

作者	生姜 DrGinger
脚本	生姜 DrGinger
视频	崔崔 CuiCui
开源学习资源	https://github.com/Visualize-ML
平台	https://www.youtube.com/@DrGinger_Jiang https://space.bilibili.com/3546865719052873 https://space.bilibili.com/513194466

01

Classic Probability Model

古典概率模型

计算可能性的大小

当我们在日常生活中思考某个(未来)事件是否会发生时，最关心的往往就是这个事件发生的可能性有多大。

就好比我们用尺子来测量桌子的长度，用温度计来测量天气的冷暖，概率就是用来测量一个未来事件发生的可能性大小的数学“工具”。

比如，我们想知道明天会不会下雨，买的彩票中奖的可能性，某只股票涨跌趋势，或者早上出门能不能及时赶上公交车，这些都是在无形中对未来事件的可能性进行“估量”。

当我们把概率应用到这些待测的事件上，就能得到一个数字化的测量值，用来表达该事件发生的可能性有多大——它可以很接近 0 (几乎不可能)，也可以很接近 1 (几乎必然)。

本章让我们从古典概率模型入手，试着量化可能性。

⚠ 注意，开始本章学习之前，特别请大家复习本册附录中有关基础数学相关内容，特别是集合运算。

1.1 随机试验



本节你将掌握的核心技能：

- ▶ 随机试验三个特征：可重复、结果明确但不可预测。
- ▶ 每一个可能结果(样本点)构成所有结果的集合(样本空间)。
- ▶ 用平面坐标图、三维坐标图、树形图等可视化样本空间结构。
- ▶ 用 Python 编程生成随机数、模拟随机试验、绘图展示模拟结果。

本节将介绍随机、随机试验、样本点、样本空间这些基本概率概念。很多概率统计概念都显得特别抽象，加上“不知所云”的符号公式，特别容易让人觉得枯燥。为了帮助大家理解掌握各种概率统计概念，尤其针对多元统计，本书中我们会看到丰富的可视化方案。本节中，请大家格外注意，我们是如何利用几何图形可视化不同随机试验的样本点、样本空间。

随机、随机试验

随机 (randomness) 指的是一种结果不可预测的状态。比如，抛硬币正面或反面朝上，掷色子得到几点，明天晴天还是雨天，这些都带有随机性。

随机试验 (random experiment) 是指在相同条件下可以重复进行，每次试验的结果不确定，但所有可能结果事先已知，且每次试验只有一个结果出现的试验。

值得强调的是随机试验的主要性质：

- ▶ **可重复性**：随机试验可以在相同条件下重复进行任意次。
- ▶ **明确性**：随机试验的所有可能结果在试验开始前都是明确的。
- ▶ **不确定性**：每次随机试验的结果是随机的，在单次随机试验前无法预测具体会发生哪个结果。

举两个例子来描述以上性质。

如图 1 所示，抛一枚硬币是一个随机试验，可以在相同条件下重复进行 (**可重复性**)，结果只有反面 (tail)、正面 (head) 两种可能 (**明确性**)；但是，某次抛硬币之前，我们不可能明确知道具体哪一面朝上 (**不确定性**)。

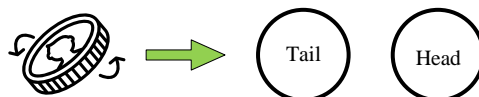


图 1. 抛一枚硬币

如图 2 所示，掷一颗骰子也是一个随机试验。可以在相同条件下重复掷骰子 (**可重复性**)，结果点数有 1、2、3、4、5、6 六种可能 (**明确性**)。但是，某次掷一颗骰子，具体得到哪个点数，我们事先肯定无法得知 (**不确定性**)。

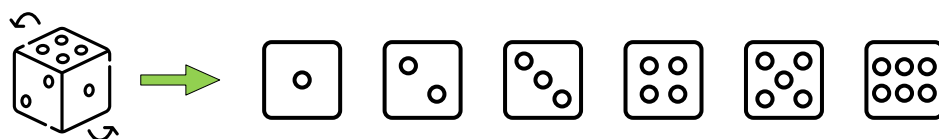


图 2. 掷一颗色子

样本空间、样本点

样本点 (sample point) 是某个随机试验的一个可能结果，每一个样本点都表示试验中某种具体情况的发生。

比如，抛一枚硬币，出现“正面”就是一个样本点，“反面”也是一个样本点。再如，掷一颗色子，点数 1、2、3、4、5、6 分别都是样本点。

样本空间 (sample space) 是指某个随机试验中所有可能结果 (即所有**样本点**) 的集合；样本空间集合常用 Ω 表示。样本空间完整地列出了在该随机试验下可能出现的每一种结果。

如图 3 所示，样本点的集合 $\Omega = \{\text{反面}, \text{正面}\}$ 为抛一枚硬币这个随机试验的**样本空间**。

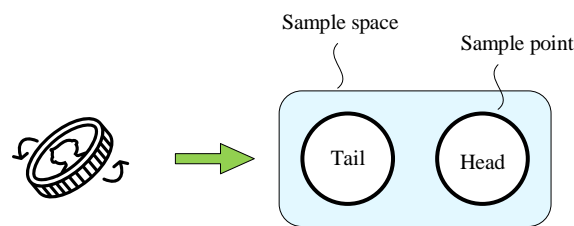


图 3. 抛一枚硬币的样本空间

如果把“反面”记作 0，“正面”记作 1，抛一枚硬币这个随机试验的样本空间是

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

显然这个集合的元素个数为 2，即基数 $\text{card}(\{0, 1\}) = 2$ ；也就是说，这个随机试验有两种可能结果，即样本点为 2。

+ 集合基数 (cardinality) 是指一个集合中元素的个数，用于衡量集合的大小。对于有限集合，基数就是集合中元素的数量；对于无限集合，基数则需要通过集合论中的概念来定义，用来比较不同大小的无限集合。

! 注意，不同随机试验有不同的样本空间，虽然样本空间都可以记作 Ω 。样本空间的样本点可以是有限个 (比如抛硬币、掷色子)，也可以是可数无限个 (比如所有正整数)；样本空间的样本点，可以是不可数无限个，比如等待公交车的时间 (从 0 到 10 分钟)。

代码 1 模拟单次抛硬币这一随机试验，共进行 10 次，每次独立生成正面或反面结果，最后输出全部结果序列。下面聊聊其中关键语句。

a 导入一个名为 random 的模块。Python 中的模块就像是一个装满工具的箱子，random 这个模块提供了生成随机数、随机选择元素等功能。只要导入它，我们就可以使用其中的函数，比如 random.choice() 来随机选择列表里的内容。

b 创建了一个空列表，用来存储每次抛硬币的结果。列表就像一个可以装很多东西的容器，你可以往里面不断添加新数据。此时 results 是空的，等下我们会把每次的“正面”或“反面”结果放进去。

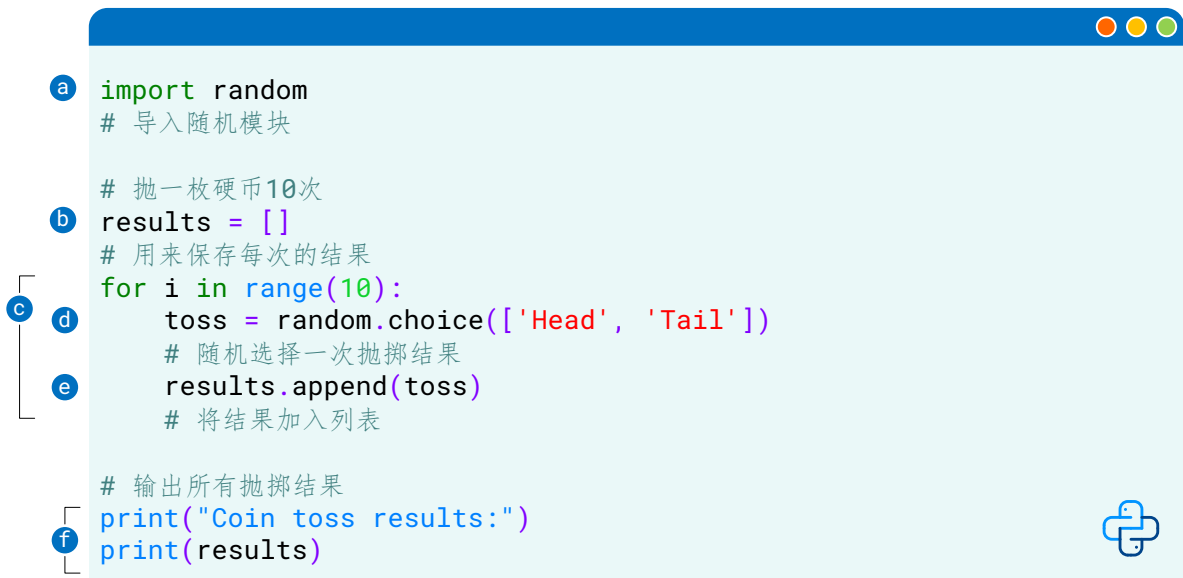
c 表示要重复执行接下来的代码 10 次。`range(10)` 会生成从 0 到 9 的一串数字（总共有 10 个数），`for i in range(10)` 就是让变量 `i` 依次取这 10 个值。虽然这里我们并不直接用 `i`，但它帮助我们循环执行 10 次抛硬币的操作。

d 这一行是代码的核心部分。`random.choice()` 是随机模块里的一个函数，它的作用是从给定的序列（如列表）中随机选取一个元素。这里我们给它的列表是 `['Head', 'Tail']`，也就是“正面”和“反面”两个可能的结果。程序运行时，它会随机返回其中一个，并把结果存入变量 `toss`。

e 把刚刚生成的那次抛掷结果加入列表 `results` 中。`.append()` 是列表的一个内置方法，用来在列表末尾添加新元素。例如，如果原来列表是 `['Head']`，调用 `append('Tail')` 后，它就变成 `['Head', 'Tail']`。执行 10 次循环后，`results` 就会保存 10 次抛硬币的全部结果。

f 中 `print()` 是 Python 的打印函数，用来把内容显示在屏幕上。

代码 1. 一枚硬币抛 10 次 |  PS_Ch01_01_01.ipynb



```

a import random
# 导入随机模块

# 抛一枚硬币10次
b results = []
# 用来保存每次的结果
c for i in range(10):
d     toss = random.choice(['Head', 'Tail'])
# 随机选择一次抛掷结果
e     results.append(toss)
# 将结果加入列表

# 输出所有抛掷结果
f print("Coin toss results:")
print(results)
  
```

为了方便可视化，如图 4 所示，我们用 1 (蓝色圆点) 代表正面，用 0 (红色圆点) 代表反面，灰色连线为了视觉上连贯。每个子图的结果虽然调用同一个随机数发生器，但是由于随机种子不固定，因此子图最终呈现出彼此各异的试验结果。

具体到某组 10 次试验，可能连续出现多个正面，也可能正面远少于反面，这种波动正是随机性的固有特征。10 幅子图对比观察发现这些结果恰好呈现了有限样本下的多样性：没有两组图案完全一致，有些组正反相当，有些组明显偏向一侧。从这些结果中，大家可以回顾本节前文提到的随机试验的三个主要性质：1) 随机试验可重复性；2) 所有可能结果明确；3) 每次结果不确定性。

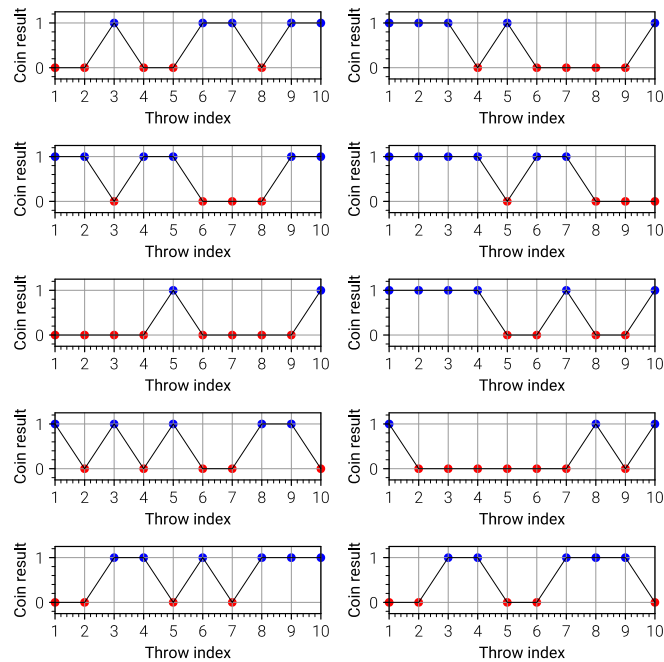


图 4.10 组 10 次抛单枚硬币结果

抛两枚硬币

一个随机试验连续抛掷两枚硬币，那么样本空间 Ω 就是 {(反面,反面), (正面,反面), (反面,正面), (正面,正面)}，对应

$$\Omega = \left\{ \begin{pmatrix} 0,1 \\ 0,0 \end{pmatrix}, \begin{pmatrix} 1,1 \\ 1,0 \end{pmatrix} \right\} \quad (2)$$

这个样本空间有 4 个样本点，即基数 $\text{card}(\Omega) = 4$ ；具体来说，连续抛掷两次硬币的 4 个不同的结果。(2) 相当于计算了笛卡儿积 (Cartesian product)。

? 请大家试着使用 `itertools.product()` 计算两个列表 (list) 的笛卡儿积。

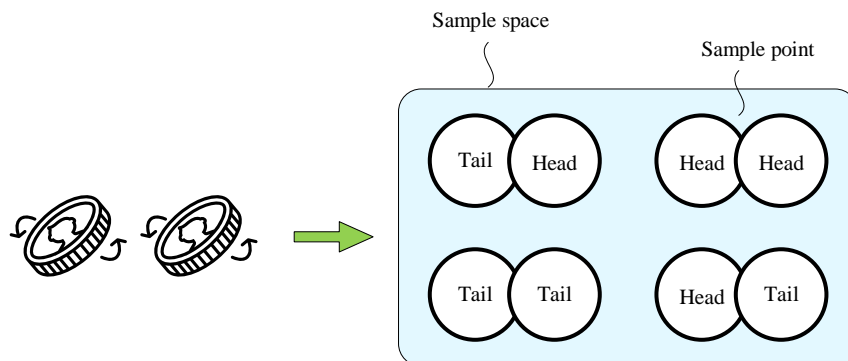


图 5. 连续抛两枚硬币的样本空间

我们可以把“连续抛两枚硬币”的所有可能结果用平面坐标系中的点来表示。如图 6 所示，**反面** (tail) 记作 0，**正面** (head) 记作 1。抛第一枚硬币的结果作为横坐标，抛第二枚硬币的结果作为纵坐标。

这样，样本空间中所有可能的结果就可以表示为以下四个坐标点：

- ▶ (0, 0)：第一枚反面，第二枚反面；
- ▶ (0, 1)：第一枚反面，第二枚正面；
- ▶ (1, 0)：第一枚正面，第二枚反面；
- ▶ (1, 1)：第一枚正面，第二枚正面。

这四个点就构成了整个样本空间，可以在平面坐标系中画出一个 2×2 的方格，帮助我们直观理解每种可能的结果。

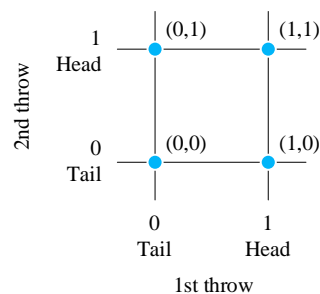


图 6. 连续抛两枚硬币的样本空间，平面坐标系



PS_Ch01_01_02.ipynb 模拟一次抛两枚硬币，连续抛 10 次。

树形图

我们还可以用**树形图** (dendrogram) 来展示随机试验的样本空间。

树形图是一种用来表示分层结构或逐步分支过程的图形。它从一个起点 (通常称为“根”) 开始，按照一定的规则向外延伸出多个分支，每个分支点代表一种可能的选择或结果。随着每一次分支的进行，图形逐层展开，最终形成一个类似树状结构的图形。下面让我们用树形图展示概率试验中的样本空间，以及获得特定样本点的路径。

图 7 展示“连续抛一枚硬币 4 次”这个随机试验。每个节点代表一次抛掷的状态，左边分支代表正面 (1)，右边分支代表反面 (0)。整棵树从根节点 (表示开始) 出发，向下延伸，每一层表示一次新的抛掷，直到达到设定的次数。

⚠ 注意区分：这里讨论的随机试验是“连续抛一枚硬币 4 次”，它本身是一次试验，样本空间包含 16 个样本点 (如“正正正正”)。而“抛一枚硬币，重复进行 4 次”则意味着将基本试验 (抛一枚硬币) 独立做了 4 遍，两者在概念上完全不同。

如图 7 所示，具体来说：

- ▶ 根节点是“开始”状态，还未抛掷。
- ▶ 第 1 层：第 1 次抛硬币，对应 $2 (=2^1)$ 结果。
- ▶ 第 2 层：第 2 次抛硬币，对应 $4 (=2^2)$ 结果。
- ▶ 第 3 层：第 3 次抛硬币，对应 $8 (=2^3)$ 结果。
- ▶ 第 4 层：第 4 次抛硬币，对应 $16 (=2^4)$ 结果，即“连续抛一枚硬币 4 次”这个随机试验样本空间大小。

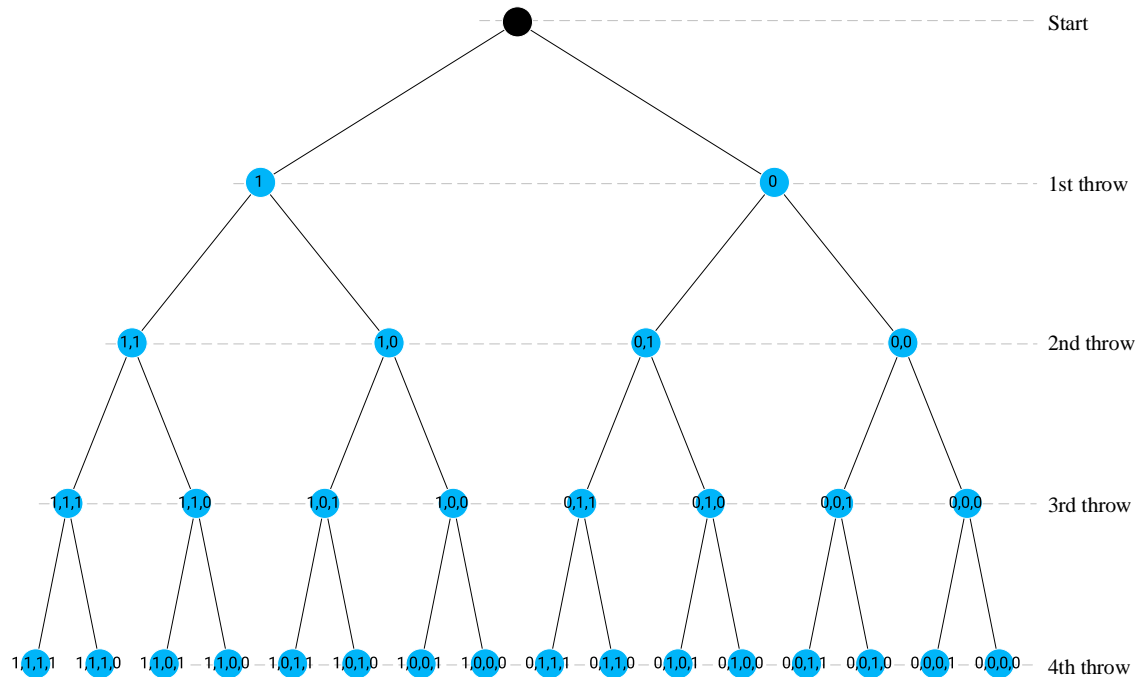


图 7. 连续抛四枚硬币的样本空间，二叉树

图 7 是否让大家想到了**二项式定理** (binomial theorem)?

二项式定理描述了二项式幂的代数展开。比如下式

$$(x + y)^4 = 1x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + 1y^4 \quad (3)$$

? 当 $n = 2, 3, 4, 5$ 时，请大家展开 $(x + y)^n$ 。

为了帮助大家建立图 7 二叉树和 (3) 的联系，让我们把 x 换成 head，把 y 换成 tail，把上式写成

$$(\text{head} + \text{tail})^4 = 1 \times \text{head}^4 + 4 \times \text{head}^3 \times \text{tail} + 6 \times \text{head}^2 \times \text{tail}^2 + 4 \times \text{head} \times \text{tail}^3 + 1 \times \text{tail}^4 \quad (4)$$

让我们逐项分析上式：

- ▶ $1 \times \text{head}^4$ ：4 次抛硬币结果都是“正面”的样本点有 1 个，如图 8 所示；
- ▶ $4 \times \text{head}^3 \times \text{tail}$ ：3 个“正面”、1 个“反面”的样本点有 4 个，如图 9 所示；
- ▶ $6 \times \text{head}^2 \times \text{tail}^2$ ：2 个“正面”、2 个“反面”的样本点有 6 个，如图 10 所示；

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

► $4 \times \text{head} \times \text{tail}^3$ ：3 个“正面”、1 个“反面”的样本点有 4 个，请大家自行画出；

► $1 \times \text{tail}^4$ ：4 个“反面”的样本点有 4 个，请大家自行画出。

抛硬币 4 次，所有可能的结果共有 16 种，从“正正正正”到“反反反反”一应俱全。其中，“恰好 4 次都是正面”这一事件，对应的样本点只有 1 个，即全正面序列。

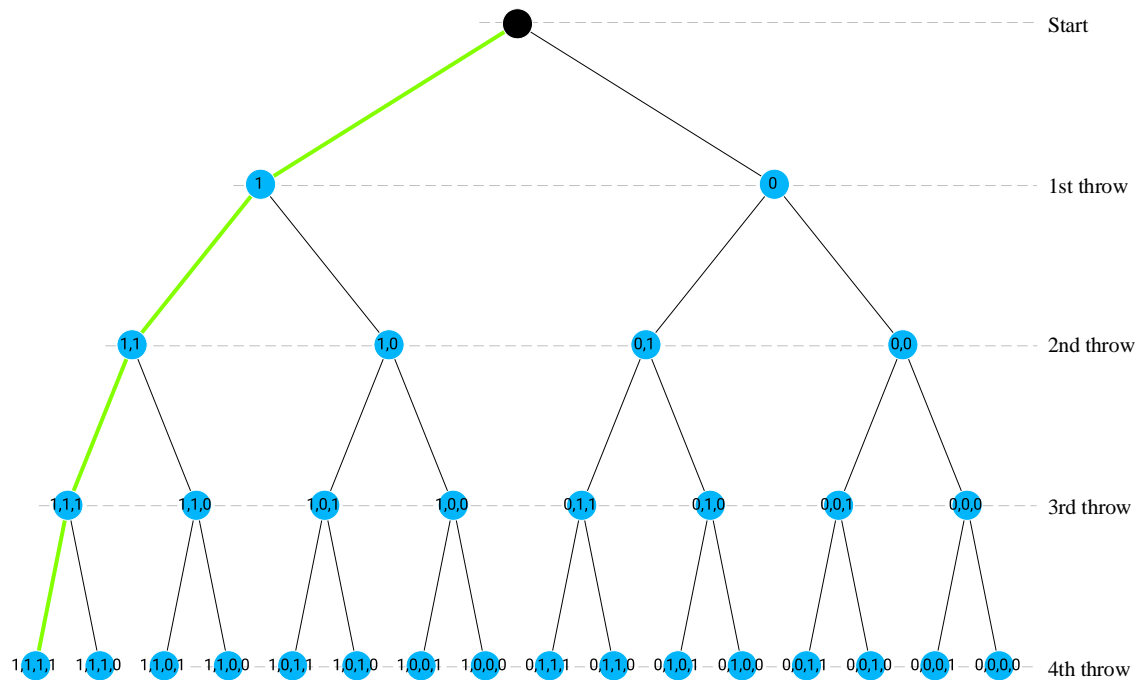


图 8. 连续抛四枚硬币，四枚硬币结果都是正面

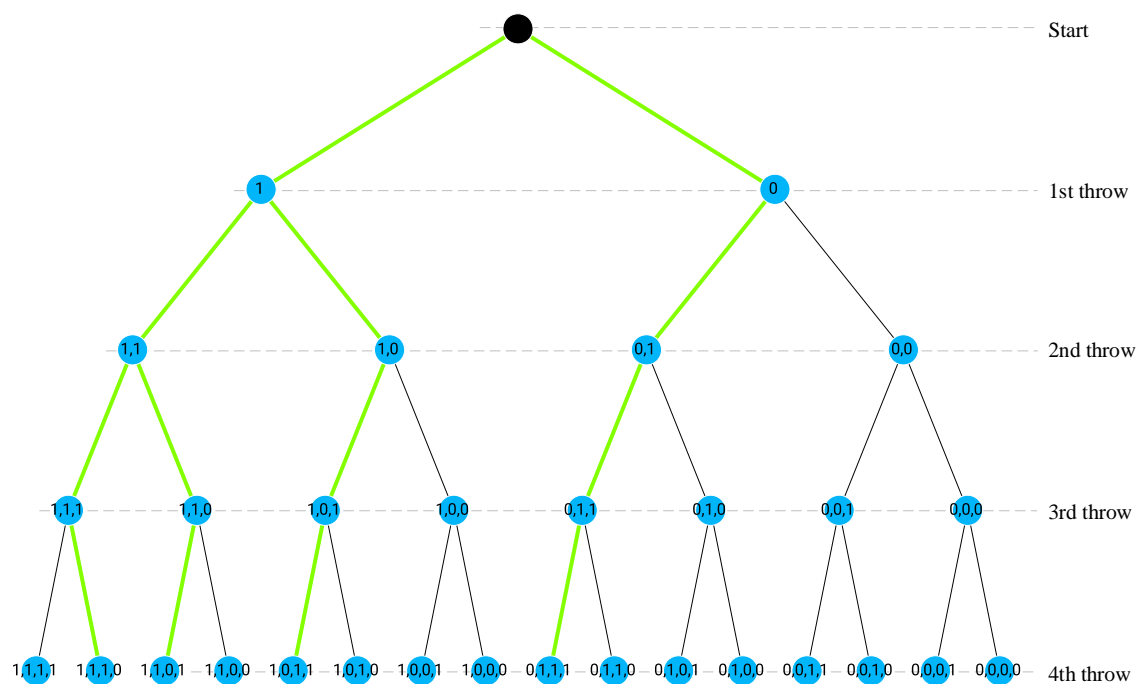


图 9. 连续抛四枚硬币，三枚硬币结果都是正面

而“恰好 2 次正面”则对应 6 个不同的样本点。为什么会是 6？因为要从 4 个位置中选出 2 个来放正面，其余放反面。这个“选法有多少种”就是组合数，记作

$$C_4^2 \quad (5)$$


组合数的计算公式是：

$$C_n^k = \frac{n!}{k!(n-k)!} \quad (6)$$

注意， n 、 k 均为正整数， $n \geq k$ 。其中 $n!$ 表示从 1 乘到 n 的阶乘，即

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 \quad (7)$$

特别约定 $0! = 1$ 。

 请大家自行计算 C_4^0 、 C_4^1 、 C_4^2 、 C_4^3 、 C_4^4 ，并且试着在图 7 中分别找到这些组合数分别对应路径。

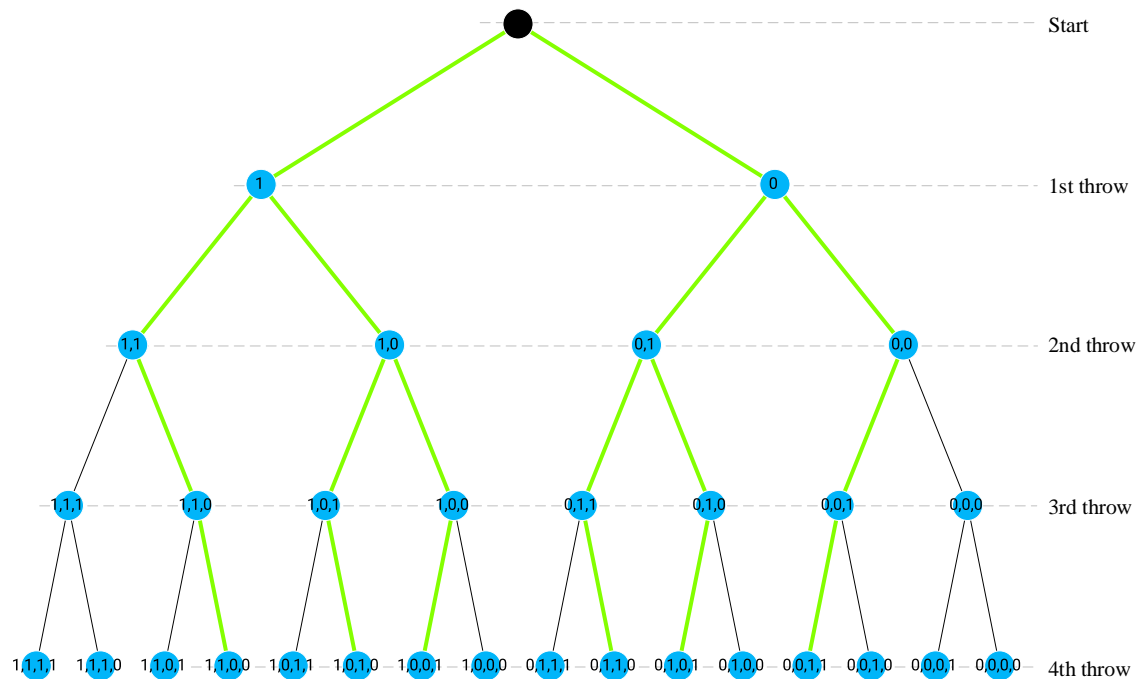


图 10. 连续抛四枚硬币，两枚硬币结果为正面


图 11 所示为用**环形树形图** (circular dendrogram) 可视化连续抛 4 枚硬币随机试验的样本空间，以及结果路径。

如图 11 所示，从 start 开始，我们看到第一层 2 条分叉，每一条分叉代表第一枚硬币的一个结果 (head 或 tail)。

然后，从每一条分叉的末端，再各自延伸出 2 条第二层分叉，代表第二枚色子的结果，同样是 head 或 tail。依次原理，随着抛硬币数量不断增多，整个树形图不断“开枝散叶”。

整个图形呈放射状展开，就像一棵环状的“样本树”，每一个叶子节点都代表一个具体的结果，也就是一个样本点。每一条从根到叶的路径，都是一次完整的掷两颗色子的试验结果。图 11 特别用粗黑线展示了四枚硬币都是正面的路径。

图 11 既清晰地表达了样本空间的结构，也帮助我们理解每一个样本点是如何由一次试验自然生成的。

 请大家在图 11 上再绘制一层“枝叶”，用来代表第五枚色子的结果。

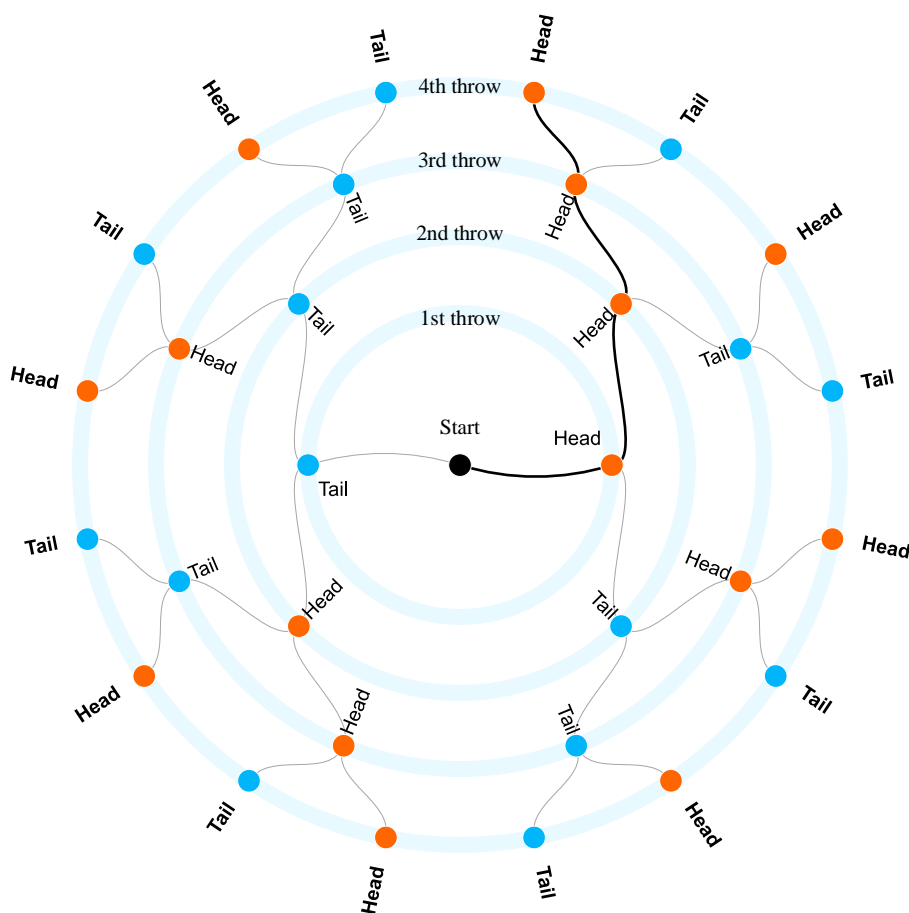


图 11. 连续抛 4 枚硬币的样本空间，环形树形图

连续掷色子

掷一颗色子的样本空间是

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad (8)$$

如图 12 所示，这个样本空间有 6 个样本点，即 $\text{card}(\Omega) = 6$ 。

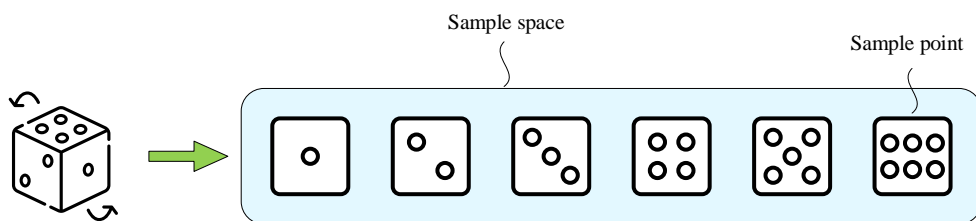


图 12. 掷一颗色子的样本空间



PS_Ch01_01_03.ipynb 模拟一次抛一颗色子，连续抛 10 次。

当我们掷两颗色子时，每一颗色子都有 6 种可能的结果，从 1 到 6。这样一来，掷两颗色子 (这个随机试验本身) 所有可能的组合就是 $6 \times 6 = 36$ 种，也就是说，样本空间包含 36 个样本点。

举个例子，如果第一颗色子点数为 6，第二颗色子点数也是 6，我们把这个样本点写成 (6, 6)；这样，掷两颗色子这个随机试验的样本空间是

$$\Omega = \left\{ \begin{array}{cccccc} (1,6) & (2,6) & (3,6) & (4,6) & (5,6) & (6,6) \\ (1,5) & (2,5) & (3,5) & (4,5) & (5,5) & (6,5) \\ (1,4) & (2,4) & (3,4) & (4,4) & (5,4) & (6,4) \\ (1,3) & (2,3) & (3,3) & (4,3) & (5,3) & (6,3) \\ (1,2) & (2,2) & (3,2) & (4,2) & (5,2) & (6,2) \\ (1,1) & (2,1) & (3,1) & (4,1) & (5,1) & (6,1) \end{array} \right\} \quad (9)$$

这个样本空间对应的集合基数为即 $\text{card}(\Omega) = 36$ 。图 13 所示为平面直角坐标系中看这个样本空间。横轴代表第一颗色子的点数，纵轴代表第二颗色子的点数。这样，每个样本点就在一个坐标格点上，组成一个 6 行 6 列的正方形网格，图像直观地展现了样本空间的结构。

比如，坐标点 (1, 1) 表示两颗色子的点数均为 1；坐标点 (6, 6) 两颗色子的点数均为 6。

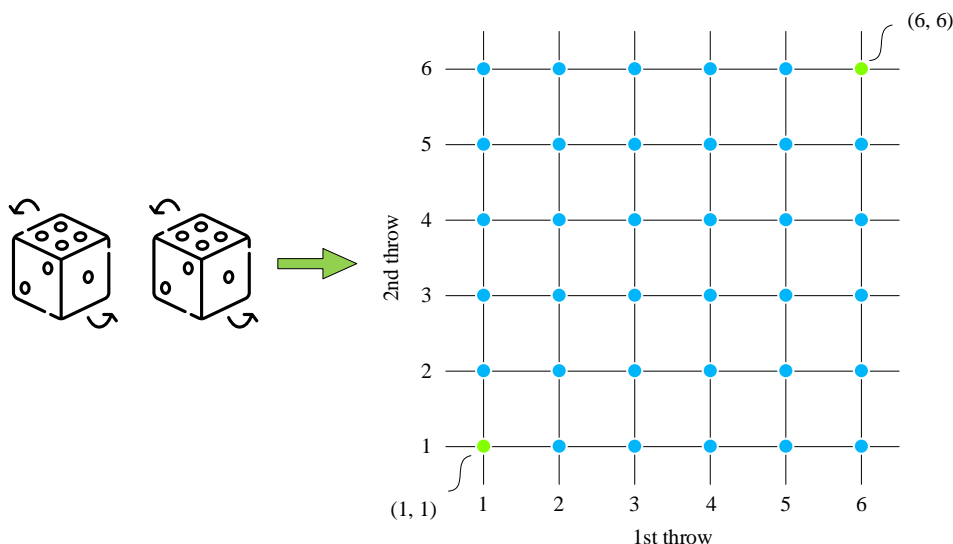


图 13. 掷两颗色子的样本空间，平面直角坐标系



PS_Ch01_01_04.ipynb 模拟一次抛两颗色子，连续抛 10 次。

我们也可以用环形树形图来形象地表示这个样本空间。如图 14 所示，从 start 开始，我们看到第一层 6 条分叉，每一条分叉代表第一颗色子的一个结果 (1 到 6)。

然后，从每一条分叉的末端，再各自延伸出 6 条第二层分叉，代表第二颗色子的结果，同样是 1 到 6。这样，我们便获得掷两颗色子 (随机试验) 的样本空间。

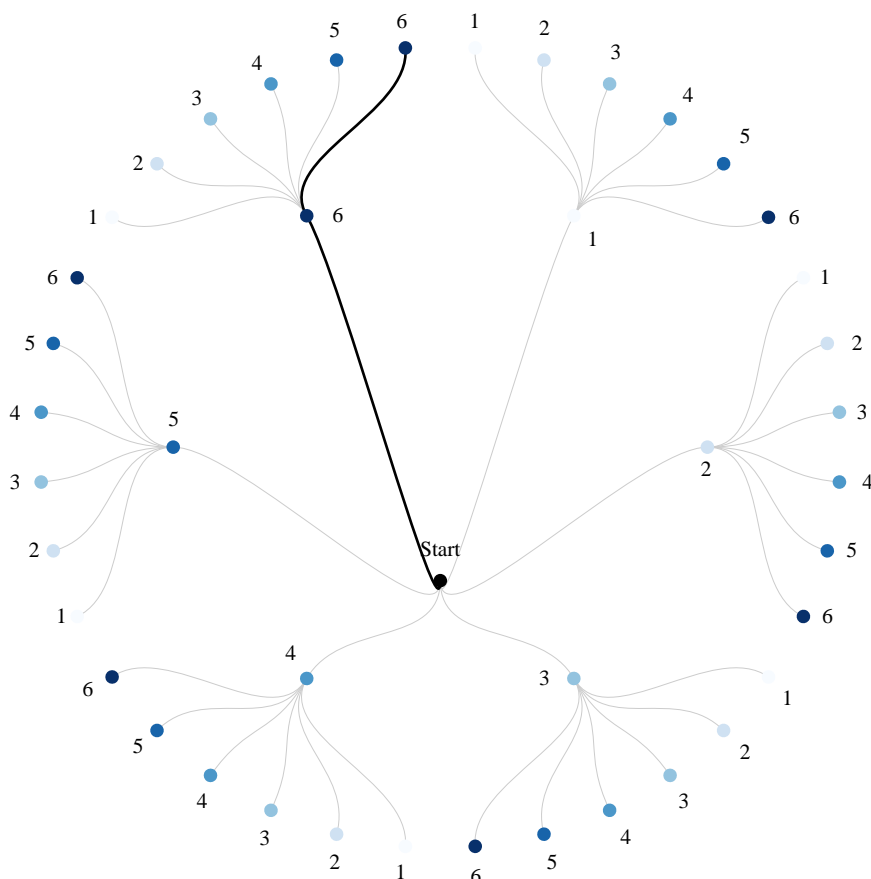


图 14. 掷两颗色子的样本空间，环形树形图

进一步，掷三颗色子（随机试验）让样本空间变得更大。掷三颗色子结果对应的样本空间一共有 216 ($6 \times 6 \times 6$) 个样本点。

如图 15 所示，在三维直角坐标系中，把第一颗色子的点数看作 x_1 轴上的坐标，第二颗色子的结果作为 x_2 轴上的坐标，第三颗色子的点数作为 x_3 轴上的坐标，那么这个随机试验所有样本点就对应于立方体中的规则分布的散点。

也就是说，掷三颗色子的每一个样本点，比如 (6, 6, 6) 表示三颗色子每颗色子都掷出了 6 点，是样本空间中的一个坐标点。

⚠ 在掷三颗色子的随机试验中，我们虽然在三维坐标系中画出了样本空间，但要注意，这个样本空间并不是充满整个立方体的实心体积，而只是立方体中分布着的 216 个有规律的点。这些点的坐标都是由三个 1~6 整数构成，比如 (6, 6, 6)，分别代表三颗色子的点数结果。因此，这种样本空间是离散的。但是，本书后续大家会发现，有些随机试验中，可能结果对应连续的量，比如时间、温度、身高、长度等等。离散样本空间 (discrete sample space) 是由有限个或可数无限个散点组成的集合；例如，掷两颗色子的所有点数组组合，共有 36 个离散的样本点。而连续样本空间 (continuous sample space) 则像一个实心的几何体。

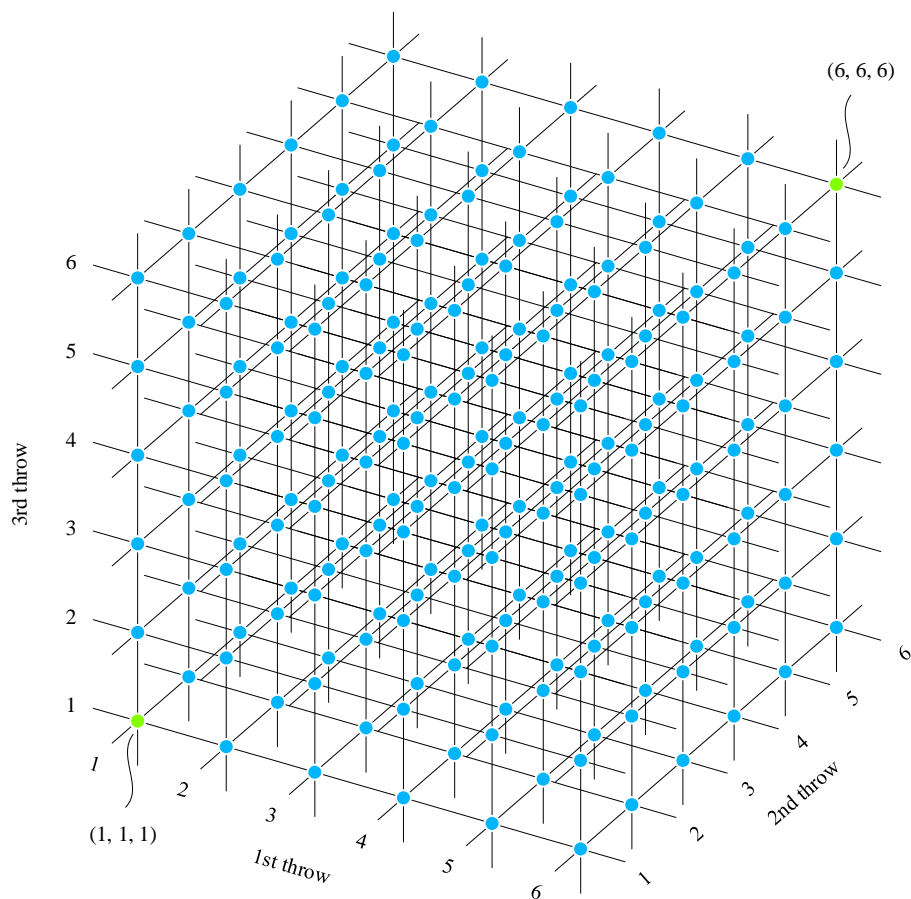


图 15. 掷三颗色子的样本空间，三维直角坐标系

图 16 所示为用环形树形图展示掷三颗色子的样本空间。这棵“树”可以看作是图 14 继续生长一层枝叶的结果。

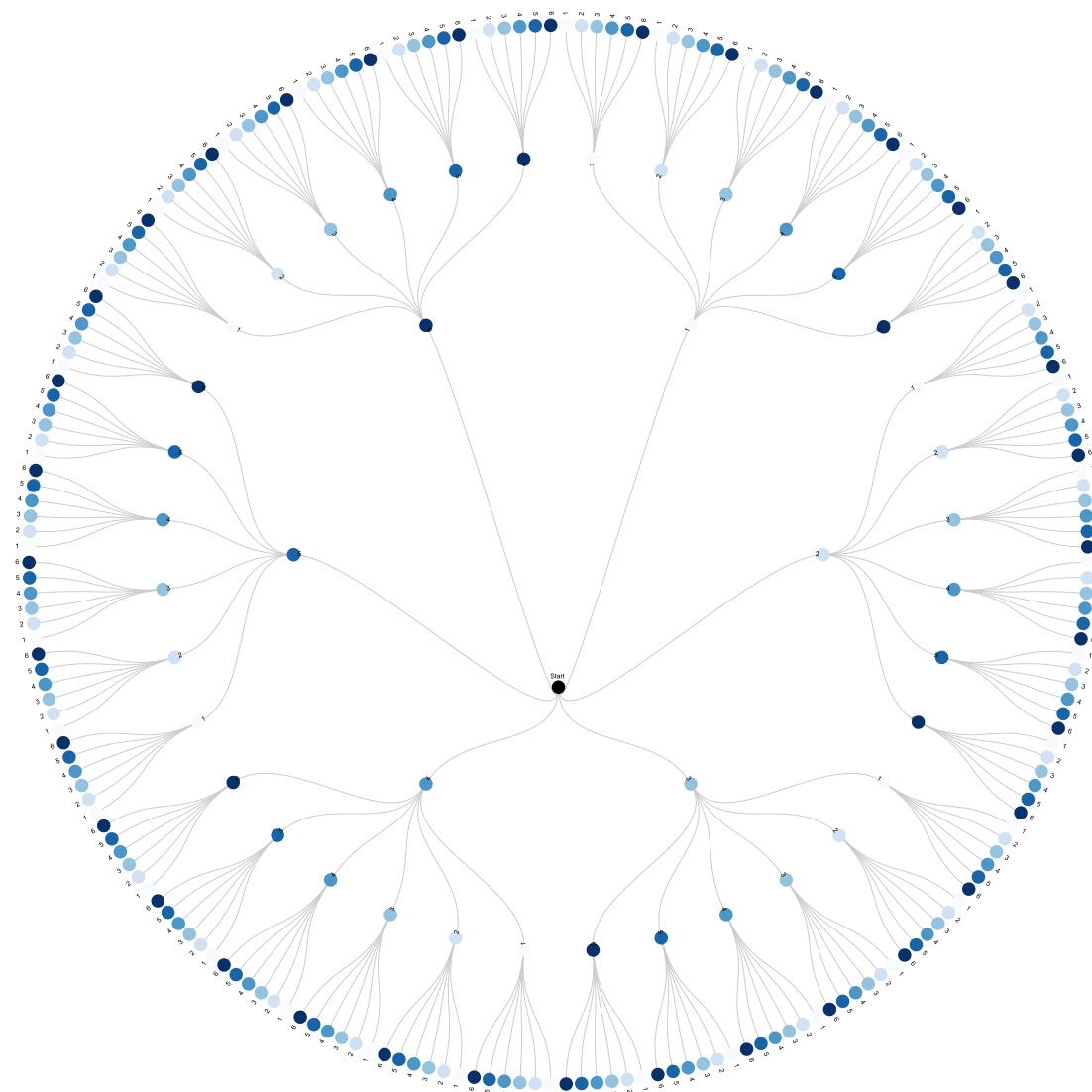


图 16. 掷三颗色子的样本空间，环形图形图

反过来看，把图 16 最外层的枝叶全部剪去；也就是说，我们不再关心第三颗色子的结果。图 16 这棵树就变成了图 14 这个“树”。

同样角度看图 15，如果同样不再关心第三颗色子的结果，这个立方体散点便降维到平面散点 (如图 17)，相当于一个维度被“折叠”，仿佛从未发生过。图 17 为掷两颗色子这个随机试验的样本空间。

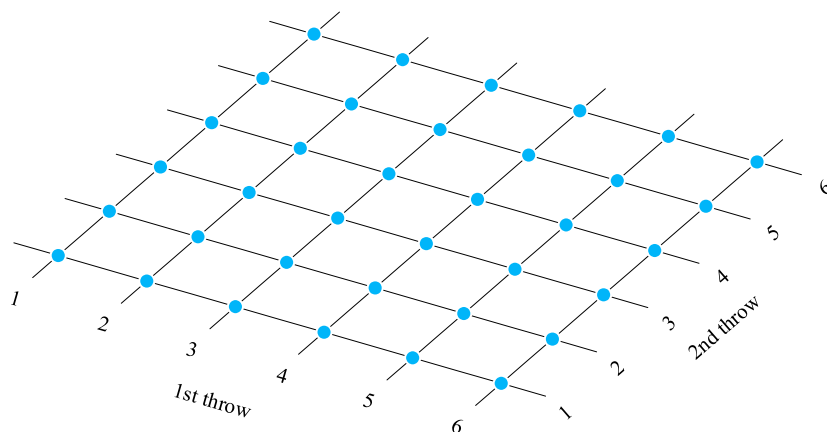


图 17. 不再关心第三颗色子结果

再进一步，如果我们也不再关心第二颗掷色子的结果，平面散点便被折叠成如图 18 所示的一维散点。图 18 就是掷一颗色子这个随机试验的样本空间。

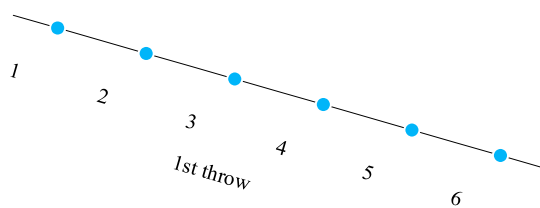


图 18. 不再关心第二、三颗色子结果

通过以上几个例子，不同的随机试验不仅试验过程不同，它们的样本空间的结构和大小也会不同。而且展示样本空间的可视化方案也丰富多样。不管怎样，理解样本空间，特别是能够在头脑中形成不同样本空间的几何形状，是后面计算概率的基础。



请大家用 DeepSeek/ChatGPT 等工具完成本节如下习题。

Q1. 把一枚硬币的正反面分别涂色，正面为红色，反面为蓝色。抛一枚硬币的样本空间是什么？连续抛两枚硬币的样本空间是什么？

Q2. 图 13 所示为用平面直角坐标系展示掷两颗色子随机试验样本空间，请在图中找到：

- ▶ 第一枚色子点数为 6 的所有样本点。注意，第二枚色子点数不重要。
- ▶ 第二枚色子点数为 6 的所有样本点。
- ▶ 两枚色子点数为 6 的所有样本点。
- ▶ 两枚色子点数之和为 8 的所有样本点。
- ▶ 两枚色子点数之和不小于 8 的所有样本点。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

- ▶ 第一枚色子点数和第二枚点数之差为 3 的所有样本点。举例，第一枚色子点数为 6，第二枚色子点数为 3，两者之差为 3。
- ▶ 两枚色子点数之差为 3 的所有样本点。
- ▶ 两枚色子点数之差不小于 3 的所有样本点。
- ▶ 第一、二枚色子点数均在 $[2, 5]$ 区间 (左右包含)。
- ▶ 第一枚点数为偶数。
- ▶ 第一枚点数不是奇数。
- ▶ 第一枚点数为奇数。
- ▶ 两枚色子点数均为偶数。
- ▶ 两枚色子点数之和为偶数。

Q3. 下图所示为连续抛 5 枚硬币随机试验样本空间的水平树形图。根据此图，请回答

- ▶ 图中一共有多少条路径 (样本点)?
- ▶ 请指出 5 枚硬币都是反面的路径。
- ▶ 请指出 3 枚硬币为反面、2 枚硬币为正面的所有路径。
- ▶ 请指出至少 3 枚硬币为反面的所有路径。
- ▶ 请指出连续抛出 3 次正面的所有路径。注意，“正正正反反”、“反正正正反”满足条件；但是“反正正正正”不满足。
- ▶ 请指出连续抛出至少 3 次正面的所有路径。注意，“正正正反反”、“反正正正正”都满足条件。

