标题 **title**

作者 *author*

2023 年 8 月 28 日

前言

目录

第一部分

科学的逻辑

第二部分

信息与数据

第三部分

决策与优化

第四部分

逻辑与博弈

第五部分

认知逻辑

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。

§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) 考虑先后掷两个骰子的情况。样本空间为

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

样本点为 (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 $A_i = \{(i,j) : 1 \leq j \leq 6\}$. “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) 考虑随机选一个圆周上的点的情况。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，一个点的长度当然是 0，所以选到一个点的概率是 0。计算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

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

这里体积应该按照广义上来理解，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 $1/2$.

几何概型的定义非常微妙，因为我们并不知道如何定义“体积”。我们来看一个有趣的例子。

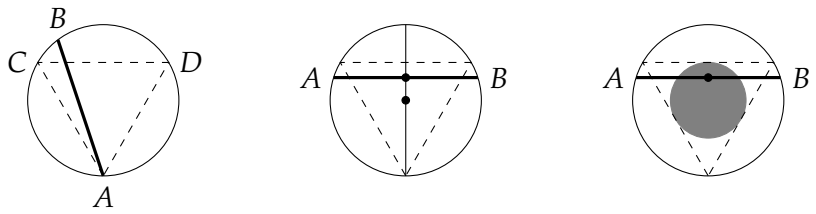
例 C.3 (Bertrand 悖论) 考虑一个圆，它的半径为 1. 现在我们随机地在圆上取一个弦，那么这个弦的长度超过 $\sqrt{3}$（即圆内接正三角形的边长）的概率是多少？我们给出三种答案。

解答 *1*. 不妨固定弦的其中一个点 A，那么另一个点 B 可以在圆上等可能取。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*）：



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？这个问题的答案非常微妙，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 $\mathscr{F}$，关于这一定义的哲学讨论，可以见第一章. 接下来，我们给出 Kolmogorov 概率论的公理化定义。

定义 C.2 (概率空间，概率测度) 设 Ω 是一个集合，$\mathscr{F}$ 是 Ω 上的一个 σ-代数。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。

下面我们来看一下之前的古典概型与几何概型都是如何构造概率空间的。

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。我们知道，对于一个长方体

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

它的体积为 $\prod_i (b_i - a_i)$. Ω 上事件域至少应该包含所有长方体，所以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。概率测度 $\Pr$ 的第一个要求是，让正方体的概率等于它的体积。第二个要求是平移不变性，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 $\Pr(A + x) = \Pr(A)$. 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空间上。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。我们还是看先后掷两个骰子的例子。如果掷完第一个骰子，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，我们很快就会觉得不对劲。此时，第一个骰子完全没有随机性！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.3 (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，从朴素的直觉来看，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 $\Pr(B|A) = \Pr(B) = \frac{1}{6}$. 如果单看数学计算，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！从朴素的直觉来说，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.4 (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，三个事件可能放在一起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. $\Pr(|\Omega) = \Pr(\cdot)$，
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。

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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A, B \in \mathscr{F}$ 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

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

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 $\Pr(Q)$。此时，事件交的概率也经常以逗号的形式写出，例如，$\Pr(A_1A_2)$ 我们会记为 $\Pr(\text{第一次取出的球是白球}, \text{第二次取出的球是白球})$. 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。然而，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 $\{X\in B\}$，这里 $B\subseteq\mathbb{N}^2$. 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.5 (随机变量，随机元) 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 S 以及其上的 σ-代数 $\mathscr{F}_S$，$X:\Omega\to S$ 是一个映射。如果对任意的 $A\in\mathscr{F}_S$，$\{X\in A\}\in\mathscr{F}$，则称 X 是一个**随机元**。如果 $S=\mathbb{R}^n$ 且 $\mathscr{F}_S=\mathscr{B}(\mathbb{R}^n)$，则称 X 是一个 n **维随机向量**。

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，它是一个映射是因为，它的值是随机的，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 $\{X\in\mathscr{B}(\mathbb{R})\}$ 表示事件“X 取合理的值”。随机变量的要求其实就是，“X 取合理的值”是一个我们可以定义概率的事件。

我们下面讨论一些随机变量的基本性质。

命题 C.3 设 $(\Omega,\mathscr{F},\Pr)$ 是一个概率空间，$X:\Omega\to\mathbb{R}$ 是一个随机变量. 假设函数 $g:\mathbb{R}\to\mathbb{R}$ 是一个连续函数，则 $g(X)=g\circ X$ 也是一个随机变量。更一般地，如果 $X:\Omega\to\mathbb{R}^n$ 是一个随机向量，$g:\mathbb{R}^n\to\mathbb{R}^m$ 是一个连续函数，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，我们可以先构造一个随机变量，然后对它进行一些连续的操作，例如加减乘除、取指数、取对数、取幂等等，这样得到的新的函数也是一个随机变量。

实际上，不连续的函数进行复合也可以构造新的随机变量，这样的函数可以用 **Borel 函数**来刻画。实际上，Borel 函数就是样本空间 $\mathbb{R}^n$ 上的随机向量。

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 $\Pr(A)$。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 $\Pr(X \in B)$，基本上只需要确定 $\Pr(X \in (a,b))$。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.6 (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 $X \sim F$.

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 $X \sim F$.

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，即对任意的 $x \in \mathbb{R}$，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，即对任意的 $x \in \mathbb{R}$，都有 $F(x-) = \lim_{y \uparrow x} F_X(y)$ 存在.

实际上，分布函数也可以由上面四条给出定义，这是因为，满足上面四条性质的函数恰好是某个随机变量的分布函数：

定理 C.3 设 F 是 $\mathbb{R}^n \to \mathbb{R}$ 的函数，满足命题 C.4 的四条性质.

在概率空间 $([0,1]^n, \mathscr{B}([0,1]^n), \lambda)$ 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，

- $\Pr(X \leq a) = F(a)$,

- $\Pr(X < a) = F(a-)$,
- $\Pr(X > a) = 1 - F(a)$,
- $\Pr(X \geq a) = 1 - F(a-)$,
- $\Pr(X = a) = F(a) - F(a-)$。

如果我们就限制在空间 $([0,1]^n, \mathscr{B}([0,1]^n), \lambda)$ 上，随机变量几乎就等同于分布函数。现在，我们将分布函数与概率测度联系在一起：

定理 C.4 *设 $F : \mathbb{R}^n \to \mathbb{R}$ 是一个分布函数，则在 $\mathbb{R}^n$ 以及 $\mathscr{B}(\mathbb{R}^n)$ 上，存在唯一的概率测度 $\Pr$，使得对任意 $a_i \leq b_i$，*

$$\Pr\left(\prod_{i=1}^{n}(a_i, b_i]\right) = \prod_{i=1}^{n}(F(b_i) - F(a_i)).$$

特别地，分布函数

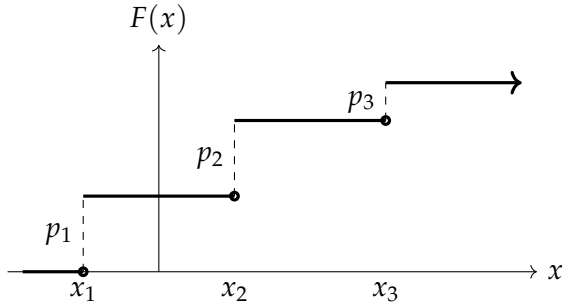
$$F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

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

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 $x_1, x_2, \ldots$ 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，这一部分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；积分学我们会在后面附录 C.3 以数学期望的形式介绍。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.7 (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? 答案是否定的，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.5 (Lebesgue 微分定理) 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.5 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。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.5，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**。

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，他们会写 $p(X = x)$ 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.5，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 $\mathrm{d}\Pr = p\mathrm{d}\lambda$；而概率刻画的是宏观的，计算任何一个集合的概率，也就是 $\Pr(A)$.

> **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.5 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**。

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。

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	$\mathsf{Exp}(\lambda)$	$p(x)=\lambda e^{-\lambda x},\ x\geq 0$	$\lambda>0$
双指数	$\mathsf{DExp}(\lambda)$	$p(x)=\frac{\lambda}{2}e^{-\lambda\lvert x\rvert},\ x\in\mathbb{R}$	$\lambda>0$
Laplace	$\mathsf{Lap}(\mu,\lambda)$	$p(x)=\frac{\lambda}{2}e^{-\lambda\lvert x-\mu\rvert},\ x\in\mathbb{R}$	$\mu\in\mathbb{R},\ \lambda>0$
正态（Gauss）	$\mathcal{N}(\mu,\sigma^2)$	$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x\in\mathbb{R}$	$\mu\in\mathbb{R},\ \sigma>0$

表 C.2: 本书中用到的连续型分布

> **注.** 从定理 C.5 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，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。

C.2.4 随机向量

C.2.5 随机变量（向量）的函数

§C.3 随机变量的数字特征，期望

§C.4 多元正态分布（Gauss 向量）

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.

[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.

[Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.

[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.

索引