

# 03

Classical Probability

## 古典概率模型

归根结底，概率就是量化的生活常识



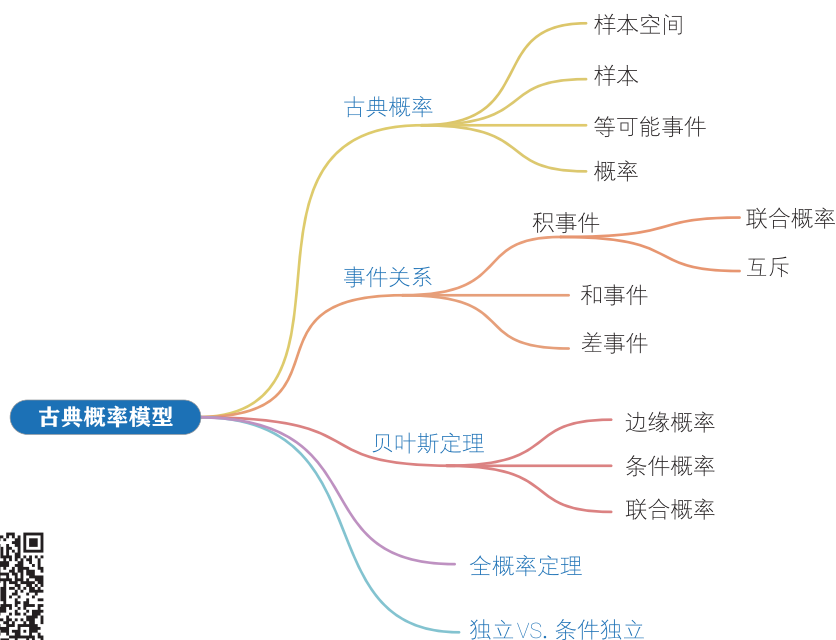
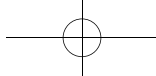
真是耐人寻味，一门以赌博为起点的学科本应该是人类知识体系中最重要研究对象。

***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年



- ◀ `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

自然界的随机无处不在，没有两朵完全一样的鸢尾花，没有两片完全一样的雪花，也没有两条完全一样的人生轨迹。鸢尾花书《数学要素》一册中曾提过，在微观、少量、短期尺度上，我们看到的更多的是不确定、不可预测、随机；但是，站在宏观、大量、更长的时间尺度上，我们可以发现确定、模式、规律。

而概率则试图量化随机事件发生的可能性。概率的研究和应用深刻影响着人类科学发展进程，本节介绍孟德尔和道尔顿两个例子。

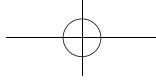
### 孟德尔的豌豆试验

**孟德尔** (Gregor Mendel, 1822—1884年) 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，他发现培育得到的子代豌豆都是圆粒 ○，如图3.1所示。

实际情况是，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 ○ 基因来讲是隐性。

如图3.1所示，当第一代杂交圆粒豌豆 ○ 自花传粉或者彼此交叉传粉后，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比例背后的数学原理就是本章要介绍的古典概率模型。

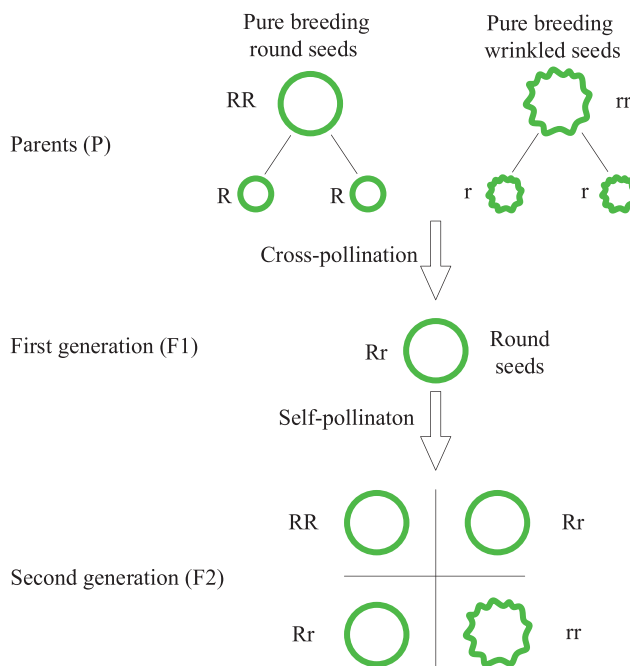


图3.1 孟德尔的豌豆试验

## 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年) 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，母亲对袜子的颜色不是很满意，她觉得“樱桃红”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认为袜子是“棕灰色”。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，因此须有一对色盲等位基因，才会表现出异常。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。

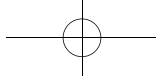
下面，我们从概率的角度分几种情况来思考红绿色盲的遗传规律。

### 情况A

如图3.2所示，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)。它与生男、生女的概率一致。

给定后代为男性 ♂，则发病比例为100%。给定后代为女性 ♀，则发病比例为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)。

条件概率的概念在概率论和统计学中非常重要，它允许我们在一些已知信息的情况下对事件的发生概率进行更精确的估计和预测。例如，在医学诊断中，医生可以根据病人的症状和体征，计算出某种疾病在不同条件下的发病率，从而帮助医生判断病人是否患有这种疾病。

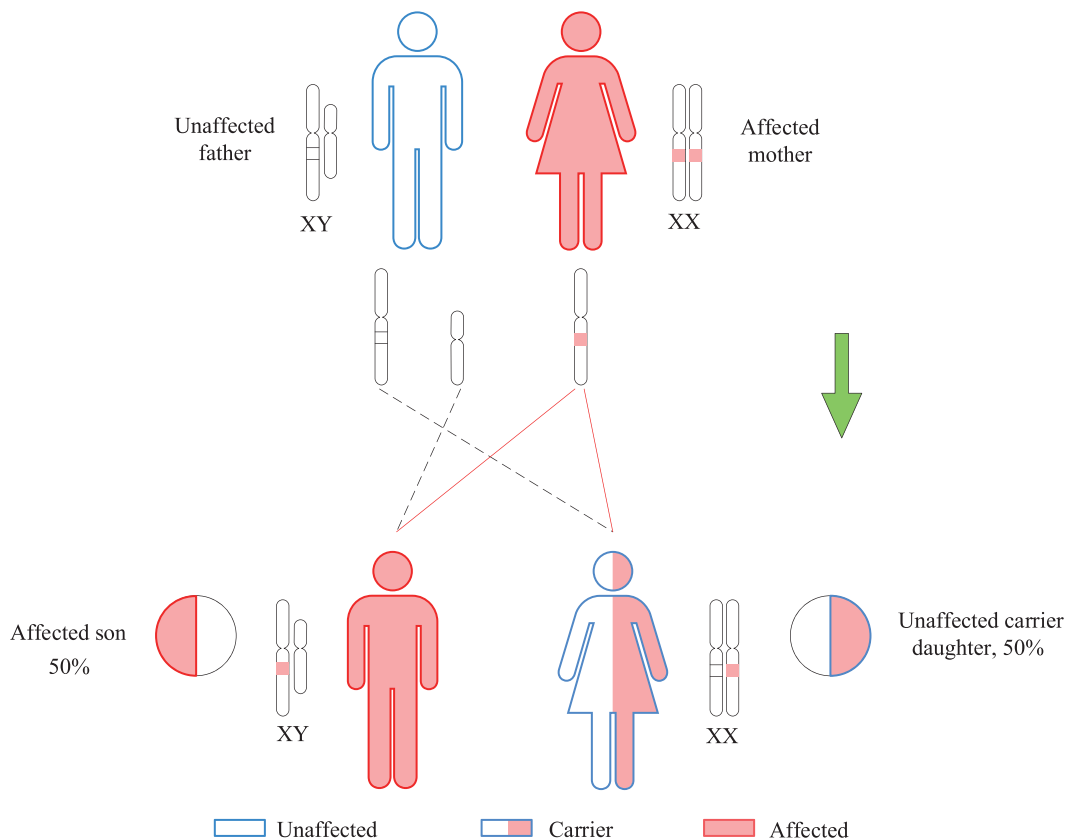


图3.2 红绿色盲基因遗传机制 (情况A)

### 情况B

如图3.3所示，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%概率为色盲基因的携带者。这些数值也都是条件概率。

### 情况C

如图3.4所示，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，如果不分男女，则后代发病的概率为50%。

其中，儿子 ♂ 有50%概率正常，50%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，则后代有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并计算后代患病概率。

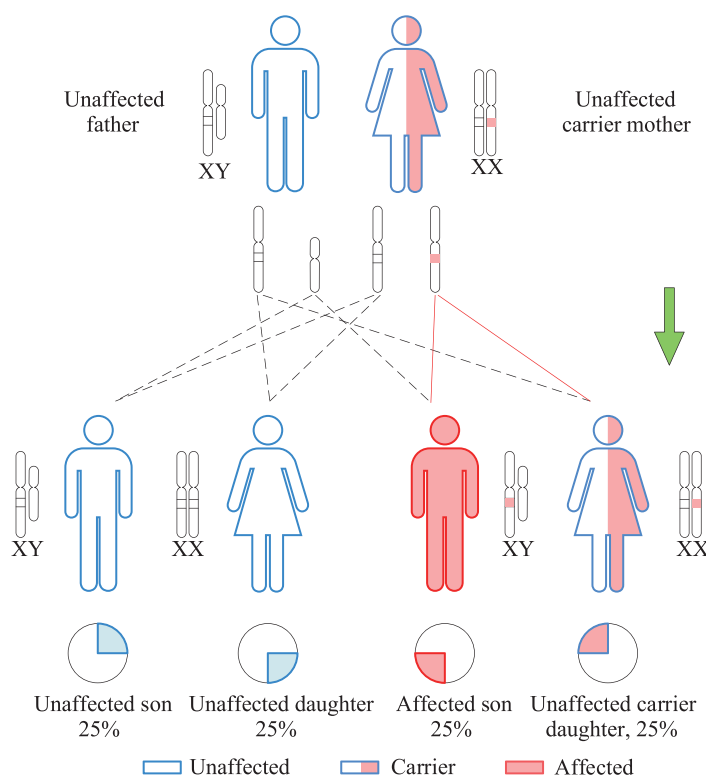
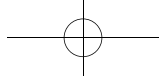


图3.3 红绿色盲基因遗传机制 (情况B)

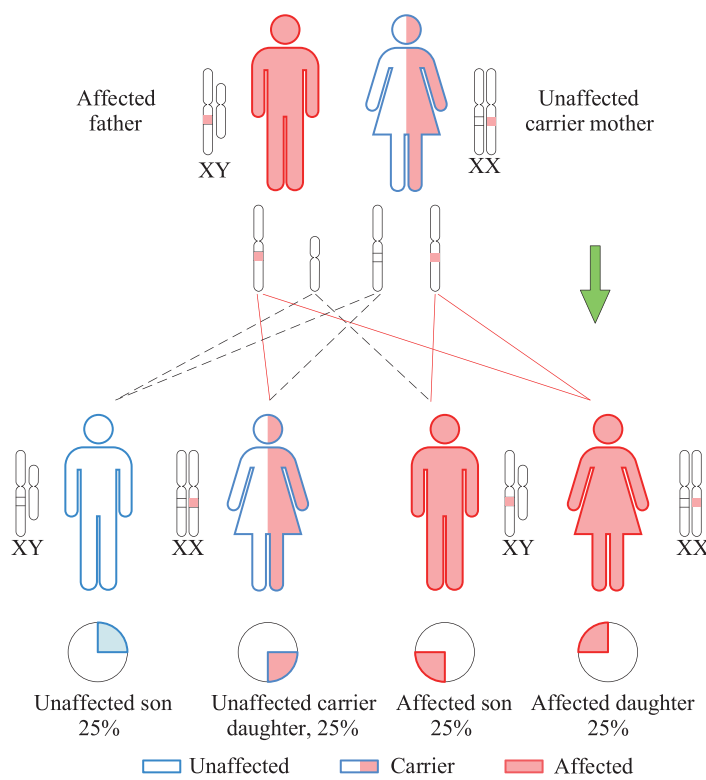


图3.4 红绿色盲基因遗传机制 (情况C)

建议大家学完本章所有内容之后，回过头来再仔细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。古典概型广泛应用集合运算，本节一边讲解概率论，一边回顾集合运算。



《数学要素》一册第4章介绍过集合相关概念，建议大家回顾。

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)  $\Omega$ 。样本空间  $\Omega$  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  $\Omega$  的概率为1，即

$$\Pr(\Omega) = 1 \quad \Omega \quad (3.1)$$



注意：本书表达概率的符号  $\Pr$  为正体。再次请大家注意，不同试验的样本空间  $\Omega$  不同。

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”或“切丝、切条”。

给定样本空间  $\Omega$  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

$$\Pr \left( \begin{array}{c} \text{Event} \\ A \\ \hline \text{Probability} \end{array} \right) \geq 0 \quad A \quad \Omega \quad (3.2)$$

大家看到任何概率值时一定要问一句，**它的样本空间是什么？**

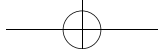
空集  $\emptyset$  不包含任何样本点，也称作**不可能事件** (impossible event)，因此对应的概率为0，即

$$\Pr(\emptyset) = 0 \quad (3.3)$$

### 等可能

设样本空间  $\Omega$  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

$$\Pr(A) = \frac{n_A}{n} \quad (3.4)$$



其中： $n_A$ 为含于事件 $A$ 的试验结果数量。

等可能事件是指在某一试验中，每个可能结果发生的概率相等的事件。简单来说，就是等可能事件每个结果发生的可能性是一样的。例如，对于一枚硬币的抛掷，假设正面和反面的出现概率是相等的，因此正面出现和反面出现是等可能事件。同样地，掷一个六面骰子，假设每个面出现的概率都是相等的，因此每个面的出现也是等可能事件。

## 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个样本均分为三类——setosa ( $C_1$ )、versicolour ( $C_2$ )、virginica ( $C_3$ )。如图3.5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

$$\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \frac{50}{150} = \frac{1}{3} \quad (3.5)$$

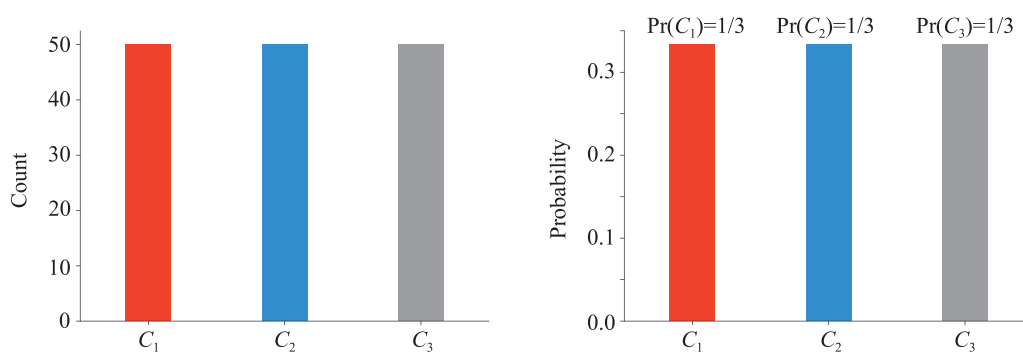


图3.5 鸢尾花150个样本数据均分为三类

## 抛一枚硬币

抛一枚硬币，1代表正面，0代表反面。抛一枚硬币可能结果的样本空间为

$$\Omega = \{0, 1\} \quad (3.6)$$

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

$$\Pr(0) = \Pr(1) = \frac{1}{2} \quad (3.7)$$

把  $\{0, 1\}$  标记在数轴上，用火柴梗图可视化上述概率值，我们便得到图3.6。

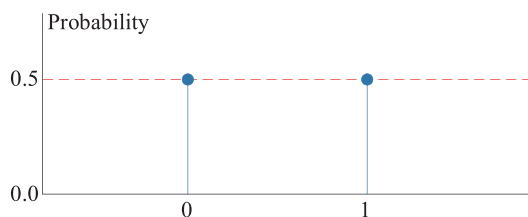


图3.6 抛一枚硬币结果和对应的理论概率值

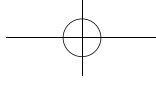


图3.7所示为反复抛一枚硬币，正面 (1)、反面 (0) 平均值随试验次数的变化。可以发现，平均结果不断靠近1/2，也就是说正反面出现的概率几乎相同。

从另外一个角度，式(3.7)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)，也称为公式概率或数学概率，是一种基于理论推导的概率计算方法。它一般基于假设所有可能的结果是等可能的，并使用数学公式计算概率。

而图3.7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)。试验概率是一种基于实际试验的概率计算方法。它通过多次重复试验来统计某个事件发生的频率，然后将频率作为概率的估计值。

理论概率可以作为试验概率的基础，即在假设所有可能结果是等可能的情况下，理论概率可以预测事件发生的概率，而试验概率则可以验证这一预测是否准确。



第15章介绍如何完成蒙特卡罗模拟 (Monte Carlo simulation)。

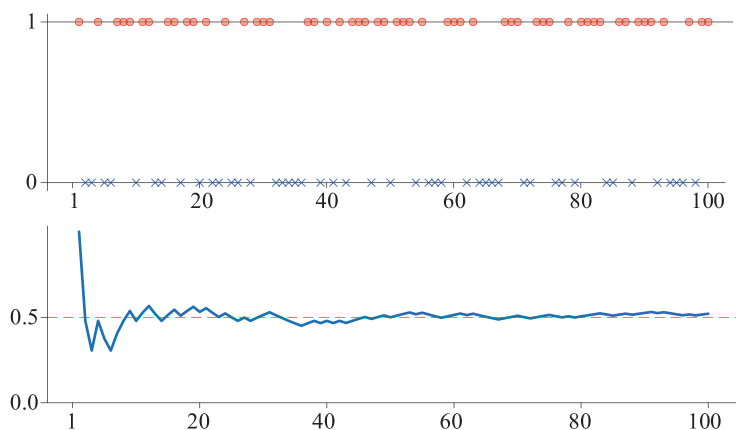


图3.7 抛硬币100次试验结果变化

## 掷骰子

如图3.8所示，掷一枚骰子试验可能结果的样本空间为

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad (3.8)$$

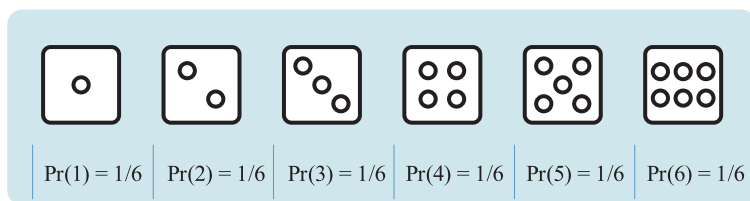
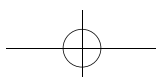


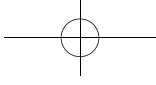
图3.8 投骰子试验

试验中，假设获得每一种点数的可能性相同。掷一枚骰子共六种结果，每种结果对应的概率为

$$\Pr(1) = \Pr(2) = \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = \frac{1}{6} \quad (3.9)$$

同样用火柴梗图把上述结果画出来，得到图3.9。这也是抛一枚骰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。这种情况下，要想估算不同结果的概率值，一般就只能通过试验。

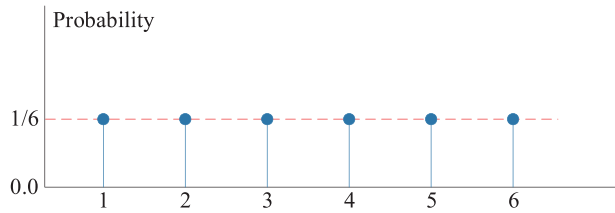


图3.9 抛一枚骰子结果和对应的理论概率值

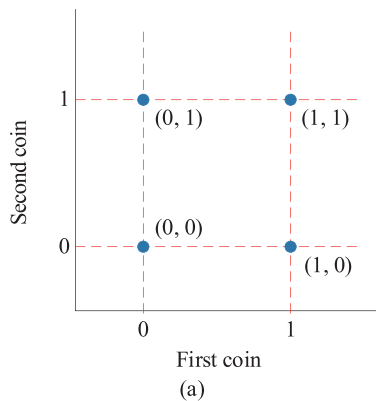
## 抛两枚硬币

下面看两个稍复杂的例子——每次抛两枚硬币。

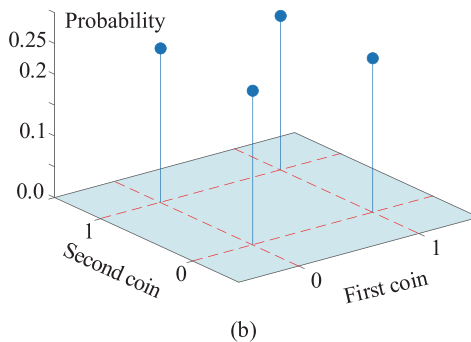
比如，如果第一枚硬币为正面、第二枚硬币为反面，结果记作  $(1, 0)$ 。这样，样本空间由以下四个点构成，即

$$\Omega = \left\{ \begin{pmatrix} 0,0 \\ 0,1 \\ 1,0 \\ 1,1 \end{pmatrix} \right\} \quad (3.10)$$

图3.10 (a) 所示为用二维坐标系展示试验结果。图3.10 (a) 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 3.10 (b) 所示。



(a)



(b)

图3.10 抛两枚硬币结果和对应的理论概率值

## 抛两枚骰子

同理，每次抛两枚骰子，样本空间  $\Omega$  的等可能试验结果数量为  $6 \times 6$ ，有

$$\Omega = \left\{ \begin{pmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{pmatrix} \right\} \quad (3.11)$$

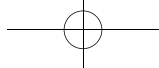


图3.11 (a) 所示为上述试验的样本空间。图3.11 (b) 中，假设骰子质地均匀，每个试验结果对应的概率均为 $1/36$ 。

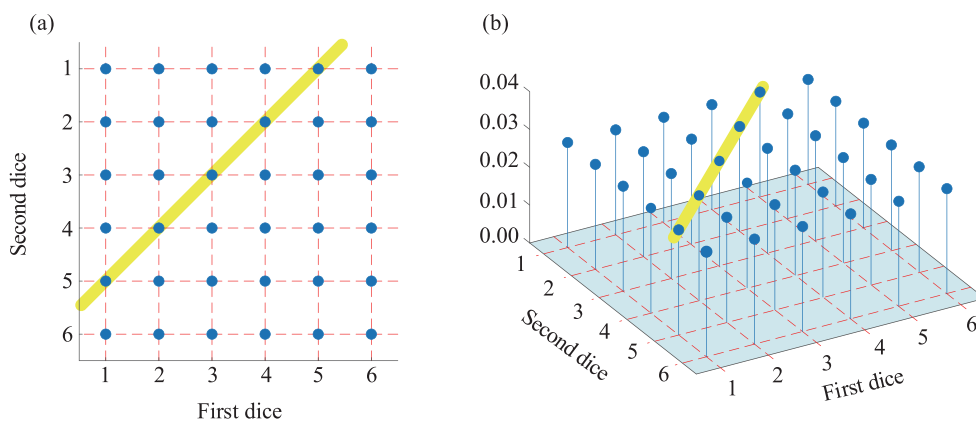


图3.11 抛两枚骰子结果和对应的理论概率值

## 抛两枚骰子：点数之和为6

下面，我们看一种特殊情况。如图3.12所示，如果我们关心两个骰子点数之和为6，会发现一共有五种结果满足条件。这五种结果为 $1+5$ 、 $2+4$ 、 $3+3$ 、 $4+2$ 、 $5+1$ 。则该事件对应概率为

$$\Pr(\text{sum} = 6) = \frac{5}{6 \times 6} \approx 0.1389 \quad (3.12)$$

图3.11 (a) 中黄色背景所示样本便代表抛两枚骰子点数之和为6的事件。

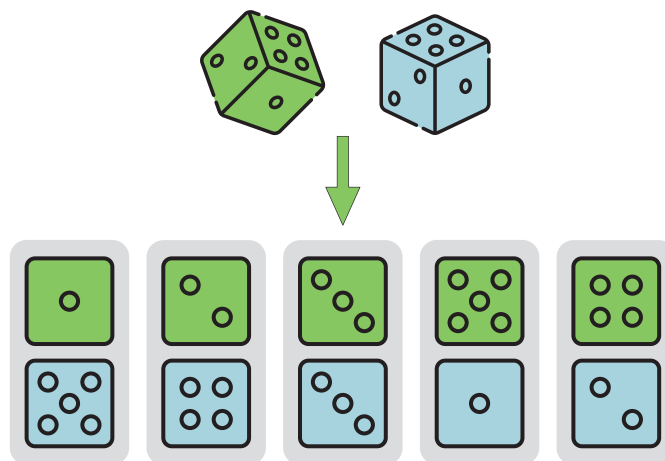


图3.12 投两枚骰子，点数之和为6

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3.13所示为“点数之和为6”事件概率随抛掷次数变化曲线。

比较式(3.12)和图3.13，通过古典概率模型得到的理论结论和试验结果相互印证。



图3.13中的横轴为对数刻度。《数学要素》一册第12章介绍过对数刻度，大家可以进行回顾。

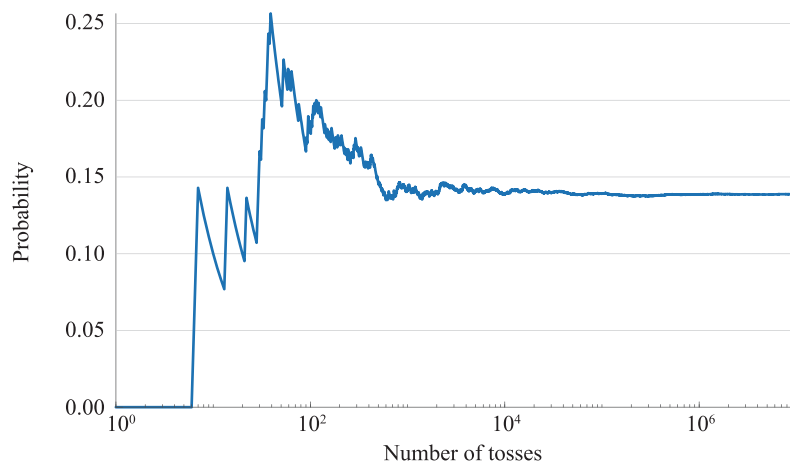
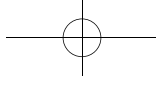


图3.13 骰子“点数之和为6”事件概率随抛掷次数变化



代码Bk5\_Ch03\_01.py模拟抛骰子试验并绘制图3.13。请大家把这个代码改写成一个Streamlit App，并用抛掷次数作为输入。

## 抛两枚骰子：点数之和的样本空间

接着上一个例子，如果我们对抛两枚骰子“点数之和”感兴趣，首先要知道这个事件的样本空间。如图3.14所示，彩色等高线对应两枚骰子点数之和。由此，得到两个骰子点数之和的样本空间为{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}。

而等高线上灰色点 ● 的横纵坐标代表满足条件的骰子点数。计算某一条等高线上点 ● 的数量，再除 $36 (= 6 \times 6)$ 便得到不同“点数之和”对应的概率值。

图3.14 (b) 所示为样本空间所有结果概率值的火柴梗图。观察图3.14 (b)，容易发现结果非等概率；但是，这些概率值也是通过等概率模型推导得到的。

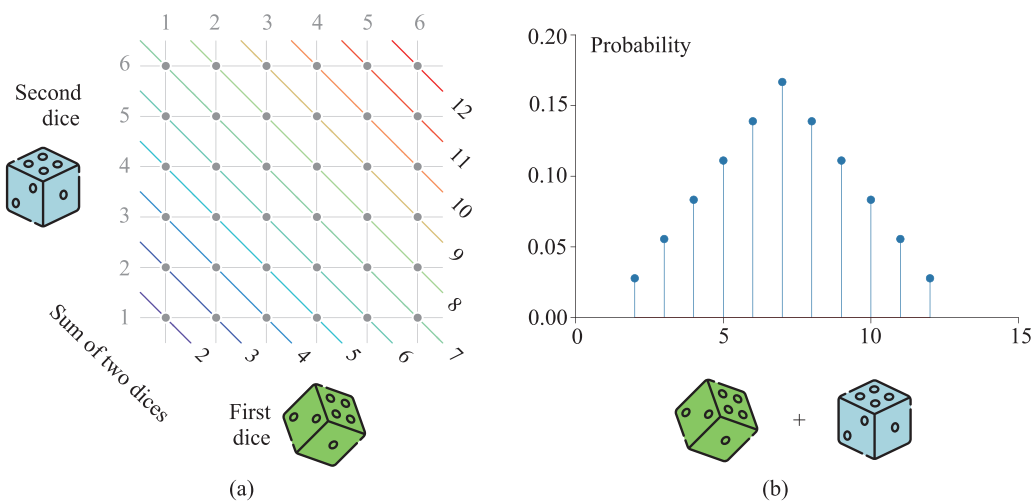
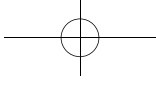


图3.14 两个骰子点数之和



## 更多“花样”

接着上面抛两枚骰子计算点数之和的试验，我们玩出更多“花样”！

如表3.1所示，抛两枚骰子，我们可以只考虑第一只骰子的点数、第一只骰子点数平方值，也可以计算两个骰子的点数平均值、乘积、商、差、差的平方等。

这些不同的花式玩法至少告诉我们以下几层信息。

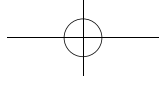
- 抛两枚色子，第一枚骰子和第二枚骰子的结果可以独立讨论；换个视角来看，一次试验中，第一枚、第二枚骰子点数结果相互不影响。
- 第一枚和第二枚骰子的点数结果还可以继续运算。
- 用文字描述这些结果太麻烦了，我们需要将它们代数化！比如，定义第一个色子结果为 $X_1$ ，第二个骰子点数为 $X_2$ ，两个点数数学运算结果为 $Y$ 。这便是下一章要探讨的**随机变量** (random variable)。
- 显然表3.1中每种花式玩法有各自的样本空间  $\Omega$ 。样本空间的样本并非都是等概率。但是，样本空间中所有样本的概率之和都是1。



表3.1所示为基于抛两枚骰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，并计算每个样本对应的概率值。我们将在下一章揭晓答案。

表3.1 基于抛两枚骰子试验结果的更多花式玩法

| 随机变量                                | 描述               | 例子   |     |      |     |      |      |     |      |     |      |     |     |
|-------------------------------------|------------------|------|-----|------|-----|------|------|-----|------|-----|------|-----|-----|
| $X_1$                               | 第一个骰子点数          | 1    | 2   | 3    | 4   | 5    | 6    | 1   | 2    | 3   | 4    | 5   | 6   |
| $X_2$                               | 第二个骰子点数          | 1    | 1   | 1    | 1   | 1    | 1    | 2   | 2    | 2   | 2    | 2   | 2   |
| $Y = X_1$                           | 只考虑第一个骰子<br>点数   | 1    | 2   | 3    | 4   | 5    | 6    | 1   | 2    | 3   | 4    | 5   | 6   |
| $Y = X_1^2$                         | 第一个骰子点数<br>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，<br>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 =  X_1 - X_2 $                   | 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 |



## 抛三枚骰子

为了大家习惯“多元”思维，我们再进行进一步将一次抛掷骰子的数量提高至三枚。

第一枚点数定义为 $X_1$ ，第二枚点数为 $X_2$ ，第三枚点数为 $X_3$ 。

图3.15 (a) 所示为抛三枚骰子点数的样本空间，这显然是个三维空间。比如，坐标点 (3, 3, 3) 代表三枚骰子的点数都是3。

图3.15 (a) 这个样本空间有216 ( $= 6 \times 6 \times 6$ ) 个样本。假设这三个骰子质量均匀，获得每个点数为等概率，则图3.15 (a) 中每个样本对应的概率为 $1/216$ 。

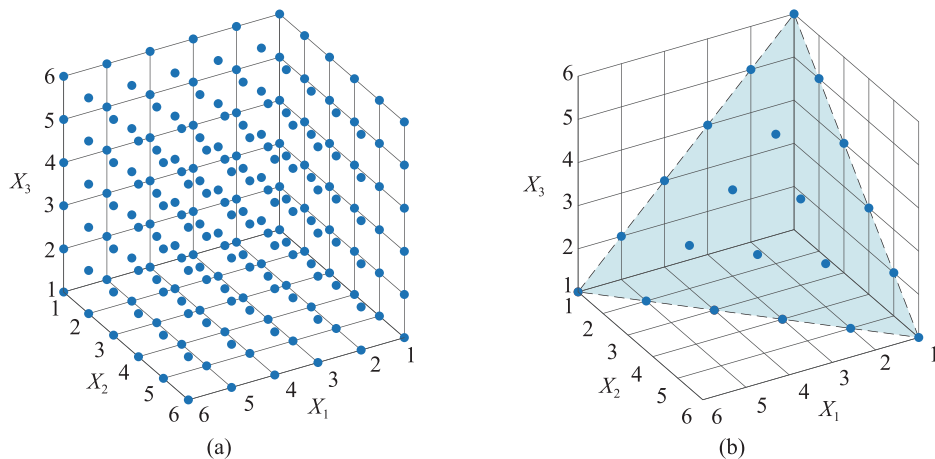


图3.15 抛三枚骰子点数的样本空间

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3.15 (b) 所示，一共有21个样本点，容易发现这些样本在同一个斜面上。相对图3.15 (a) 所示的样本空间，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3.16所示，将散点投影在平面上得到图3.16 (b)。能够完成这种投影是因为 $X_1 + X_2 + X_3 = 8$ 这个等式关系。

通过这个例子，相信大家已经发现多元统计中几何思维的重要性。

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) 和Dirichlet分布 (第7章)。

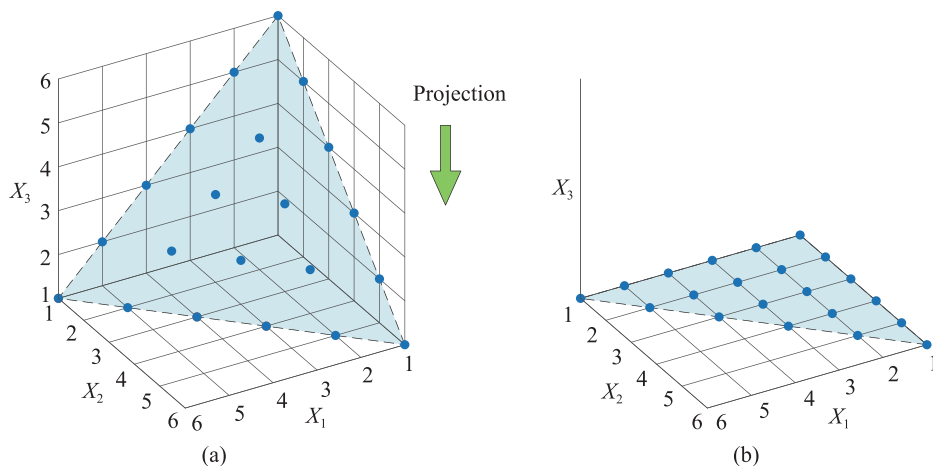


图3.16 将事件 $A$ 的样本点投影到平面上

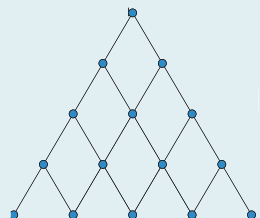
## 3.3 回顾：杨辉三角和概率

### 杨辉三角

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



《数学要素》一册第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。

$$\begin{aligned}(a+b)^0 &= 1 \\(a+b)^1 &= a+b \\(a+b)^2 &= a^2+2ab+b^2 \\(a+b)^3 &= a^3+3a^2b+3ab^2+b^3 \\(a+b)^4 &= a^4+4a^3b+6a^2b^2+4ab^3+b^4\end{aligned}\quad (3.13)$$


其中： $a$ 和 $b$ 均不为0。

### 抛硬币

把二项式展开用在理解抛硬币的试验。 $(a+b)^n$ 中 $n$ 代表一次抛掷过程中的硬币数量， $a$ 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，则 $a=b=1/2$ 。举个例子，如果硬币质地均匀，每次抛10 ( $n$ ) 枚硬币，正好出现6次正面朝上对应的概率为

$$\Pr(\text{heads} = 6) = C_{10}^6 \frac{1}{2^{10}} = \frac{210}{1024} = \frac{210}{1024} \approx 0.20508 \quad (3.14)$$

每次抛10枚硬币，至少出现6次正面的概率为

$$\Pr(\text{heads} \geq 6) = \frac{C_{10}^6 + C_{10}^7 + C_{10}^8 + C_{10}^9 + C_{10}^{10}}{2^{10}} = \frac{210+120+45+10+1}{1024} = \frac{386}{1024} \approx 0.37695 \quad (3.15)$$

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，并且计算两个事件的当前概率。图3.17所示为两事件概率随抛掷次数变化的曲线。这也是试验概率、理论概率的相互印证。

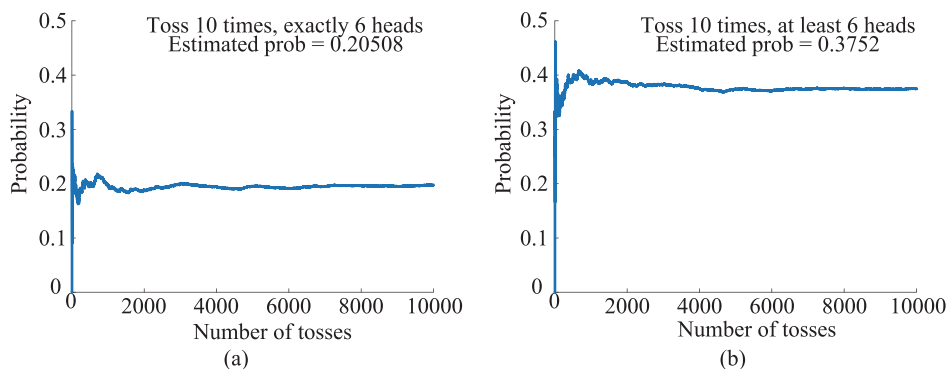
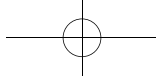


图3.17 试验概率随抛掷次数变化：(a) 正好出现6次正面；(b) 至少出现6次正面



Bk5\_Ch03\_02.py完成上述两个试验并绘制图3.17。

## 回忆二叉树

《数学要素》一册第20章还介绍过杨辉三角和二叉树的联系，如图3.18所示。

站在二叉树中间节点处，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”的概率。

假设，向上走、向下走的概率均为 $1/2$ 。图3.18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。

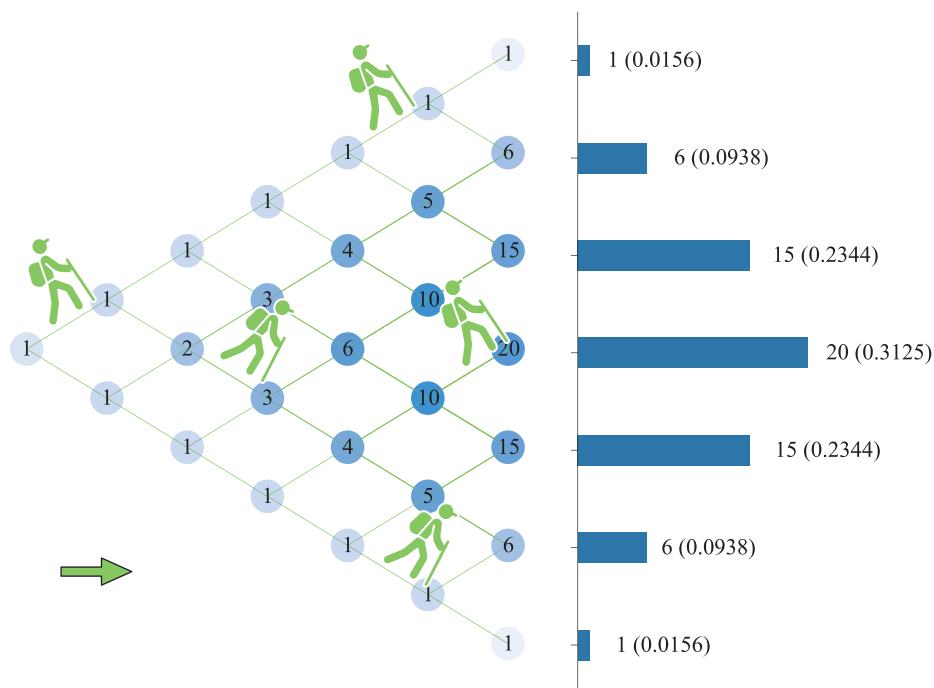
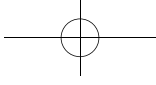


图3.18 杨辉三角逆时针旋转90度得到一个二叉树 (图片基于《数学要素》一册第20章)

## 3.4 事件之间的关系：集合运算

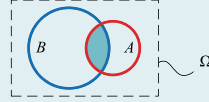
### 积事件

事件 $A$ 与事件 $B$ 为样本空间 $\Omega$ 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。



$\Pr(A \cap B)$  表示 $A$ 和 $B$ **积事件的概率** (probability of the intersection of events  $A$  and  $B$ 或joint probability of  $A$  and  $B$ )。  $\Pr(A \cap B)$  也叫作 $A$ 和 $B$ 的**联合概率** (joint probability)。  $\Pr(A \cap B)$  也常记作  $\Pr(A, B)$ ，即

$$\Pr\left(\underbrace{A \cap B}_{\text{Joint}}\right) = \Pr\left(\underbrace{A, B}_{\text{Joint}}\right) \quad \text{图 3.16} \quad (3.16)$$

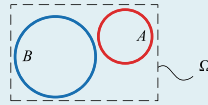


## 互斥

如果事件 $A$ 与事件 $B$ 两者交集为空，即  $A \cap B = \emptyset$ ，则称**事件 $A$ 与事件 $B$ 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，即

$$\underbrace{A \cap B}_{\text{Joint}} = \emptyset \Rightarrow \Pr\left(\underbrace{A \cap B}_{\text{Joint}}\right) = \Pr\left(\underbrace{A, B}_{\text{Joint}}\right) = 0 \quad \text{图 3.17} \quad (3.17)$$

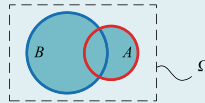


## 和事件

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。  $\Pr(A \cup B)$  表示事件 **$A$ 和 $B$ 和事件的概率** (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

$$\underbrace{\Pr(A \cup B)}_{\text{Union}} = \Pr(A) + \Pr(B) - \underbrace{\Pr(A \cap B)}_{\text{Joint}} \quad \text{图 3.18} \quad (3.18)$$



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

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \quad \text{图 3.19} \quad (3.19)$$

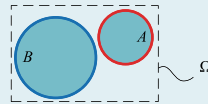
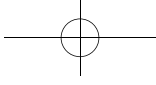


表3.2总结了常见集合运算维恩图。

表3.2 常见集合运算和维恩图

| 符号          | 解释                          | 维恩图 |
|-------------|-----------------------------|-----|
| $\Omega$    | 必然事件，即整个样本空间 (sample space) |     |
| $\emptyset$ | 不可能事件，即空集 (empty set)       |     |



续表

| 符号                                                                                                             | 解释                                                                                                                               | 维恩图 |
|----------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|-----|
| $A \subset B$                                                                                                  | 事件 $B$ 包含事件 $A$ (event $A$ is a subset of event $B$ ),<br>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 $B$ ),<br>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                |     |
| $A \cap B = \emptyset$                                                                                         | 事件 $A$ 和事件 $B$ 互斥 (events $A$ and $B$ are disjoint), 两个<br>事件互不相容 (two events are mutually exclusive),<br>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 $B$ ),<br>即当事件 $A$ 和事件 $B$ 至少有一个发生时, 事件 $A \cup B$<br>发生                        |     |
| $A - B$                                                                                                        | 事件 $A$ 与事件 $B$ 的差事件 (the difference between two<br>events $A$ and $B$ ),<br>即事件 $A$ 发生、事件 $B$ 不发生, 则 $A - B$ 发生                  |     |
| $A \cup B = \Omega$ 且 $A \cap B = \emptyset$<br>也可以记作 $\bar{A} = B = \Omega - A$<br>(complement of event $A$ ) | 事件 $A$ 与事件 $B$ 互为逆事件 (complementary events),<br>对立事件 (collectively exhaustive),<br>即对于任意一次试验, 事件 $A$ 和事件 $B$ 有且仅有一个<br>发生        |     |

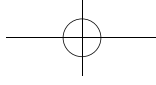
## 3.5 条件概率：给定部分信息做推断

**条件概率** (conditional probability) 是在给定部分信息的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的概念，本书大多数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哪里会用到“条件概率”。

在抛两个骰子试验中，事件 $A$ 为其中一个骰子点数为5，事件 $B$ 为点数之和为6。给定事件 $B$ 发生的条件下，事件 $A$ 发生的概率为多少？



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 (setosa、versicolor、virginica)? 对应的概率大概是多少?

根据150个鸢尾花样本数据，如果某一朵鸢尾花的花萼长度为5厘米，它的花萼宽度最可能为多宽?

## 条件概率

$A$ 和 $B$ 为样本空间 $\Omega$ 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，即

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (3.20)$$

其中： $\Pr(A \cap B)$ 为事件 $A$ 和 $B$ 的联合概率； $\Pr(B)$ 为事件 $B$ 的边缘概率。

$\Pr(B)$ 、 $\Pr(A \cap B)$ 都是在 $\Omega$ 中计算得到的概率值。



注意：我们也可以这样理解 $\Pr(A|B)$ ， $B$ 实际上是“新的样本空间”—— $\Omega_B$ ！  
 $\Pr(A|B)$ 是在 $\Omega_B$ 中计算得到的概率值。

$\Omega_B$ 是 $\Omega$ 的子集，两者的联系正是 $\Pr(B)$ ，即 $B$ 在 $\Omega$ 中对应的概率。 $\Pr(B)$ 也可以写成“条件概率”的形式，即 $\Pr(B|\Omega)$ 。

类似地，事件 $A$ 发生的条件下事件 $B$ 发生的条件概率为

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \quad (3.21)$$

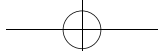
其中： $\Pr(A)$ 为 $A$ 事件边缘概率， $\Pr(A) > 0$ 。

类似地， $\Pr(B|A)$ 也可以理解为 $B$ 在“新的样本空间” $\Omega_A$ 中的概率。

## 联合概率

利用，联合概率 $\Pr(A \cap B)$ 可以整理为

$$\Pr(A \cap B) = \Pr(A, B) = \Pr(A|B) \cdot \Pr(B) \quad (3.22)$$



式 (3.22) 相当于“套娃”。首先在  $\Omega_B$  中考虑  $A$  (实际上是  $A \cap B$ )，然后把  $A \cap B$  再放回  $\Omega$  中。也就是说，把  $\Pr(A|B)$  写成  $\Pr(A \cap B|B)$  也没问题。因为， $A$  只有  $A \cap B$  这部分在  $B (\Omega_B)$  中。

同样， $\Pr(A \cap B)$  也可以写成

$$\underbrace{\Pr(A \cap B)}_{\text{Joint}} = \underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(B|A)}_{\text{Conditional}} \underbrace{\Pr(A)}_{\text{Marginal}} \quad (3.23)$$

## 举个例子

掷一颗骰子，一共有六种等概率结果，即  $\Omega = \{1, 2, 3, 4, 5, 6\}$ 。

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3.19所示，事件  $B \cap C$  发生的概率  $\Pr(B \cap C) = \Pr(B, C) = 1/3$ 。

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

$$\Pr(C|B) = \frac{\Pr(B \cap C)}{\Pr(B)} = \frac{\Pr(B, C)}{\Pr(B)} = \frac{1/3}{1/2} = \frac{2}{3} \quad (3.24)$$

图3.19所示也告诉我们一样的结果。请大家回顾本章最初给出的孟德尔豌豆试验和道尔顿红绿色盲，手算其中的条件概率。

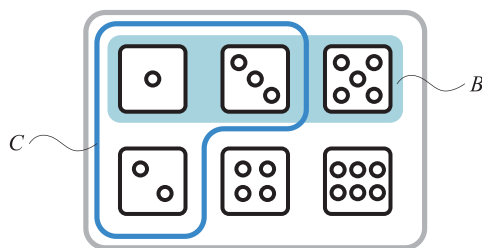


图3.19 事件  $B$  发生条件下事件  $C$  发生的条件概率

## 推广

式 (3.22) 可以继续推广， $A_1, A_2, \dots, A_n$  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，即

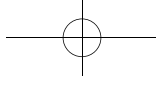
$$\begin{aligned} \Pr(A_1 \cap A_2 \cap \dots \cap A_n) &= \Pr(A_1, A_2, A_3, \dots, A_{n-1}, A_n) \\ &= \Pr(A_n | A_1, A_2, A_3, \dots, A_{n-1}) \Pr(A_{n-1} | A_1, A_2, A_3, \dots, A_{n-2}) \dots \Pr(A_2 | A_1) \Pr(A_1) \end{aligned} \quad (3.25)$$

式 (3.25) 也叫作条件概率的**链式法则** (chain rule)。

比如， $n = 4$  时，式 (3.25) 可以写成

$$\begin{aligned} \underbrace{\Pr(A_1, A_2, A_3, A_4)}_{\text{Joint}} &= \underbrace{\Pr(A_4 | A_1, A_2, A_3)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_1, A_2, A_3)}_{\text{Joint}} \\ &= \underbrace{\Pr(A_4 | A_1, A_2, A_3)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_3 | A_1, A_2)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_1, A_2)}_{\text{Joint}} \\ &= \underbrace{\Pr(A_4 | A_1, A_2, A_3)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_3 | A_1, A_2)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_2 | A_1)}_{\text{Conditional}} \Pr(A_1) \end{aligned} \quad (3.26)$$





大家可以把式(3.26)想成多层套娃。式(3.26)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，即

$$(A_1 \cap A_2 \cap A_3 \cap A_4) \subset (A_1 \cap A_2 \cap A_3) \subset (A_1 \cap A_2) \subset A_1 \quad (3.27)$$

## 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) 提出的。毫不夸张地说，贝叶斯定理撑起机器学习、深度学习算法的半天天。



本书后续内容将见缝插针地讲解贝叶斯定理和应用，特别是在贝叶斯分类(第18、19两章)、贝叶斯推断(第20、21、22三章)这两个主题中。

贝叶斯定理的基本思想是根据**先验概率** (prior) 和新的**证据** (evidence) 来计算**后验概率** (posterior)。在实际应用中，我们通常根据一些已知的先验知识，来计算事件的先验概率。然后，当我们获取新的证据时，就可以利用贝叶斯定理来计算事件的后验概率，从而更新我们的信念或概率。



托马斯·贝叶斯 (Thomas Bayes) | 英国数学家 | 1702—1761年  
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 • 贝叶斯派 • 贝叶斯推断 • 朴素贝叶斯分类 • 贝叶斯回归

贝叶斯定理描述的是两个条件概率的关系，如

$$\underbrace{\Pr(A|B)}_{\text{Conditional}} \underbrace{\Pr(B)}_{\text{Marginal}} = \underbrace{\Pr(B|A)}_{\text{Conditional}} \underbrace{\Pr(A)}_{\text{Marginal}} = \underbrace{\Pr(A \cap B)}_{\text{Joint}} = \underbrace{\Pr(A, B)}_{\text{Joint}} \quad (3.28)$$