3 Classical Probability 古典概率模型

归根结底，概率就是量化的生活常识

> 真是耐人寻味，一门以赌博游戏为起点的学科理应成为人类知识的最重要研究对象。
>
> ***It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.***
>
> —— 皮埃尔-西蒙·拉普拉斯 (Pierre-Simon Laplace) | 法国著名天文学家和数学家 | 1749 ~ 1827

- ◄ `numpy.array()` 构造一维序列，严格来说不是行向量
- ◄ `numpy.cumsum()` 计算累计求和
- ◄ `numpy.linspace()` 在指定的间隔内，返回固定步长的数据
- ◄ `numpy.random.gauss()` 产生服从正态分布的随机数
- ◄ `numpy.random.randint()` 产生随机整数
- ◄ `numpy.random.seed()` 确定随机数种子
- ◄ `numpy.random.shuffle()` 将序列的所有元素重新随机排序
- ◄ `numpy.random.uniform()` 产生服从均匀分布的随机数

3.1 无处不在的概率

概率的研究和应用深刻影响着人类科学发展进程，这里介绍孟德尔和道尔顿两个例子。

孟德尔的豌豆实验

孟德尔 (Gregor Mendel, 1822 ~ 1884) 之前，生物遗传机制主要是基于猜测，而不是试验。

在修道院蔬菜园里，孟德尔对不同豌豆品种进行了大量异花授粉试验。比如，孟德尔把纯种圆粒豌豆 和纯种皱粒豌豆 杂交，观察到产生的后代豌豆都是圆粒 ，如图 1 所示。

实际上，决定皱粒 的基因没有被呈现出来，因为决定皱粒 的基因相对于圆粒 基因来讲为隐形。

如图 1 所示，当第一代杂交圆粒豌豆 自花传粉或者彼此交叉传粉，它们的后代籽粒显示出 3:1 的固定比例，即 3/4 的圆粒 和 1/4 的皱粒 。

从精确的 3:1 的比例来看，孟德尔不仅仅推断出基因中离散遗传单位的存在，而且意识到这些离散的遗传单位在豌豆中成对出现，并且在形成配子的过程中分离。3:1 的比例背后的数学原理就是本章要介绍的古典概率模型。

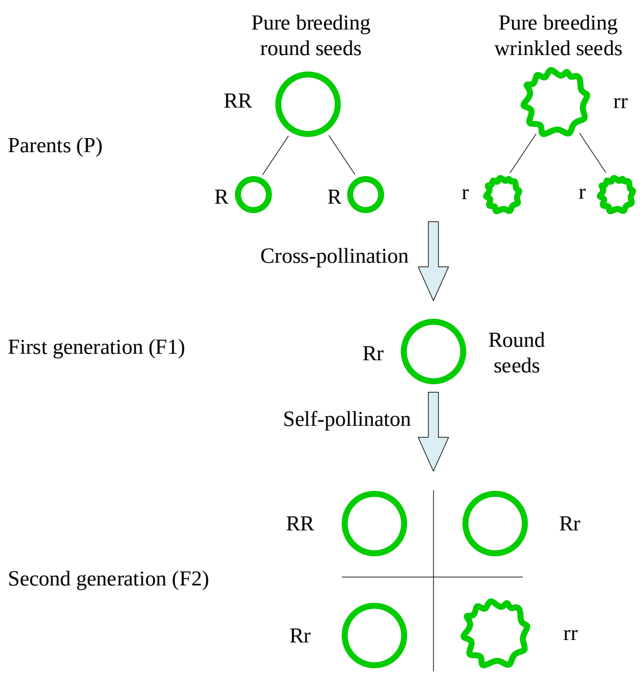


图 1. 孟德尔的豌豆试验

道尔顿发现红绿色盲

18 世纪英国著名的化学家**道尔顿** (John Dalton, 1766 ~ 1844) 偶然发现红绿色盲。道尔顿给母亲选了一双“棕灰色”的袜子作为圣诞礼物。但是，母亲对袜子的颜色略有难色，她觉得“樱桃红色”过于艳丽。

道尔顿十分疑惑，他问了家里的亲戚，发现只有弟弟和自己对袜子颜色意见一致。道尔顿意识到红绿色盲必然通过某种方式遗传。

现代人已经研究清楚，红绿色盲的遗传方式是 X 连锁隐性遗传。男性 ♂ 仅有一条 X 染色体，因此只需一个色盲基因就表现出色盲。

女性 ♀ 有两条 X 染色体，因此需有一对色盲等位基因，才会表现异常。而只有一个致病基因的女性 ♀ 只是红绿色盲基因的携带者，个体表现正常。

下面，我们从概率的角度分几种情况来思考红绿色盲的遗传规律。

情况 A

一个女性 ♀ 红绿色盲患者和一个正常男性 ♂ 生育。后代中，儿子 ♂ 都是红绿色盲；女儿 ♀ 虽表现正常，但从母亲 ♀ 获得一个红绿色盲基因，因此女儿 ♀ 都是红绿色盲基因的携带者。

不考虑性别的话，后代中发病可能性为 50%。这个可能性就是**概率** (probability)。它和生男、生女的概率一致。

给定后代为男性 ♂，发病比例为 100%。给定后代为女性 ♀，发病比例为 0%，但是携带红绿色盲基因的比例为 100%。反过来，给定后代发病这个条件，可以判定后代 100%为男性 ♂。这就是本章后文要介绍的**条件概率** (conditional probability)。

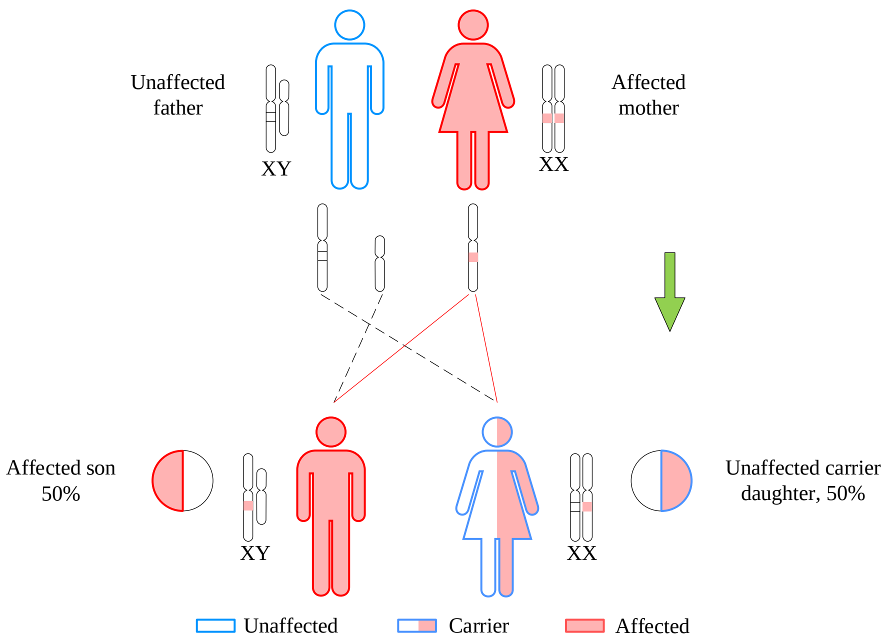


图 2. 红绿色盲基因遗传机制，情况 A

情况 B

一个女性♀红绿色盲基因携带者和一个正常男性♂生育。后代中，整体考虑，后代患病的概率为 25%。

其中，儿子♂中，50%概率为正常，50%概率为红绿色盲。女儿都不是色盲，但有 50%概率是色盲基因的携带者。这些数值也都是条件概率。

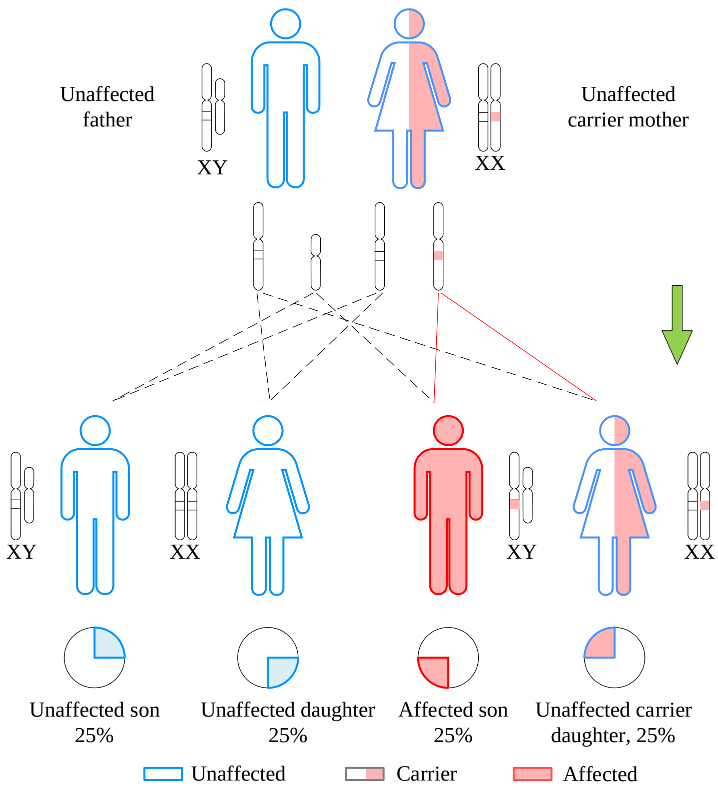


图 3. 红绿色盲基因遗传机制，情况 B

情况 C

一个女性♀红绿色盲基因的携带者和一个男性♂红绿色盲患者生育。整体考虑来看，不分男女的话，后代发病的概率为 50%。

其中，儿子♂50%概率正常，50%的概率为红绿色盲。女儿♀有 50%概率为红绿色盲，50%概率是色盲基因的携带者。

换一个条件，如果已知后代为红绿色盲患者，后代 50%概率为男性♂，50%概率为女性♀。

除了以上三种情况，请大家思考还有哪些组合情况。

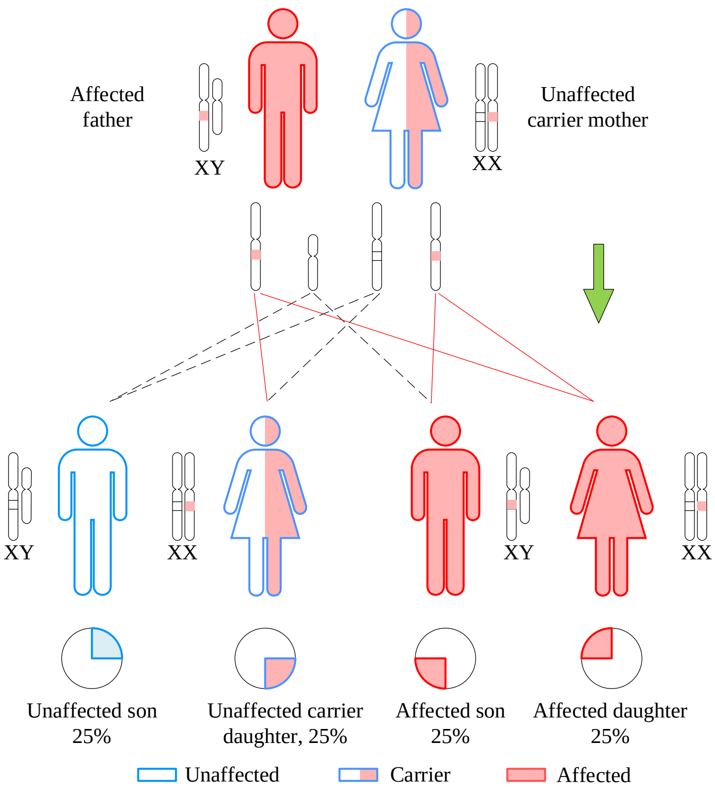


图 4. 红绿色盲基因遗传机制，情况 C

建议大家学完本章所有内容之后，再回头琢磨孟德尔和道尔顿这两个例子。

3.2 古典概率：离散均匀概率律

概率模型是对不确定现象的数学描述。本章的核心是古典概型。古典概型，也叫**等概率模型** (equiprobability)，是最经典的一种概率模型。古典模型中基本事件为有限个，并且每个基本事件为等可能。古典概率论广泛应用集合运算，本节一边讲解概率论，一边回顾集合相关知识。

给定一个随机试验，所有的结果构成的集合为**样本空间** (sample space) Ω。样本空间 Ω 中的每一个元素为一个**样本** (sample)。不同的随机试验有各自的样本空间。样本空间作为集合，也可以划分成不同**子集** (subset)。

概率

整个样本空间 Ω 的概率为 1，即：

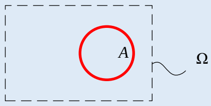
$$\Pr(\Omega) = 1 \tag{1}$$

样本空间概率为 1，从这个视角来看，本书后续内容似乎都围绕着如何将 1“切片切块”、“切丝切条”。

⚠ 注意，本书表达概率的符号 Pr 为正体。再次请大家一定注意，不同试验的样本空间 Ω 不同。

给定样本空间 Ω 的一个**事件** (event) A，Pr(A) 为**事件 A 发生的概率** (the probability of event A occurring 或 probability of A)，Pr(A) 满足：

$$\underbrace{\Pr\left(\overbrace{A}^{\text{Event}}\right)}_{\text{Probability}} \geq 0 \tag{2}$$



大家看到任何概率值时一定要注意，它的样本空间是什么。

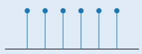
空集 ∅ 不包含任何样本点，也称作**不可能事件** (impossible event)，因此对应的概率为 0：

$$\Pr(\varnothing) = 0 \tag{3}$$

等可能

设样本空间 Ω 由 n 个**等可能事件** (equally likely events 或 events with equal probability) 构成，事件 A 的概率为：

$$\Pr(A) = \frac{n_A}{n} \tag{4}$$



其中，n_A 为含于事件 A 的试验结果数量。这实际上便是等概率模型。

以鸢尾花数据为例

举个例子，从 150 (n) 个鸢尾花数据中取一个样本点，任何一个样本被取到的概率为 1/150 (1/n)。

再举个例子，鸢尾花数据集的 150 个样本均分为 3 类——setosa (C_1)、versicolour (C_2)、virginica (C_3)。如图 5 所示，从 150 个样本中取出任一样本，样本标签为 C_1、C_2、C_3 对应的概率相同，都是：

$$\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \frac{50}{150} = \frac{1}{3} \tag{5}$$

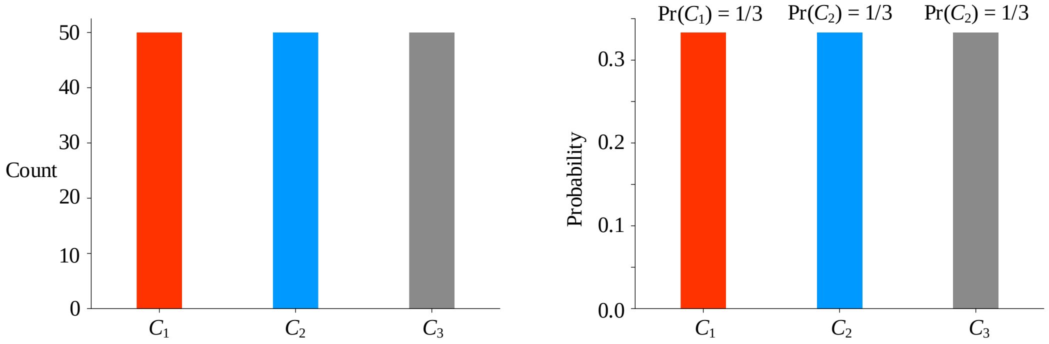


图 5. 鸢尾花 150 个样本数据均分为三类

抛一枚硬币

再举个例子，抛一枚硬币，1 代表正面，0 代表反面。

抛一枚硬币的可能结果的样本空间为：

$$\Omega = \{0,1\} \tag{6}$$

假设硬币质地均匀，获得正面和反面的概率相同，均为 1/2，即：

$$\Pr(0) = \Pr(1) = \frac{1}{2} \tag{7}$$

把 {0, 1} 标记在数轴上，将对应的概率以火柴梗图可视化，我们便得到图 6。

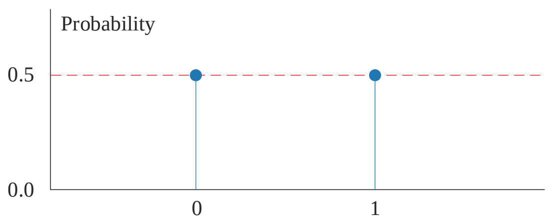


图 6. 抛一枚硬币结果和对应的理论概率值

图 7 所示为反复抛一枚硬币，正面 (1)、反面 (0) 平均值随试验次数变化。可以发现平均结果不断靠近 1/2，也就是说正反面出现的概率几乎相同。

从另外一个角度，(7) 给出的是用古典概率模型 (等可能事件和枚举法) 得出的理论概率 (theoretical probability)。而图 7 是采用试验得到的统计结果，印证了概率模型结果。根据大量的、重复的统计试验结果计算随机事件中各种可能发生结果的概率，称为试验概率 (experimental probability)。概率和统计的关系可见一斑。

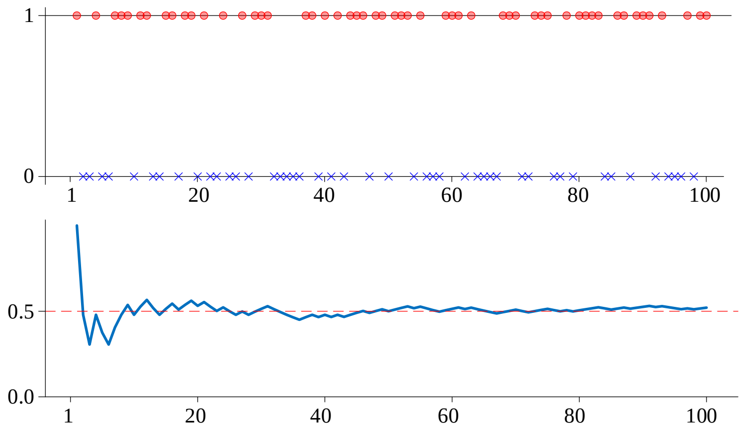


图 7. 抛硬币 100 次试验结果变化

掷色子

如图 8 所示，掷一枚色子试验得到的所有结果集合为：

$$\Omega = \{1,2,3,4,5,6\} \tag{8}$$

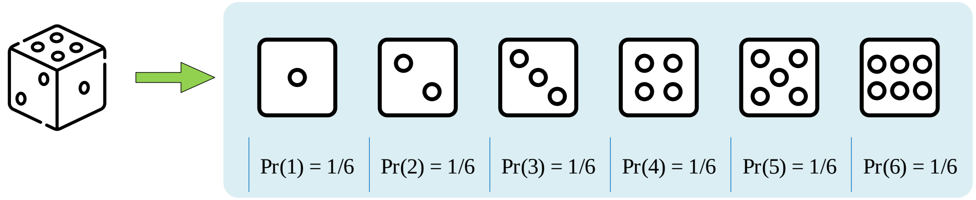


图 8. 投色子试验

试验中，假设获得每一种点数的可能性相同。掷一枚色子共 6 种结果，每种结果对应的概率为：

$$\Pr(1) = \Pr(2) = \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = \frac{1}{6} \tag{9}$$

同样用火柴梗图把上述结果画出来，得到图 9。这也是抛一枚色子得到不同点数对应概率的理论值。

然而实际情况可能并非如此。想象一种特殊情况，某一枚特殊的色子，它的质地不均匀，可能产生点数 6 的概率略高于其他点数。要想估算不同结果的概率值，一般只能通过试验。

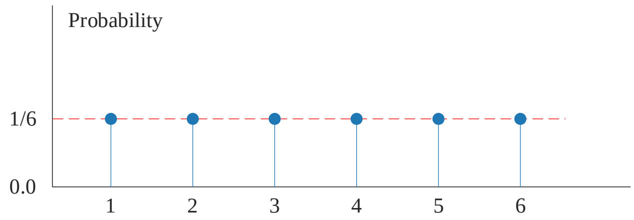


图 9. 抛一枚色子结果和对应的理论概率值

抛两枚硬币

下面看两个稍复杂的例子——每次抛两枚硬币。

比如说，如果第一枚硬币正面、第二枚硬币反面，结果记做 (1, 0)。这样，样本空间由以下 4 个点构成：

$$\Omega = \begin{Bmatrix} (0,0) & (0,1) \\ (1,0) & (1,1) \end{Bmatrix} \tag{10}$$

图 10 (a) 所示为用二维坐标系展示试验结果。图中横轴代表第一枚硬币点数，纵轴为第二枚硬币对应点数。

假设，两枚硬币质地均匀，抛一枚硬币获得正、反面的概率均为 1/2。而抛两枚硬币对应结果的概率如图 10 (b) 所示。

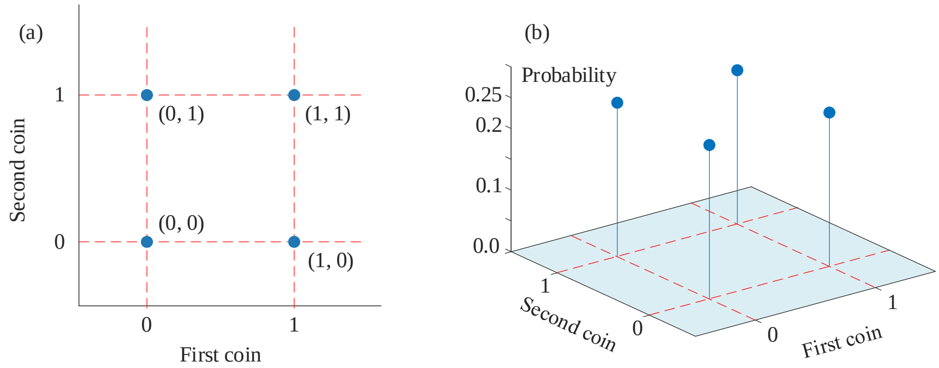


图 10. 抛两枚硬币结果和对应的理论概率值

抛两枚色子

每次抛 2 枚色子，样本空间 Ω 的等可能试验结果数量为 6 × 6：

$$\Omega = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (2,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (3,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (4,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \end{Bmatrix} \tag{11}$$

图 11 (a) 所示为上述试验的样本空间。图 11 (b) 中，每个试验结果对应的概率均为 1/36。

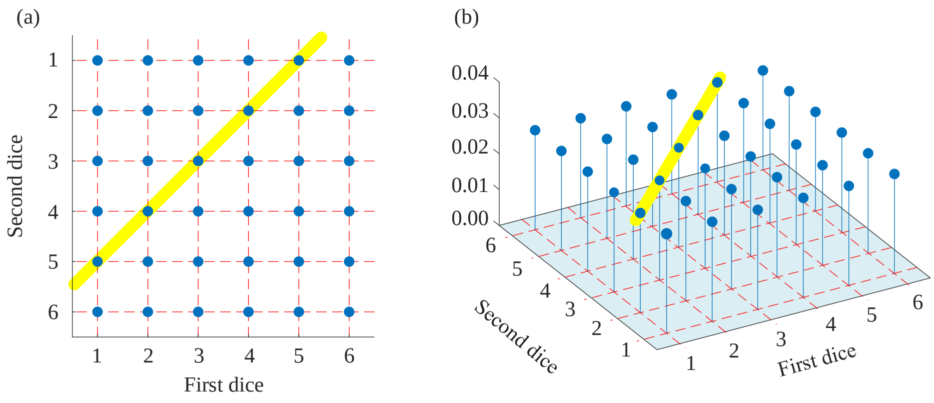


图 11. 抛两枚色子结果和对应的理论概率值

抛两枚色子：点数之和为 6

下面，我们看一种特殊情况。如图 12 所示，这 5 种结果为 1 + 5、2 + 4、3 + 3、4 + 2、5 + 1。该事件对应概率为：

$$\Pr(\text{sum}=6)=\frac{5}{6\times 6}\approx 0.1389 \tag{12}$$

图 11 (a) 中黄色背景所示样本也代表抛两枚色子点数之和为 6 的事件。

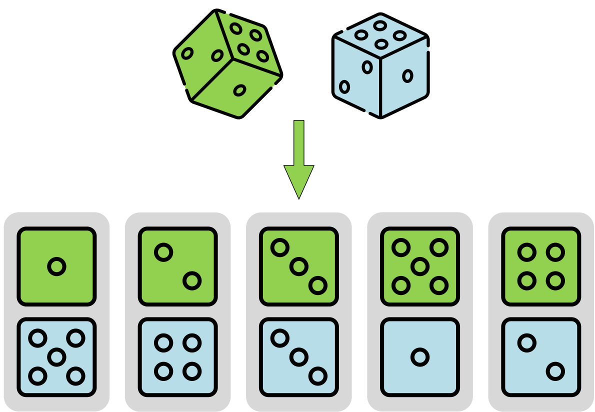


图 12. 投两个色子，点数之和为 6

编写代码进行 10,000,000 次试验，累计“点数之和为 6”事件发生次数，并且计算该事件当前概率。图 13 所示“点数之和为 6”事件概率随抛掷次数变化曲线。

比较 (12) 和图 13，通过古典概率模型得到的结论和试验统计结果相互印证。

⚠ 注意，图 13 横轴为对数刻度。《数学要素》第 12 章介绍过对数刻度，大家可以回顾。

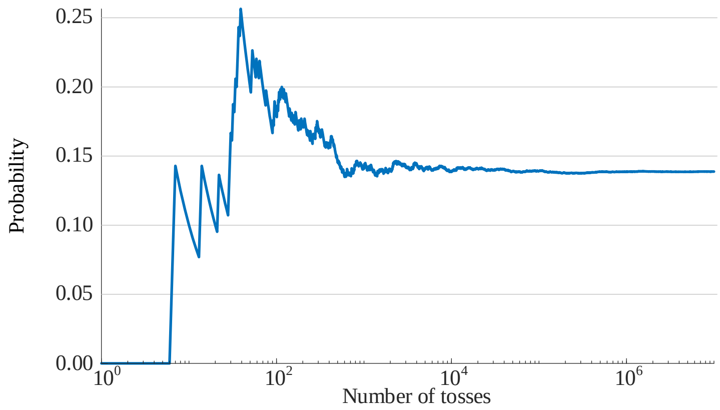


图 13. “色子点数之和为 6”事件概率随抛掷次数变化

代码 Bk5_Ch03_01.py 模拟抛色子试验并绘制图 7。

抛两枚色子：点数之和的样本空间

接着上一个例子，如果我们对抛两枚色子点数之和感兴趣。首先要知道这个事件的样本空间。如图 14 所示，彩色等高线对应不同两枚色子点数之和，由此得到两个色子点数之和的样本空间为 {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}。

而等高线线上灰色点 ● 的数量代表满足条件的色子点数情况。查点的数量，再除 36 (= 6 × 6) 便得到样本概率值，如图 14 (b) 所示。观察图 14 (b)，大家可以发现结果非等概率；但是，这些概率值也是通过等概率模型推导得到。

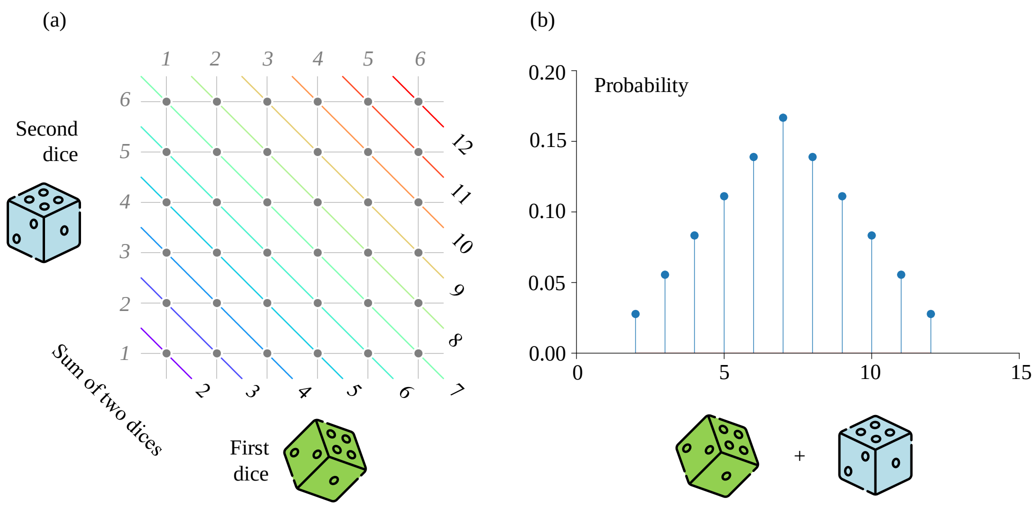


图 14. 两个色子点数之和

更多“花样”

接着上面抛两枚色子的试验，我们再玩更多“花样”！

比如，抛两枚色子，我们可以只考虑第一只色子的点数、第一只色子点数平方值，也可以计算两个色子的点数平均值、乘积、商、差、差的平方等等。

这些不同的花式玩法至少告诉我们一下几层信息：

- 抛两枚色子，第一枚色子和第二枚色子的结果可以独立讨论；换个视角，第一、二枚色子点数相互不影响；
- 第一枚和第二枚色子的点数结果还可以继续运算；
- 用文字描述这些结果太麻烦了，我们需要将它们代数化！比如，我们定义第一个色子结果为 X_1，第二个色子点数为 X_2，这便是下一章要探讨的随机变量 (random variable)。表 1 所示为基于抛两枚色子试验结果的更多花式玩法。请大家试着找到每种运算的样本空间，并试着计算每个样本对应的概率值。我们将在下一章揭晓答案。
- 第四：显然表 1 中各种花式玩法的样本空间 Ω 是变化的。但是，样本空间中所有样本的概率之和也都是 1。

表 1. 基于抛两枚色子试验结果的更多花式玩法

随机变量	描述	例子											
X_1	第一个色子点数	1	2	3	4	5	6	1	2	3	4	5	6
X_2	第二个色子点数	1	1	1	1	1	1	2	2	2	2	2	2
$Y = X_1$	只考虑第一个色子点数	1	2	3	4	5	6	1	2	3	4	5	6
$Y = X_1^2$	第一个色子点数平方	1	4	9	16	25	36	1	4	9	16	25	36
$Y = X_1 + X_2$	点数之和	2	3	4	5	6	7	3	4	5	6	7	8
$Y = \frac{X_1 + X_2}{2}$	点数平均值	1	1.5	2	2.5	3	3.5	1.5	2	2.5	3	3.5	4
$Y = \frac{X_1 + X_2 - 7}{2}$	中心化点数之和，再求平均	−2.5	−2	−1.5	−1	−0.5	0	−2	−1.5	−1	−0.5	0	0.5
$Y = X_1 X_2$	点数之积	1	2	3	4	5	6	2	4	6	8	10	12
$Y = \frac{X_1}{X_2}$	点数之商	1	2	3	4	5	6	0.5	1	1.5	2	2.5	3
$Y = X_1 - X_2$	点数之差	0	1	2	3	4	5	-1	0	1	2	3	4
$Y = \lvert X_1 - X_2 \rvert$	点数之差的绝对值	0	1	2	3	4	5	1	0	1	2	3	4
$Y = (X_1 - 3.5)^2 + (X_2 - 3.5)^2$	中心化点数平方和	12.5	8.5	6.5	6.5	8.5	12.5	8.5	4.5	2.5	2.5	4.5	8.5

抛三枚色子

为了大家习惯“多元”思维，我们再进一步将一次抛掷色子的数量提高至三枚。第一枚点数定义为 X_1，第二枚点数 X_2，第三枚点数 X_3。

图 15 (a) 所示为抛三枚色子点数的样本空间，这显然是个三维空间。比如，坐标点 (3, 3, 3) 代表三枚色子的点数都是 3。

图 15 (a) 这个样本空间有 216 (= 6 × 6 × 6)个样本。假设这三个色子质量均匀，获得每个点数为等概率，则图 15 (a) 中每个样本对应的概率为 1/216。

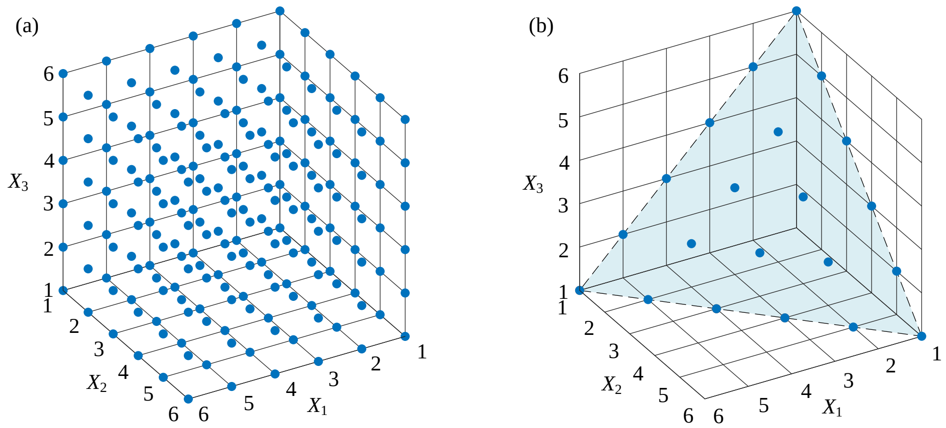


图 15. 抛三枚色子点数的样本空间

定义事件 A 为三枚色子的点数之和为 8，即 $X_1 + X_2 + X_3 = 8$。事件 A 对应的样本集合如所图 15 (b) 所示，一共有 21 个样本点，容易发现这些样本在同一个空间斜面上。相对图 15 (a) 这个样本空间，事件 A 的概率为 21/216。

大家可能已经发现，实际上，我们可以用水平面来可视化事件 A 的样本集合。如图 16 所示，将散点可以投影在平面上得到图 16 (b)。能够完成这种投影的原因是因为 $X_1 + X_2 + X_3 = 8$ 这个等式关系。

这种投影思路将会用到本书后续要介绍的多项分布 (第 5 章)、Beta 分布 (第 7 章)。

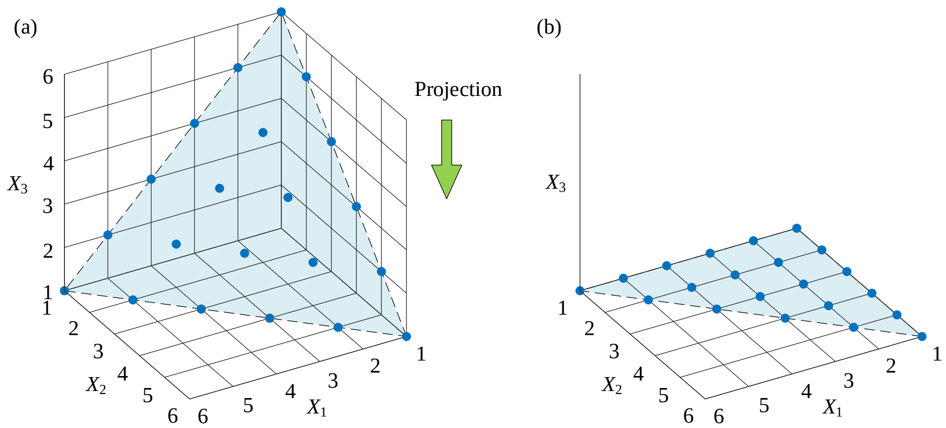


图 16. 将事件 A 的样本点投影到平面上

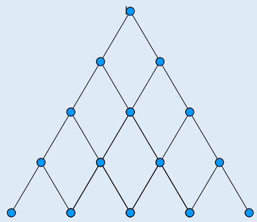
3.3 回顾：杨辉三角和概率

杨辉三角

《数学要素》第 20 章介绍过杨辉三角和古典概率模型的联系，本节稍作回顾。

杨辉三角又叫**帕斯卡三角** (Pascal's triangle)，是二项式系数的一种写法。$(a+b)^n$ 展开后，按单项 a 的次数从高到低排列得到：

$$\begin{aligned}(a+b)^0 &= 1\\(a+b)^1 &= a+b\\(a+b)^2 &= a^2+2ab+b^2\\(a+b)^3 &= a^3+3a^2b+3ab^2+b^3\\(a+b)^4 &= a^4+4a^3b+6a^2b^2+4ab^3+b^4\end{aligned} \tag{13}$$



其中，a 和 b 均不为 0。单项式系数可以用组合数写成图 17。

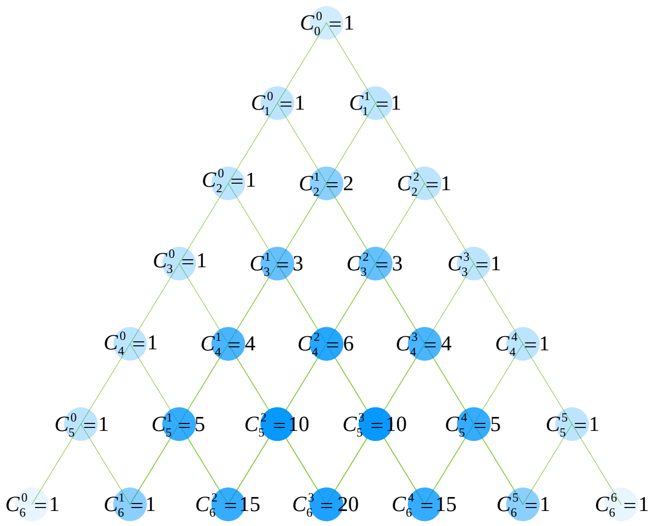


图 17. 用组合数来写杨辉三角

抛硬币

把二项式展开用在理解抛硬币的试验。$(a+b)^n$ 中 n 代表一次抛掷中硬币数量，a 可以理解为“硬币正面朝上”对应概率，b 为“硬币反面朝上”对应概率。如果硬币质地均匀，$a = b = 1/2$。

举个例子，每次抛 10 (n) 枚硬币，正好出现 6 次正面对应概率为：

$$\Pr(\text{heads}=6)=C_{10}^{6}\frac{1}{2^{10}}=\frac{210}{1024}=\frac{210}{1024}\approx 0.20508 \tag{14}$$

每次抛 10 枚硬币，至少出现 6 次正面的概率为：

$$\Pr(\text{heads}\geq 6)=\frac{C_{10}^{6}+C_{10}^{7}+C_{10}^{8}+C_{10}^{9}+C_{10}^{10}}{2^{10}}=\frac{210+120+45+10+1}{1024}=\frac{386}{1024}\approx 0.37695 \tag{15}$$

编写代码，一共抛 10000 次，每次抛 10 枚硬币。分别累计“正好出现 6 次正面”、“至少出现 6 次正面”两个事件的次数，并且计算两个事件当前概率。图 18 所示两事件概率随抛掷次数变化曲线。

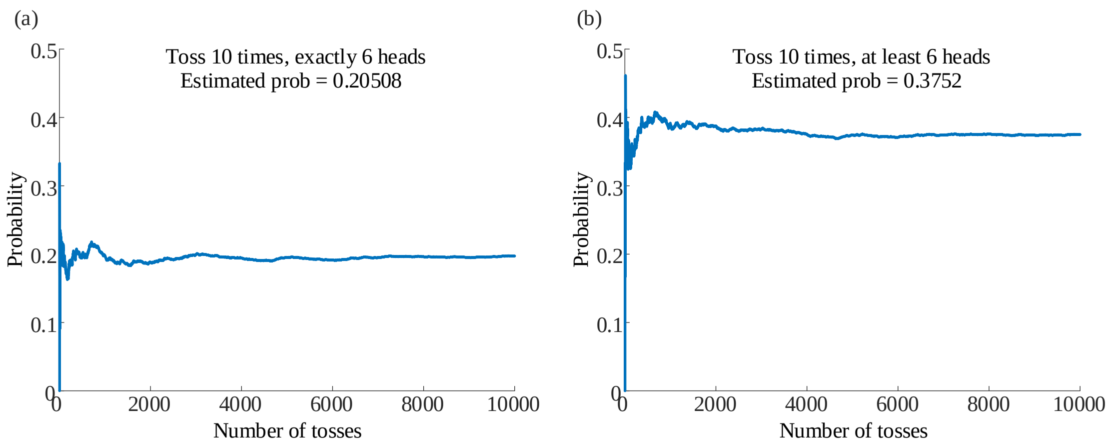


图 18. 试验概率随抛掷次数变化：a) 正好出现 6 次正面；b) 至少出现 6 次正面

Bk5_Ch03_02.py 完成上述两个试验并绘制图 18。

回忆二叉树

《数学要素》第 20 章还介绍过杨辉三角和二叉树的联系，如图 19 所示。站在中间节点处，向上走、还是向下走对应的概率便分别对应“硬币正面朝上”、“硬币反面朝上”概率。

假设，向上走、向下走的概率均为 1/2。图 19 右侧的直方图展示了两组数，分别是达到终点不同节点的路径数量、概率值。请大家回忆如何用组合数计算这些概率值。

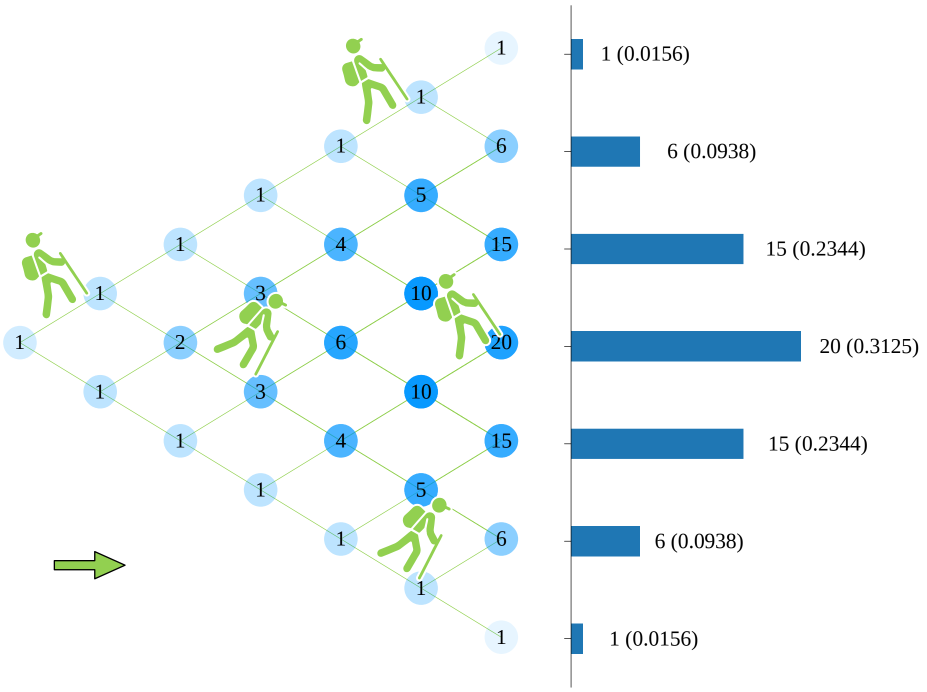


图 19. 杨辉三角逆时针旋转 90 度得到一个二叉树，图片基于《数学要素》第 20 章

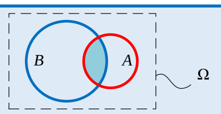
3.4 事件之间的关系：集合运算

积事件

事件 A 与事件 B 为样本空间 Ω 中的两个事件， $A\cap B$ 代表 A 和 B 的**积事件** (the intersection of events A and B)，指的是某次试验时，事件 A 和事件 B 同时发生。

$\Pr(A\cap B)$ 代表 A 和 B **积事件概率** (probability of the intersection of events A and B 或 joint probability of A and B)。$\Pr(A\cap B)$ 也叫做 A 和 B **联合概率** (joint probability)。$\Pr(A\cap B)$ 也常记做 $\Pr(A,B)$：

$$\Pr\Big(\underset{\text{Joint}}{A\cap B}\Big)=\Pr\Big(\underset{\text{Joint}}{A,B}\Big) \tag{16}$$

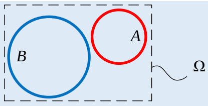


互斥

如果事件 A 与事件 B 为两者交集为空 $A\cap B=\varnothing$，则称**事件 A 和事件 B 互斥** (events A and B are disjoint)，或称 **A 和 B 互不相容** (two events are mutually exclusive)。

白话说，事件 A 与事件 B 不可能同时发生，也就是说 $\Pr(A \cap B)$ 为 0：

$$\underbrace{A \cap B}_{\text{Joint}} = \varnothing \quad \Rightarrow \quad \Pr\left(\underbrace{A \cap B}_{\text{Joint}}\right) = \Pr\left(\underbrace{A, B}_{\text{Joint}}\right) = 0 \tag{17}$$

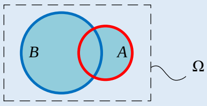


和事件

事件 $A \cup B$ 为 A 和 B 的**和事件** (union of events A and B)。具体来说，当事件 A 和事件 B 至少有一个发生时，事件 $A \cup B$ 发生。$\Pr(A \cup B)$ 代表事件 A 和 B **和事件概率** (probability of the union of events A and B 或 probability of A or B)。

$\Pr(A \cup B)$ 和 $\Pr(A \cap B)$ 之间关系为：

$$\underbrace{\Pr(A \cup B)}_{\text{Union}} = \Pr(A) + \Pr(B) - \underbrace{\Pr(A \cap B)}_{\text{Joint}} \tag{18}$$



如果事件 **A 和 B 互斥** (events A and B are mutually exclusive)，即 $A \cap B = \varnothing$。对于这种特殊情况，$\Pr(A \cup B)$ 为：

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \tag{19}$$

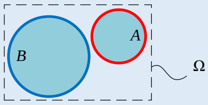
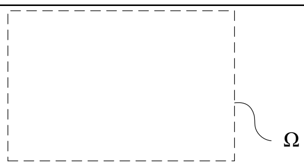
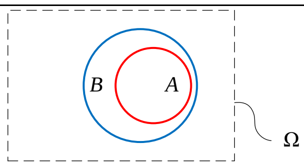
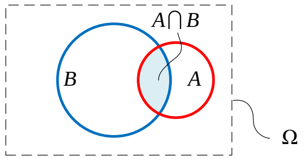
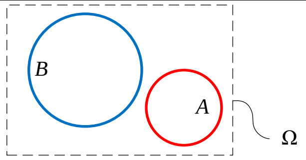
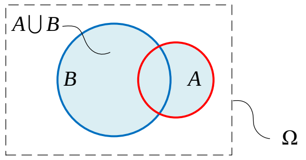
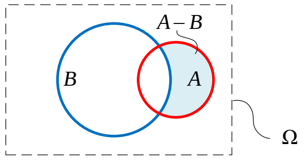
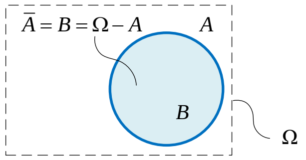


表 2 总结常见集合运算维恩图。

表 2. 常见集合运算和维恩图

符号	解释	维恩图
Ω	必然事件，即整个样本空间 (sample space)	
$\varnothing$	不可能事件，即空集 (empty set)	
$A \subset B$	事件 B 包含事件 A (event A is a subset of event B) 即，事件 A 发生，事件 B 必然发生	

$A\cap B$	事件 A 和事件 B 的积事件 (the intersection of events A and B) 即，某次试验时，当事件 A 和事件 B 同时发生时，事件 $A\cap B$ 发生	
$A\cap B=\varnothing$	事件 A 和事件 B 互斥 (events A and B are disjoint)，两个事件互不相容 (two events are mutually exclusive) 即，事件 A 和事件 B 不能同时发生	
$A\cup B$	事件 A 和事件 B 的和事件 (the union of events A and B) 即，当事件 A 和事件 B 至少有一个发生时，事件 $A\cup B$ 发生	
$A-B$	事件 A 与事件 B 的差事件 (the difference between two events A and B) 即，事件 A 发生、事件 B 不发生， $A-B$ 发生	
$A\cup B=\Omega$ 且 $A\cap B=\varnothing$ 也可以记做 $\bar{A}=B=\Omega-A$ (complement of event A)	事件 A 与事件 B 互为逆事件 (complementary events)，对立事件 (collectively exhaustive) 即，对于任意一次试验，事件 A 和事件 B 有且仅有一个发生	

3.5 条件概率：给定部分信息做推断

条件概率 (conditional probability) 是在给定部分信息基础上对试验结果的一种推断。条件概率是机器学习、数学科学中至关重要概念，本书很多内容都是围绕条件概率展开，请大家格外留意。

三个例子

下面给出三个例子说明什么我们哪里会用到“条件概率”。

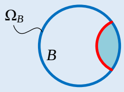
在抛两个色子试验中，事件 A 为其中一个色子点数为 5，事件 B 为点数之和为 6。给定事件 B 发生条件下，事件 A 发生的概率多少？

给定花萼长度为 5 厘米，花萼宽度为 2 厘米，根据 150 个鸢尾花样本数据，鸢尾花样本最可能是哪一类？对应的概率大概是多少？

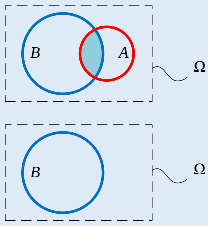
根据 150 个鸢尾花样本数据，如果花萼长度为 5 厘米，花萼宽度最可能多宽？

条件概率

A 和 B 为样本空间 Ω 中的两个事件，其中 $\Pr(B) > 0$。那么，**事件 B 发生的条件下事件 A 发生的条件概率** (conditional probability of event A occurring given B occurs 或 probability of A given B) 可以通过下式计算得到：



$$\underbrace{\Pr\left(A\middle|\overset{\text{Given }B}{B}\right)}_{\text{Conditional}} = \frac{\overbrace{\Pr(A\cap B)}^{\text{Joint}}}{\underbrace{\Pr(B)}_{\text{Marginal}}} \tag{20}$$

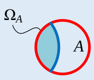


其中，$\Pr(A\cap B)$ 为 A 和 B 事件的联合概率，$\Pr(B)$ 也叫 B 事件边缘概率。

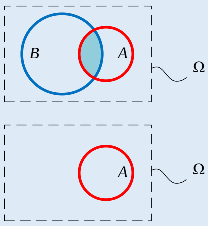
注意，我们也可以这么理解 $\Pr(A|B)$，B 实际上是新的"样本空间"Ω_B！$\Pr(A|B)$ 是在 Ω_B 中计算得到的概率值。而 $\Pr(B)$、$\Pr(A\cap B)$ 都是在 Ω 中计算得到的概率值。

而 Ω_B 是的子集 Ω，两者的联系正是 $\Pr(B)$，即 B 在 Ω 中对应的概率。$\Pr(B)$ 也可以写成"条件概率"的形式 $\Pr(B\,|\,\Omega)$。

类似地，事件 A 发生的条件下事件 B 发生的条件概率为：



$$\underbrace{\Pr\left(B\middle|\overset{\text{Given }A}{A}\right)}_{\text{Conditional}} = \frac{\overbrace{\Pr(A\cap B)}^{\text{Joint}}}{\underbrace{\Pr(A)}_{\text{Marginal}}} \tag{21}$$

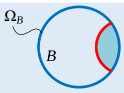


其中，$\Pr(A\cap B)$ 为 A 和 B 事件的联合概率，$\Pr(B)$ 也叫 B 事件边缘概率。

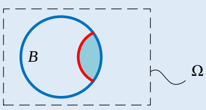
类似地，$\Pr(B|A)$ 也可以理解为 A 在新的"样本空间"Ω_A 中的概率。

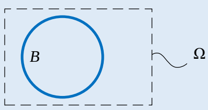
联合概率

利用 (20)，联合概率 $\Pr(A\cap B)$ 可以整理为：



$$\underbrace{\Pr(A\cap B)}_{\text{Joint}} = \underbrace{\Pr(A,B)}_{\text{Joint}} = \underbrace{\Pr(A|B)}_{\text{Conditional}}\cdot\underbrace{\Pr(B)}_{\text{Marginal}} \tag{22}$$





上式相当于“套娃”。首先在 Ω_B 中考虑 A (实际上是 $A \cap B$)，然后把在放回 Ω 中。也就是说，把 $\Pr(A \mid B)$ 写成 $\Pr(A \cap B \mid B)$ 也没问题。因为，A 只有 $A \cap B$ 这部分在 B (Ω_B) 中。

同样，$\Pr(A \cap B)$ 也可以写成：

$$\underbrace{\Pr(A \cap B)}_{\text{Joint}} = \underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(B \mid A)}_{\text{Conditional}} \underbrace{\Pr(A)}_{\text{Marginal}} \tag{23}$$

举个例子

掷一颗色子，一共有 6 种等概率结果 $\Omega = \{1, 2, 3, 4, 5, 6\}$。

事件 B 为“点数为奇数”，事件 C 为“点数小于 4”。事件 B 的概率 $\Pr(B) = 1/2$，事件 C 的概率 $\Pr(C) = 1/2$。

如图 20 所示，$B \cap C$ 事件发生的概率 $\Pr(B \cap C) = \Pr(B, C) = 1/3$。

这样的话，在事件 B (点数为奇数) 条件下，事件 C (点数小于 4) 发生的条件概率为：

$$\Pr(C \mid B) = \frac{\Pr(B \cap C)}{\Pr(B)} = \frac{\Pr(B, C)}{\Pr(B)} = \frac{1/3}{1/2} = \frac{2}{3} \tag{24}$$

图 20 也告诉我们一样的结果。请大家回顾本章最初给出孟德尔豌豆实验和道尔顿红绿色盲，计算其中的条件概率。

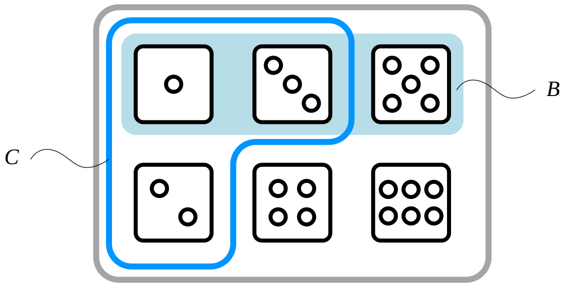


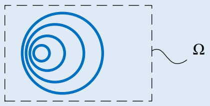
图 20. 事件 B 发生条件下事件 C 发生的条件概率

推广

(22) 可以继续推广，$A_1, A_2, \ldots, A_n$ 为 n 个事件，它们的联合概率可以展开写成一系列条件概率的乘积：

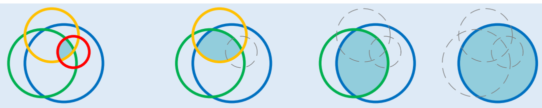
$$\begin{aligned}\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) &= \Pr(A_1, A_2, A_3, \cdots, A_{n-1}, A_n) \\ &= \Pr(A_n \mid A_1, A_2, A_3, \cdots, A_{n-1}) \Pr(A_{n-1} \mid A_1, A_2, A_3, \cdots, A_{n-2}) \cdots \Pr(A_2 \mid A_1) \Pr(A_1)\end{aligned} \tag{25}$$

这也叫做条件概率的**链式法则** (chain rule)。比如，$n = 4$ 时，上式写成为：



$$\begin{aligned}\underbrace{\Pr\left(A_1,A_2,A_3,A_4\right)}_{\text{Joint}} &= \underbrace{\Pr\left(A_4\middle|A_1,A_2,A_3\right)}_{\text{Conditional}}\cdot\underbrace{\Pr\left(A_1,A_2,A_3\right)}_{\text{Joint}}\\ &= \underbrace{\Pr\left(A_4\middle|A_1,A_2,A_3\right)}_{\text{Conditional}}\cdot\underbrace{\Pr\left(A_3\middle|A_1,A_2\right)}_{\text{Conditional}}\cdot\underbrace{\Pr\left(A_1,A_2\right)}_{\text{Joint}}\\ &= \underbrace{\Pr\left(A_4\middle|A_1,A_2,A_3\right)}_{\text{Conditional}}\cdot\underbrace{\Pr\left(A_3\middle|A_1,A_2\right)}_{\text{Conditional}}\cdot\underbrace{\Pr\left(A_2\middle|A_1\right)}_{\text{Conditional}}\Pr\left(A_1\right)\end{aligned} \tag{26}$$

大家可以把上式想成多层套娃。上式配图假设事件相互之间完全包含，这样方便理解。实际上，事件求积的过程已经将多余的部分“切掉”：



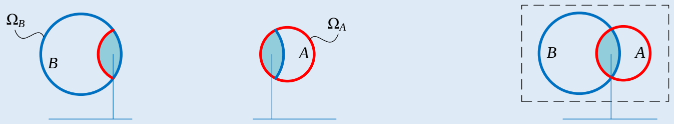
$$\left(A_1\cap A_2\cap A_3\cap A_4\right)\subset\left(A_1\cap A_2\cap A_3\right)\subset\left(A_1\cap A_2\right)\subset A_1 \tag{27}$$

3.6 贝叶斯定理：条件概率、边缘概率、联合概率关系

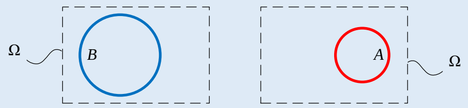
贝叶斯定理 (Bayes' theorem) 是由**托马斯·贝叶斯** (Thomas Bayes) 提出。贝叶斯定理可以说撑起机器学习、数据科学经典算法的半边天。本书后续将见缝插针地讲解贝叶斯定理和应用。

托马斯·贝叶斯 (Thomas Bayes) | 英国数学家 | 1702 ~ 1761
贝叶斯统计的开山鼻祖，以贝叶斯定理闻名于世。关键词：● 贝叶斯定理 ● 朴素贝叶斯分类 ● 贝叶斯回归 ● 贝叶斯派

贝叶斯定理描述的是两个条件概率的关系：



$$\underbrace{\Pr\left(A\middle|B\right)}_{\text{Conditional}}\underbrace{\Pr\left(B\right)}_{\text{Marginal}} = \underbrace{\Pr\left(B\middle|A\right)}_{\text{Conditional}}\underbrace{\Pr\left(A\right)}_{\text{Marginal}} = \underbrace{\Pr\left(A\cap B\right)}_{\text{Joint}} = \underbrace{\Pr\left(A,B\right)}_{\text{Joint}} \tag{28}$$



其中：

- $\Pr(A|B)$ 是指在 B 发生条件下 A 发生的**条件概率** (conditional probability)；也就是说，$\Pr(A|B)$ 的样本空间为 Ω_B；
- $\Pr(B|A)$ 是指在 A 发生条件下 B 发生的条件概率；也就是说，$\Pr(B|A)$ 的样本空间为 Ω_A；
- $\Pr(A)$ 是 A 的**边缘概率** (marginal probability)，不考虑事件 B 的因素，样本空间为 Ω；
- $\Pr(B)$ 是 B 的边缘概率，不考虑事件 A 的因素，样本空间为 Ω；
- $\Pr(A \cap B)$ 是事件 A 和 B 的联合概率，样本空间为 Ω。

图 21 给出理解贝叶斯原理的图解法。

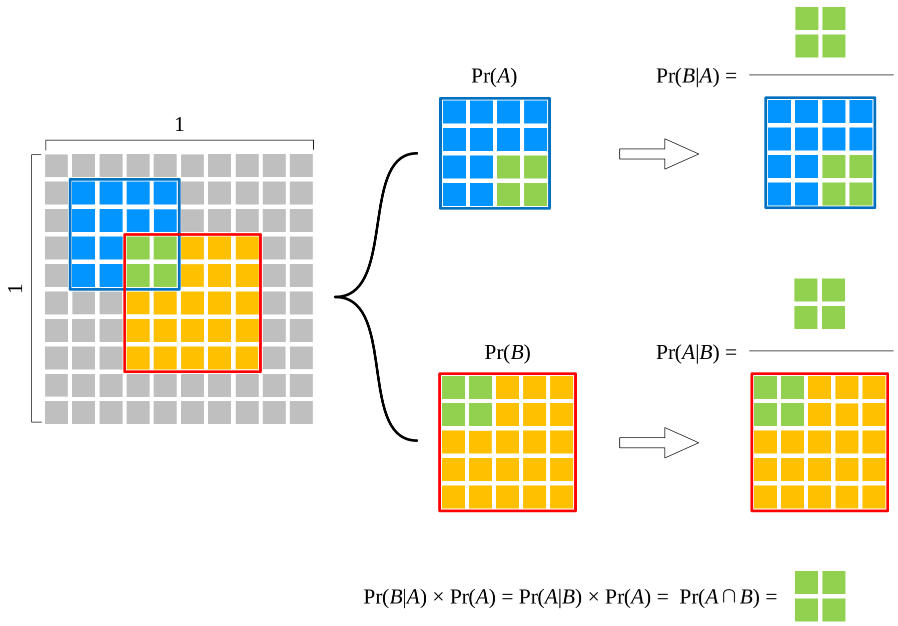


图 21. 贝叶斯原理图解

抛色子试验

现在，我们就用抛色子的试验来解释本节介绍的这几个概率计算。

根据本章前文内容，抛一枚色子可能得到 6 种结果，构成的空间全集为 $\Omega = \{1, 2, 3, 4, 5, 6\}$。6 个色子点数中的每一种结果的概率相同，$\Pr(1) = \Pr(2) = \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = 1/6$。

设"色子点数为偶数"事件为 A，因此 $A = \{2, 4, 6\}$，对应概率为 $\Pr(A) = 3/6 = 0.5$。

A 事件的补集 B 对应事件"色子点数为奇数"，$B = \{1, 3, 5\}$，事件 B 的概率为 $\Pr(B) = 1 – \Pr(A) = 0.5$。

事件 A 和 B 交集 $A \cap B$ 为空集 $\varnothing$，因此：

$$\Pr\left(A \cap B\right) = \Pr\left(A, B\right) = 0 \tag{29}$$

而 A 和 B 两者的并集 $A \cup B = \Omega$，因此对应的概率为 1：

$$\Pr(A \cup B) = 1 \tag{30}$$

C 事件被定为“色子点数小于 4”，因此 $C = \{1, 2, 3\}$，事件 C 的概率 $\Pr(C) = 0.5$。

图 23 展示的是 A、B 和 C 事件的关系。

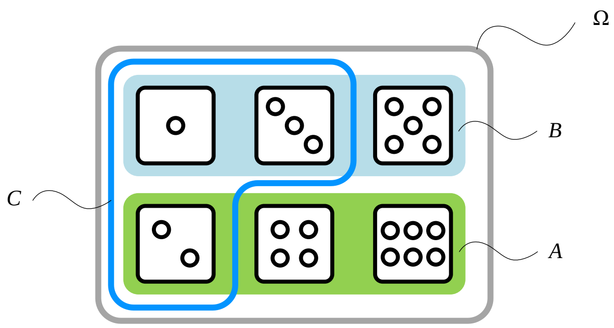


图 22. A、B、C 事件定义

如图 23 (a) 所示，事件 A 和 C 的交集 $A \cap C = \{2\}$，因此 $A \cap C$ 的概率：

$$\Pr(A \cap C) = \Pr(A, C) = \frac{1}{6} \tag{31}$$

如图 23 (b) 所示，事件 B 和 C 的交集 $B \cap C = \{1, 3\}$，因此 $B \cap C$ 的概率：

$$\Pr(B \cap C) = \Pr(B, C) = \Pr(\{1\}) + \Pr(\{3\}) = \frac{1}{3} \tag{32}$$

A 和 C 的并集 $A \cup C = \{1, 2, 3, 4, 6\}$，对应的概率为：

$$\Pr(A \cup C) = \Pr(A) + \Pr(C) - \Pr(A, C) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6} \tag{33}$$

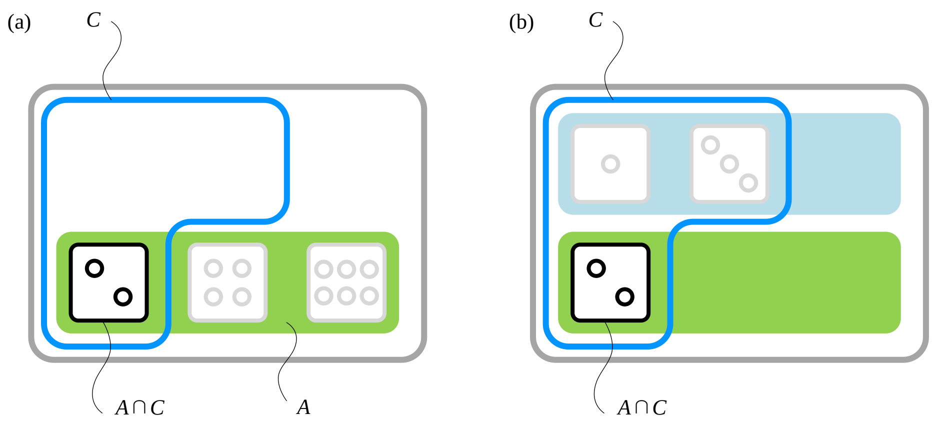


图 23. 条件概率 $\Pr(C|A)$ 和条件概率 $\Pr(A|C)$

简单来说，条件概率 $\Pr(C|A)$ 代表在 A 事件发生的条件下，C 事件发生概率。用贝叶斯公式可以求解 $\Pr(C|A)$：

$$\Pr(C|A) = \frac{\Pr(A,C)}{\Pr(A)} = \frac{1/6}{1/2} = \frac{1}{3} \tag{34}$$

类似的，在 C 事件发生的条件下，A 事件发生的条件概率 $\Pr(A|C)$ 为：

$$\Pr(A|C) = \frac{\Pr(A,C)}{\Pr(C)} = \frac{1/6}{1/2} = \frac{1}{3} \tag{35}$$

请大家自行计算图 24 所示的 $\Pr(C|B)$ 和 $\Pr(B|C)$ 这两个条件概率。

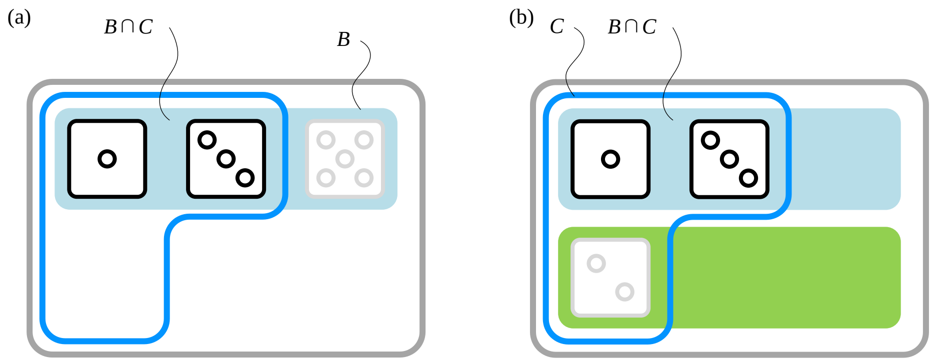
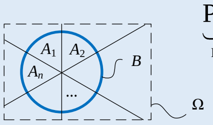


图 24. 条件概率 $\Pr(C|B)$ 和 $\Pr(B|C)$

3.7 全概率定理：穷举法

假设 $A_1, A_2, \ldots, A_n$ 互不相容，形成对样本空间 Ω 的**分割** (partition)，也就是每次试验事件 $A_1, A_2, \ldots, A_n$ 中有且仅有一个发生。

假定 $\Pr(A_i) > 0$，对于空间 Ω 中任意事件 B，下式成立：



$$\begin{aligned}\underbrace{\Pr(B)}_{\text{Marginal}} &= \sum_{i=1}^{n} \underbrace{\Pr(A_i \cap B)}_{\text{Joint}} = \Pr(A_1 \cap B) + \Pr(A_2 \cap B) + \cdots + \Pr(A_n \cap B) \\ &= \sum_{i=1}^{n} \underbrace{\Pr(A_i, B)}_{\text{Joint}} = \Pr(A_1, B) + \Pr(A_2, B) + \cdots + \Pr(A_n, B)\end{aligned} \tag{36}$$

上式就叫做**全概率定理** (law of total probability)。这本质上就是穷举法，也叫枚举法。

图 25 给出的例子是三个互不相容事件 A_1、A_2、A_3 对 Ω 形成分割。通过全概率定理，即穷举法，$\Pr(B)$ 可以通过下式计算得到：

$$\underbrace{\Pr(B)}_{\text{Marginal}} = \underbrace{\Pr(A_1, B)}_{\text{Joint}} + \underbrace{\Pr(A_2, B)}_{\text{Joint}} + \underbrace{\Pr(A_3, B)}_{\text{Joint}} \tag{37}$$

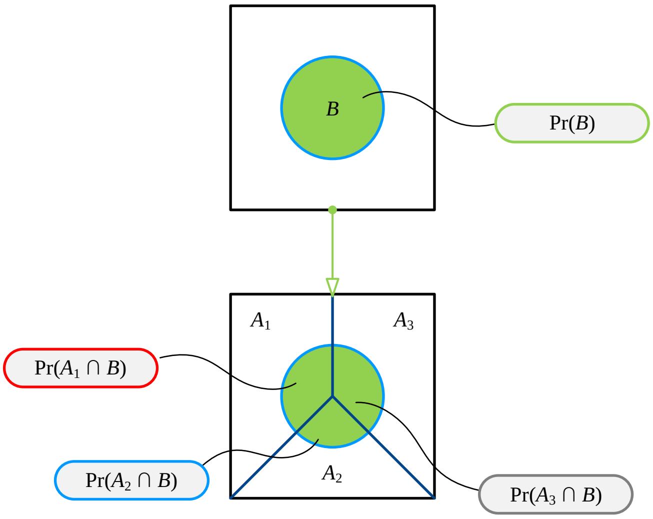


图 25. A_1, A_2, A_3 对空间 Ω 分割

引入贝叶斯定理

利用贝叶斯定理，以为 A_1, A_2, …, A_n 条件，展开 (36)：

$$\begin{aligned}\Pr(B) &= \sum_{i=1}^{n} \underbrace{\Pr(A_i, B)}_{\text{Joint}} = \sum_{i=1}^{n} \underbrace{\Pr(B|A_i)}_{\text{Conditional}} \underbrace{\Pr(A_i)}_{\text{Marginal}} \\ &= \Pr(B|A_1)\Pr(A_1) + \Pr(B|A_2)\Pr(A_2) + \cdots + \Pr(B|A_n)\Pr(A_n)\end{aligned} \tag{38}$$

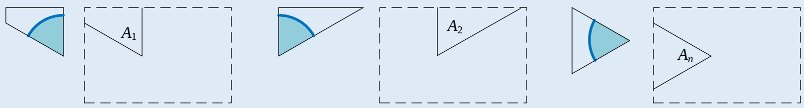


图 26 所示为分别给定 A_1, A_2, A_3 条件下，事件 B 发生的情况。

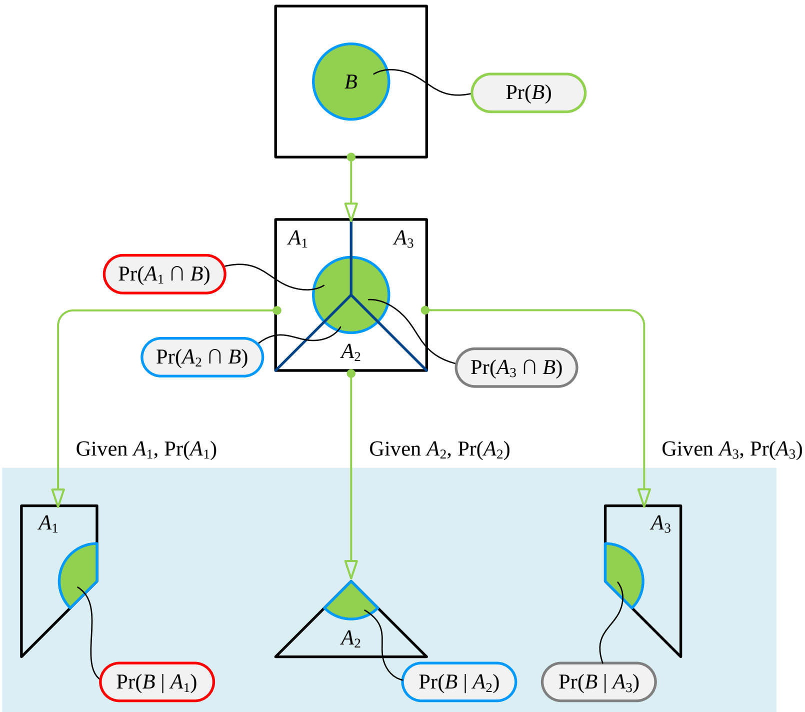


图 26. 分别给定 A_1, A_2, A_3 条件下，事件 B 发生的情况

反过来，根据贝叶斯定理，在给定事件 B 发生条件下 ($\Pr(B) > 0$)，任意事件 A_i 发生的概率为：

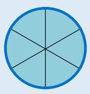
$$\Pr\left(A_i \mid B\right) = \frac{\Pr\left(A_i, B\right)}{\Pr\left(B\right)} = \frac{\Pr\left(B \mid A_i\right) \cdot \Pr\left(A_i\right)}{\Pr\left(B\right)} \tag{39}$$

利用贝叶斯定理，以为 B 条件，进一步展开 (36)：

$$\begin{aligned} \Pr\left(B\right) &= \sum_{i=1}^{n} \underbrace{\Pr\left(A_i, B\right)}_{\text{Joint}} = \sum_{i=1}^{n} \underbrace{\Pr\left(A_i \mid B\right)}_{\text{Conditional}} \underbrace{\Pr\left(B\right)}_{\text{Marginal}} \\ &= \Pr\left(A_1 \mid B\right)\Pr\left(B\right) + \Pr\left(A_2 \mid B\right)\Pr\left(B\right) + \cdots + \Pr\left(A_n \mid B\right)\Pr\left(B\right) \end{aligned} \tag{40}$$

(40) 等式左右消去 $\Pr(B)$ ($\Pr(B) > 0$)，得到：

$$\sum_{i=1}^{n} \Pr\left(A_i \mid B\right) = \Pr\left(A_1 \mid B\right) + \Pr\left(A_2 \mid B\right) + \cdots + \Pr\left(A_n \mid B\right) = 1 \tag{41}$$



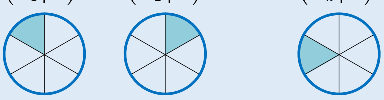


图 27 所示为给定 B 条件下，事件 A_1、A_2、A_3 发生的情况。

看到这里，对贝叶斯定理和全概率定理还是一头雾水的读者不要怕，本书后续会利用不同实例反复讲解这两个定理。

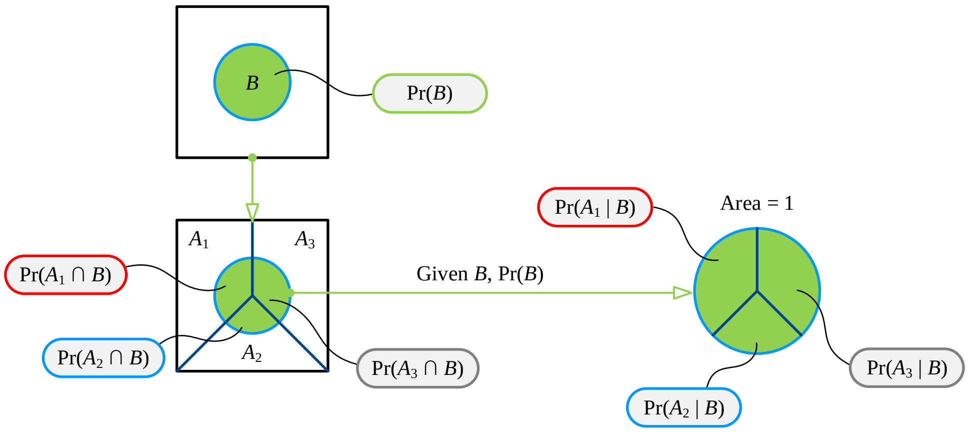


图 27. 给定 B 条件下，事件 A_1、A_2、A_3 发生的情况

3.8 独立、互斥、条件独立

独立

上一节介绍的条件概率 $\Pr(A|B)$ 刻画了在事件 B 发生的条件下，事件 A 发生的可能性。

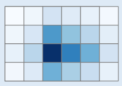
有一种特殊的情况，事件 B 发生与否，不会影响事件 A 发生的概率，也就是如下等式成立：

$$\underbrace{\Pr\left(A|B\right)}_{\text{Conditional}} = \underbrace{\Pr\left(A\right)}_{\text{Marginal}} \quad \Leftrightarrow \quad \underbrace{\Pr\left(B|A\right)}_{\text{Conditional}} = \underbrace{\Pr\left(B\right)}_{\text{Marginal}} \tag{42}$$

如果 (42) 给出的等式成立，则称**事件 A 和事件 B 独立** (events A and B are independent)。

如果 A 和 B 独立，联立 (28) 和 (42) 可以得到：

$$\Pr\left(A \cap B\right) = \underbrace{\Pr\left(A, B\right)}_{\text{Joint}} = \underbrace{\Pr\left(A\right)}_{\text{Marginal}} \cdot \underbrace{\Pr\left(B\right)}_{\text{Marginal}} \tag{43}$$



如果一组事件 A_1、A_2… A_n，它们两两相互独立，则下式成立：

$$\Pr\left(A_1 \cap A_2 \cap \cdots \cap A_n\right) = \Pr\left(A_1, A_2, \cdots, A_n\right) = \Pr\left(A_1\right) \cdot \Pr\left(A_2\right) \cdots \Pr\left(A_n\right) = \prod_{i=1}^{n} \Pr\left(A_i\right) \tag{44}$$

抛三枚色子

接着本章前文“抛三枚色子”的例子。大家应该清楚，一次性抛三枚色子，这三枚色子点数互不影响，也就是“独立”。

如图 28 所示，第一枚色子的点数 (X_1) 取不同值 (1 ~ 6) 时，相当于把样本空间这个立方体切成 6 个“切片”。每个切片都有 36 个点，因此每个切片对应的概率均为：

$$\frac{6 \times 6}{6 \times 6 \times 6} = \frac{1}{6} \tag{45}$$

也就相当于把概率“1”，均分为 6 份。而 1/6 对应第一枚色子的点数 (X_1) 取不同值的概率。

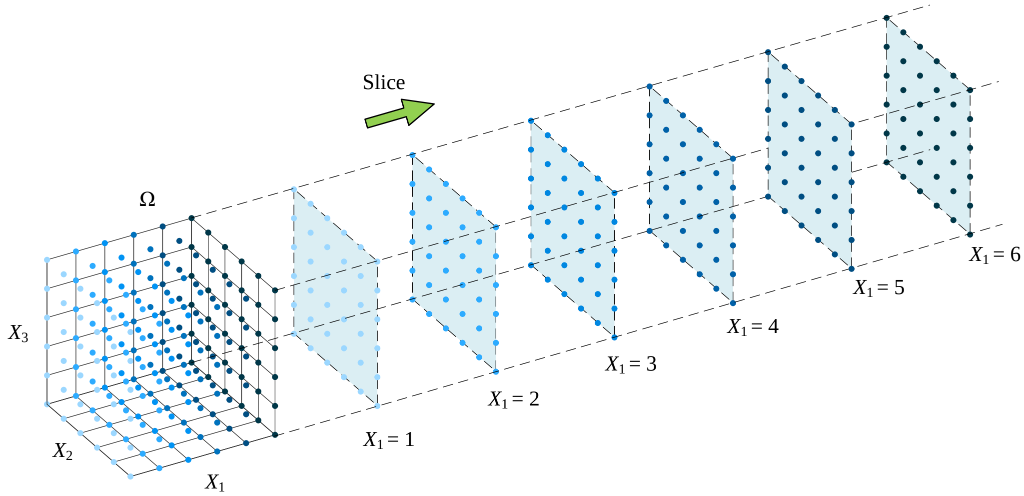


图 28. X_1 视角下的“抛三枚色子结果”

(3, 3, 3) 这个结果在整个样本空间中对应的概率为 1/216。如图 29 所示，1/216 这个数值可以有四种不同的求法：

$$\frac{1}{216} = \underbrace{\frac{1}{6}}_{X_1=3} \times \underbrace{\frac{1}{36}}_{(X_2,X_3)=(3,3)} = \underbrace{\frac{1}{6}}_{X_2=3} \times \underbrace{\frac{1}{36}}_{(X_1,X_3)=(3,3)} = \underbrace{\frac{1}{6}}_{X_3=3} \times \underbrace{\frac{1}{36}}_{(X_1,X_2)=(3,3)} = \underbrace{\frac{1}{6}}_{X_1=3} \times \underbrace{\frac{1}{6}}_{X_2=3} \times \underbrace{\frac{1}{6}}_{X_3=3} \tag{46}$$

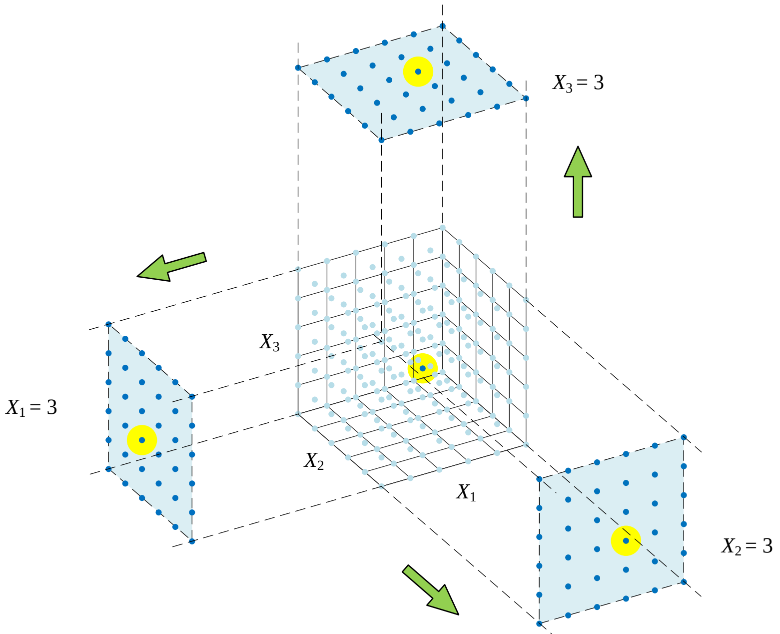


图 29. (3, 3, 3) 结果在样本空间和三个各方向切片上的位置

再换个角度，图 29 中立方体代表概率为 1，而 X_1、X_2、X_3 这三个维度独立，并将“1”均匀地切分成 216 份：

$$\left(\underbrace{\overbrace{\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}}^{=1}}_{X_1=1\sim6}\right)\times\left(\underbrace{\overbrace{\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}}^{=1}}_{X_2=1\sim6}\right)\times\left(\underbrace{\overbrace{\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}}^{=1}}_{X_3=1\sim6}\right)=1 \tag{47}$$

上式体现的就是乘法分配律。从向量角度来看，上式相当于三个向量的张量积，撑起一个如所示的三维数组。再次强调，之所以能用这种方式计算联合概率，就是因为“独立”。

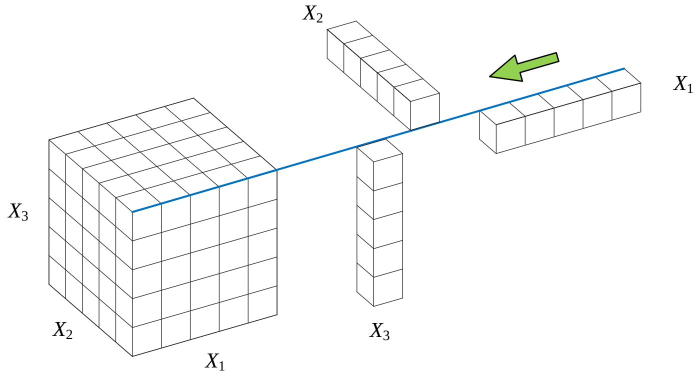


图 30. 三个向量的张量积

请大家格外注意，互斥不同于独立。表 3 对比一般情况、互斥、独立之间的主要特征。

表 3. 比较一般情况、互斥、独立

A 和 B	Pr(A and B) Pr($A \cap B$)	Pr(A or B) Pr($A \cup B$)	Pr(A | B)	Pr(B | A)
一般情况 Pr(A) > 0 Pr(B) > 0	Pr(A) × Pr(B | A) Pr(B) × Pr(A | B)	Pr(A) + Pr(B) – Pr($A \cap B$)	Pr($A \cap B$)/Pr(B)	Pr($A \cap B$)/Pr(A)
互斥	0	Pr(A) + Pr(B)	0	0
独立	Pr(A) × Pr(B)	Pr(A) + Pr(B) – Pr(A) × Pr(B)	Pr(A)	Pr(B)

条件独立

在给定事件 C 发生条件下，如果如下等式成立，则称**事件 A 和事件 B 在 C 发生条件下条件独立** (events A and B are conditionally independent given an event C)：

$$\Pr\left(A \cap B \middle| C\right) = \Pr\left(A, B \middle| C\right) = \Pr\left(A \middle| C\right) \cdot \Pr\left(B \middle| C\right) \tag{48}$$

请大家格外注意，A 和 B 相互独立，无法推导得到 A 和 B 条件独立。而 A 和 B 条件独立，也无法推导得到 A 和 B 相互独立。

本书后文还会深入讨论并比较独立、条件独立。

本章的核心公式如下。请大家回忆公式的含义，并在空白处把配图画出来。

$$\begin{aligned} &\Pr\left(A, B\right) = \Pr\left(A \middle| B\right) \cdot \Pr\left(B\right) = \Pr\left(B \middle| A\right) \cdot \Pr\left(A\right) \\ &\Pr\left(B\right) = \sum_{i=1}^{n} \Pr\left(C_i \cap B\right) \\ &\Pr\left(A, B\right) = \Pr\left(A\right) \cdot \Pr\left(B\right) = \Pr\left(B\right) \cdot \Pr\left(A\right) \end{aligned} \tag{49}$$

古典概率有效地解决抛硬币、抛色子、口袋里摸球这些简单的概率问题，等概率模型、全概率定理、贝叶斯定理等重要的概率概念也随之产生。建议大家在学习时，已经要注意“多维”，让自己习惯多特征随机变量。

随着概率统计的研究不断深入，这些数学工具的应用场景也开始变得更加广泛多样。然而基于集合论的古典概率模型显得力不从心。引入随机变量、概率分布等概念，实际上就是将代数思想引入概率统计，以便于对更复杂的问题抽象建模、定量分析。这是下一章要讲解的内容。