AI 中的数学

邓小铁　李翰禹

2024 年 9 月 28 日

目录

第零章　引言 **1**

第一部分　AI 的逻辑 **2**

第一章　合情推理 **3**

§1.1 命题逻辑的演绎推理 4

§1.2 合情推理的数学模型 12

§1.2.1 合情推理的基本假设，似然 14

§1.2.2 似然与概率 19

§1.2.3 先验与基率谬误 21

§1.3 合情推理的归纳强论证 23

§1.3.1 归纳强论证 23

§1.3.2 有效论证和归纳强论证的比较 28

§1.4 先验模型的存在性 33

§1.5 章末注记 35

§1.6 习题 36

第二章 Markov 链与模型 37
§2.1 Markov 链 38
§2.2 Markov 奖励过程（MRP） 49
§2.3 Markov 决策过程（MDP） 55
§2.4 隐 Markov 模型（HMM） 64
§2.4.1 评估问题 67
§2.4.2 解释问题 69
§2.5 扩散模型 72
§2.5.1 采样逆向过程 77
§2.5.2 训练逆向过程 78
§2.6 章末注记 81
§2.7 习题 81

第二部分 信息与数据 82

第三章 熵与 Kullback-Leibler 散度 83
§3.1 熵 84
§3.1.1 概念的导出 84
§3.1.2 概念与性质 89
§3.2 Kullback-Leibler 散度 98
§3.2.1 定义 98
§3.2.2 两个关于信息的不等式 101
§3.3 编码理论 102
§3.3.1 熵与编码 102
§3.3.2 K-L 散度、交叉熵与编码 106

§3.4 在机器学习中的应用：语言生成模型 108
§3.5 附录：Shannon 定理的证明 110
§3.6 习题 113
§3.7 章末注记 115

第四章 高维几何，Johnson-Lindenstrauss 引理 117
§4.1 高维几何 119
§4.1.1 高维球体 119
§4.1.2 Stein 悖论 123
§4.1.3 为什么我们要正则化？远有潜龙，勿用 130
§4.2 集中不等式 131
§4.3 J-L 引理的陈述与证明 138
§4.4 J-L 引理的应用 143
§4.5 习题 146
§4.6 章末注记 147

第五章 差分隐私 148
§5.1 数据隐私问题 149
§5.2 差分隐私的定义与性质 153
§5.3 差分隐私的应用 161
§5.3.1 随机反应算法 161
§5.3.2 全局灵敏度与 Laplace 机制 164
§5.3.3 DP 版本 Llyod 算法 167
§5.4 习题 170
§5.5 章末注记 170

第三部分　决策与优化 **171**

第六章　凸分析 **172**

§6.1 决策与优化的基本原理 173
§6.1.1 统计决策理论 173
§6.1.2 优化问题 176
§6.1.3 例子：网格搜索算法 183
§6.2 凸函数 188
§6.3 凸集 194
§6.3.1 基本定义和性质 195
§6.3.2 分离超平面定理 199
§6.4 习题 202
§6.5 章末注记 202

第七章　对偶理论 **202**

§7.1 约束的几何意义 205
§7.2 条件极值与 Lagrange 乘子法 213
§7.3 Karush–Kuhn–Tucker 条件 217
§7.4 Lagrange 对偶 222
§7.4.1 原始规划与对偶规划 222
§7.4.2 对偶的几何意义 227
§7.4.3 弱对偶定理 229
§7.4.4 Slater 条件，强对偶定理 230
§7.5 应用：支持向量机（SVM） 235
§7.6 习题 238
§7.7 章末注记 238

第八章 不动点理论 **239**

§8.1 Banach 不动点定理 240

§8.2 Brouwer 不动点定理 251

§8.3 习题 258

§8.4 章末注记 258

第四部分 博弈与逻辑 **259**

第九章 逻辑与博弈 **260**

§9.1 博弈的基本语言：以井字棋为例 262

§9.2 输赢博弈 264

§9.2.1 博弈的不同维度 264

§9.2.2 Zermelo 定理与 AlphaGo Zero 266

§9.3 正则形式博弈 273

§9.3.1 定义 274

§9.3.2 理性与均衡 276

§9.3.3 生成对抗网络 279

§9.3.4 混合策略 282

§9.4 随机博弈（Markov 博弈） 288

§9.5 习题 299

§9.6 章末注记 299

第五部分 认知与逻辑 **300**

第十章 共同知识，Bayes 博弈，Aumann 知识算子 **301**

§10.1 “泥泞的孩童”谜题 304

§10.2 不完全信息博弈（Bayes 博弈） 309
§10.3 电子邮件博弈 319
§10.4 Aumann 知识算子 324
§10.5 习题 333
§10.6 章末注记 333

第十一章 模态逻辑，知识的逻辑 334
§11.1 知识逻辑的形式语言 336
§11.2 Kripke 语义 341
§11.3 模态可定义性 348
§11.4 知识逻辑的基本模型与性质 351
§11.4.1 知识逻辑的 Kripke 模型与公理 351
§11.4.2 Kripke 模型与 Aumann 结构 358
§11.4.3 “泥泞的孩童”再回顾：形式化解法 360
§11.5 对不一致达成一致 362
§11.5.1 模型 363
§11.5.2 定理及其证明 366
§11.6 习题 368
§11.7 章末注记 368

第六部分 附录：预备知识 369

附录 A 线性代数基础 370
§A.1 线性空间 370
§A.2 线性映射 377
§A.3 矩阵 383

§A.4 双线性型与二次型 392
§A.5 带内积的线性空间 398
§A.6 行列式 407
§A.7 算子范数与谱理论 411

附录 B 微分学基础 **419**
§B.1 点集拓扑 419
§B.1.1 度量空间，范数 419
§B.1.2 开集与闭集 424
§B.1.3 紧致性，收敛性，完备性 428
§B.1.4 连续映射 432
§B.1.5 与实数序有关的性质 437
§B.2 一元函数的微分学 440
§B.2.1 导数与微分的定义 441
§B.2.2 微分学基本定理 446
§B.3 多元函数的微分学 448
§B.3.1 微分、偏导数与导数的定义 448
§B.3.2 微分学基本定理 458
§B.3.3 隐函数定理 461

附录 C 概率论基础 **467**
§C.1 从朴素概率论到公理化概率论 467
§C.1.1 Kolmogorov 概率论 467
§C.1.2 条件概率，独立性 473
§C.2 随机变量，分布函数 479
§C.2.1 基本定义 479

§C.2.2 离散型随机变量 485
§C.2.3 连续型随机变量 486
§C.2.4 随机向量，条件分布，独立性 491
§C.2.5 随机变量（向量）的函数 498
§C.3 随机变量的数字特征，条件数学期望 502
§C.3.1 数学期望，Lebesgue 积分 502
§C.3.2 数学期望的性质 509
§C.3.3 随机变量的内积空间 513
§C.3.4 特征函数 516
§C.3.5 条件数学期望 518
§C.4 多元正态分布（Gauss 向量） 525
§C.5 大数定律 527

第一部分

AI 的逻辑

第二部分

信息与数据

第三部分

决策与优化

第四部分

博弈与逻辑

第五部分

认知与逻辑

第六部分

附录：预备知识

附录 C　概率论基础

本附录主要介绍 Kolmogorov 概率论，讨论只局限在数学层面，不涉及概率论的哲学讨论. 本附录的连续型随机变量（向量）的讨论需要微积分的基本知识，关于微分学的部分，可以参见附录 B；积分学（主要是 Lebesgue 积分）我们会在附录 C.3 以数学期望的形式介绍.

§C.1　从朴素概率论到公理化概率论

§C.1.1　Kolmogorov 概率论

朴素的概率论通常讨论两种极端的情况，一个是可以用数数的方式来计算概率的情况，比如说掷骰子，另一个是用面积的方式来计算概率的情况，比如在随机选一个圆周上的点. 这两个情况分别对应了*古典概型*和*几何概型*.

我们先给一些术语. 考虑一个随机试验，它的所有可能结果组成的集合称为**样本空间**，记为 Ω. 样本空间的元素称为**样本点**，通常记为 ω. 样本空间的某些子集被称为**事件**. 我们来看看这些概念在朴素的概率论中都具体是什么.

例 C.1 (古典概型) 先后掷两个骰子，样本空间为

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

样本点 (i,j) 表示第一个骰子掷出 i 点，第二个骰子掷出 j 点.

“第一个骰子掷出 i 点”这个事件可以表示为 $A_i = \{(i,j) : 1 \le j \le 6\}$. “第一个骰子掷出 i 点，第二个骰子掷出 j 点”这个事件可以表示为 $B_{ij} = \{(i,j)\}$. □

例 C.2 (几何概型) 在圆周上随机选点. 如果用弧度来表示圆周上的点，那么样本空间为

$$\Omega = [0, 2\pi).$$

样本点为 ω，表示选出点的弧度.

事件 $A = [0,\pi)$ 表示选出了上半圆周，事件 $B = [0,\pi/2) \cup [\pi, 3\pi/2)$ 表示选出了右上或左下的 $1/4$ 圆周. □

那么，如何定义概率呢？朴素地说，概率是某个事件出现的可能性占总可能的比例.

对于古典概型，我们简单认为每个样本点出现的概率都是相同的，也就是说，如果用 p_ω 表示样本点 ω 出现的概率，那么对任意 $\omega \in \Omega$，都有 $p_\omega = 1/|\Omega|$. 于是，对于任意事件 A，它发生的概率为

$$\sum_{\omega \in A} p_\omega = \frac{|A|}{|\Omega|}.$$

例如在上面掷骰子的例子中，$p_\omega = 1/36$，A 发生的概率为 $1/6$，B 发生的概率为 $1/36$.

对于几何概型，不能再用古典概型的方式定义概率. 一段长为 2π 的圆弧上，有不可数个点. 如果选到每个点的概率相等，那么这个概率必须是 0，否则所有点的概率和是无穷大. 更麻烦的是，我们也不能用古典概型的方式计算某个事件的概率！例如，选到上半圆周的概率，就是把所有上半圆周上的点的概率加起来，任意多个 0 相加依然还是 0，所以这样的定义出来的概率永远是零，这样是不可行的.

朴素的直觉告诉我们，选到上半圆周的概率是 $1/2$，因为上半圆周刚好占了半个圆周. 所以几何概型的概率定义利用了*体积*的概念. 事件 A 的概率定义为

$$\frac{\text{事件 } A \text{ 对应的体积}}{\text{样本空间 } \Omega \text{ 对应的体积}}.$$

这里体积是广义上的，一维集合的体积就是长度，二维集合的体积就是面积，三维集合的体积就是体积，以此类推.

例如在上面圆周的例子中，A 对应的体积（长度）为 π，Ω 对应的体积（长度）为 2π，所以 A 发生的概率为 $1/2$. 同理，B 的概率也是 $1/2$.

几何概型的定义看似合理，却并不严谨：我们并不知道如何定义“体积”. 我们来看一个有趣的例子.

例 C.3 (Bertrand 悖论) 考虑一个圆，它的半径为 1. 现在我们随机地在圆上取一个弦，那么这个弦的长度超过 $\sqrt{3}$（即圆内接正三角形的边长）的概率是多少?

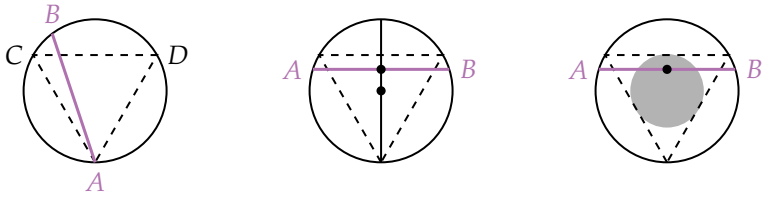
我们给出三种答案，这三种答案对应了我们对“随机”的不同理解.

解答 1. 不妨固定弦的其中一个点 A，另一个点 B 在圆上等可能选取. 以 A 为顶点作圆内接正三角形 ACD，弦的长度超过 $\sqrt{3}$ 等价于 B 在弧 CD 上，所以概率为 $1/3$.

解答 2. 弦长只与它到圆心的距离有关系，与方向无关. 弦长超过 $\sqrt{3}$ 等价于它到圆心的距离小于 $1/2$，所以概率为 $1/2$.

解答 3. 弦被它的中点唯一确定，弦长大于 $\sqrt{3}$ 等价于中点落在一个半径为 $1/2$ 的同心小圆内，所以概率为同心小圆面积比上大圆面积，即 $(1/2)^2 = 1/4$.

三种解答的示意图见下（从左到右分别是解答 1 到 3）：



□

同样的事件因为我们对“随机”的理解不同，得到了不同的概率！因此，我们需要一个更加严格的定义来描述概率.

首先注意到，概率应该是一个函数，它的值域是 $[0,1]$. 那么，它的定义域应该是什么呢？我们已经看到，概率应该定义在事件上，而非样本点上. 那么，概率可以定义在任意事件上吗？这个问题很微妙，我们不在这里讨论. 这里只是指出，我们关心的并不总是任意事件，而是一类被 σ-代数所刻画的事件.

定义 C.1 (σ-代数) 设 Ω 是一个集合，$\mathscr{F}$ 是 Ω 的子集的集合. 如果 $\mathscr{F}$ 满足

1. $\Omega \in \mathscr{F}$;
2. 如果 $A \in \mathscr{F}$，则 A 的补集 $\Omega \setminus A \in \mathscr{F}$;

3. 如果 $A_1, A_2, \ldots \in \mathscr{F}$，则 $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

则称 $\mathscr{F}$ 是 Ω 上的一个 **σ-代数**. □

在样本空间中，我们要求事件也形成一个 σ-代数，这样的 σ-代数称为**事件域**，记为 $\mathscr{F}$. 在数学上，σ-代数包括了绝大部分我们可以构造的事件，这是因为，容易验证，σ-代数中的事件对可数交、可数并和补运算都是封闭的，并且包含了样本空间和空集. 关于这一定义的哲学讨论，可以见第一章.

样本空间连同它的事件域，被称为**可测空间**.

定义 C.2 (可测空间) 设 Ω 是一个集合，$\mathscr{F}$ 是 Ω 上的一个 σ-代数. 则称 $(\Omega, \mathscr{F})$ 是一个**可测空间**.

设 $S \subseteq \Omega$，如果 $S \in \mathscr{F}$，则称 S 是 **$\mathscr{F}$-可测的**. □

定义可测空间与 $\mathscr{F}$-可测的概念，主要是为了区分一个集合到底是不是我们所关心的事件，我们只关心 $\mathscr{F}$-可测的集合.

接下来，我们给出 Kolmogorov 概率论的公理化定义.

定义 C.3 (概率空间，概率测度) 设 $(\Omega, \mathscr{F})$ 是一个可测空间. 如果函数 $\Pr : \mathscr{F} \to [0, 1]$ 满足

1. 正则性：$\Pr(\Omega) = 1$;
2. 可列可加性：如果 $A_1, A_2, \cdots \in \mathscr{F}$ 是两两不相交的事件，则

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i),$$

则称 $(\Omega, \mathscr{F}, \Pr)$ 是一个**概率空间**，$\Pr$ 称为**概率测度**或**概率**. □

容易证明，概率有如下性质：

命题 C.1 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，则：

1. $\Pr(\varnothing) = 0$；
2. 单调性：对任意的 $A, B \in \mathscr{F}$，如果 $A \subseteq B$，则 $\Pr(A) \leq \Pr(B)$；
3. 有限可加性：对两两不相交的 $A_1, A_2, \ldots, A_n \in \mathscr{F}$，有

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

他们的证明都不困难，我们略去.

下面，我们回到古典概型与几何概型，看看如何对他们构造概率空间.

对于古典概型来说，我们容易写出它的概率空间. 此时事件域恰好为所有 Ω 的子集的集合，概率测度的定义也就是我们之前的定义：$\Pr(A) = |A|/|\Omega|$.

对于几何概型来说，概率空间最大的困难在于事件域和概率测度的定义. 为了简化讨论，我们集中在 $\Omega = [0,1]^n$，也就是 n 维立方体的情况.

先考虑事件域. 首先，事件域至少要包含长方体

$$\prod_i (a_i, b_i) = \{x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : a_i < x_i < b_i\}.$$

这是我们可以构造的最基本的集合了. 我们就定义事件域为包含所有长方体的最小 σ-代数 $\mathscr{B}([0,1]^n)$. 换言之，如果还有一个 σ-代数 $\mathscr{F}$ 包含所有长方体，那么 $\mathscr{B}([0,1]^n) \subseteq \mathscr{F}$. 我们将这一 σ-代数称为 **Borel 代数**.

Borel 代数包含了绝大部分我们要讨论的集合，例如开集、闭集、单点集、有限集、可数集等，可以简单归纳为“合理的集合”.

事件域的定义已经给出，我们还需要定义概率测度 Pr，它应该满足以下两个要求：

- 让正方体的概率等于它的体积. 按照朴素的直觉，长方体 $\prod_i (a_i, b_i)$ 的体积应该是 $\prod_i (b_i - a_i)$，也就是

$$\Pr\left(\prod_i (a_i, b_i)\right) = \prod_i (b_i - a_i).$$

- 平移不变性. 也就是说，如果 $A \in \mathscr{B}([0,1]^n)$，那么对任意的 $x \in \mathbb{R}^n$，定义 $A + x = \{y \in \mathbb{R}^n : y = x + z, z \in A\}$，只要 $A \in \mathscr{B}([0,1]^n)$，就有 $\Pr(A + x) = \Pr(A)$.

一个惊人的事实是，这样的概率测度是存在且唯一的，我们称之为 **Lebesgue 测度**，常记为 λ.

注意，Borel 代数和 Lebesgue 测度的定义可以不局限在 $[0,1]^n$，他们可以定义在与实数相关的各种集合上. 在本附录中，我们最主要是用的是 $\mathbb{R}^n$ 上的相关定义，例如 $\mathscr{B}(\mathbb{R}^n)$ 就是包含所有 n 维开长方体（每条边是开区间）的最小 σ-代数，λ 就是定义在 $\mathscr{B}(\mathbb{R}^n)$ 上的 Lebesgue 测度. $\mathbb{R}^n$ 上的 Lebesgue 测度其实是概率测度的扩展（而非概率测度），因为此时不再要求有正则性（即 $\lambda(\Omega) = 1$），但额外要求 $\lambda(\varnothing) = 0$.

§C.1.2 条件概率，独立性

接下来，我们讨论条件概率与独立性. 我们还是看掷两个骰子的例子. 掷完第一个骰子，我们马上观察结果，然后再掷第二个骰子. 问第

一个骰子是 i，第二个是 j 的概率是多少？如果继续套用原来的概率空间，很快就会觉得不对劲. 此时，第一个骰子的结果完全没有随机性！所以朴素的直觉告诉我们，这里的概率应该有另一个依赖于第一次投骰子结果的定义，这样的概率就是*条件概率*.

我们直接给出一般情况下条件概率的定义.

定义 C.4 (条件概率) 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A, B \in \mathscr{F}$ 是两个事件，且 $\Pr(A) > 0$. 则称

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

是事件 B 在事件 A 发生的条件下发生的**条件概率**. □

以上定义要求 A 发生概率为正，然而，A 是零概率的时候也是可能有条件概率的. 例如，从 $[0,1] \times [0,1]$ 中均匀地随机选一个点 (X, Y)，观察它的横坐标 X. 不管什么样的 x，$X = x$ 的概率都是 0. 然而，直觉上，条件在 $X = x$ 上，$Y > 1/2$ 的概率不仅存在，而且应该是 $1/2$. 在附录 C.2 中，我们会针对一类特殊的事件，给出此时条件概率的定义.

我们继续看投两个骰子的例子. 假设事件 A 是“第一个骰子是 i“，事件 B 是“第二个骰子是 j”. 我们可以计算出 $\Pr(B|A) = \Pr(B) = 1/6$. 如果单看计算的结果，这是一个非常神奇的式子：条件在 A 上和不条件在 A 上概率是一样的！从直觉来说，这件事情却并不神秘，因为第一个骰子的结果和第二个骰子的结果是不应该有关系的. 我们把这种现象称为**独立性**. 更一般地，对任意事件 A, B，如果 $\Pr(A) > 0$，那么

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

最后一个式子并不要求 $\Pr(A) > 0$，因此我们用它作为独立性的定义，这样定义可以不依赖条件概率.

定义 C.5 (独立性) 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A, B \in \mathscr{F}$ 是两个事件. 如果 $\Pr(A \cap B) = \Pr(A)\Pr(B)$，则称事件 A 和 B **相互独立**.

一般地，给定一个事件族 $\mathscr{A} \subseteq \mathscr{F}$，如果对任意的有限个不同的 $A_1, A_2, \ldots, A_n \in \mathscr{A}$，都有

$$\Pr\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \Pr(A_i),$$

则称事件族 $\mathscr{A}$ 中的事件是**相互独立**的. □

我们在定义中还给出了多个事件相互独立的定义，这一定义是说不管挑出其中多少有限个事件，他们都应该满足交的概率等于概率的积. 这并不等价于任意两个事件都相互独立，我们看下面的例子.

例 C.4 两个人进行石头剪刀布游戏，每个人独立等概率地出剪刀石头布.

考虑下面三个事件：$A = \{\text{甲出了石头}\}$，$B = \{\text{乙出了剪刀}\}$，$C = \{\text{甲赢}\}$.

容易算出，$\Pr(A \cap B) = \Pr(A)\Pr(B) = 1/9$，$\Pr(A \cap C) = \Pr(A)\Pr(C) = 1/9$，$\Pr(B \cap C) = \Pr(B)\Pr(C) = 1/9$，所以 A, B, C 两两独立.

但是 A, B, C 不是相互独立的：$\Pr(A \cap B \cap C) = 1/9 \neq 1/27 = \Pr(A)\Pr(B)\Pr(C)$. □

这个例子说明，三个事件的独立性远比他们任意两个之间的独立性要复杂，三个事件放在一起可能才会出现不独立的情况. 对于一般情

况，这样的现象更加普遍，所以我们多个事件的独立性定义是要求任意有限个事件都独立，而不是任意两个事件都独立.

最后，我们给出条件概率的一些性质.

命题 C.2 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，那么

1. 对任意 $A \in \mathscr{F}$ 满足 $\Pr(A) > 0$，$\Pr(\cdot|A)$ 也是一个概率测度；
2. $\Pr(\cdot|\Omega) = \Pr(\cdot)$，
3. 对任意 $A \in \mathscr{F}$ 满足 $\Pr(A) > 0$，$\Pr(A|A) = 1$.

以上性质的证明都很简单，我们就不给出了.

接下来我们给两个在 Bayes 概率论以及随机过程中很重要的性质.

定理 C.1 (全概率公式) 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A_1, A_2, \ldots \in \mathscr{F}$ 是一列两两不相交的事件，且 $\Pr(A_i) > 0$，$\bigcup_{i=1}^{\infty} A_i = B$，则对任意的 $C \in \mathscr{F}$，有

$$\Pr(C|B) = \sum_{i=1}^{\infty} \Pr(C|A_i) \Pr(A_i).$$

特别地，对于有限个 A_i，这一定理也成立.

证明. 注意到

$$\Pr(C) = \Pr(C \cap B) = \Pr\left(C \cap \bigcup_{i=1}^{\infty} A_i\right) = \Pr\left(\bigcup_{i=1}^{\infty} (C \cap A_i)\right) = \sum_{i=1}^{\infty} \Pr(C \cap A_i).$$

最后一个等号是因为 $C \cap A_i$ 两两不相交. 另一方面，

$$\Pr(C \cap A_i) = \Pr(C|A_i) \Pr(A_i),$$

所以

$$\Pr(C) = \sum_{i=1}^{\infty} \Pr(C|A_i)\Pr(A_i).$$

对于有限个 A_i，只需要把无穷求和改成有限求和，利用有限可加性即可. □

全概率公式是一种分而治之的思想，它把一个复杂的事件分解成若干个简单的事件，然后再把简单的事件的概率加起来. 我们来看一个例子.

例 C.5 从装有 w 个白球和 b 个黑球的盒子中随机地取出一个球，不放回，再取出一个球. 问第二个球是白球的概率是多少?

设事件 A 是“第一个球是白球”，事件 B 是“第二个球是白球”. 我们有

$$\begin{aligned}\Pr(B) &= \Pr(B|A)\Pr(A) + \Pr(B|\bar{A})\Pr(\bar{A}) \\ &= \frac{w-1}{w+b-1}\cdot\frac{w}{w+b} + \frac{w}{w+b-1}\cdot\frac{b}{w+b} \\ &= \frac{w}{w+b}.\end{aligned}$$

这里 $\bar{A}$ 指的是 A 的补集，即“第一个球是黑球”. □

定理 C.2 (贝叶斯公式) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间，$A, B \in \mathcal{F}$ 且 $\Pr(A), \Pr(B) > 0$，则

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

这一公式的证明几乎是显然的，我们略去.

一个特别重要的推论被称为链式法则，它是 *Bayes* 网络的基础.

推论 C.1 (链式法则) 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A_1, A_2, \ldots, A_n \in \mathscr{F}$，且 $\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) > 0$，则

$$\begin{aligned}&\Pr(A_1 \cap A_2 \cap \cdots \cap A_n)\\ =&\Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_1 \cap A_2)\cdots\Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).\end{aligned}$$

我们也看一个例子.

例 C.6 (Pólya 的罐子) 一个罐子装有 w 个白球和 b 个黑球，随机取出一个，观察它的颜色，放回，再放回相同颜色的 c 个球，再随机取一次，重复上述操作，如此反复 n 次，问每一次都取到白球的概率是多少?

设事件 A_i 是"第 i 次取出的球是白球". 我们有

$$\begin{aligned}\Pr(A_1) &= \frac{w}{w+b},\\ \Pr(A_2|A_1) &= \frac{w+c}{w+b+c},\\ \Pr(A_3|A_1 \cap A_2) &= \frac{w+2c}{w+b+2c},\\ &\cdots\\ \Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) &= \frac{w+nc}{w+b+nc}.\end{aligned}$$

所以

$$\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) = \frac{w}{w+b} \cdot \frac{w+c}{w+b+c} \cdots \frac{w+nc}{w+b+nc}. \qquad \square$$

注. 在概率论中，我们经常要讨论事件的交，所以我们通常会把 $A \cap B$ 简记为 AB. 此外，事件不相交我们也称之为**互斥**. 事件 A 的补事件，即 $\Omega \setminus A$. 我们会记为 $\bar{A}$ 或 A^c.

另外，我们也经常要讨论一个关于 ω 的陈述 $Q(\omega)$ 定义的事件 $\{\omega \in \Omega : Q(\omega)\}$，在 Pólya 的罐子的例子中，事件 A_1 其实就是由陈述 $Q(\omega)$："ω 中第一次取出的球是白球"定义的事件. 在这种情况下，我们将这一事件简记为 $\{Q\}$，它的概率就是 $\Pr(\{Q\})$ 或者简记为 $\Pr(Q)$.

最后，事件交的概率也经常以逗号的形式写出. 例如，在 Pólya 的罐子的例子中，我们会把概率 $\Pr(A_1A_2)$ 记为

$$\Pr(\text{第一次取出的球是白球}, \text{第二次取出的球是白球}).$$

这样的记号更直观，并且在随机变量部分会经常使用.

§C.2 随机变量，分布函数

接下来，我们讨论随机变量. 从某种意义上说，随机变量是另一种刻画概率测度的手段. 不过，随机变量能够更加直观、定量描述概率空间中的事件，所以这是一个更加容易使用的概念.

§C.2.1 基本定义

为了理解随机变量的概念，我们依然从古典概型入手.

例 C.7 继续考虑先后投两个骰子的情况，假设它的概率空间是 $(\Omega, \mathscr{F}, \Pr)$，他们的定义我们在附录 C.1.1 的末尾已经讨论过了.

我们可以定义一个从样本空间 Ω 到 $\mathbb{N}$ 的函数 $S(i,j) = i + j$，也就

是两个点数的和. 我们来看看 S 与事件域的关系. $\{S = s\} = \{(i,j) \in \Omega : i+j = s\}$，所以 S 将原本的事件精简成了一个数字. 这个过程丢弃了一些事件，例如 S 无法表达事件 $\{(1,2)\}$，实际上，它无法区分 $(1,2)$ 和 $(2,1)$，它把这两个样本点都看成了 3. 但是，S 仍然保留了很多信息，例如，S 可以区分事件 $\{(1,1)\}$ 和 $\{(2,2)\}$，它们分别对应 2 和 4. 总结来说，S 将原本更精细的事件域压缩成了更粗糙的事件域.

有了上面的感觉，我们可以看一个更抽象的函数. 定义一个从样本空间 Ω 到 $\mathbb{N}^2$ 的函数 X，它的定义为 $X(i,j) = (i,j)$. 换句话说，它把样本点看成一个 $\mathbb{N}^2$ 的元素. $\mathscr{F}$ 中的所有事件都可以表达为 $\{X \in B\}$，$B \subseteq \mathbb{N}^2$. 所以 X 完全刻画了整个事件域. □

上面例子中的 S 和 X 都是随机变量的例子. 我们给出随机变量的定义.

定义 C.6 (随机变量，随机向量，Borel 函数) 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}$ 是一个函数. 如果对任意的 $x \in \mathbb{R}$，$\{X \in \mathscr{B}(\mathbb{R})\} \in \mathscr{F}$，则称 X 是一个**随机变量**.

一般地，考虑一个集合 $\mathbb{R}^n$ 以及其上的 Borel 代数 $\mathscr{B}(\mathbb{R}^n)$，$X : \Omega \to \mathbb{R}^n$ 是一个映射. 如果对任意的 $A \in \mathscr{B}(\mathbb{R}^n)$，集合 $\{X \in A\}$ $\mathscr{F}$-可测，即 $\{X \in A\} \in \mathscr{F}$，则称 X 是一个 n **维随机向量**，简称**随机向量**. 如果 $(\Omega, \mathscr{F}) = (\mathbb{R}^m, \mathscr{B}(\mathbb{R}^m))$，则称 X 是一个 **Borel 函数**. □

下面对这个定义做一些说明. 首先，随机变量是一个映射，而不是一个数字，这一点经常会被人误解. 直观上说，随机变量的值是随机的，这个随机性是因为背后有一个未知的力量在抛硬币，我们把从抛硬币到观测值这一整个东西称之为随机变量.

定义的后面还涉及了 σ-代数相关的东西，我们也给一个简要说明.

Borel 代数包含了“合理的集合”，所以 $\{X \in B\}$（$B \in \mathscr{B}(\mathbb{R})$）表示事件“$X$ 落在合理的值集上”. 随机变量的要求其实就是，“X 落在合理的值集上”是一个我们可以定义概率（即可测）的事件.

我们下面讨论一些随机变量的基本性质.

定理 C.3 *设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}^n$ 是一个随机向量，$g : \mathbb{R}^n \to \mathbb{R}^m$ 是一个 Borel 函数，则 $g(X) = g \circ X$ 也是一个随机向量.*

这一性质告诉我们了一种构造随机变量的方式：对一个随机变量进行一些 Borel 函数的操作. 下面的性质告诉我们，Borel 函数包含了绝大部分我们关心的函数，因此在实际中，我们不需要担心一个映射作用完之后是否还是随机变量.

命题 C.3 *下面函数是 Borel 函数：*

1. *所有的连续函数；*
2. *给定 $A \in \mathscr{B}(\mathbb{R}^n)$，示性函数 $I_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$*
3. *两个 Borel 函数的复合函数.*

接下来，我们进入分布函数的讨论. 我们说过，随机变量某种意义上给出了概率测度的另一种刻画方式，而这一桥梁就是由分布函数给出的.

考虑概率空间 $(\Omega, \mathscr{F}, \Pr)$，以及一个随机变量 $X : \Omega \to \mathbb{R}$. 要刻画概率测度 $\Pr$，我们需要给出所有的事件 $A \in \mathscr{F}$ 的概率 $\Pr(A)$. 如果 A 可以被写成 $\{X \in B\}$ 的形式，那么我们可以用 $\Pr(X \in B)$

来刻画 $\Pr(A)$. 而我们之前说过，要确定 $\Pr(X \in B)$，至少要先确定 $\Pr(X \in (a,b))$. 这一概率还是有两个未定元 a,b，所以更简便的方式是确定 $F_X(b) = \Pr(X \in (-\infty, b])$，容易证明，开区间的概率完全可以由 $F_X(b)$ 给出，所以 F_X 完全刻画了 $\Pr$. 更一般地，我们有如下定义.

定义 C.7 (分布函数) 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}$ 是一个随机变量. 定义函数 $F_X : \mathbb{R} \to \mathbb{R}$ 为 $F_X(x) = \Pr(X \leq x)$. 我们称 F_X 是 X 的**分布函数**，记作 $X \sim F_X$.

如果 $X : \Omega \to \mathbb{R}^n$ 是一个随机向量，定义函数 $F_X : \mathbb{R}^n \to \mathbb{R}$ 为 $F_X(x) = \Pr(X \leq x)$，这里 $X \leq x$ 是指对任意的 $i = 1, 2, \ldots, n$，都有 $X_i \leq x_i$. 我们称 F_X 是 X 的**分布函数**，记作 $X \sim F_X$. □

容易验证，分布函数具有如下的性质：

命题 C.4 *设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}$ 是一个随机变量，F_X 是它的分布函数，则*

1. *F_X 是一个非减函数；*
2. *$\lim_{x \to -\infty} F_X(x) = 0$，$\lim_{x \to +\infty} F_X(x) = 1$；*
3. *F_X 是右连续的，即对任意 $x \in \mathbb{R}$，都有 $\lim_{y \downarrow x} F_X(y) = F_X(x)$；*
4. *F_X 在每一点处的左极限存在，即对任意 $x \in \mathbb{R}$，都有 $F(x-) = \lim_{y \uparrow x} F_X(y)$ 存在.*

实际上，分布函数也可以由命题 C.4 的前三条性质给出定义，这是因为，满足前三条性质的函数恰好是某个随机变量的分布函数：

定理 C.4 *设 F 是 $\mathbb{R} \to \mathbb{R}$ 的函数，满足命题 C.4 的前三条性质.*

在概率空间 $([0,1],\mathscr{B}([0,1]),\lambda)$ 上，存在一个随机变量 X，使得 $F_X = F$.

所以，我们今后也称呼满足命题 C.4 四条性质的函数为分布函数.

我们看一个分布函数计算概率的简单例子.

例 C.8 考虑 $\mathbb{R}$ 上的分布函数 F，它由随机变量 X 定义. 那么，

- $\Pr(X \leq a) = F(a)$,
- $\Pr(X < a) = F(a-)$,
- $\Pr(X > a) = 1 - F(a)$,
- $\Pr(X \geq a) = 1 - F(a-)$,
- $\Pr(X = a) = F(a) - F(a-)$. □

对于 $\mathbb{R}^n \to \mathbb{R}$ 型的分布函数 F，我们也有类似的讨论. 此时有多个维度，所以我们需要引入一个差分算子 $\Delta_{a_i b_i}$，它的作用是对第 i 维作差：

$$\begin{aligned}&\Delta_{a_i b_i} F(x_1, x_2, \ldots, x_n)\\ =&F(x_1, x_2, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_n) - F(x_1, x_2, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n).\end{aligned}$$

例如，对于区间 $(a,b] = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : a_i < x_i \leq b_i\}$，我们有

$$\Pr(X \in (a,b]) = \Delta_{a_1 b_1} \Delta_{a_2 b_2} \cdots \Delta_{a_n b_n} F_X(x_1, x_2, \ldots, x_n).$$

容易证明，分布函数具有如下的性质：

命题 C.5 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}^n$ 是一个随机向量，F_X 是它的分布函数，则

1. 对任意 $a_i \le b_i$，$i = 1,2,\dots,n$，都有 $\Delta_{a_i b_i} F_X(x_1, x_2, \dots, x_n) \ge 0$；
2. 所有 x_i 趋于正无穷时，F_X 趋于 1；任意一个 x_i 趋于负无穷时，F_X 趋于 0；
3. F_X 对所有的 x_i 都是右连续的，即当 $y \downarrow x$（即对所有分量都有 $y_i \downarrow x_i$）时，都有 $F_X(y) \to F_X(x)$.

同样，以上三条性质就决定了一个分布函数. 我们有如下的定理：

定理 C.5 设 F 是 $\mathbb{R}^n \to \mathbb{R}$ 的函数，满足命题 C.5 的三条性质.

在概率空间 $([0,1]^n, \mathscr{B}([0,1]^n), \lambda)$ 上，存在一个随机向量 X，使得 $F_X = F$.

因此，我们今后也称呼满足命题 C.5 三条性质的函数为分布函数.

> **注.** 定理 C.4 和定理 C.5 其实还发挥着另一个重要的作用. 随机变量和随机向量的定义是非常抽象的，所以我们并不能很直接验证随机变量的存在性. 然而，分布函数却是极其容易构造的. 所以利用分布函数的存在性我们可以确保随机变量的存在性.

如果我们就限制在空间 $([0,1]^n, \mathscr{B}([0,1]^n), \lambda)$ 上，随机向量几乎就等同于分布函数. 在更一般的情况下，两个随机向量 X, Y 的分布函数相同时，我们称 X, Y 同分布，记为 $X \overset{d}{=} Y$.

现在，我们将分布函数与概率测度联系在一起：

定理 C.6 设 $F : \mathbb{R}^n \to \mathbb{R}$ 是一个分布函数，则在可测空间 $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ 上，存在唯一的概率测度 $\Pr$，使得对任意 $a_i \le b_i$，

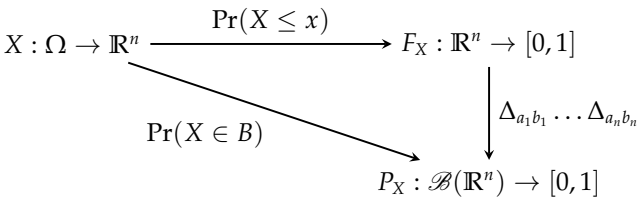
$$\Pr\left(\prod_{i=1}^{n}(a_i, b_i]\right) = \Delta_{a_1 b_1}\Delta_{a_2 b_2}\cdots\Delta_{a_n b_n} F(x_1, x_2, \dots, x_n).$$

特别地，分布函数

$$F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

对应的概率测度就是我们之前讨论的 $[0,1]$ 上的 Lebesgue 测度.

总结来说，随机向量 X、概率测度 P_X 和分布函数 F_X 的关系如图：

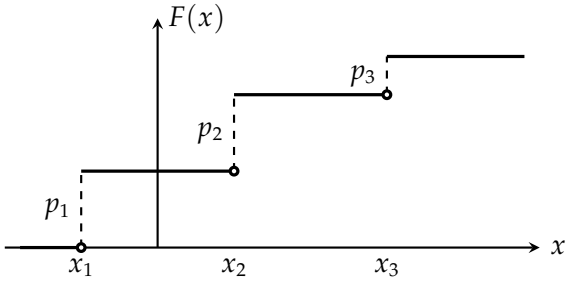


这张图的每一个箭头都可以反过来，但是反过来的这些关系都比较不直观，所以我们不再讨论.

根据上面的讨论，分布函数的特性决定了随机变量的特性. 根据分布函数的不同性质，我们可以将随机变量分为不同的类型. 下面我们将讨论一些重要的类别.

§C.2.2 离散型随机变量

我们首先讨论离散型随机变量. 离散型随机变量的分布函数 F 称之为离散型分布，它是一个阶梯函数，它的函数值只在有限或者可数个点 $x_1, x_2, \dots$ 上发生跳变，在 x_i 的跳变为 $p_i = F(x_i) - F(x_i-)$. 这一分布函数对应的概率测度 Pr 我们称之为离散型测度，这种测度集中在 x_i 上，即 $\Pr(X = x_i) = p_i$. 分布函数形如下图：



离散型分布可以由分布列给出，分布列是一个序列 $p_1, p_2, \ldots$，其中 $p_i = \Pr(X = x_i)$，且 $\sum_{i=1}^{\infty} p_i = 1$.

表 C.1 列举了一些本书中用到的离散型分布，他们都是整数取值，所以我们记 $p_i = \Pr(X = i)$.

名称	符号	分布列	参数
离散均匀	$\mathcal{U}[n]$	$p_i = 1/n,\ i = 1, \ldots, n$	$n \in \mathbb{N}$
Bernoulli	$B(1,p)$	$p_1 = p,\ p_0 = 1-p$	$p \in [0,1]$
对称 Bernoulli	—	$p_1 = p_{-1} = 1/2$	—
二项	$B(n,p)$	$p_k = \binom{n}{k} p^k (1-p)^{n-k}$	$n \in \mathbb{N},\ p \in [0,1]$

表 C.1: 本书中用到的离散型分布

§C.2.3 连续型随机变量

我们再来讨论*连续型随机变量*，连续型随机变量的分布函数 F 称为*连续型分布*，对应的概率测度 Pr 称之为*绝对连续测度*. 从名字上就可以看出，测度才是定义连续型随机变量的关键. 我们给出绝对连续测度的定义.

定义 C.8 (绝对连续测度) $\mathbb{R}$ 上的测度 Pr 称为**绝对连续测度**，如果对任意 $\epsilon > 0$，存在 $\delta > 0$ 使得任意 $A \in \mathscr{B}(\mathbb{R})$ 满足 $\lambda(A) < \delta$，都有 $\Pr(A) < \epsilon$. □

直观上说，绝对连续测度的意思是当体积 $\lambda(\cdot)$ 发生微小变化的时候（变化量为 $\lambda(A)$），测度 $\Pr(\cdot)$ 也只发生微小的变化（变化量为 $\Pr(A)$），这和通常函数连续的定义并没有太大的区别.

那么，绝对连续测度对应的是连续分布函数吗？并非如此！不过，绝对连续测度对应的分布函数有相当漂亮的一种刻画方式：

定理 C.7 (微积分基本定理) *设 $F : \mathbb{R} \to \mathbb{R}$ 是绝对连续测度对应的分布函数，那么*

$$\lambda(\{x \in \mathbb{R} : F'(x) \text{ 不存在}\}) = 0.$$

定义函数：

$$f(x) = \begin{cases} F'(x), & F'(x) \text{ 存在}, \\ 0, & \text{其他}. \end{cases}$$

则 f 是一个非负可积函数，且对任意的 $a < b$，都有

$$F(b) - F(a) = \int_a^b f(x)\mathrm{d}x. \tag{C.1}$$

此处的积分可以理解为 Riemann 积分或者后面附录 C.3 中的 Lebesgue 积分.

定理 C.7 意味着，绝对连续测度对应的分布函数几乎处处可以求导，并且所得到的导函数积分回去还是原来的分布函数，也就是微积分基本定理成立. 这样的函数我们称之为*绝对连续函数*.

那么，这个 f 应该如何理解呢？先不管定理 C.7，回到绝对连续测

度，仿照导数的定义，考虑极限

$$\frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x) = \lim_{\lambda(A)\to 0, x\in A} \frac{\Pr(A)}{\lambda(A)},$$

也就是点 x 附近 $\Pr(\cdot)$ 的微小变化相对于 $\lambda(\cdot)$ 的微小变化.

那么，给定一个集合 A，要如何求 $\Pr(A)$？按照微积分的朴素直观，我们应该将 Pr 微小的变化转变为 λ 微小的变化，也就是积分：

$$\Pr(A) = \int_{x\in A} \frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x)\mathrm{d}\lambda(x).$$

我们可以把(C.1)改写成如上的形式：

$$\Pr((a,b]) = \int_{x\in(a,b]} f(x)\mathrm{d}x.$$

在一维的情况下，x 的微小变化就是 $\lambda(x)$ 的微小变化，所以 $\mathrm{d}x = \mathrm{d}\lambda(x)$. 综合这两点，我们容易相信，

$$f(x) = \frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x) \iff \mathrm{d}\Pr = f(x)\mathrm{d}\lambda.$$

所以，f 应该理解为“密度”. 打个比方，λ 是物体的体积，Pr 是物体的质量，那么 f 就是这个物体每个很小的部分上的体积质量除以体积，也就是密度. 所以，我们将 f 称之为**概率密度函数**，或者简称**密度**. 通常，X 的密度记作 p_X.

那么，概率测度和密度的区别是什么呢？对于刚接触概率论的人来说，似乎很难理解他们之间的区别. 比如说，有时候会写 $p(X=x)$ 甚至 $\Pr(X=x)$ 来表示密度在 x 处的值 $p(x)$，又或者，用 $\int \Pr(X=x)\mathrm{d}x$

来表示对密度的积分. 这些当然都是不对的，我们下面慢慢论述.

首先，根据定理 C.7，F 是连续函数，所以根据例 C.8，$\Pr(X = x) = F(x) - F(x-) = F(x) - F(x) = 0$. 所以 $\Pr(X = x)$ 根本就是零，它和密度函数没有任何关系，所以上面这些写法都是错的.

那么，要怎么理解密度 $p(\cdot)$ 和概率测度 $\Pr(\cdot)$ 的区别呢？当然，从定义的角度他们就完全不同：一个是从实数到实数的映射，一个是从实数的集合到实数的映射. 但是这样的区别对于初学者来说并不直观. 最直观的区别就在于密度这一词：虽然铅很重（密度大），但是几亿倍于铅体积的棉花却应该比铅重. 所以，密度是微观的，刻画很小部分集合的概率值，也就是 $\mathrm{d}\Pr = p_X \mathrm{d}\lambda$；而概率刻画的是宏观的，计算任何一个集合的概率，也就是 $\Pr(X \in A)$.

> **注.** 上面的记号 $\mathrm{d}\Pr / \mathrm{d}\lambda$ 并不是随意写出来的，我们叫它导数也不是随意的. 在测度论中，定理 C.7 可以被推广为 **Radon-Nikodym 定理**，这一定理直接保证了形如 $\mathrm{d}\Pr / \mathrm{d}\lambda$ 的函数的存在性，这一函数被称之为 **Radon-Nikodym 导数**.

利用密度，我们可以很容易计算概率：

定理 C.8 *设 X 是一个连续型随机变量，f 是它的密度函数，则对任意的 $B \in \mathscr{B}(\mathbb{R})$，都有*

$$\Pr(X \in B) = \int_{x \in B} f(x) \mathrm{d}x.$$

在表 C.2 中，我们给出本书中用到的一些连续型分布的密度函数.

> **注.** 从定理 C.7 来看，密度函数的定义似乎是唯一的，但是从积分的角度，如果密度函数在几个点上的值发生了变化，并不影响整个积分的值，

名称	符号	密度函数	参数
连续均匀	$\mathcal{U}(a,b)$	$p(x)=\frac{1}{b-a},\ x\in[a,b]$	$a<b$
指数	$\mathsf{Exp}(\lambda)$	$p(x)=\lambda \mathrm{e}^{-\lambda x},\ x\geq 0$	$\lambda>0$
双指数	$\mathsf{DExp}(\lambda)$	$p(x)=\frac{\lambda}{2}\mathrm{e}^{-\lambda\lvert x\rvert},\ x\in\mathbb{R}$	$\lambda>0$
Laplace	$\mathsf{Lap}(\mu,\lambda)$	$p(x)=\frac{\lambda}{2}\mathrm{e}^{-\lambda\lvert x-\mu\rvert},\ x\in\mathbb{R}$	$\mu\in\mathbb{R},\ \lambda>0$
正态（Gauss）	$\mathcal{N}(\mu,\sigma^2)$	$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x\in\mathbb{R}$	$\mu\in\mathbb{R},\ \sigma>0$

表 C.2: 本书中用到的连续型分布

从而也不影响求概率. 比如均匀分布 $\mathcal{U}(a,b)$，端点 a,b 的值到底是 0 还是 $1/(b-a)$ 并不重要，取任何一个值都是可以的.

注. 密度函数通常是需要分段写出的，比如，$U(a,b)$ 的密度函数，严格来说应该写为

$$p(x)=\begin{cases}0, & x<a,\\ \frac{1}{b-a}, & x\in[a,b],\\ 0, & x>b.\end{cases}$$

为了简化记号，我们可以用**示性函数**来表示这一分类. 设 $A\subseteq\mathbb{R}$，定义函数

$$I_A(x)=\begin{cases}1, & x\in A,\\ 0, & x\notin A.\end{cases}$$

则 $U(a,b)$ 的密度函数可以写为

$$p(x)=\frac{1}{b-a}I_{[a,b]}(x).$$

更一般地，示性函数中的字母 A 可以是任意一个事件，而关于事件的那些记号都可以在 A 这里写出. 示性函数在概率论中有着核心的作用，

我们在后面将会经常用到示性函数.

§C.2.4 随机向量，条件分布，独立性

我们前面已经说过，随机向量就是 $\Omega \to \mathbb{R}^n$ 的映射. n 维的随机向量可以看成 n 个随机变量的组合，可以写作 $X = (X_1, \ldots, X_n)^{\mathsf{T}}$. 通常，我们将 X 的分布函数称为 $X_1, \ldots, X_n$ 的*联合分布*，将 X_i 的分布函数称为 X 的*边缘分布*.

关于随机变量的分类可以完全平行移植到随机向量上. 下面我们分别讨论.

离散型随机向量指的是它对应的概率测度集中在有限或可数个点上. 这样的分布依然可以用分布列给出：$\Pr(X_1 = x_1, \ldots, X_n = x_n) = p_{x_1,\ldots,x_n}$，其中 x_i 取遍所有可能的值.

本书中使用的离散型随机向量只有多项分布，符号为 $PN(n, p_1, \ldots, p_k)$，分布列为

$$\Pr(X_1 = i_1, \ldots, X_n = i_n) = \frac{n!}{i_1! \cdots i_k!} p_1^{i_1} \cdots p_k^{i_k},$$

其中 $n \in \mathbb{N}$，$p_i \geq 0$，$\sum_{i=1}^{k} p_i = 1$.

连续型随机向量指的是它对应的概率测度是绝对连续的. 连续型随机向量的分布函数依然由绝对连续函数刻画：

定理 C.9 *设 $F : \mathbb{R}^n \to \mathbb{R}$ 是绝对连续测度对应的分布函数，那么存在一个非负可积函数 $f : \mathbb{R}^n \to \mathbb{R}$，使得对任意的 $(x_1, \ldots, x_n) \in \mathbb{R}^n$，都有*

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n.$$

此时，f 称为 X 的**概率密度函数**，或者简称**密度**. 通常，X 的密度记作 p_X.

类似随机变量的讨论，密度函数依然可以被写做导数的形式. 假设 Pr 是绝对连续测度，它对应的密度是 p，那么

$$\frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x) = p(x) \iff \mathrm{d}\Pr = p(x)\mathrm{d}\lambda.$$

这里，我们需要再给出一些 $\mathrm{d}\lambda$ 和 $\mathrm{d}x$ 关系的讨论. $\mathrm{d}\lambda$ 应该理解为 Lebesgue 测度的微小变化，然而我们并不假定这一变化是如何产生的. $\mathrm{d}x$ 理解为 x 的微小变化. x 的微小变化自然就产生了 λ 的微小变化，即 $\lambda(\mathrm{d}x)$. 所以，在 x 处，$\mathrm{d}\lambda$ 和 $\mathrm{d}x$ 之间的关系应该是 $\mathrm{d}\lambda = \lambda(\mathrm{d}x)$，于是 $\mathrm{d}\lambda$ 应该理解为 $\mathrm{d}x$ 形成的长方体的体积.

同样，密度给出了概率计算的一个重要工具：

定理 C.10 设 X 是一个 n 维连续型随机向量，f 是它的密度函数，则对任意的 $B \in \mathscr{B}(\mathbb{R}^n)$，都有

$$\Pr(X \in B) = \int_{x \in B} f(x)\mathrm{d}x.$$

利用联合密度，可以计算边缘密度：

定理 C.11 设 $X = (X_1, \ldots, X_n)$ 是一个 n 维连续型随机向量，，则对任意的 $1 \leq i \leq n$，都有

$$p_{X_i}(x_i) = \int_{(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)\in\mathbb{R}^{n-1}} p_X(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \cdots \mathrm{d}x_n.$$

这一命题当然也可以自然推广到求随机向量的边缘密度，例如利

用 $X=(X_1,X_2,X_3)$ 的联合密度计算 (X_1,X_2) 的边缘密度：

$$p_{X_1,X_2}(x_1,x_2)=\int_{x_3\in\mathbb{R}}p_X(x_1,x_2,x_3)\mathrm{d}x_3.$$

连续型随机变量的一个重要的例子是多元正态分布，或者（非退化）Gauss 向量. 它的密度函数为

$$p(x)=\frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)},$$

其中 $\mu\in\mathbb{R}^n$，Σ 是一个 $n\times n$ 的正定矩阵. 这一分布的符号是 $\mathcal{N}(\mu,\Sigma)$.

关于 Gauss 向量的性质，我们将在附录 C.4 中讨论.

接下来，我们讨论条件分布.

对于离散型随机向量 $X=(X_1,X_2)$，它的分布完全由分布列给出. 我们可以定义 X_1 在给定 X_2 的条件下的分布列：

$$\Pr(X_1=x_1|X_2=x_2)=\frac{\Pr(X_1=x_1,X_2=x_2)}{\Pr(X_2=x_2)}.$$

由此给出了随机变量 X_1 在给定 X_2 的条件下的*条件分布列*，继而给出了*条件分布*. 这一定义也可以推广到 X_i 是随机向量的情况.

然而，对于一般的随机向量，特别是连续型随机向量，这一定义是行不通的. 比如，如果 $X=(X_1,X_2)$ 是连续型随机向量，那么 $\Pr(X_2=x_2)=\Pr(X_1=x_1,X_2=x_2)=0$，所以条件概率的分子和分布概率都是零，这样的定义是没有意义的.

转换思路，去尝试定义所谓的*条件分布函数*：$\Pr(X_1\leq x_1|X_2=x_2)$.

考虑 $\Pr(X_1 \le x_1 | x_2 < X_2 \le x_2 + \epsilon)$，再令 $\epsilon \downarrow 0$，我们有如下计算：

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \Pr(X_1 \le x_1 | x_2 < X_2 \le x_2 + \epsilon) \\
=&\lim_{\epsilon \downarrow 0} \frac{\Pr(X_1 \le x_1, X_2 \le x_2 + \epsilon) - \Pr(X_1 \le x_1, X_2 \le x_2)}{\Pr(x_2 < X_2 \le x_2 + \epsilon)} \\
=&\lim_{\epsilon \downarrow 0} \frac{F_X(x_1, x_2 + \epsilon) - F_X(x_1, x_2)}{F_{X_2}(x_2 + \epsilon) - F_{X_2}(x_2)}.
\end{aligned}$$

如果上面的极限存在，我们就定义它是 X_1 在给定 X_2 的条件下的**条件分布**.

如果 X 是连续性随机变量，我们还可以继续算下去：

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \frac{F_X(x_1, x_2 + \epsilon) - F_X(x_1, x_2)}{F_{X_2}(x_2 + \epsilon) - F_{X_2}(x_2)} \\
=&\frac{\partial F_X(x_1, x_2)}{\partial x_2} \frac{1}{p_{X_2}(x_2)} \\
=&\int_{-\infty}^{x_1} \frac{\partial^2 F_X(y, x_2)}{\partial x_2 \partial y} \frac{1}{p_{X_2}(x_2)} \mathrm{d}y \\
=&\int_{-\infty}^{x_1} \frac{p_{X_1, X_2}(y, x_2)}{p_{X_2}(x_2)} \mathrm{d}y.
\end{aligned}$$

对照定理 C.7，我们知道 $p_{X_1,X_2} / p_{X_2}$ 具有密度函数的形式，所以连续性随机向量所定义的条件分布也是连续型分布，密度函数被 $p_{X_1,X_2} / p_{X_2}$ 通常记作 $p_{X_1|X_2}$，称为 X_1 在给定 X_2 的条件下的**条件密度**.

以上讨论也可以自然推广到 X_i 是随机向量的情况，我们就不给出了.

最后，我们讨论随机向量之间的独立性. 随机向量之间的独立性完全由事件的独立性刻画，所以我们有如下定义：

定义 C.9 (随机向量的独立性) 设 $X_1, \ldots, X_n$ 是 n 个随机向量，第 i 个

的维数是 n_i. 如果对任意的 $1 \leq i_1, \ldots, i_k \leq n$，以及任意的 $B_{i_1} \in \mathscr{B}(\mathbb{R}^{n_{i_1}}), \ldots, B_{i_k} \in \mathscr{B}(\mathbb{R}^{n_{i_k}})$，都有

$$\Pr(X_{i_1} \in B_{i_1}, \ldots, X_{i_k} \in B_{i_k}) = \Pr(X_{i_1} \in B_{i_1}) \cdots \Pr(X_{i_k} \in B_{i_k}),$$

则称 $X_1, \ldots, X_n$ 是**独立的.**

特别地，如果 $X_1, \ldots, X_n$ 是一维的，那么这定义了随机变量之间的独立性. □

这一定义中包含了无穷多个需要验证的等式，利用分布函数，我们可以将独立性的验证转化为一个等式的验证：

定理 C.12 *设 $X_1, \ldots, X_n$ 是 n 个随机向量，第 i 个的维数是 n_i，F_i 是 X_i 的分布函数，F 是 $(X_1, \ldots, X_n)$ 的联合分布函数. $X_1, \ldots, X_n$ 独立的充分必要条件是*

$$F(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n),$$

其中 $x_i \in \mathbb{R}^{n_i}$.

对于离散型随机向量，它的分布函数完全由分布列决定，所以定理 C.12 等价于如下命题：

命题 C.6 *设 $X_1, \ldots, X_n$ 是 n 个离散型随机向量，第 i 个的维数是 n_i，p_i 是 X_i 的分布列，p 是 $(X_1, \ldots, X_n)$ 的联合分布列. $X_1, \ldots, X_n$ 独立的充分必要条件是*

$$p(x_1, \ldots, x_n) = p_1(x_1) \cdots p_n(x_n),$$

其中 $x_i \in \mathbb{R}^{n_i}$.

对于连续型随机向量，它的分布函数完全由密度决定，所以定理 C.12 等价于如下命题：

命题 C.7 设 $X_1,\dots,X_n$ 是 n 个连续型随机向量，第 i 个的维数是 n_i，p_i 是 X_i 的密度函数. 假设他们的联合分布具有密度函数 p. $X_1,\dots,X_n$ 独立的充分必要条件是

$$p(x_1,\dots,x_n)=p_1(x_1)\cdots p_n(x_n),$$

其中 $x_i\in\mathbb{R}^{n_i}$.

上面两个命题都有更简单的形式：

推论 C.2 设 $X_1,\dots,X_n$ 是 n 个连续型（离散型）随机向量，第 i 个的维数是 n_i，假设他们的联合分布具有密度函数（分布列）p. $X_1,\dots,X_n$ 独立的充分必要条件存在函数 $f_1,\dots,f_n$ 使得

$$p(x_1,\dots,x_n)=f_1(x_1)\cdots f_n(x_n),$$

其中 $x_i\in\mathbb{R}^{n_i}$.

利用这一命题，判断独立性的时候，我们只要尝试将联合密度（分布列）分解成若干个函数的乘积即可.

对于连续型随机向量，这一判据特别要注意密度函数的分段情况. 比如，考虑 $X=(X_1,X_2)$，其密度函数为

$$p(x_1,x_2)=\begin{cases}8x_1x_2, & 0\le x_1\le x_2, 0\le x_2\le 1,\\ 0, & \text{其他.}\end{cases}$$

如果忽略了 x_i 的取值范围，我们很容易以为 $p(x_1, x_2)$ 可以写成 $f(x_1)f(x_2)$，所以他们独立. 然而事实并不是这样的！计算 X_1 的边缘密度：

$$p_1(x_1) = \int_{x_2 \in \mathbb{R}} p(x_1, x_2) \mathrm{d}x_2 = \begin{cases} 4x_1(1 - x_1^2), & 0 \leq x_1 \leq 1, \\ 0, & \text{其他}. \end{cases}$$

再计算 X_2 的边缘密度：

$$p_2(x_2) = \int_{x_1 \in \mathbb{R}} p(x_1, x_2) \mathrm{d}x_1 = \begin{cases} 4x_2^3, & 0 \leq x_2 \leq 1, \\ 0, & \text{其他}. \end{cases}$$

显然，$p_1(x_1) \cdot p_2(x_2) \neq p(x_1, x_2)$，所以 X_1, X_2 不独立.

如果使用示性函数来书写密度函数，这一问题更不容易被忽视，在上面的例子中，$p(x_1, x_2) = 8x_1x_2I_{0 \leq x_1 \leq x_2 \leq 1}(x_1, x_2)$，示性函数显然是拆不成分别只关于 x_1 和 x_2 的函数乘积的.

自然，利用条件分布，我们可以给出独立的另一种刻画：

命题 C.8 设 X_1, X_2 是两个随机变量，他们的联合分布是离散型或连续型的. X_1, X_2 独立的充分必要条件是对任意的 x_1, x_2，都有

$$\Pr(X_1 \leq x_1 | X_2 = x_2) = \Pr(X_1 \leq x_1).$$

如果 X_1, X_2 是离散型的，那么这一条件可以改写为

$$\Pr(X_1 = x_1 | X_2 = x_2) = \Pr(X_1 = x_1).$$

如果 X_1, X_2 是连续型的，那么这一条件可以改写为

$$p_{X_1|X_2}(x_1|x_2) = p_{X_1}(x_1).$$

注意，上述判据并不需要真的把等式右边的量算出来，我们只需要判断刻画条件分布的量（条件分布函数、条件分布列或条件密度）中，是不是只出现了 x_1 而没有出现 x_2.

§C.2.5 随机变量（向量）的函数

我们前面说过，如果 X 是随机向量，g 是一个 Borel 函数，那么 $g(X) = g \circ X$ 也是一个随机向量. 这里，记号 $g \circ X$ 将 X 看成一个映射，于是得到的是一个复合函数；而记号 $g(X)$ 则更直观，它表示把 X 看成一个数学对象（随机向量），然后对它进行函数运算，得到另一个同类型的数学对象（随机向量）. 我们将始终采取后者的记号，但请务必注意，符号 $g(X)$ 中的 X 绝对不应该理解为一个数，而应该理解为一个随机向量.

随机变量的函数最直接的问题就是，它的分布是什么？我们只关注离散型和连续型随机向量的情况.

对于离散型随机向量，它的分布完全由分布列决定，很容易得到如下命题：

定理 C.13 设 X 是一个离散型随机向量，g 是一个函数，那么 $Y = g(X)$ 也是一个离散型随机向量，它的分布列为

$$\Pr(Y = y) = \sum_{x \in g^{-1}(y)} \Pr(X = x).$$

对于连续型随机向量，它的分布完全由密度决定. 我们现在来推导连续型随机向量的函数的密度.

设 X 是一个 n 维连续型随机向量，$g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$，即 $\mathbb{R}^n \to \mathbb{R}^n$ 的连续可微函数. 为了简便起见，我们假设 g 是单射，并且反函数也连续可微. 设 $Y = g(X)$，可以证明，Y 是一个连续型随机向量.

我们现在来计算 Y 的密度. 考虑 Y 取值的一个微小的区域 $\mathrm{d}y$，$\mathrm{d}P_Y = p_Y \lambda(\mathrm{d}y)$ 是 Y 在 $\mathrm{d}y$ 上的概率，同样区域的概率也可以用 X 去计算：

$$\mathrm{d}P_X = p_X \lambda(\mathrm{d}x), \quad Y \in \mathrm{d}y \iff X \in \mathrm{d}x,$$

当然，这里 $\mathrm{d}y$ 和 $\mathrm{d}x$ 由函数 g 联系在一起，因为 $Y = g(X)$，所以 $\mathrm{d}y/\mathrm{d}x = g'(X)$，注意，这相当于微元 $\mathrm{d}y$ 和微元 $\mathrm{d}x$ 的有向体积的比. 最后，根据概率相等，可以写出如下的等式：

$$\mathrm{d}P_Y = \mathrm{d}P_X \iff p_Y \lambda(\mathrm{d}y) = p_X \lambda(\mathrm{d}x). \tag{C.2}$$

考虑到密度是计算体积而非有向体积，根据 Jacobi 行列式的几何意义（见附录 B.3.1），

$$p_Y(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| p_X(x) = \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1} p_X(x) = |\det g'(x)|^{-1} p_X(x).$$

这就得到了 Y 的密度函数.

如果 g 不是单射，那么上面的 (C.2) 需要考虑 g 每一个单射的局部. 例如，如果 $g(x) = x^2$，那么 g 在 $(0, +\infty)$ 上和 $(-\infty, 0)$ 上都是单射，一个 y 对应了两个 x. 在这种情况下，每一个 y 所对应的 x 都贡献了概率，

所以 (C.2) 需要写成

$$\mathrm{d}P_Y = \sum_{g(x)=y} \mathrm{d}P_X(x) \iff p_Y\lambda(\mathrm{d}y) = \sum_{g(x)=y} p_X(x)\lambda(\mathrm{d}x)(x). \quad \text{(C.3)}$$

总结以上讨论，我们得到连续型随机向量的函数的密度的计算公式：

定理 C.14 设 X 是一个连续型随机向量，$g \in C(\mathbb{R}^n, \mathbb{R}^n)$，即 $\mathbb{R}^n \to \mathbb{R}^n$ 的连续函数，假设 $\lambda(\{x \in \mathbb{R}^n : \det g'(x) \neq 0\}) = 0$，则 $Y = g(X)$ 也是一个连续型随机向量，它的密度函数为

$$p_Y(y) = \begin{cases} \sum_{g(x)=y} |\det g'(g^{-1}(y))|^{-1} p_X(g^{-1}(y)), & \det g'(g^{-1}(y)) \neq 0, \\ 0, & \text{其他.} \end{cases}$$

其中求和号中 $g^{-1}(y)$ 是根据相应的 x，用反函数定理（定理 *B.18*）求出局部反函数.

这一定理的表述比较宽泛，我们可以给一个具体的例子来理解.

例 C.9 设 X 是一个连续型随机变量，$g(x) = x^2$，我们来计算 $Y = X^2 = g(X)$ 的密度. 直接计算定理 C.14 中的公式，我们有

$$\begin{aligned} &\sum_{g(x)=y} |\det g'(g^{-1}(y))|^{-1} p_X(g^{-1}(y)) \\ =&\sum_{x^2=y} \frac{1}{2|x|} p_X(x) \\ =&\frac{1}{2\sqrt{y}} p_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} p_X(-\sqrt{y}). \end{aligned}$$

这就给出了 Y 的密度. □

一般来说，定理 C.14 中的公式并不好记，最实用的还是根据 X 和 Y 在算相同的概率这一事实直接写出 (C.3)，然后根据具体的 g 来计算. 比如上面的例子，我们可以直接写出

$$p_Y\lambda(\mathrm{d}y) = p_X(\sqrt{y})\lambda(\mathrm{d}x)(\sqrt{y}) + p_X(-\sqrt{y})\lambda(\mathrm{d}x)(-\sqrt{y}).$$

两边除以 $\lambda(\mathrm{d}y)$，再利用 $\mathrm{d}y/\mathrm{d}x = 1/(2x)$，就得到了 Y 的密度.

最后，如果映射 g 并不是保持维度的，例如 $g : \mathbb{R}^n \to \mathbb{R}^m$，但 $m < n$[1]，那么我们可以将 g 补全到 n 维映射，也就是说，我们可以定义一个新的函数 $G : \mathbb{R}^n \to \mathbb{R}^n$ 满足

$$G(x_1, \dots, x_n) = (g_1(x_1, \dots, x_n), \dots, g_m(x_1, \dots, x_n), x_{m+1}, \dots, x_n)^{\mathsf{T}}.$$

然后，利用这一函数计算出 $g(X)$ 和 $(X_{m+1}, \dots, X_n)$ 的联合概率密度，再求出 $g(X)$ 的边缘密度.

我们看一个简单的例子.

例 C.10 (卷积) 设 X, Y 是随机变量，我们来计算 $Z = X + Y$ 的密度. 我们可以将 Z 看成是 $g(X, Y) = (X + Y, Y)$ 的第一个维度. 映射 $(x, y) \mapsto (x + y, y)$ 显然是双射，所以 (C.3) 退化为 (C.2)，我们有

$$p_{Z,Y}(z, y) = \left|\frac{\partial(z, y)}{\partial(x, y)}\right|^{-1} p_{X,Y}(x, y) = p_{X,Y}(z - y, y).$$

[1]如果 $m > n$，那么 $g(X)$ 一定不会是连续型随机变量，因为它的每个维之间一定会产生相互的关联，所以我们不讨论这种情况.

于是，Z 的边缘密度为

$$p_Z(z) = \int_{y\in\mathbb{R}} p_{X,Y}(z-y,y)\mathrm{d}y.$$

这被称为 X 和 Y 的**卷积**. □

最后，对随机向量作用函数是不会影响独立性的：

命题 C.9 *设 $X_1,\dots,X_n$ 是 n 个随机向量，第 i 个的维数是 n_i，g_i 是 $\mathbb{R}^{n_i}\to\mathbb{R}^{m_i}$ 的 Borel 函数，$Y_i = g_i(X_i)$，如果 $X_1,\dots,X_n$ 相互独立，那么 $Y_1,\dots,Y_n$ 也相互独立.*

§C.3 随机变量的数字特征，条件数学期望

分布函数或者随机变量依然是一个映射，研究起来还是会比较复杂. 我们希望能够用一些数字来刻画随机变量的特征，这样可以进一步简化问题. 在这一节中，我们将介绍随机变量的数字特征，以及条件数学期望.

§C.3.1 数学期望，Lebesgue 积分

数学期望在数学上是很直观的，我们可以从一个赌博的例子入手来找一些感觉.

例 C.11 在一个地下赌场，有赌徒甲乙两人. 这是一个公平的赌局，每局甲乙获胜概率都是 $1/2$，每局各出赌注 50 块. 谁先赢到三局，就可以

赢得全部的赌注. 赌博进行了三轮，甲赢了两局，乙赢了一局. 这时，突然有消息说警察马上就要来查封赌场，甲乙于是决定将目前的所有赌资进行分割. 他们应该如何分割呢?

再赌两盘就会决出胜负，赌博一共会有三种可能：

1. 第四盘甲赢，于是甲赢的所有赌注，这样的概率是 $1/2$;
2. 第四盘乙赢，第五盘甲赢，于是甲赢的所有赌注，这样的概率是 $(1/2)\times(1/2)=1/4$;
3. 乙连赢两盘，于是乙赢的所有赌注，这样的概率是 $(1/2)\times(1/2)=1/4$.

现在的赌资是 $100\times 3=600$ 块，甲有 $1/2+1/4=3/4$ 的概率会拿到全部，乙有 $1/4$ 的概率会拿到全部. 于是，按照概率去平分的话，甲应该拿走 450 块，乙应该拿走 150 块. □

这个例子说明了期望的一种理解方式：在面对随机性的时候，我们按照概率的权重分配. 比如，上面的例子中，设 X 是甲赢的赌注，那么 X 的分布列为 $\Pr(X=0)=1/4$，$\Pr(X=600)=3/4$，所以 $\mathbb{E}[X]=0\times 1/4+600\times 3/4=450$.

以上的例子给了我们定义随机变量期望的基础：定义示性函数的数学期望. 设 A 是一个事件，那么 I_A 是一个随机变量：

$$I_A(\omega)=\begin{cases}1, & \omega\in A,\\ 0, & \omega\notin A.\end{cases}$$

我们称之为事件 A 的示性函数. 示性函数的分布列是

$$\Pr(I_A = 1) = \Pr(A), \quad \Pr(I_A = 0) = \Pr(A^c) = 1 - \Pr(A).$$

所以，示性函数的数学期望，按照上面的逻辑，应该是

$$\mathbb{E}[I_A] = 1 \times \Pr(A) + 0 \times \Pr(A^c) = \Pr(A).$$

示性函数建立了概率和数学期望的联系. 下面，我们来定义一般随机变量的数学期望，这一定义的过程反映了一种数学的思想：用简单东西的极限去研究复杂的东西.

第一步，定义示性函数的数学期望[2]. $\mathbb{E}[I_A] = \Pr(A)$.

第二步，定义简单随机变量的数学期望. 简单随机变量是形如 $X = \sum_{k=1}^{n} x_k I_{A_k}$ 的随机变量，其中 $x_k \in \mathbb{R}$，A_k 是事件. 定义

$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k \Pr(A_k).$$

这一定义与第一步是相容的：因为 $I_A = 1 \cdot I_A$，所以 $\mathbb{E}[I_A] = 1 \cdot \Pr(A) = \Pr(A)$.

第三步，定义非负随机变量的数学期望. 非负随机变量是指 $X(\omega) \geq 0$ 对任意 ω 成立的随机变量 X. 考虑一列简单随机变量 $\{X_n\}_{n=1}^{\infty}$，它满足对于每一个 $\omega \in \Omega$ 都有当 $n \to \infty$ 时 $X_n(\omega) \uparrow X(\omega)$. 容易验证，$\mathbb{E}[X_n]$ 也是单调递增的，所以根据命题 B.13，$\mathbb{E}[X_n]$ 有有限的极限或者趋于正无穷，我们都记为 $\lim_{n\to\infty} \mathbb{E}[X_n]$.

[2]从逻辑上说，示性函数的数学期望是被定义出来的，而不是被算出来的，因为此时我们还完全没有定义什么是数学期望.

定义 C.10 (数学期望（Lebesgue 积分），非负情形) 称

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$$

为随机变量 X 的**数学期望**或 **Lebesgue 积分**. □

可以证明，这一定义不依赖于 $\{X_n\}_{n=1}^{\infty}$ 的选取，因而是良定义的. 此外，容易看出，这一定义与第二步是相容的，所以第三步扩展了第二步的定义.

第四步，定义一般随机变量的数学期望. 考虑随机变量 X，定义 $X^+ = \max\{X, 0\}$，$X^- = -\min\{X, 0\}$，也就是 X 的正数部分和负数部分，那么 $X = X^+ - X^-$. 我们有如下定义：

定义 C.11 (数学期望（Lebesgue 积分），一般情形) 称随机变量 X 的数学期望**存在**，如果 $\mathbb{E}[X^+]$ 和 $\mathbb{E}[X^-]$ 至少有一个有限. 此时，定义

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

为随机变量 X 的**数学期望**或 **Lebesgue 积分**.

如果 $\mathbb{E}[X^+]$ 和 $\mathbb{E}[X^-]$ 都是有限的，那么称 X 有**有限期望**或**可积的**. □

当我们强调积分的时候，$\mathbb{E}[X]$ 也会写为

$$E[X] = \int_{\Omega} X \mathrm{d}\Pr.$$

以上定义适用于任何一种概率空间和概率测度. 容易看出来，这一定义也适用于 $\mathbb{R}^n$ 上的 Lebesgue 测度，我们唯一需要改变的就是示性

函数的 Lebesgue 积分的定义：对任意 $A \in \mathscr{B}(\mathbb{R}^n)$，定义

$$\int_{\mathbb{R}^n} I_A(\omega)\lambda(\mathrm{d}\omega) = \lambda(A).$$

然后对简单函数定义积分，再对非负函数定义积分，最后对一般函数定义积分.

对于 $\mathbb{R}^n$ 上的 Lebesgue 积分，我们一般省略 λ[3]，直接写成

$$\int_{\mathbb{R}^n} f(x)\mathrm{d}x.$$

这与我们所熟知的积分符号就完全一致了.

上面定义随机变量期望的过程中，最难以理解的是第三步，也就是非负随机变量的数学期望. 我们来具体算一下它的表达式.

设 X 是一个非负随机变量，分布为 F. 我们来计算 $\mathbb{E}[X]$，与其说是计算，不如说重新推导一遍第三步的过程. 首先，我们将 X 取值范围离散化，每 $1/n$ 一段，X 的值都压到形如 k/n 的点上，这样就转化为了一个离散型随机变量：

$$X_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{k/n \le X < (k+1)/n\}}.$$

容易看出，$X_n(\omega) \uparrow X(\omega)$ 对任意 ω 成立. 于是，我们有

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

[3]对于一维的情况，见附录 C.2.3 的讨论. 在高维空间中，这样的记号其实是相当糟糕的：在微分学中，求导数时，$\mathrm{d}x$ 被理解为切空间的向量，或者一个微小的位移；求然而在积分学中，$\mathrm{d}x$ 被理解为所对应平行体的体积. 所以其实 $\lambda(\mathrm{d}x)$ 这一记号虽然复杂，但是含义更准确.

我们来计算 $\mathbb{E}[X_n]$，注意到 X_n 是一个简单随机变量，我们有

$$\begin{aligned}\mathbb{E}[X_n] &= \sum_{k=0}^{\infty} \frac{k}{n} \Pr\left(\frac{k}{n} \le X < \frac{k+1}{n}\right) \\ &= \sum_{k=0}^{\infty} \frac{k}{n} \left(F\left(\frac{k+1}{n}\right) - F\left(\frac{k}{n}\right)\right).\end{aligned}$$

按照极限的想法，当 $n \to \infty$ 时，上式的求和项相当于 $x\mathrm{d}F(x)$，这里 $\mathrm{d}F$ 表示 x 微小变化时对应的 F 的微小变化. 所以形式上我们有

$$\mathbb{E}[X] = \int_{x \ge 0} x\mathrm{d}F(x) = \int_{\mathbb{R}} x\mathrm{d}F(x),$$

这里第二个等式是因为在 $x < 0$ 的时候 F 恒等于 0，因而可以理解为 $\mathrm{d}F(x) = 0$.

如果 X 不是非负的，那么对 X^+ 和 X^- 分别计算数学期望，然后相减，就得到了一般随机变量的数学期望，它依然满足：

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\mathrm{d}F(x).$$

所以，随机变量的数学期望完全取决于它的分布函数.

对于离散型随机变量来说，F 只在点 $x_1, x_2, \ldots$ 会发生改变，其他地方都是常值，所以我们有

$$\int_{\mathbb{R}} x\mathrm{d}F(x) = \sum_{k=1}^{\infty} x_k(F(x_k) - F(x_k-)) = \sum_{k=1}^{\infty} x_k \Pr(X = x_k).$$

对于连续型随机变量来说，$\mathrm{d}F = p\mathrm{d}x$，这里 p 是对应的密度. 于是

我们有

$$\int_{\mathbb{R}} x\mathrm{d}F(x) = \int_{\mathbb{R}} xp(x)\mathrm{d}x.$$

以上就是概率论中常见的求期望的形式.

我们再介绍一个非常有用的符号，它允许我们在某个事件 A 上求积分：

$$\int_A X\mathrm{d}\Pr = \int_\Omega XI_A\mathrm{d}\Pr = \mathbb{E}[XI_A].$$

相应地，在 $\mathbb{R}^n$ 上，对我们也可以定义

$$\int_A f(x)\lambda(\mathrm{d}x) = \int_{\mathbb{R}^n} f(x)I_A(x)\lambda(\mathrm{d}x).$$

刻画随机变量的数字特征，除了可以用随机变量的期望，还可以用随机变量的函数的期望，我们列举一个重要的概念.

定义 C.12 (矩，方差，特征函数) 设 X 是一个随机变量，我们有如下定义：

- k 是一个正整数，称 $\mathbb{E}[X^k]$ 为 X 的 k **阶矩**；称 $\mathbb{E}[(X-\mathbb{E}[X])^k]$ 为 X 的 k **阶中心矩**；
- 称 $\mathbb{E}[(X-\mathbb{E}[X])^2]$ 为 X 的**方差**，记为 $\mathsf{Var}(X)$；
- 称 $f_X(t) = \mathbb{E}[\exp(itX)]$ 为 X 的**特征函数**. 一般地，如果 X 是 n 维随机向量，那么 $f_X : \mathbb{R}^n \to \mathbb{C}$，$f_X(t) = \mathbb{E}[\exp(i\langle t, X\rangle)]$ 被称为 X 的**特征函数**. □

我们将会在后面讨论他们的性质.

§C.3.2 数学期望的性质

我们已经给出了数学期望的定义，下面我们罗列一些数学期望的性质，但都不给出证明.

命题 C.10 *1. 期望的线性性：设 X,Y 是随机变量，$a,b\in\mathbb{R}$，如果 $\mathbb{E}[X]$ 和 $\mathbb{E}[Y]$ 都存在，那么 $\mathbb{E}[aX+bY]$ 存在，且*

$$\mathbb{E}[aX+bY]=a\mathbb{E}[X]+b\mathbb{E}[Y].$$

2. 单调性：设 X,Y 是随机变量，如果 $X\leq Y$，那么

$$\mathbb{E}[X]\leq\mathbb{E}[Y].$$

3. 绝对值不等式：设 X 是随机变量，那么

$$\mathbb{E}[|X|]\geq|\mathbb{E}[X]|.$$

4. 局部可积性：设 X 是随机变量，并且 $\mathbb{E}[X]$ 存在，那么对任意事件 A，$\mathbb{E}[XI_A]$ 也存在；如果 $\mathbb{E}[X]$ 有限，那么 $\mathbb{E}[XI_A]$ 也有限.

接下来，我们讨论随机变量函数的期望的求法. 假设 X 是一个 n 维随机向量，$g:\mathbb{R}^n\to\mathbb{R}$ 是一个 Borel 函数，那么 $g(X)$ 也是一个随机变量（定理 C.3）. 计算 $\mathbb{E}[g(X)]$ 有以下两种方式，我们下面分别讨论.

第一种，利用附录 C.2.5 中的方法，我们可以将 $g(X)$ 的分布写出来，然后计算期望. 我们来看一个例子.

例 C.12 设 $X \sim \mathcal{U}(0,1)$，计算 $\mathbb{E}[X^2]$. 直接算出 $Y = X^2$ 的密度函数为

$$p_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 \le y \le 1, \\ 0, & \text{其他}. \end{cases}$$

于是，

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} y p_Y(x) \mathrm{d}y = \int_0^1 \frac{y}{2\sqrt{y}} \mathrm{d}y = \frac{1}{3}.$$ □

第二种，我们从定义出发，直接计算 $\mathbb{E}[g(X)]$. 我们先考虑最简单的情况，即 g 连续并且 $0 \le g \le C$ 的情况，这里 C 是一个正常数. 我们还是试图使用第三步，用简单随机变量去逼近 $g(X)$. 我们选择离散化 X，还是一样定义

$$X_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{k/n \le X < (k+1)/n\}}.$$

可以证明[4]

$$\mathbb{E}[g(X)] = \lim_{n\to\infty} \mathbb{E}[g(X_n)].$$

用 X 的分布函数 F 写出来 $\mathbb{E}[X_n]$ 就是

$$\mathbb{E}[X_n] = \sum_{k=0}^{\infty} g\left(\frac{k}{n}\right)\left(F\left(\frac{k+1}{n}\right) - F\left(\frac{k}{n}\right)\right).$$

[4]注意，这里 $g(X_n)$ 未必单调趋于 $g(X)$ 了，所以这里我们其实跳了一个比较重要的步骤，即不单调趋于的时候极限也可以拿到期望外面. 由于这一步的证明比较技术，而且对本书的讨论不是特别重要，所以这里略去.

取极限，写成积分的形式，我们有：

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\mathrm{d}F(x).$$

利用逼近的思想，我们可以将上述结论推广到 g 是任意的 Borel 函数的情况，于是我们有：

定理 C.15 设 X 是一个 n 维随机向量，每一维都可积，$g:\mathbb{R}^n \to \mathbb{R}$ 是一个 *Borel* 函数，那么 $\mathbb{E}[g(X)]$ 存在，且

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x)\mathrm{d}F_X(x).$$

特别地，如果 X 是一个离散型随机变量，取值为 $x_1, x_2, \ldots$，那么

$$\mathbb{E}[g(X)] = \sum_{k=1}^{\infty} g(x_k)\Pr(X = x_k).$$

如果 X 是一个连续型随机变量，密度为 p_X，那么

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x)p_X(x)\mathrm{d}x.$$

例 C.13 我们重新算一次上面的例 C.12，这次我们用定理 C.15 来计算. 设 $X \sim \mathcal{U}(0,1)$，我们有

$$E[X^2] = \int_0^1 x^2\mathrm{d}x = \frac{1}{3}.$$ □

从这两个例子就可以看出，以上两种方法，通常来说第二种会更加容易计算一些，因为它只需要做一次积分，而第一种方法还需要算变量

替换的 Jacobi 行列式.

接下来，我们讨论示性函数的性质.

命题 C.11 *1. 设 A 是一个事件，那么 $\mathbb{E}[I_A] = \Pr(A)$，$\mathsf{Var}(I_A) = \Pr(A)(1-\Pr(A))$.*

2. 设 A, B 是两个事件，那么 $I_A I_B = I_{AB}$，特别地，$I_A^2 = I_A$.

这些性质的证明都比较容易，这里就不给出了.

利用示性函数，我们可以重写事件独立性的定义：

命题 C.12 *设 A, B 是两个事件，那么 A 和 B 独立的充分必要条件是*

$$\mathbb{E}[I_A I_B] = \mathbb{E}[I_A]\mathbb{E}[I_B].$$

如果我们还记得随机变量的期望是如何定义的，那么我们可以发现，命题 C.12 的结论可以推广到随机变量的情形：

定理 C.16 *设 X, Y 是两个相互独立的随机变量，那么 $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

需要注意的是，这一命题的逆命题不一定成立.

最后，我们给一个重要的不等式. 我们说函数 $g : \mathbb{R} \to \mathbb{R}$ 是*凸函数*，如果对任意 $x, y \in \mathbb{R}$，$t \in [0,1]$，都有

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y).$$

关于凸函数的更多讨论，见第 6.2 节. 我们有如下不等式：

定理 C.17 (Jensen 不等式) *设 X 是一个随机变量，$g : \mathbb{R} \to \mathbb{R}$ 是一个凸函数，那么*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

§C.3.3 随机变量的内积空间

我们指出，随机变量利用期望可以定义内积，从而定义内积空间，关于内积空间的讨论，见附录 A.5. 在附录 C.4 中，这一事实非常重要.

我们定义内积如下：

定义 C.13 (协方差) 设 X, Y 是两个随机变量，称

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

为 X 和 Y 的**协方差**. □

容易验证，在差一个常数的意义下，协方差是一个对称正定的双线性型：

命题 C.13 $\mathrm{Cov}(\cdot, \cdot)$ 具有以下性质：

1. 对称性：任意随机变量 X, Y，$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$；
2. 单边线性性：任意随机变量 X, Y，$a, b \in \mathbb{R}$，$\mathrm{Cov}(aX + bY, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)$；
3. 正定性：任意随机变量 X，$\mathrm{Cov}(X, X) \geq 0$，且 $\mathrm{Cov}(X, X) = 0$ 当且仅当存在常数 C 使得 $\Pr(X = C) = 1$.

于是，在差一个常数的意义下，协方差是一个随机变量空间的内积. 按照内积空间的性质，随机变量的范数自然就是它的方差.

> **注.** 在命题 C.13 中，我们使用了 $\Pr(X = C) = 1$ 这样的表达. 在概率论中，如果一个事件是概率 1 发生的，我们称之为几乎必然发生. 在涉及与数学期望有关的性质的时候，我们通常只能在几乎必然的意义下成立，而

不能在一般意义下成立. 比如说，“在差一个常数的意义下，协方差是一个随机变量空间的内积”这句话其实并不准确，严格来说，应该是“在差一个常数和几乎必然相等的意义下，协方差是一个随机变量空间的内积”. 也就是说，如果 $\|X\| = 0$，那么 X 几乎必然为常数.

协方差与独立性密切关联：

命题 C.14 设 X, Y 是两个随机变量，如果 X 和 Y 相互独立，那么 $\mathsf{Cov}(X, Y) = 0$.

我们称 $\mathsf{Cov}(X, Y) = 0$ 的两个随机变量是不相关的，用内积空间的术语，不相关的意思就是随机变量正交. 不相关的随机变量不一定是独立的，但是独立的随机变量一定是不相关的.

协方差的概念可以推广到多个随机变量上：

定义 C.14 (协方差矩阵) 设 $X_1, \dots, X_n$ 是 n 个随机变量，称他们的 Gram 矩阵为**协方差矩阵**，记为 Σ，其中

$$\Sigma_{ij} = \mathsf{Cov}(X_i, X_j).$$

如果 X 和 Y 分别是 m 维和 n 维随机向量，那么符号 $\mathsf{Cov}(X, Y)$ 表示的是 $m \times n$ 的矩阵 $(\mathsf{Cov}(X_i, X_j))_{ij}$，称为 X 和 Y 的**协方差矩阵**. 特别地，如果 $X = Y$，那么我们记 $\mathsf{Cov}(X, X)$ 为 $\mathsf{Var}(X)$，称为 X 的**协方差矩阵**. □

根据 Gram 矩阵的性质（命题 A.8），X 的协方差矩阵是一个对称半正定矩阵.

类似地，我们也可以定义随机向量的数学期望：

定义 C.15 (随机向量的数学期望) 设 $X = (X_1, \dots, X_n)^{\mathsf{T}}$ 是一个 n 维随机向量，称

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])^{\mathsf{T}}$$

为 X 的**数学期望.** □

接下来，我们按照线性代数的思路，研究线性变换对于期望以及协方差矩阵的影响.

首先是期望，很容易证明如下的结论：

定理 C.18 设 X 是一个 n 维随机向量，A 是一个 $m \times n$ 的矩阵，那么 $\mathbb{E}[AX] = A\mathbb{E}[X]$.

接下来是协方差矩阵. 利用 Gram 矩阵与二次型的关系，我们容易写出如下的结论：

定理 C.19 设 X 是一个 n 维随机向量，A 是一个 $m \times n$ 的矩阵，那么

$$\mathsf{Var}(AX) = A\mathsf{Var}(X)A^{\mathsf{T}}.$$

证明. 考虑向量 t，和 n 维随机向量 Y，$t^{\mathsf{T}}Y$ 是一个随机变量，我们可以得到一个二次型

$$g(t) = \mathsf{Var}(t^{\mathsf{T}}Y) = \mathsf{Cov}(t^{\mathsf{T}}Y, t^{\mathsf{T}}Y) = t^{\mathsf{T}}\mathsf{Var}(Y)t.$$

当 $Y = AX$ 时，我们有

$$g(t) = \mathsf{Var}(t^{\mathsf{T}}AX) = \mathsf{Var}((A^{\mathsf{T}}t)X) = t^{\mathsf{T}}A\mathsf{Var}(X)A^{\mathsf{T}}t.$$

所以，对任意 t 都有 $t^{\mathsf{T}}\mathsf{Var}(AX)t = t^{\mathsf{T}}A\mathsf{Var}(X)A^{\mathsf{T}}t$，所以 $\mathsf{Var}(AX) = A\mathsf{Var}(X)A^{\mathsf{T}}$. □

上面的计算可以有一个线性代数的理解. 假如说 $X_1, \dots, X_n$ 是线性无关的，那么 $t^{\mathsf{T}}X$ 可以理解为某个向量在 $X_1, \dots, X_n$ 下的基表示，于是 t 是坐标. 而 $t^{\mathsf{T}}AX = (A^{\mathsf{T}}t)X$，因此 A^{T} 应该理解为某个线性映射 F 在 $X_1, \dots, X_n$ 下的矩阵. Gram 矩阵是二次型 $f(x) = \|x\|^2$ 在 $X_1, \dots, X_n$ 下的矩阵，因此在 F 的作用下，二次型的矩阵表示会做一个相应的合同变换，即 $A\mathsf{Var}(X)A^{\mathsf{T}}$.

§C.3.4 特征函数

在这一部分，我们讲述随机变量的特征函数，它是分布的另一种刻画方式.

显然，特征函数由分布函数决定. 反过来，特征函数也可以唯一决定分布！

定理 C.20 具有相同特征函数的随机变量（向量）具有相同的分布函数.

特征函数其实可以求出随机变量的分布函数：

定理 C.21 (逆转公式) 设 X 是随机变量，它的特征函数为 f_X，分布函数为 F_X，那么

1. 对于 F 的任意两个连续点 $a < b$，

$$F_X(b) - F_X(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{\mathrm{e}^{-ita} - \mathrm{e}^{-itb}}{it} f_X(t)\mathrm{d}t.$$

2. 如果 $\int_{\mathbb{R}} |f_X(t)|\mathrm{d}t < +\infty$，那么 X 具有密度 p_X，且

$$p_X(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-itx} f_X(t)\mathrm{d}t.$$

这一公式也有随机向量的版本：

定理 C.22 (逆转公式，随机向量版本) 设 X 是 n 维随机向量，它的特征函数为 f_X，分布函数为 F_X，那么

1. 对于 F 的两个点 $a < b$，满足

$$\Pr(X_1 = c_1, \ldots, X_{k-1} = c_{k-1}, X_k \in (a_k, b_k], X_{k+1} = c_{k+1}, \ldots, X_n = c_n) = 0,$$

其中 $c_i \in \{a_i, b_i\}$，我们有

$$\begin{aligned} &F_X(b) - F_X(a) \\ =&\lim_{T_1,\ldots,T_n\to\infty} \frac{1}{(2\pi)^n}\int_{-T_1}^{T_1}\cdots\int_{-T_n}^{T_n}\prod_{k=1}^{n}\frac{\exp(-it_k a_k) - \exp(-it_k b_k)}{it_k} f_X(t)\mathrm{d}t. \end{aligned}$$

2. 如果 $\int_{\mathbb{R}^n} |f_X(t)|\mathrm{d}t < +\infty$，那么 X 具有密度 p_X，且

$$p_X(x) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} \mathrm{e}^{-it^{\mathsf{T}}x} f_X(t)\mathrm{d}t.$$

特征函数特别适合处理独立随机变量的和：

定理 C.23 设 $X_1, \ldots, X_n$ 是 n 个相互独立的随机变量，它们的特征函数分别为 $f_1, \ldots, f_n$，那么 $X_1 + \cdots + X_n$ 的特征函数为 $f_1 \ldots f_n$.

比起卷积，用特征函数来算独立随机变量的和，方便得多.

特征函数也可以用来判定随机变量的独立性：

定理 C.24 设 $X_1,\dots,X_n$ 是 n 个随机变量，它们的特征函数分别为 $f_1,\dots,f_n$，随机向量 $X=(X_1,\dots,X_n)^{\mathsf{T}}$，它的特征函数为 f，那么 $X_1,\dots,X_n$ 相互独立的充分必要条件是

$$f(t_1,\dots,t_n)=f_1(t_1)\cdots f_n(t_n).$$

特征函数的导数可以用来计算随机变量的矩：

定理 C.25 设 X 是一个随机变量，它的特征函数为 f_X，那么对任意正整数 k，

$$\mathbb{E}[X^k]=\frac{f_X^{(k)}(0)}{i^k}.$$

总结起来，我们之前可以用分布列和密度函数来计算或者判定随机变量的各种性质和特征，现在都可以用特征函数来处理了.

§C.3.5 条件数学期望

数学期望的定义，从本质上说，就是对所有的取值做加权平均. 但是，有时候我们并不需要对所有的取值做加权平均，而只需要对某些取值做加权平均. 这时候，我们就需要引入条件数学期望的概念. 我们从一个直观的例子出发.

例 C.14 一个罐子里有 4 个红球，2 个灰球，4 个白球. 红球，灰球和白球的分数分别是 $4,2,1$. 现在随机抽一个球，抽球人戴着黑白滤镜的眼镜观察球的颜色，他不能分辨红球和灰球，但是可以区分这两种球和白球. 那么，在他观察过这个球之后，期望上得到的分数是多少?

和条件概率有类似的情况，此时并不完全是纯随机的，因为抽球人可以区分一些东西. 于是，样本空间可以分成两个部分，一个是 $A_1 = \{r, g\}$，即抽到的球是红球或灰球；另一个是 $A_2 = \{w\}$，即抽到的球是白球. 在第一种情况下，期望上的分数是

$$4 \cdot \Pr(\{s\}|A_1) + 2 \cdot \Pr(\{g\}|A_1) = 3.$$

在第二种情况下，期望上的分数是

$$1 \cdot \Pr(\{w\}|A_2) = 1.$$

□

更一般地，考虑样本空间 Ω，事件 $A_1, \ldots, A_n$，它们两两互斥，且 $\bigcup_{i=1}^{n} A_i = \Omega$，这形成了 Ω 的一个分割，记为 $\mathscr{A}$. 我们再假设 $\Pr(A_i) > 0$，我们有如下定义：

定义 C.16 (基于分割的条件数学期望) 设 X 是一个随机变量，$\mathscr{A} = \{A_1, \ldots, A_n\}$ 是 Ω 的一个分割，满足 $\Pr(A) > 0$ 对任意 $A \in \mathscr{A}$ 成立. X 在 $\mathscr{A}$ 上的**条件数学期望**是一个随机变量，记为 $\mathbb{E}[X|\mathscr{A}]$，它的定义为

$$\mathbb{E}[X|\mathscr{A}](\omega) = \sum_{i=1}^{n} \frac{\mathbb{E}[XI_{A_i}]}{\Pr(A_i)} I_{A_i}(\omega).$$

□

这个定义就是在说，当 ω 落在分割的某个集合 A_i 上时，我们按照 A_i 上的条件概率算期望. 记号 $\frac{\mathbb{E}[XI_{A_i}]}{\Pr(A_i)}$ 也记为 $\mathbb{E}[X|A_i]$，它的含义可以从 $X = I_B$ 来理解：

$$\frac{\mathbb{E}[I_B I_{A_i}]}{\Pr(A_i)} = \frac{\Pr(A_i B)}{\Pr(A_i)} = \Pr(B|A_i).$$

这一计算对示性函数解释了“按照 A_i 上的条件概率算期望”. 按照随机变量数学期望的定义，这一理解可以推广到一般随机变量.

条件数学期望是一个随机变量，意思就是我们能够消除某些不确定性. 在求数学期望的时候，我们完全不知道样本 ω 落在哪里，所以只能对整个 Ω 有一个预期. 在求对分割的条件数学期望的时候，我们能够知道 ω 落在了某个 A_i 中，因此我们的不确定性只在于 A_i 上，所以我们可以只对 A_i 中的 ω 有一个预期.

下面我们推广这一定义. 注意到，分割 $\mathscr{A}$ 其实生成了一个 Ω 的 σ-代数，即 $\sigma(\mathscr{A})$，它是包含 $A_1, \ldots, A_n$ 的最小 σ-代数. 容易验证，这一 σ-代数里的集合都是若干个 A_i 的并形成的. 分割里的事件代表了我们可以感知到的最小事件，而 σ-代数里的事件代表了我们可以感知到的事件的集合.

取 $A \in \sigma(\mathscr{A})$，要如何计算 X 在 A 上的期望呢？我们有两种方式，第一种，直接计算：$\mathbb{E}[XI_A]$. 第二种，我们将 A 写成 $A = \bigcup_{i=1}^{k} A_{n_i}$. 在每个 A_i 上，我们知道期望是 $\mathbb{E}[XI_A|A_i]$. 而落到 A_i 上的概率是 $\Pr(A_i)$，于是，按照数学期望加权平均的直觉，X 在 A 上的期望应该是

$$\sum_{i=1}^{k} \mathbb{E}[XI_A|A_i] \Pr(A_i).$$

这正好就是随机变量 $\mathbb{E}[XI_A|\mathscr{A}]$ 的数学期望 $\mathbb{E}[\mathbb{E}[XI_A|\mathscr{A}]]$.

对任意 $A \in \sigma(\mathscr{A})$，这两种计算方式都应该相等：

$$\mathbb{E}[XI_A] = \mathbb{E}[\mathbb{E}[XI_A|\mathscr{A}]]. \tag{C.4}$$

这给了我们一般情况下的条件数学期望的定义：

定义 C.17 (基于 σ-代数的条件数学期望) 设 X 是一个非负随机变量，$\mathscr{G}$ 是 Ω 的一个 σ-代数，随机变量 $\mathbb{E}[X|\mathscr{G}]$ 被称为 X 关于 $\mathscr{G}$ 的**条件数学期望**，如果它满足

1. 对任意 $B \in \mathscr{B}(\mathbb{R})$，$\{\mathbb{E}[X|\mathscr{G}] \in B\}$ 是 $\mathscr{G}$-可测的；
2. 对任意 $A \in \mathscr{G}$，$\mathbb{E}[XI_A] = \mathbb{E}[\mathbb{E}[X|\mathscr{G}]I_A]$.

设 X 是一个一般的随机变量，如果

$$\min\{\mathbb{E}[X^+|\mathscr{G}], \mathbb{E}[X^-|\mathscr{G}]\} < +\infty,$$

那么 X 关于 $\mathscr{G}$ 的**条件数学期望**是一个随机变量，记为 $\mathbb{E}[X|\mathscr{G}]$，定义为

$$\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X^+|\mathscr{G}] - \mathbb{E}[X^-|\mathscr{G}]. \qquad \square$$

这一定义分成两部分，这类似于我们在定义数学期望时候做的事情：先定义非负的情况，再定义一般情况. 对于非负随机变量的定义，第一条要求是说，“$\mathbb{E}[X|\mathscr{G}]$ 落在合理的值集上”这件事情是可以用 $\mathscr{G}$ 中事件描述的，这和随机变量的定义是类似的；而第二条则反映了“条件”的性质，也就是我们刚刚讨论的 (C.4) 式.

定义中的 $\mathscr{G}$ 可以理解为我们观测样本的能力. $\mathscr{G}$ 越大，则越能确定 ω 具体的范围，所以条件期望就越像 $X(\omega)$；$\mathscr{G}$ 越小，则越不能确定 ω 具体的范围，所以条件期望就越像 $\mathbb{E}[X]$.

注意，基于 σ-代数的条件数学期望和基于分割的条件数学期望是一致的，所以这一定义是合理的.

最后，随机向量也是可以诱导条件数学期望的：

定义 C.18 (随机向量诱导的 σ-代数) 设 X 是一个 n 维随机向量，那么 X **诱导的** σ**-代数**是 Ω 的一个 σ-代数，记为 $\sigma(X)$，它的元素为 $\{X \in B\}$，其中 $B \in \mathscr{B}(\mathbb{R}^n)$. □

我们说过，$\{X \in B\}$ 表示“X 落在合理的值集上”. 在之前定义随机变量的时候，我们要求取合理的值集是一个事件，这里则是更加简单粗暴，我们直接定义 $\{X \in B\}$ 是一个事件. 接下来，我们可以定义随机向量诱导的条件数学期望：

定义 C.19 (随机向量诱导的条件数学期望) 设 X 是一个随机变量，Y 是一个随机向量，那么 X 关于 Y 的**条件数学期望**是一个随机变量，记为 $\mathbb{E}[X|Y]$，定义为 $\mathbb{E}[X|\sigma(Y)]$. □

我们之前定义过条件分布 $\Pr(X \leq x|Y = y)$，利用这一分布，我们可以求出一个条件数学期望 $\mathbb{E}[X|Y = y]$. 下面的命题表明，这一定义和定义 C.19 是相容的：

命题 C.15 *设 X 是一个随机变量，Y 是一个 n 维随机向量，那么存在一个 Borel 函数 $g : \mathbb{R}^n \to \mathbb{R}$，使得对任意 $\omega \in \Omega$，有*

$$\mathbb{E}[X|Y](\omega) = g(Y(\omega))$$

并且

$$\mathbb{E}[X|Y = y] = g(y).$$

我们不满足于 $\mathbb{E}[X|Y = y]$，而是费尽周章定义条件期望 $\mathbb{E}[X|Y]$，是因为他通常来说更好用，特别是在随机过程中，它能给出很多公式直观上的含义. 这一点在第二章中会有很多体现.

接下来我们讨论条件数学期望的性质，我们依然只列举而不证明.

命题 C.16 设 $(\Omega, \mathscr{F}, \Pr)$ 是概率空间，$\mathscr{G} \subseteq \mathscr{F}$ 是 Ω 的一个 σ-代数，那么

1. 期望的线性性：设 X, Y 是随机变量，$a, b \in \mathbb{R}$，如果 $\mathbb{E}[X|\mathscr{G}]$ 和 $\mathbb{E}[Y|\mathscr{G}]$ 都存在，那么 $\mathbb{E}[aX + bY|\mathscr{F}]$ 存在，且

$$\mathbb{E}[aX + bY|\mathscr{G}] = a\mathbb{E}[X|\mathscr{G}] + b\mathbb{E}[Y|\mathscr{G}].$$

2. 单调性：设 X, Y 是随机变量，如果 $X \leq Y$，那么

$$\mathbb{E}[X|\mathscr{G}] \leq \mathbb{E}[Y|\mathscr{G}].$$

3. 绝对值不等式：设 X 是随机变量，那么

$$\mathbb{E}[|X||\mathscr{G}] \geq |\mathbb{E}[X]|\mathscr{G}|.$$

4. 如果 $\mathscr{G} = \{\varnothing, \Omega\}$，那么

$$\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X].$$

5. 望远性：设 X 是随机变量，如果 $\mathscr{G}_1, \mathscr{G}_2 \subseteq \mathscr{F}$ 都是 Ω 的 σ-代数，且 $\mathscr{G}_1 \subseteq \mathscr{G}_2$，那么

$$\mathbb{E}[\mathbb{E}[X|\mathscr{G}_2]|\mathscr{G}_1] = \mathbb{E}[\mathbb{E}[X|\mathscr{G}_1]|\mathscr{G}_2] = \mathbb{E}[X|\mathscr{G}_1].$$

6. 重期望公式：设 X 是随机变量，那么

$$\mathbb{E}[\mathbb{E}[X|\mathscr{G}]] = \mathbb{E}[X].$$

7. 设 X, Y 是随机变量，如果 $\sigma(Y) \subseteq \mathscr{G}$，那么

$$E[XY|\mathscr{G}] = Y\mathbb{E}[X|\mathscr{G}].$$

我们主要需要解释的是望远性. 可以把 σ-代数理解成观测的能力，这一代数越大，观测的越细致. 望远性的意思就是，如果我们用两次观测能力强弱不同的 σ-代数观测 X，那么最终的结果只取决于最粗糙的那个 σ-代数.

另外，重期望公式本质上就是期望版本的全概率公式（定理 C.1）. 从基于分割的条件数学期望的角度来看，这件事会更明显. 假设我们有一个分割 $\mathscr{A} = \{A_1, \ldots, A_n\}$，并且 $\Pr(A_i) > 0$，那么

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathscr{A}]] = \sum_{i=1}^{n} \mathbb{E}[X|A_i] \Pr(A_i).$$

最后，性质 7 是在说，如果 Y 是 $\mathscr{G}$-可测的（也就是我们用 $\mathscr{G}$ 可以完全确定 Y），那么求条件期望的时候 Y 就相当于一个常数，可以提到期望的外面.

§C.4 多元正态分布（Gauss 向量）

在这一节中，我们利用附录 C.3.3 和附录 C.3.4 中的工具，来研究多元正态分布.

多元正态分布的定义在附录 C.2.4 中已经给出，首先，我们不加证明地给出它的特征函数：

定理 C.26 设 $\mu \in \mathbb{R}^n$，Σ 是一个 $n \times n$ 的对称正定矩阵，那么随机向量 $X \sim \mathcal{N}(\mu, \Sigma)$ 的特征函数为

$$f_X(t) = \exp\left(it^{\mathsf{T}}\mu - \frac{1}{2}t^{\mathsf{T}}\Sigma t\right).$$

利用 (4.4)，我们可以计算出多元正态分布的期望和协方差矩阵：

命题 C.17 设 $X \sim \mathcal{N}(\mu, \Sigma)$，那么

$$\mathbb{E}[X] = \mu, \quad \mathsf{Var}(X) = \Sigma.$$

现在我们将这一定义推广. 注意到，Σ 就是 X 的协方差矩阵，所以定理 C.26 中的 Σ 并不要求正定，只要半正定就可以定义一个特征函数了. 我们将这一定义推广到半正定矩阵的情形：

定义 C.20 (Gauss 向量) 设 $\mu \in \mathbb{R}^n$，Σ 是一个 $n \times n$ 的对称半正定矩阵，如果随机向量 X 的特征函数为

$$f_X(t) = \exp\left(it^{\mathsf{T}}\mu - \frac{1}{2}t^{\mathsf{T}}\Sigma t\right),$$

那么称 X 是一个 **Gauss 向量**，记为 $X \sim \mathcal{N}(\mu, \Sigma)$. □

如果 Σ 退化，X 不能写出密度，所以也不是连续型随机向量. 但是，利用特征函数，我们依然可以研究 X 的性质. 特别是命题 C.17，对于 Gauss 向量仍然成立.

Gauss 向量可以完全由它的期望和协方差矩阵刻画. 首先，Gauss 向量的独立性等价于不相关性：

定理 C.27 设 $X=(X_1,\dots,X_n)\sim\mathcal{N}(\mu,\Sigma)$，那么 $X_1,\dots,X_n$ 相互独立的充分必要条件是 $X_1,\dots,X_n$ 两两不相关，即 Σ 是一个对角矩阵.

需要注意的是，如果 X 是正态分布，Y 是正态分布，这并不意味着 (X,Y) 是 Gauss 向量，因而并不能用不相关来作为独立性的判据. 因此，在一般情况下，我们必须要验证 $(X_1,\dots,X_n)$ 是 Gauss 向量，然后才能断言不相关等价于独立.

当然，这一判据可以推广到多个 Gauss 向量的情形：

推论 C.3 设 $X_1,\dots,X_n$ 是 n 个 *Gauss* 向量，它们相互独立的充分必要条件是 $X_1,\dots,X_n$ 两两不相关，即协方差矩阵 $\mathrm{Cov}(X_i,X_j)=O$，$i\neq j$.

其次，利用定理 C.18 和定理 C.19，我们可以得到如下的结论：

定理 C.28 设 $X\sim\mathcal{N}(\mu,\Sigma)$，$A$ 是一个 $m\times n$ 的矩阵，那么 $AX\sim\mathcal{N}(A\mu,A\Sigma A^{\mathsf{T}})$.

取特定的 A，我们可以得到一个实用的推论：Gauss 向量的子向量仍然是 Gauss 向量，也就是说，取 $X=(X_1,\dots,X_n)^{\mathsf{T}}\sim\mathcal{N}(\mu,\Sigma)$，那么对任意的 $1\leq k\leq n$，$i_1,\dots,i_k\in\{1,\dots,n\}$，$(X_{i_1},\dots,X_{i_k})^{\mathsf{T}}$ 也是 Gauss 向量.

§C.5 大数定律

接下来我们讨论大数定律. 大数定律来自人类对随机现象的直观认识，它表明，如果重复抛一枚公平的硬币，那么正面朝上的次数会趋于总次数的一半.

首先，我们要定义，对于一系列随机的实验，什么叫做“趋于”. 比如说，抛一万次硬币，当然有正的概率会出现一万次正面朝上，但是这件事情在现实中是不太可能的. 我们要定义一个概率的概念，来描述这种“趋于”.

定义 C.21 (依概率收敛) 设 $X_1, X_2, \ldots$ 是一系列随机变量，X 是一个随机变量. 如果对任意 $\varepsilon > 0$，有

$$\lim_{n\to\infty} \Pr(|X_n - X| > \varepsilon) = 0,$$

那么称 X_n **依概率收敛到** X，记为 $X_n \xrightarrow{P} X$. □

换句话说，依概率收敛的意思就是，当 n 趋于无穷，X_n 和 X 有任何固定偏差的概率趋于 0.

接下来，我们给出大数定律的一个形式化描述：

定理 C.29 (Khinchin 大数定律) 设 $X_1, X_2, \ldots$ 是一系列独立同分布的随机变量，$\mathbb{E}[X_i]$ 存在，那么

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X_1].$$

这一定律说明，独立重复试验的结果会趋于它们的期望.

证明. 由于 $X_1, X_2, \ldots$ 是独立同分布的，所以 $\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots$ 都是相等的，记为 μ. 令

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

我们计算 S_n 的特征函数：

$$\begin{aligned} f_{S_n}(t) &= \mathbb{E}[e^{itS_n}] \\ &= \mathbb{E}[e^{it/n(X_1+\cdots+X_n)}] \\ &= \mathbb{E}[e^{it/nX_1}\cdots e^{it/nX_n}] \\ &= \mathbb{E}[e^{it/nX_1}]\cdots\mathbb{E}[e^{it/nX_n}] \\ &= \left(\mathbb{E}[e^{it/nX_1}]\right)^n \\ &= \left(f_{X_1}\left(\frac{t}{n}\right)\right)^n. \end{aligned}$$

根据定理 C.25，我们有

$$f_{X_1}(t) = 1 + it\mu + o(t), \quad t \to 0.$$

于是

$$f_{S_n}(t) = \left(1 + \frac{it\mu}{n} + o\left(\frac{t}{n}\right)\right)^n \to e^{it\mu}, \quad n \to \infty.$$

因此，根据定理 C.20，S_n 的分布趋于一个在 μ 处集中的退化分布，根据定义，这意味着对任意 x，当 $n \to \infty$ 时，

$$\Pr(S_n \le x) \to \begin{cases} 0, & x < \mu, \\ 1, & x \ge \mu. \end{cases}$$

于是，对于任意 $\varepsilon > 0$，当 $n \to \infty$ 时，

$$\begin{aligned}\Pr(|S_n - \mu| > \varepsilon) &\leq \Pr(S_n < \mu - \varepsilon) + \Pr(S_n > \mu + \varepsilon) \\ &= \Pr(S_n < \mu - \varepsilon) + 1 - \Pr(S_n \leq \mu + \varepsilon) \\ &\to 0.\end{aligned}$$

这就完成了证明. □

参考文献

[Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[Inf] Information | Etymology, origin and meaning of information by etymonline. https://www.etymonline.com/word/information. (accessed 2023-07-10).

[Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.

[McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.

[Uff22] Jos Uffink. Boltzmann's Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.

[李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.