数增加而增加. 这显然不符合保险公司总是希望发展更多客户的常识. 因此方差不能代表真正的风险.

实际上, 保险公司不很关心项目的随机性大小, 因为随机性中包含正反两方面. 保险公司关心盈亏, 关心会不会破产, 考虑其 VaR

$$P(S < d) = P(\sum_{i=1}^{n} (X_i - \mu) > u - d + n(c - \mu))$$

$$= P\left(\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i-\mu) > \frac{u-d+n(c-\mu)}{\sqrt{n}\sigma}\right)$$
$$\longrightarrow 1-\Phi\left(\frac{u-d+n(c-\mu)}{\sqrt{n}\sigma}\right),$$

最后的极限是中心极限定理, $\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i-\mu)$ 的分布趋于标准正态分布, 而

$$\frac{u-d+n(c-\mu)}{\sqrt{n}\sigma}$$

随客户人数增加趋于无穷, 所以保险公司项目获利低于预警值 d 的可能性趋于 0. 这个模型才比较合理地解释了保险公司的风险.

上面谈论风险, 但并没有能够真正定义风险. 什么是风险? 怎么度量风险? 这是一个困难的问题. 风险是多种多样的, 在不同场合显示不同特征, 对不同的人含义不同, 其度量也是很主观的. 当我们说用方差来度量风险的时候, 只是描述了风险的波动特征, 同样, 把损失超过某个可承受范围的概率当作风险时, 也只是描述了风险的不可承受特征. 因此, 一直到现在, 对于怎么定义风险尚未有定论.

第六章　结语

最后, 我们回到现实世界, 试图解释抽象的概率和真实世界的概率之间的关系, 也许听起来依然是喃喃自语.

6.1　不确定性与随机性

在前面, 我们把不确定和随机当作同义词使用, 这两个词都是日常用语, 不存在定义, 不可能精确, 在日常使用时也无法区分. 但在我们学习了前面这些知识之后, 作为本讲义的理解, 从稍微专业一点的角度, 这两个词还是应该有细微的区别, 不确定性是强调结果不是确定的, 无法预测, 这个词的含义只能到此为止, 而随机性的含义要更特殊一些, 是指具有某种规律的不确定性, 确切地说, 随机性是具有某种给定分布的不确定性, 为什么? 因为当我们说某个现象具有随机性时, 实际上总是假设概率空间存在, 而概率空间中的概率测度是样本空间这个集合上的分布, 它规定哪些结果出现的机会大哪些出现的机会小, 这正是分布的含义, 随机性就是指这个分布.

简单地说, 不确定的现象分为有随机性的现象和没有随机性的现象, 有随机性的现象通过假设概率空间存在进行进一步研究, 而对于没有随机性的现象无法用概率论的方法研究. 但是一个现象是否有随机性是无法通过现有的知识判断的, 只能人为地假设. 究竟什么样的现象可以认为是不确定但又不是随机现象? 举个也许不是很恰当的例子, 甲乙两个人玩游戏, 甲选一个数字 0 或者 1 写在纸片上, 乙来猜. 对于乙来说这个数字是不确定的, 但也很难说是随机的, 因为很难说甲是按照某种分布来写的.

在现实中, 不确定现象从没有随机性到有随机性的现象各种各样, 五花八门. 为了应用概率论研究这些现象, 必须假设随机性, 然后通过各种手段来估计或者推断分布. 这些假设有时是自然的, 有时不是那么自然, 由此获得的结论有时应用的还好, 有时

毫无价值. 为了理解和运用数学方法的方便, 我们总是考虑随机变量及其分布, 也就是通过映射把样本空间里被关注的结果用数量来表示, 这样就等于把样本空间从符号转化为数, 把一般的分布转化为分布函数.

6.2　概率的含义

概率在数学上意义是明确的, 就是一个符合几条公理的概念. 因为很多经典的概率模型可以纳入这个公理体系, 且可以解释很多已知的现象, 所以, 自 1933 年问世之后, 概率这个概念被数学界广泛地接受, 概率论迅速地成为数学的一个重要分支. 由大数定律, 在可重复随机现象情况下概率是频率的极限, 也就是说, 概率是一个可以检验的量, 这使得概率论成为数学连接自然科学的一个桥梁.

但是, 现下这个课程的关注点不全是数学意义的概率, 更重要的是通过它来考察和探索现实世界的概率. 在现实世界中, 随机现象或者说不确定性是客观存在且无处不在, 更需要注意的是, 其中可重复的随机现象是少数, 人们所关心的大多数随机现象都不是可重复的, 例如地震, 台风, 疾病, 死亡, 事故等等. 概率这个词以及其含义是人类在探索随机现象的漫长过程中逐步形成的, 它比数学中的概率概念要宽泛的多, 现实中概率这个概念的准确含义是一个非常有争议的问题, 涉及到怎么合理地用概率来解释现实世界中的随机现象, 学术界有很多不同的观点. 确立概率公理体系的著名数学家 Kolmogorov 在二十世纪 60 年代的一次数学会议上曾经表示概率公理化可能“过于成功”了使得数学家或者科学家不再很关心现实世界中概率究竟是什么这个问题了. 这至少说明 Kolmogorov 不认为现实世界中概率的意义已经明确了. 下文是作者为上海高中教材《数学必修第三册》中第 12 章概率初步所写的一个关于概率意义的课后阅读, 可以被认为是大多数学者所持有的观点.

概率, 经验概率与主观概率

概率是赋予事件的一个数, 它表达该事件有多大的可能性发生. 在等可能的假设下, 掷硬币得正面的概率是 $1/2$, 掷骰子得 6 的概率是 $1/6$. 但是如果没有等可能假设. 这些概率是很难知道的. 例如, 假设投掷的一个硬币由一个金属薄片和一个木头薄片黏贴组成. 这样的硬币显然不再是等可能的, 掷这个硬币金属面朝上的概率 p 和木头片朝上的概率 q 依然是存在的且 $\mathrm{p}+\mathrm{q}=1$, 但理论上没有人能够算出来这两个概率究竟是多少. 这时候因试验可以任意多次重复, 仍可以用频率或者经验概率来进行估计, 这正是大数定律所断言的.

现在换个问题. 一个具体的人, 例如 20 岁的小明, 能够健康活到 70 岁的可能性有多大? 因为不能预测, 健康活到 70 岁是一个随机事件. 它有没有一个概率呢? 因为小明这个人是不能重复的, 大数定律失效. 这时即使说概率, 也只是一种心理预期, 无法加以检验. 这是作为数学的概率论无法解决的问题, 称之为主观概率, 因为它只是反映说话者的主观判断.

尽管小明问自己能健康活到 70 岁的可能性多大在数学上是没有意义的, 但我们仍然可以做一点有意义的事情. 就是把其他人看成是小明的某种重复, 从而求得这个事件的“频率”, 也就是求健康活到 70 岁的人数在总人数中的比例. 这个比例是统计数据, 对具体的个人而言是没有意义的, 但是它仍然有其统计意义, 在某些场合是非常有用的, 例如保险公司可以据此来计算人寿保险的保险费率.

从这里可以明白, 生活中经常说的话, 如“学某某专业好找工作”或者“锻炼让人长寿”等等, 都是统计意义上的断言, 对具体个人没有什么特别的意义.

参考文献

[1] Bernoulli, Jakob, Ars Conjectandi, 1713

[2] DIACONIS, PERSI and SKYRMS, BRIAN, The ten great ideas about chance, Princeton University Press, 2018

[3] Dirac, P.A.M., The principles of Quantum Mechanics, Oxford University Press, 1958

[4] Feller, W., Probability Theory and its Application, Vol. I(1959: Third edition), Vol. II(1970), Wiley & Son

[5] Shafer, Glenn, The origins and legacy of Kolmogorov's Grundbegriffe, 2003

[6] 上海市高中数学, 必修第三册第 12 章, 选择性必修第二册

[7] 王梓坤, 概率论基础及其应用, 北京师范大学出版社, 北京, 2007

[8] 应坚刚, 何萍, 概率论 (第二版), 复旦大学出版社, 2016