

第一节 · 概率

司继春

上海对外经贸大学统计与信息学院

首先我们回顾一下**概率空间** (probability space) 的定义。我们知道, 在概率论中, 概率空间为一个三元组: $(\Omega, \mathcal{F}, \mathcal{P})$, 其中 Ω 为样本空间, \mathcal{F} 为所有事件的集合, \mathcal{P} 为概率测度。下面我们分别探讨三元组的每个元素。

1 样本空间

样本空间 (sample space) Ω 为我们关心的随机试验的所有结果的集合, 而 Ω 中的元素 $\omega \in \Omega$ 称之为**样本点** (sample point)。例如:

1. 随机从一堆扑克牌中抽取一张扑克, 其花色的样本空间为:

$$\Omega_1 = \{\heartsuit, \spadesuit, \clubsuit, \diamondsuit\}$$

而样本点为 \heartsuit 、 \spadesuit 、 \clubsuit 以及 \diamondsuit 。

2. 某银行一天所接待的所有客户数, 其样本空间为 $\Omega_2 = \mathbb{N}$, 样本点为自然数。
3. 随机从人群中抽取一个人, 其身高的样本空间为 $\Omega_3 = \mathbb{R}^+$, 而样本点为实数。

注意在上面的三个例子中, 样本空间有细微差别。 Ω_1 的元素个数为有限个, 而 Ω_2 和 Ω_3 的元素个数有无穷多个。其中, Ω_1 与 Ω_2 都可以与自然数 \mathbb{Z} 或者 \mathbb{N} 的子集建立起一一对应的关系, 我们称之为**可数集** (countable set), 而像 Ω_3 这样不能与自然数 \mathbb{Z} 或者其子集建立起一一对应关系的, 我们称之为**不可数集** (uncountable set)。

在概率论中, 显示的定义样本空间是非常重要的, 同一个问题, 如果定义的样本空间不同, 可能会得到完全不同的结果。

例 1. (贝特朗悖论) 考虑一个内接于圆的等边三角形。若随机选方圆上的个弦, 则此弦的长度比三角形的边较长的概率是多少?

根据不同的假设, 有三种解法:

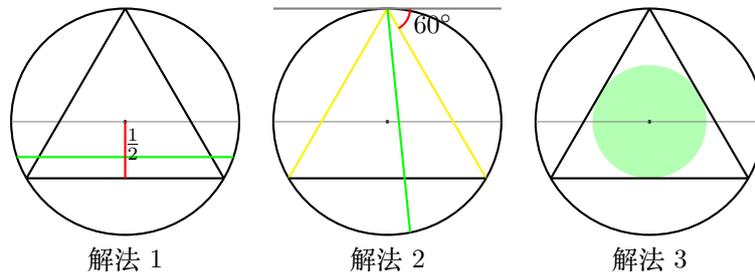


图 1: 贝特朗悖论

解法一，如图所示，在垂直于三角形任意一边的直径上随机取一个点，并通过该点做一条垂直于该直径的弦，在该点位于半径中点的时候弦长度等于三角形的边长度，所以概率为 $\frac{1}{2}$ 。

解法二，如图所示，通过三角形任意一个顶点做圆的切线，因为等边三角形内角为 60° ，所以左边右边的角都是 60° 。由该顶点做一条弦，弦的另一端在圆上任意一点。由图可知弦与切线成 60° 角和 120° 角之间的时候弦长度大于三角形边长，所以概率为 $\frac{1}{3}$ 。

解法三，如图所示，当弦的中点在阴影标记的圆内时，弦的长度大于三角形的边长，而大圆的弦中点一定在圆内，大圆的面积是 πr^2 ，小圆的面积是 $\frac{1}{4}\pi r^2$ ，所以概率为 $\frac{1}{4}$ 。

同一个问题为什么会得到三种不同的答案呢？原因在于，圆内“取弦”时规定尚不够具体，不同的“等可能性假定”导致了不同的样本空间：第一种解法中，假设弦的中点在直径上均匀分布；第二种解法中，假设弦的另一端在圆周上均匀分布；第三种解法中，假设弦的中点在大圆内均匀分布。

因而在定义概率时，第一步必须明确地指出样本空间是什么。

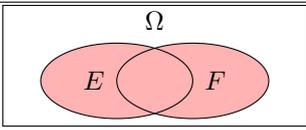
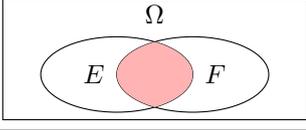
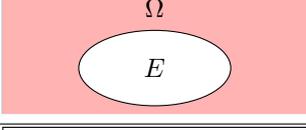
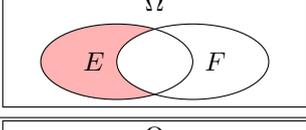
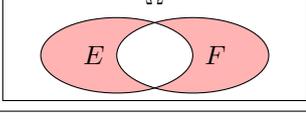
2 事件

我们称样本空间 Ω 的子集（包含 Ω 本身）为**事件**（event）。如果 A 为 Ω 的一个子集，如果随机试验的结果为 A 中的一个样本点，我们称之为发生了事件 A 。在通常情况下，当我们称概率时，指的是事件发生的概率。

我们首先回忆集合的运算与性质。表1列出了常见集合运算的定义。对于表1中列出的集合运算，我们有以下的运算法则：

1. 交换律: $A \cup B = B \cup A$, $A \cap B = B \cap A$
2. 结合律: $A \cup (B \cap C) = (A \cup B) \cap C$, $A \cap (B \cup C) = (A \cap B) \cup C$
3. 分配率: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
4. 德摩根律: $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$

表 1: 集合的运算

运算	符号	定义	图示
并	$E \cup F$	$\{\omega \in E \text{ OR } \omega \in F\}$	
交	$E \cap F$	$\{\omega \in E \text{ AND } \omega \in F\}$	
补集	E^c	$\{\omega \in \Omega, \omega \notin E\}$	
差	$E \setminus F$	$E \cap F^c$	
对称差	$E \Delta F$	$(E \setminus F) \cup (F \setminus E)$	

此外，以上运算法则容易推广至可数个集合的运算，比如德摩根律：

$$\left(\bigcup_{i=1}^{\infty} A_i \right)^c = \bigcap_{i=1}^{\infty} A_i^c$$

以及：

$$\left(\bigcap_{i=1}^{\infty} A_i \right)^c = \bigcup_{i=1}^{\infty} A_i^c$$

如果两个事件 A 和 B 满足 $A \cap B = \emptyset$ ，我们称之为**互斥事件** (disjoint or exclusive)。如果对于一系列事件 A_1, A_2, \dots ，对于任意 i 和 j ，有 $A_i \cap A_j = \emptyset$ ，则称之为**两两互斥事件**。如果 A_1, A_2, \dots 为两两互斥事件，且 $\bigcup_{i=1}^{\infty} A_i = \Omega$ ，则 A_1, A_2, \dots 为样本空间的一个**划分** (partition)。

对于一个一般的样本空间，有数量巨大的子集。我们希望挑出那些我们需要研究的子集，同时剔除那些性质不是十分良好的子集，这就诞生了 σ -代数的概念。

定义 1. (σ -代数) 如果样本空间 Ω 的一系列子集的集合 \mathcal{F} 满足：

1. $\emptyset \in \mathcal{F}$
2. 若 $A \in \mathcal{F}$ ，则 $A^c \in \mathcal{F}$

3. 若 $A_1, A_2 \dots \in \mathcal{F}$, 则 $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

我们称 \mathcal{F} 为一个 σ -代数, 或者 σ -域。

注意 σ -代数 \mathcal{F} 为一个集合, 其成员为样本空间 Ω 的子集, 即事件, 所以 \mathcal{F} 为事件的集合。以上定义中 3 要求如果可数个集合在 \mathcal{F} 中, 那么这可数个集合的并集也要求在 \mathcal{F} 中。而同时, 结合 2 中关于补集的要求, 根据德摩根律:

$$\left(\bigcup_{i=1}^{\infty} A_i \right)^c = \bigcap_{i=1}^{\infty} A_i^c$$

因而 2 和 3 结合起来要求可数个集合的交集也需要在 \mathcal{F} 中。

例 2. 对于上述定义的 Ω_1 , 如果我们关心单个样本点: $\{\heartsuit\}, \{\clubsuit\}, \{\spadesuit\}, \{\diamondsuit\}$, 现构建对应的 σ -代数。根据要求, 首先集合中应该包含:

$$\{\emptyset, \{\heartsuit\}, \{\clubsuit\}, \{\spadesuit\}, \{\diamondsuit\}, \Omega_1\} \triangleq F_1$$

同时, F_1 中的元素的并集也需要包含在其中, 因而:

$$\left\{ \begin{array}{l} \emptyset, \quad \{\heartsuit\}, \quad \{\clubsuit\}, \quad \{\spadesuit\}, \quad \{\diamondsuit\}, \quad \Omega_1, \\ \{\heartsuit, \clubsuit\}, \quad \{\heartsuit, \spadesuit\}, \quad \{\heartsuit, \diamondsuit\}, \quad \{\clubsuit, \spadesuit\}, \quad \{\clubsuit, \diamondsuit\}, \quad \{\spadesuit, \diamondsuit\}, \end{array} \right\} \triangleq F_2$$

进而, F_1 中的元素的补集也需要包含在其中, 因而:

$$\left\{ \begin{array}{l} \emptyset, \quad \{\heartsuit\}, \quad \{\clubsuit\}, \quad \{\spadesuit\}, \quad \{\diamondsuit\}, \quad \Omega_1, \\ \{\heartsuit, \clubsuit\}, \quad \{\heartsuit, \spadesuit\}, \quad \{\heartsuit, \diamondsuit\}, \quad \{\clubsuit, \spadesuit\}, \quad \{\clubsuit, \diamondsuit\}, \quad \{\spadesuit, \diamondsuit\}, \\ \{\heartsuit, \clubsuit, \spadesuit\}, \quad \{\heartsuit, \clubsuit, \diamondsuit\}, \quad \{\heartsuit, \spadesuit, \diamondsuit\}, \quad \{\clubsuit, \spadesuit, \diamondsuit\} \end{array} \right\} \triangleq F_3$$

仔细观察, 发现 F_3 已经满足定义中的要求, 因而 $\mathcal{F} = F_3$ 即我们要构建的 σ -代数。

若我们关心的样本空间为 $\Omega = \mathbb{R}$, 我们令 \mathcal{S} 为所有开区间的集合: $\mathcal{S} = \{(a, b) \mid -\infty < a < b < +\infty\}$, 那么包含 \mathcal{S} 的最小 σ -代数我们记为 \mathcal{B} , 并称之为 **Borel σ -代数** 或 **Borel 域**, 而 \mathcal{B} 中的元素成为 **Borel 集 (Borel set)**。注意由于:

$$(a, b) = \bigcap_{i=1}^{\infty} \left(a, b + \frac{1}{i} \right)$$

因而所有的左开右闭区间也都是 Borel 集。同理可证所有的左闭右开区间 $[a, b)$ 、闭区间 $[a, b]$ 及其可数并、交都为 Borel 集。

3 概率

现在, 经过以上准备之后, 我们可以定义概率了。

定义 2. (Kolmogorov axioms) 给定一个样本空间 Ω 以及相应的 σ -代数 \mathcal{F} , 函数 $\mathcal{P} : \mathcal{F} \rightarrow [0, 1]$ 若满足:

1. 对于所有的事件 $A \in \mathcal{F}$, $\mathcal{P}(A) \geq 0$
2. $\mathcal{P}(\Omega) = 1$
3. 若 $A_1, A_2, \dots \in \mathcal{F}$ 为两两互斥事件, 则 $\mathcal{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$ (可数可加性或可列可加性)

则我们称 \mathcal{P} 为**概率函数**或**概率测度**。

以上概率的定义通常称之为概率的公理化定义 (Axioms of Probability), 或者柯尔莫哥洛夫公理 (Kolmogorov Axioms)。注意以上的定义并没有限定概率函数的形式, 只要满足以上三个条件的函数 \mathcal{P} 都可以被定义为概率函数。

例 3. (抛硬币的概率) 现在我们进行一项抛硬币的随机试验。记正面为 H , 反面为 T , 那么我们关心的样本空间为 $\Omega = \{H, T\}$ 。假设硬币质地均匀, 即正面和反面的概率相等, 那么

$$\mathcal{P}(\{H\}) = \mathcal{P}(\{T\})$$

包含 $\{H\}, \{T\}$ 的最小 σ -代数为 $\mathcal{F} = \{\emptyset, \Omega, \{H\}, \{T\}\}$ 。根据概率的定义, $\mathcal{P}(\Omega) = \mathcal{P}(\{H\} \cup \{T\}) = \mathcal{P}(\{H\}) + \mathcal{P}(\{T\}) = 1$, 因而 $\mathcal{P}(\{H\}) = \mathcal{P}(\{T\}) = 0.5$ 。因而 $(\Omega, \mathcal{F}, \mathcal{P})$ 即组成了概率空间。

如果假设硬币是不均匀的, 且获得正面的概率为 0.1, 那么同样根据概率定义, $\mathcal{P}'(\{H\}) = 0.1, \mathcal{P}'(\{T\}) = \mathcal{P}'(\Omega) - \mathcal{P}'(\{H\}) = 1 - 0.1 = 0.9$, 从而 $(\Omega, \mathcal{F}, \mathcal{P}')$ 组成了新的概率空间。

以上我们使用抛硬币的例子构建了样本空间有限时的概率空间, 然而当样本点逐渐增多时, \mathcal{F} 的元素个数也会相应增加, 逐个检验 \mathcal{F} 中元素是否满足概率定义的三个条件变的更为复杂。而对于一般的有限或者可数个样本点的情形 (或统称为离散的样本空间), 我们可以使用如下的方法定义概率函数。

定理 1. 令 $\Omega = \{s_1, s_2, \dots, s_n\}$ 为有限集, 令 \mathcal{F} 为 S 子集的任意 σ -代数。令 p_1, p_2, \dots, p_n 为非负实数且 $\sum_{i=1}^n p_i = 1$ 。对于任意集合 $A \in \mathcal{F}$, 定义

$$\mathcal{P}(A) = \sum_{\{s_i, i \in A\}} p_i \quad (1)$$

则 \mathcal{P} 为 \mathcal{F} 上的概率函数。对于可数集 $\Omega = \{s_1, s_2, \dots\}$ 可类似构建概率函数。

Proof. 定义 (2) 中第 1 条显然满足, 对于第 2 条, 根据上述定义, $\mathcal{P}(\Omega) = \sum_{i=1}^n p_i = 1$, 满足。对于第三条, 对于两两互斥事件 A_1, \dots, A_k :

$$\mathcal{P}\left(\bigcup_{i=1}^k A_i\right) = \sum_{\{s_i, i \in \bigcup_{i=1}^k A_i\}} p_i = \sum_{i=1}^k \sum_{\{s_j, j \in A_i\}} p_j = \sum_{i=1}^k \mathcal{P}(A_i)$$

因而上述定义的概率函数 \mathcal{P} 满足 Kolmogorov 公理。 \square

例 4. 如果我们重复、独立的进行抛硬币试验 (**伯努利试验**) N 次, 且 $\mathcal{P}(\{H\}) = p$, 那么 N 次试验中得到正面的次数的集合为: $\Omega = \{0, 1, \dots, N\}$, 对应的概率为: $\mathcal{P}(\{k\}) = \binom{N}{k} p^k (1-p)^{N-k}$, $\mathcal{P}(\{k\}) \geq 0$ 且 $\sum_{k=0}^N \mathcal{P}(\{k\}) = 1$, 因而根据定理 (1), 使用 (1) 式定义的 \mathcal{P} 即定义了样本空间 Ω 上任意 σ -代数的所有子集的概率函数。

我们关心在一个小时之内到达某银行的客户数, 客户数为可数集, 样本空间为 $\Omega = \{0, 1, 2, \dots\}$ 。取一个非常大的自然数 n , 我们可以把一个小时分解为等长的 n 段, 即 $(0, \frac{1}{n}], (\frac{1}{n}, \frac{2}{n}] \dots (\frac{n-1}{n}, 1]$, 当 n 很大时, 一个区间段内有两个客户到达的概率几乎可以忽略不计。假设每段时间客户到达的概率相等, 且反比于 n , 不妨假设为 $\frac{\lambda}{n}$, 那么一小时内总的人数:

$$\mathcal{P}^*(\{k\}) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

令 $n \rightarrow \infty$, 则 $\frac{\binom{n}{k}}{n^k} = \frac{n!}{k!(n-k)!n^k} \rightarrow \frac{1}{k!}$, $\left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}$, 因而

$$\mathcal{P}(\{k\}) = \frac{\lambda^k}{k!} e^{-\lambda} > 0$$

且 $\sum_{k=0}^{\infty} \mathcal{P}(\{k\}) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1$ ($e^{\lambda x}$ 在 $x=0$ 处泰勒展开可得), 因而根据定理 (1), 使用 (1) 式定义的概率函数 \mathcal{P} 即定义了样本空间 Ω 上任意 σ -代数的概率函数。

3.1 分布函数与概率定义

尽管在样本空间为可数的情况下定义概率函数相对简单, 然而当我们考虑的样本空间为不可列时, 概率函数的定义变得尤为困难。

例 5. (勒贝格不可测集) 如果我们选取样本空间 $\Omega = [0, 1]$, 令 Ω 中的所有有理数集合为 Q' , 由于有理数为可数集合, 因而可以写成 $Q' = \{q_1, q_2, \dots\}$ 。对于 $(0, 1)$ 之间的任意实数 a , 定义集合

$$S_a = \left\{ \begin{array}{ll} a+q & \text{if } a+q < 1 \\ a+q-1 & \text{if } a+q \geq 1 \end{array} \forall q \in Q' \right\}$$

那么可知 $\bigcup_{a \in (0,1)} S_a = [0, 1]$ 。由于 S_a 也是可数集, 因而可以将其写为:

$$S_a = \{s_{a1}, s_{a2}, \dots\}$$

令 T_1 为所有 S_a 中的 s_{a1} , T_2 为所有 S_a 中的 s_{a2} , 因而我们有可数个 T_k , $\bigcup_{k=1}^{\infty} T_k = [0, 1]$, 且 T_k 两两不相交。每个 T_k 地位相等因而 $\mathcal{P}(T_k) = \mathcal{P}(T_{k'})$ 。若 $\mathcal{P}(T_k) > 0$, 则:

$$1 = \mathcal{P}([0, 1]) = \mathcal{P}\left(\bigcup_{k=1}^{\infty} T_k\right) = \sum_{k=1}^{\infty} \mathcal{P}(T_k) = \infty$$

如果 $\mathcal{P}(T_k) = 0$, 则:

$$1 = \mathcal{P}([0, 1]) = \mathcal{P}\left(\bigcup_{k=1}^{\infty} T_k\right) = \sum_{k=1}^{\infty} \mathcal{P}(T_k) = 0$$

无论如何都会得到矛盾。

因而在概率论中, 在仅仅给定样本空间的情况下, 并非任意集合都可以确定其概率。我们一般将上述性质不够良好的集合称之为(勒贝格)不可测集, 而概率空间中 \mathcal{F} 应该排除这些性质不够良好的不可测集。

定义 3. (分布函数) 如果函数 $F: \mathbb{R} \rightarrow \mathbb{R}$ 满足:

1. 单调性: $F(a) \leq F(b), a \leq b$
2. 右连续: $\lim_{x \rightarrow x_0^+} F(x) = F(x_0)$
3. $F(-\infty) = 0, F(\infty) = 1$

则称 F 为分布函数 (distribution function, d.f.)。特别的, 令

$$\delta_t(x) = \begin{cases} 0 & x < t \\ 1 & x \geq t \end{cases}$$

若 $\{a_j\}$ 为可数集, $b_j > 0, \sum_j b_j = 1$, 则 $F(x) = \sum_j b_j \delta_{a_j}(x)$ 为分布函数, 我们称之为**离散型分布函数** (discrete d.f.); 处处连续的分布函数成为**连续型分布函数** (countinuous d.f.)。

例 6. 令 $a_1 = 0, a_2 = 1, b_1 = \frac{1}{3}, b_2 = \frac{2}{3}$, 则 $F_d(x) = \sum_{j=1}^2 b_j \delta_{a_j}(x)$ 为离散型分布函数, 如图 (2.a) 所示; $F_c(x) = \frac{e^x}{1+e^x}$ (Logistic 分布) 为连续型分布函数, 如图 (2.b) 所示, 而 $F(x) = \frac{1}{3}F_d(x) + \frac{2}{3}F_c(x)$ 也为分布函数, 如图 (2.c) 所示。

定理 2. 每个分布函数都可以写为一个离散型分布函数和一个连续型分布函数的凸组合, 且该分解唯一。

在有了分布函数之后, 我们可以使用分布函数定义 \mathbb{R} 上的概率函数。例 (2) 中我们通过开区间定义了 Borel 域, 如果我们定义:

$$P((-\infty, x]) = F(x) \tag{2}$$

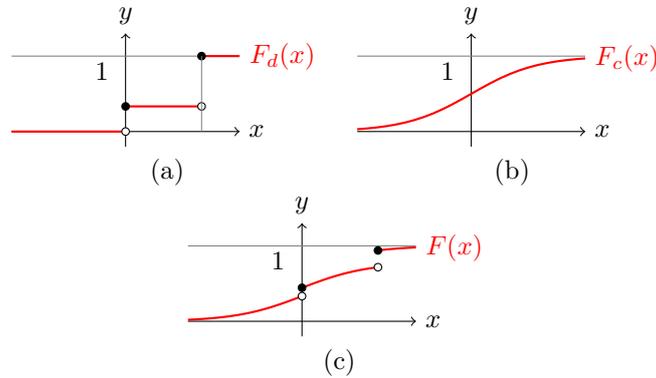


图 2: 分布函数

则对于任意的 $-\infty < a < b < +\infty$, 有¹:

$$\begin{aligned} P((a, b]) &= F(b) - F(a) \\ P((a, b)) &= F(b-) - F(a) \\ P([a, b)) &= F(b-) - F(a-) \\ P([a, b]) &= F(b) - F(a-) \end{aligned}$$

定理 3. 给定任意的分布函数 F , 式 (2) 定义了 Borel 域 \mathcal{B} 上的概率测度。

在此, 我们仅仅概括性的了解一下整个定义过程, 至于具体的理论推导, 感兴趣可以参考 Ash (2000) Ch.1.3-1.4。现在考虑一个集合 $S \subset \mathbb{R}$, 如果 S 可以写成可数个不相交的左开右闭区间的并集, 即:

$$S = \bigcup_{i=1}^{\infty} (a_i, b_i]$$

那么 $P(S) = \sum_{i=1}^{\infty} P((a_i, b_i]) = \sum_{i=1}^{\infty} (F(b_i) - F(a_i))$ 。而类似的, 对于任意的一个开集 $U \subset \mathbb{R}$, 都可以写为可数个开区间的并集, 即:

$$U = \bigcup_{i=1}^{\infty} (c_i, d_i)$$

类似地, $P(U) = \sum_{i=1}^{\infty} [F(d_i-) - F(c_i)]$, 故对于所有的开集, 我们定义了其概率。进而, 由于闭集是开集的补集, 我们可以使用开集的概率来定义闭集的概率。

在定义了开集和闭集的概率之后, 对于任意一个集合 $S \subset \mathbb{R}$, 我们可以定

¹这也就是为什么开区间、闭区间、半开半闭区间在形成 Borel 域以及定义概率函数上都是等价的, 然而一般用左开右闭区间的原因, 因为 $P((a, b]) = F(b) - F(a)$, 不像其他三个定义需要使用极限, 表达更方便。

义其**外测度** (outer measure):

$$P^*(S) = \inf_{\text{开集 } U, S \subset U} P(U)$$

和**内测度** (inner measure):

$$P_*(S) = \sup_{\text{闭集 } C, C \subset S} P(C)$$

易知 $P_*(S) \leq P^*(S)$ 。一般来说, 等号不一定成立, 然而如果等号成立, 我们就定义 $P(S) = P^*(S) = P_*(S)$, 并称 S 为**可测集** (measurable set)。可以证明, 所有的可测集是一个 σ 代数, 且包含了所有的开区间。由于 Borel 域 \mathcal{B} 是包含所有开区间的最小 σ 代数, 因而这个概率定义了 Borel 域 \mathcal{B} 上的概率函数。

至此, 我们就定义了实数集 \mathbb{R} 上的概率空间: $(\mathbb{R}, \mathcal{B}, P)$ ²。

有一些命题, 尽管其并非对于每个 $\omega \in \Omega$ 都成立, 但是 $\mathcal{P}(\{\omega : \text{命题成立}\}) = 1$, 即这个命题成立的概率为 1, 那么我们称这个命题是**几乎处处** (almost everywhere) 成立的, 或者**几乎必然** (almost sure) 成立的, 简记为 *a.e.* 或者 *a.s.*。例如:

例 7. 取 $\Omega = \mathbb{R}$, 定义

$$X_n(\omega) = \begin{cases} 0 & \text{if } |\omega| > \frac{1}{n} \\ n & \text{if } |\omega| \leq \frac{1}{n} \end{cases}$$

因而除了在一个点 $\omega = 0$ 处之外, $\lim_{n \rightarrow \infty} X_n(\omega) = 0$ 。如果分布函数连续, 单点集 $P(\{0\}) = 0$, 因而 $\lim_{n \rightarrow \infty} X_n(\omega) = 0$ 以概率 1 成立, 我们称 $\lim_{n \rightarrow \infty} X_n(\omega) = 0$ 几乎必然成立, 简记为: $\lim_{n \rightarrow \infty} X_n(\omega) = 0$ *a.e.* 或者 $\lim_{n \rightarrow \infty} X_n(\omega) = 0$ *a.s.*。

3.2 概率函数的性质

定理 4. 对于概率函数 \mathcal{P} , 有以下性质:

1. $\mathcal{P}(A) \leq 1$
2. $\mathcal{P}(A^c) = 1 - \mathcal{P}(A)$
3. $\mathcal{P}(\emptyset) = 0$
4. $\mathcal{P}(A \cup B) + \mathcal{P}(A \cap B) = \mathcal{P}(A) + \mathcal{P}(B)$
5. $A \subset B \Rightarrow \mathcal{P}(A) = \mathcal{P}(B) - \mathcal{P}(B \setminus A) \leq \mathcal{P}(B)$

²注意在此讲义中, 对于一般的概率空间, 我们使用 \mathcal{P} 作为概率函数的标记, 而对于 Borel 集上的概率函数, 我们使用 P 作为特殊标记。

$$6. \mathcal{P}(\cup_i A_i) \leq \sum_i \mathcal{P}(A_i)$$

7. 如果 C_1, C_2, \dots 为样本空间 Ω 的一个划分, 那么 $\mathcal{P}(A) = \sum_{i=1}^{\infty} \mathcal{P}(A \cap C_i)$ 。

例 8. (Bonferroni's Inequality) 根据定理 (4), 我们有:

$$\mathcal{P}(A \cap B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cup B) \geq \mathcal{P}(A) + \mathcal{P}(B) - 1$$

该不等式确定了事件 A 和 B 同时发生的概率的下界。比如, 天气预报预计明天上海下雨(A)的概率为 0.90, 北京下雨(B)的概率为 0.8。如果假设两地是否下雨是独立事件, 那么两地同时下雨的概率为 $\mathcal{P}(A \cap B) = \mathcal{P}(A) \cdot \mathcal{P}(B) = 0.72$ 。然而现实情况是, 我们并不知道两地下雨是否独立, 但是根据上面的不等式, 我们仍然可以得到两地同时下雨的概率的一个下界: $\mathcal{P}(A \cap B) \geq 0.8 + 0.9 - 1 = 0.7$, 即两地同时下雨的概率至少有 0.7。

4 条件概率与独立

在此之前我们讨论的都是无条件概率。然而现实的应用中, 我们经常碰到用已知信息推断未知信息的问题, 这就涉及到条件概率的概念。条件概率即指, 给定事件 B 发生的情况下, 事件 A 发生的概率是多少。无条件概率和有条件概率有着不同的应用, 比如如果我们想要标记山体滑坡的危险路段, 那么我们可以统计路段长时间以来发生山体滑坡的概率, 即无条件概率, 进行标记; 而当我们进行灾害预警时, 我们知道给定天气是阴雨天的情况下, 山体滑坡的概率会变的异常高, 这个时候我们就是在使用条件概率了, 即 $\mathcal{P}(\text{发生山体滑坡} | \text{阴雨天})$ 可能是灾害预警所关注的, 而不是 $\mathcal{P}(\text{发生山体滑坡})$ 。

定义 4. (条件概率) 如果 A 和 B 为 Ω 中的两个事件, 且 $\mathcal{P}(B) > 0$, 那么给定 B , 事件 A 发生的条件概率为:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)} \quad (3)$$

注意只有当事件 B 有正概率发生时, 条件概率的定义才有意义。实际上, 条件概率可以理解为我们把原始的样本空间 Ω 限定在新的样本空间 B 中, 并相应对原概率函数使用式 (3) 对概率函数进行了重新定义, 因而概率的性质在条件概率下依然成立。

定理 5. (全概率公式) 如果 C_1, C_2, \dots 为样本空间 Ω 的一个划分, 那么 $\mathcal{P}(A) = \sum_{i=1}^{\infty} \mathcal{P}(A|C_i) \cdot \mathcal{P}(C_i)$ 。特别的, 对于任意事件 B , 有 $\mathcal{P}(A) = \mathcal{P}(A|B) \cdot \mathcal{P}(B) + \mathcal{P}(A|B^c) \cdot \mathcal{P}(B^c)$ 。

Proof. 根据条件概率定义, $\sum_{i=1}^{\infty} \mathcal{P}(A|C_i) \cdot \mathcal{P}(C_i) = \sum_{i=1}^{\infty} \mathcal{P}(A \cap C_i)$, 根据定理 (4.7) 可证。 \square

如果事件 A 和 B 都有正的概率, 那么我们可以同时定义 $\mathcal{P}(A|B)$ 以及 $\mathcal{P}(B|A)$, 根据定义,

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

以上关系式我们称之为**贝叶斯法则** (Bayes' Rule)。更进一步, 根据全概率公式:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A) \cdot \mathcal{P}(A)}{\mathcal{P}(B)} = \frac{\mathcal{P}(B|A) \cdot \mathcal{P}(A)}{\mathcal{P}(B|A) \cdot \mathcal{P}(A) + \mathcal{P}(B|A^c) \cdot \mathcal{P}(A^c)}$$

贝叶斯法则被广泛应用在统计学中, 因为贝叶斯法则体现了我们认识世界的一般规律, 很好的将理论与现实、已知与未知、主观与客观联系在一起。在贝叶斯公式中, 条件概率 $\mathcal{P}(B|A)$ 是我们已知的部分, 或者我们的理论, 而条件概率 $\mathcal{P}(A|B)$ 则是我们未知的需要判断的部分, 而 $\mathcal{P}(A)$ 则是主观的先验部分。贝叶斯公式的优雅之处在于, 它告诉我们, 如何使用我们的理论 ($\mathcal{P}(B|A)$) 和我们的主观先验 ($\mathcal{P}(A)$) 结合起来, 使用更多的信息 (B), 推断我们所未知的知识 $\mathcal{P}(A|B)$ 。

例 9. 小明正在追求一个女孩小红, 但是小明不知道小红的心意。假设 A 代表小红喜欢小明, A^c 代表小红不喜欢小明, 由于小明不确定小红的心思, 因而主观先验概率 $\mathcal{P}(A) = \mathcal{P}(A^c) = 0.5$ 。为了探明小红的心意, 4月1日小明发短信跟小红表白, 小红回短信表示愿意接受。小明的理论认为, 如果小红喜欢自己, 有 99% 的可能性会表示接受; 而如果小红不喜欢自己, 有一半的可能性小红是在开愚人节玩笑。如果令 B 代表小红接受的行为, 那么小红的理论可以归结为: $\mathcal{P}(B|A) = 0.99, \mathcal{P}(B|A^c) = 0.5$ 。那么现在, 根据贝叶斯法则, 小明可以对小红喜欢自己的概率做出判断:

$$\begin{aligned} \mathcal{P}(A|B) &= \frac{\mathcal{P}(B|A) \cdot \mathcal{P}(A)}{\mathcal{P}(B|A) \cdot \mathcal{P}(A) + \mathcal{P}(B|A^c) \cdot \mathcal{P}(A^c)} \\ &= \frac{0.99 \times 0.5}{0.99 \times 0.5 + 0.5 \times 0.5} \approx 66\% \end{aligned}$$

贝叶斯法则可以扩展到更一般的形式:

定理 6. (贝叶斯法则) 如果 B_1, B_2, \dots 为样本空间 Ω 的一个划分, 令 $A \in \mathcal{B}$, 那么:

$$\mathcal{P}(B_i|A) = \frac{\mathcal{P}(A|B_i) \mathcal{P}(B_i)}{\sum_{j=1}^{\infty} \mathcal{P}(A|B_j) \mathcal{P}(B_j)}$$

例 10. 有三扇门, 其中一扇门里有奖品, 三选一, 你选择其中一扇门之后, 主持人先不揭晓答案, 而是从另外两扇门中排除掉一个没有奖品的门, 现在主持人问你, 要不要换个门, 请问你换还是不换? 这里假设第 i 扇门里面有奖品的事件为 A_i , 没有奖品为 A_i^c , 因而总共三种情况, 即 $\{A_1 A_2^c A_3^c, A_1^c A_2 A_3^c, A_1^c A_2^c A_3\}$,

三种情况是等可能的。不失一般性，假设你选择了第一扇门，而主持人打开了第三扇门。如果根据第三扇门内没有奖品这一信息，可以得出： $\mathcal{P}(A_1 A_2^c | A_3^c) = \frac{\mathcal{P}(A_1 A_2^c A_3^c)}{\mathcal{P}(A_3^c)} = \mathcal{P}(A_1^c A_2 | A_3^c) = \frac{\mathcal{P}(A_1^c A_2 A_3^c)}{\mathcal{P}(A_3^c)} = \frac{1}{2}$ ，即是否换门都可以。

然而，主持人选择哪一扇门这一动作本身就可以带来信息。记 S_i 为主持人翻开某扇门的概率。如果奖品藏在第一扇门内，那么 $\mathcal{P}(S_3 | A_1 A_2^c A_3^c) = \frac{1}{2}$ ，而 $\mathcal{P}(S_3 | A_1^c A_2 A_3^c) = 1$ ，也就是说，当你选择了第一扇门，而奖品在第二扇门里面，那么主持人一定会选择打开第二扇门。根据贝叶斯公式：

$$\begin{aligned} \mathcal{P}(A_1 A_2^c | S_3 A_3^c) &= \frac{\mathcal{P}(S_3 | A_1 A_2^c A_3^c) \cdot \mathcal{P}(A_1 A_2^c | A_3^c)}{\mathcal{P}(S_3 | A_1 A_2^c A_3^c) \cdot \mathcal{P}(A_1 A_2^c | A_3^c) + \mathcal{P}(S_3 | A_1^c A_2 A_3^c) \cdot \mathcal{P}(A_1^c A_2 | A_3^c)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}} = \frac{1}{3} \end{aligned}$$

同理 $\mathcal{P}(A_1^c A_2 A_3^c | S_3) = \frac{2}{3}$ ，因而考虑到主持人选择打开第三扇门这一行为，换第二扇门是最好的选择。

然而很多时候，两个事件发生与否并没有什么关联，因而使用条件概率并不能获得更多的信息。比如尽管从个人经验来看，我每次看国足国足都会输球，然而实际上我是否看国足的比赛与国足比赛是否能赢球，并没有任何关联，因而不能使用我是否看过去比赛去预测国足比赛是否能赢球，否则我可以通过足彩立即实现财务自由。在统计上，我们经常需要假设两类事件没有什么关联，才能简化并进行分析。比如当我们做抛硬币实验时，我们就潜在假设了每次抛出硬币与其他次抛出硬币的结果是没有任何关联的。这就引出了「统计独立性」的概念。

定义 5. (统计独立性) 如果两个事件 A 和 B 满足：

$$\mathcal{P}(A \cap B) = \mathcal{P}(A) \cdot \mathcal{P}(B)$$

那么我们称事件 A 和 B 为独立事件。

定理 7. 如果 A 和 B 为独立事件，那么以下事件对也为独立事件：

1. A 和 B^c
2. A^c 和 B
3. A^c 和 B^c 。

注意当我们考虑多于两个事件时，以上定义并不能进行扩展。

例 11. 现在抛掷两枚骰子，我们关心如下三个事件：

$$A = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), \}$$

$$B = \{\text{两枚骰子之和介于 7 和 10 之间}\}$$

$$C = \{\text{两枚骰子之和为 2, 7 或者 8}\}$$

可以计算得到: $\mathcal{P}(A) = \frac{1}{6}$, $\mathcal{P}(B) = \frac{1}{2}$, $\mathcal{P}(C) = \frac{1}{3}$, 而:

$$\begin{aligned}\mathcal{P}(A \cap B \cap C) &= \mathcal{P}(\{(4, 4)\}) \\ &= \frac{1}{36} \\ &= \mathcal{P}(A) \cdot \mathcal{P}(B) \cdot \mathcal{P}(C)\end{aligned}$$

然而:

$$\mathcal{P}(B \cap C) = \frac{11}{36} \neq \mathcal{P}(B) \cdot \mathcal{P}(C)$$

因而事件 B 和 C 并不独立。

为了更好的定义多于两个事件时的独立, 我们使用如下定义:

定义 6. 我们称一系列事件 A_1, A_2, \dots, A_n 为**相互独立的** (**mutually independent or jointly independent**), 如果对于任意的子列 $A_{i_1}, A_{i_2}, \dots, A_{i_k}$, 有:

$$\mathcal{P}\left(\bigcap_{j=1}^k A_{i_j}\right) = \prod_{j=1}^k \mathcal{P}(A_{i_j})$$

习题

练习 1. 整数集 \mathbb{Z} 是可数集还是不可数集? 有理数集 \mathbb{Q} 是可数集还是不可数集? 开区间 $(0, 1)$ 是可数集还是不可数集?

练习 2. 思考题: 标准篮球的直径为 24.6cm , 而标准篮筐的直径为 45cm , 如果篮球垂直入框, 中心落点均匀的落在篮筐内, 请问投出空心球的概率有多大? 如果篮球以 60° 角入框呢? 篮球以多大的角度入框则不可能投出空心球?

练习 3. 试用交、并、补三个运算表示 $(E \triangle F)^c$ 。

练习 4.

1. 上例中, 包含 $\{\clubsuit\}$ 的最小 σ -代数是?
2. 现在抛一枚骰子, 则结果的样本空间为 $\Omega_4 = \{1, 2, 3, 4, 5, 6\}$, 那么包含所有单个样本点的 σ -代数 \mathcal{F}_4 中有多少个事件?

练习 5. 百年一遇的自然灾害 (即每年发生的概率为 $p = 1\%$) 10 年期间至少发生 1 次的概率是多少? (假设这种自然灾害每年发生与否是独立的)

练习 6. 七个人玩桌游「炸碉堡」, 经过第一轮投票, 已知第一轮三个行动中有一个是坏人, 剩下的四个人中也有一个是坏人。如果在第二轮中由你选择三个人作为行动人, 你的目标是尽可能的选出三个好人。你有如下三个策略:

1. 从三个人中选一个, 另外四个人中选两个

2. 从三个人中选两个, 另外四个人中选一个
3. 完全从四个人中选出新的三个

请问以上三个策略中, 哪一个策略选出三个好人的概率最高?

练习 7. 给定任意一个连续分布函数 F 及其定义的概率函数 P , \mathbb{R} 上的单点集 $\{a, a \in \mathbb{R}\}$ 的概率 $P(\{a, a \in \mathbb{R}\})$ 是多少? 根据概率函数的性质, \mathbb{R} 上的任意可数集的概率 $P(\{a_i, i = 1, 2, \dots, a_i \in \mathbb{R}\})$ 是多少? 所以概率为 0 的事件一定是不可能事件么?

练习 8. 试证明定理 (4)。

练习 9. 请给出两个事件至少有一个发生的概率, $\mathcal{P}(A \cup B)$ 的一个上界和一个下界。

练习 10. 试证明以上定理。

参考文献

- [1] Ash, R.B., Doleans-Dade, C., 2000. Probability and measure theory. Academic Press.
- [2] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.
- [3] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.
- [4] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.