

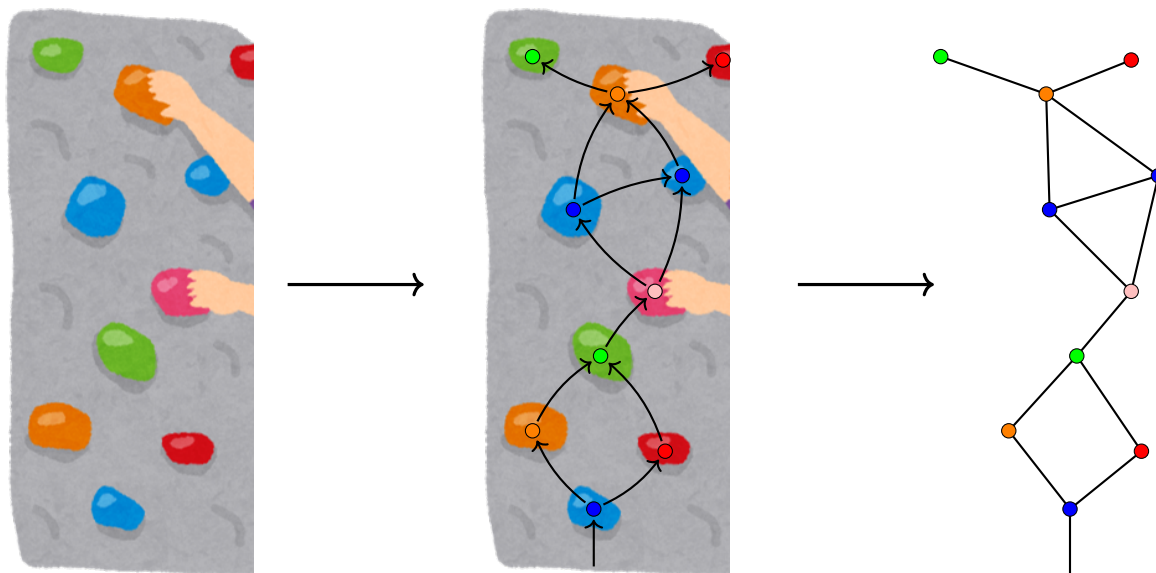
MAT120: 论证 第11讲讲义

连续概率

上一讲中，我们把概率引入为量化不确定性的一种方法，并学习了如何通过计数结果来计算古典概率。在这一讲中，我们将发展条件概率，也就是在得知额外信息之后的概率。我们还将学习一个实用工具，决策树，用来整理多步骤试验。最后，我们开始计算抛硬币时的结果个数，这将引向下一讲的二项试验。

1 引言：攀岩者

假设我们追踪某人攀爬岩壁时手的移动。我们可以把这种情形理想化，画成一种“图”，用来表示这些不同颜色的岩点是如何彼此连接起来的：



在这张图中，随着攀岩者沿着岩壁向上移动，他的手会从底部的蓝色岩点开始，沿着某一条路径向上，到达顶部的绿色岩点或红色岩点。注意，在这个过程中，有些阶段攀岩者需要在前进方式上作出选择。例如，在岩壁的最开始，攀岩者可以选择从蓝色岩点移动到橙色岩点，或者也可以选择移动到红色岩点。而在另外一些位置上，攀岩者并没有选择，路线是被完全决定的。比如，从第一个橙色岩点出发，他必须沿着路径走到下一个绿色岩点。

假设每当攀岩者作出一次选择时，他有50% 的概率选择左边的岩点，也有50% 的概率选择右边的岩点。事实上，从底部到顶部一共有12 条不同的路径。攀岩者走任意给定一条路径的概率是多少？

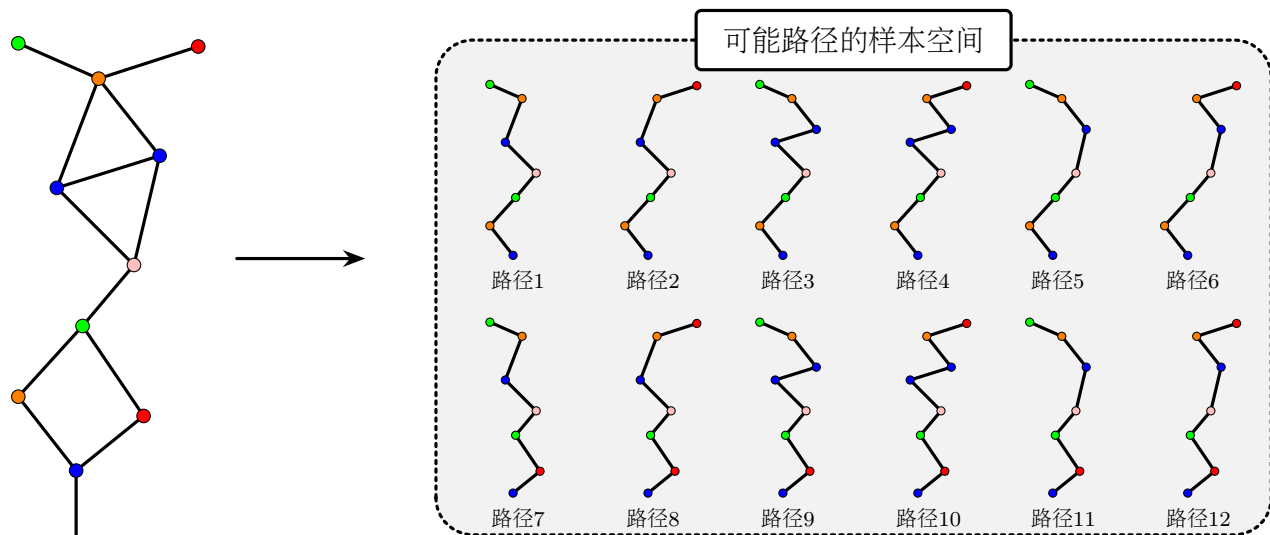
1.1 尝试用朴素计数来解决

上一讲中，我们看到过一个在古典概率情形下成立的基本公式：

$$P(\text{事件}) = \frac{\text{有利于该事件的结果个数}}{\text{结果总数}}.$$

简而言之，我们常常可以通过数出与某个一般事件对应的结果个数，再除以总可能数，来计算这个事件的概率。

对于这个攀岩者的例子，我们也可以尝试这么做。首先，我们应当描述样本空间。在这里，样本空间就是从底部到顶部的所有路径组成的空间。花一点时间思考后，你可以相信，这个岩壁上总共有12 条不同的路径。它们如下图所示：



因此，这个情形下的样本空间是集合

$$S = \{\text{路径1}, \text{路径2}, \text{路径3}, \dots, \text{路径12}\}.$$

于是，看起来很自然地会想用上面的公式来计算某一条给定路径的概率。比如，对于路径1，我们可能会猜测：

$$P(\text{路径1}) \stackrel{?}{=} \frac{\text{有利于该事件的结果个数}}{\text{结果总数}} = \frac{1}{12}.$$

事实上，这个结果并不正确。后面我们会正确地计算它，但现在我们只需要先注意到：在这些路径组成的空间里，有些路径比另一些路径更复杂。这些不同的路径**并不是**等可能的结果，因此上面的公式并不是正确的方法。

2 决策树

2.1 什么是决策树？

上一讲中，我们简要介绍了条件概率的概念，而它也出现在我们用来计算复合事件“ A and B ”概率的公式中。回忆一下，对于相关事件 A 和 B ，事件“ A and B ”的概率由下式给出：

$$P(A \text{ and } B) = P(A) P(B | A).$$

等价地，也可以写成

$$P(A \text{ and } B) = P(B) P(A | B).$$

这里的 $P(B | A)$ 就是条件概率，大致意思是“在假设 A 发生的前提下， B 的概率”。一般而言，条件概率反映的是：当我们根据某个假设更新了信息之后，概率会如何改变。

上一讲中，我们看到条件概率可以由下面的公式定义：

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} \quad (\text{前提是 } P(A) > 0).$$

在古典概率中，每个结果都是等可能的，因此上面的公式可以通过计数来计算：

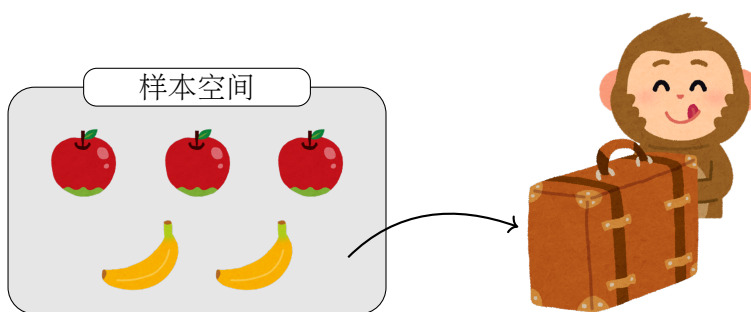
$$P(B | A) = \frac{\text{集合 } A \cap B \text{ 的大小}}{\text{集合 } A \text{ 的大小}}.$$

这是一种计算集合论中交集 $A \cap B$ 大小的方法：我们先把注意力限制在事件 A 所包含的结果上，然后再数一数这些结果里有多少也属于 B 。

当我们考虑相继概率时，条件概率会特别有用。这里所谓相继概率，是指多步骤试验中，不同步骤可能出现不同结果的情形。决策树就是一种图示，它把这类多步骤试验的所有可能结果完整地列出来。不同结果彼此“分叉”，并以树的分支表示。¹ 简单地说，决策树帮助我们整理可能结果的空间，从而使每一个单独结果的概率更容易计算。

2.2 例子：丛林与小偷

假设你在丛林里，把自己的包放在一旁无人看管。一只当地的猴子从树上爬下来，随机翻找你的包。碰巧的是，你包里装着5个准备当午餐的水果：其中有3个苹果和2根香蕉。猴子开始随机偷水果，并且每找到一个就先吃掉，然后再继续翻包。在概率术语中，这叫作不放回抽取。



练习1

- (a) 猴子第一次偷到苹果的概率是多少？
- (b) 猴子第一次偷到香蕉的概率是多少？

解答

一共有5个水果，所以我们可以直接计数：

- (a) $P(\text{第一次拿到苹果}) = \frac{3}{5}$ 。
- (b) $P(\text{第一次拿到香蕉}) = \frac{2}{5}$ 。

现在，我们关注一个相继发生的问题：

它先偷到一根香蕉，然后又偷到一个苹果的概率是多少？

¹更准确地说，一条分支是从树的起点一直到最末端的一条线。

很明显，这里应该使用条件概率，因为第二次的结果显然会受到第一次结果的影响。

为了准确表示顺序，令 B_1 表示“第一次抽到香蕉”，令 A_2 表示“第二次抽到苹果”。那么“先香蕉后苹果”就是事件“ B_1 and A_2 ”。

我们可以使用条件概率公式来计算它。这里，让事件 A 对应于“猴子拿到一个苹果”，让事件 B 对应于“猴子拿到一根香蕉”。于是，复合事件“ B and A ”就可以理解为“猴子先拿到一根香蕉然后又拿到一个苹果”。因此，我们可以用下面的公式来计算这个概率：

$$P(B_1 \cap A_2) = P(B_1) \cdot P(A_2 | B_1).$$

我们分步骤来算：

步骤1: 前面我们已经看到， $P(B_1) = \frac{2}{5}$ 。

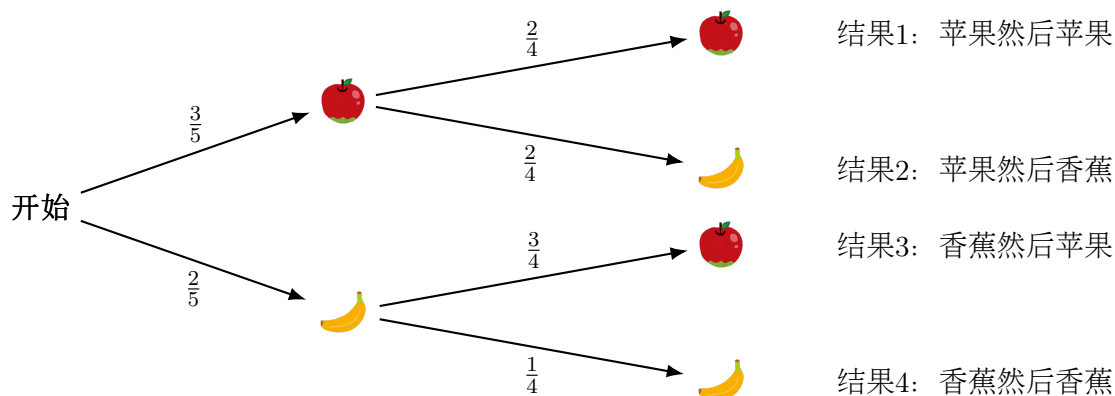
步骤2: 现在来计算 $P(A_2 | B_1)$ 。注意，如果一根香蕉已经被拿走（并被猴子吃掉），那么第二次从包里拿水果时，猴子面对的样本空间就不同了。第一次拿水果时，包里一共有5个水果；但是由于已经少了一根香蕉，现在只剩下4个水果：3个苹果和1根香蕉。这个新的样本空间大小为4，而其中有3个苹果可供猴子选择。因此，条件概率为

$$P(A_2 | B_1) = \frac{3}{4}.$$

步骤3: 现在把这两个数相乘，就得到先拿香蕉再拿苹果的概率：

$$P(B_1 \text{ and } A_2) = P(B_1) P(A_2 | B_1) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}.$$

我们可以把整个过程整理成如下所示的一棵树：



如你所见，我们之前做的计算，也就是

$$P(\text{香蕉然后苹果}) = P(B_1) \cdot P(A_2 | B_1) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10},$$

也可以通过检查上面决策树中对应结果3的分支，从图形上直接确认。这里，我们只是把第三条分支上写着的数字相乘而已。事实上，我们刚刚已经看到了一个一般技巧的第一个实例：如果我们能把一个多步骤试验的结果安排成一棵树，那么任何一个结果的概率，都可以通过把那条分支上出现的所有概率相乘得到。

利用这个观察，我们也可以很快算出树中另外三个结果的概率：

- 结果1: $P(A \text{ then } A) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$,
- 结果2: $P(A \text{ then } B) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$,
- 结果4: $P(B \text{ then } B) = \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$ 。

注意到这四个概率之和为

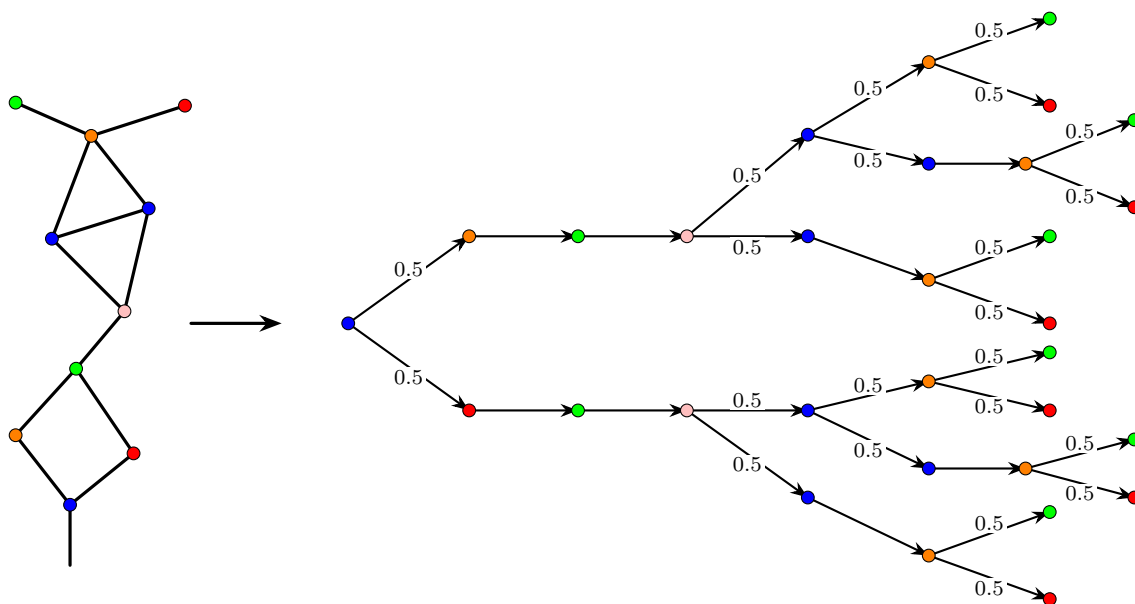
$$\frac{3}{10} + \frac{3}{10} + \frac{3}{10} + \frac{1}{10} = 1.$$

这就等于说，其中某一个结果必然会发生。

2.3 回到攀岩者

我们刚刚看到，决策树可以用来整理一个多步骤试验的所有可能结果。要找出某一条特定路径的概率，我们只需要把那条路径上写着的概率相乘即可。我们可以用这个技巧来求出攀岩者在岩壁上各条路线所对应的概率。

首先，我们把这个系统写成一棵树，然后在每一个“选择点”上标出所假设的50-50 概率。完整的树如下所示。



上面的树是按路径顺序排列的：第一条分支对应路径1，第二条分支对应路径2，以此类推。你应当注意到，有些路径比另一些路径更复杂，例如，路径12 看上去就比路径9 或路径10 更简单。这是因为，沿着岩壁上升时，有些路线只需要作出三次选择，而另外一些路线则需要作出四次选择。事实上，凡是包含“在粉色节点处向左转”这一选择的路径，都会比那些在该节点处向右转的路径更复杂。

无论如何，我们可以利用决策树的理论来计算攀岩者沿岩壁任意给定一条路径前进的概率。和前面那个“盗窃猴子”的例子一样，在这里我们只需把该分支上写着的数值相乘，就得到总体概率。结果总结如下。

每条路径的概率

我们要区分每条路径所需要作出的选择次数：

- 路径1, 2, 3, 4, 7, 8, 9 和10 都需要作出四次选择。在决策树上，这些路径对应的分支上都写着4个0.5，因此它们的总概率为

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}.$$

- 路径5, 6, 11 和12 一共只需要作出三次选择。因此，走到其中任意一条路径的概率会比前者更大：

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

我们总可以通过验证所有答案相加是否等于1 来检查计算是否正确。这里有：

$$8 \left(\frac{1}{16} \right) + 4 \left(\frac{1}{8} \right) = \frac{1}{2} + \frac{1}{2} = 1,$$

这说明计算是正确的。注意，我们在这里还无意中发现了另一件事：攀岩者有50% 的概率会走那几条看起来更简单的路径，也有50% 的概率不会。

3 计数结果

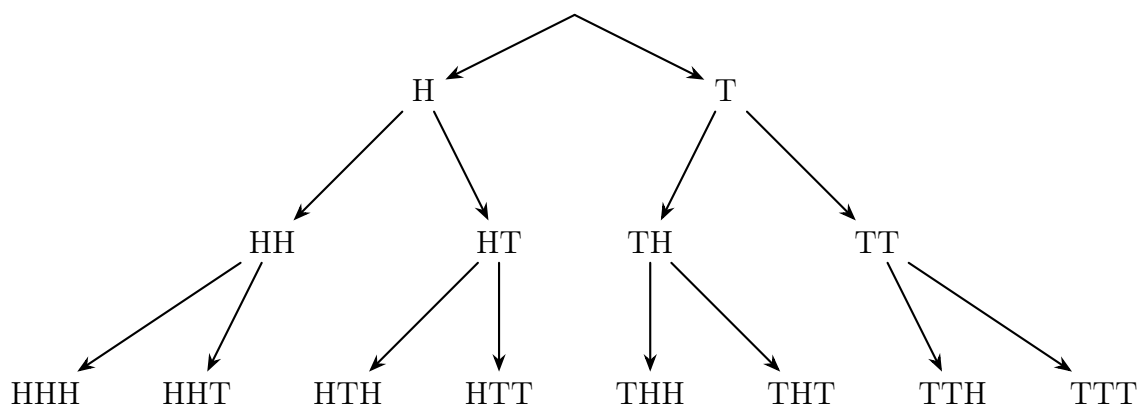
3.1 抛硬币3 次

假设我们抛一枚硬币3 次。

练习2

- (a) 恰好出现2 次反面的概率是多少？
- (b) 恰好出现1 次反面的概率是多少？

这个样本空间中的结果是等可能的，因此最好的方法就是直接数出每种情形中的有利结果个数。和前面一样，如果愿意的话，我们也可以把它们画成一棵树。为了后面方便，我们将把这棵树画成试验起点在最上方的形式。



练习2 的解答

根据上面的树，我们看到，抛3次硬币的样本空间大小为8。由于这些结果是等可能的，因此我们可以通过直接计数来计算概率。

(a) 恰好出现2次反面：有三个结果有利于这个事件，即 $\{HTT, THT, TTH\}$ ，因此

$$P(\text{恰好2次反面}) = \frac{3}{8}.$$

(b) 恰好出现1次反面：有三个结果有利于这个事件，即 $\{THH, HTH, HHT\}$ ，因此

$$P(\text{恰好1次反面}) = \frac{3}{8}.$$

3.2 计算 n 次抛硬币中反面的个数

根据前面的例子，我们可能会觉得，涉及抛硬币的概率计算似乎很容易。然而，随着次数增加，情况会迅速变得复杂。先注意我们的概率树中每一层的大小如何随着时间增长：在上面的树中，第一层有两个结果，第二层有四个结果，第三层有八个结果。如果我们再抛一次硬币，那么对于每一个已经存在的由 H 和 T 组成的字符串，都会再分出两个新结果。于是，可能结果的总数又会翻倍，因此第四层将有16个结果，也就是说，四次抛硬币一共有16种可能。在一般情况下，如果我们抛硬币 n 次，那么总结果数将是 2^n 。

假设我们想数出在 n 次抛掷之后恰好出现某个特定数量的反面，比如恰好出现 r 次反面。原则上，我们可以继续使用前面的方法，数出恰好有 r 个反面的结果个数，然后再除以 2^n 。但这个过程很快就会变得困难。例如：在6次抛硬币中，恰好有2次反面的结果有多少个？这可能意味着你要在 $2^6 = 64$ 个由 H 和 T 构成的不同结果中筛选，找出那些含有2个 T 和4个 H 的结果。²

在实际中，如果有一种巧妙的方法，能让我们不必写下大量令人眼花缭乱的 H 和 T ，再逐个检查，就能直接数出恰好有 r 个反面的结果个数，那就再好不过了。事实上，确实有这样一种巧妙的方法，而且我们其实已经见过它了。

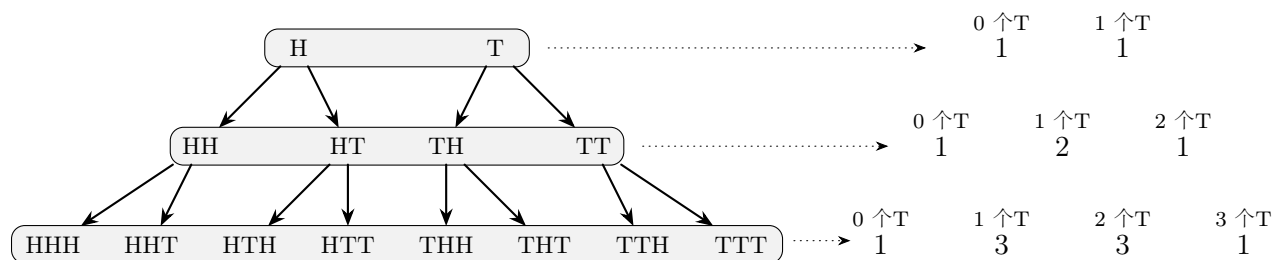
3.2.1 计算 n 次抛掷中恰好有 r 次反面

为了说明这个技巧，让我们再次观察上面的树，并把其中所有可能结果数一遍。

- 1次抛掷：如果我们抛一次硬币，那么结果集合是 $\{H, T\}$ 。其中有一个结果包含0个 T ，也有一个结果包含1个 T 。
- 2次抛掷：抛两次得到的结果为 $\{HH, HT, TH, TT\}$ 。其中有一个结果有0个 T （即结果 HH ），有2个结果有1个 T （即 HT 和 TH ），还有一个结果有2个 T （即 TT ）。
- 3次抛掷：这时结果集合大小为8，也就是上面树的最后一层。我们已经做过这个计数：我们知道有一个结果有0个 T （即 HHH ），有3个结果有1个 T （即 HHT, THH, HTH ），有3个结果有2个 T （即 HTT, THT, TTH ），并且有一个结果有3个 T （即 TTT ）。

为了使这一点更直观，我们可以把这个“反面计数器”写在树的旁边：

²如果你觉得这个过程还算轻松，那么不妨考虑 $n = 10$ 且 $r = 4$ 的情形。



如你所见，图右侧的三角形看起来非常像帕斯卡三角形的前几行。事实上，这就是帕斯卡三角形。为什么？因为在抛硬币中计算恰好有 r 次反面的结果个数，本质上就是在说：“我们有 n 个位置可供选择，我想知道从中选出 r 个位置作为反面的所有可能方式。”

如果我们把帕斯卡三角形重新画出来，并标注每个位置对应可能出现的反面次数，那么我们就可以直接读出这些数字，而不需要做任何繁琐的计数：

0个T 1										$2^0 = 1$ 个结果
0个T 1	1个T 1									$2^1 = 2$ 个结果
0个T 1	1个T 2	2个T 1								$2^2 = 4$ 个结果
0个T 1	1个T 3	2个T 3	3个T 1							$2^3 = 8$ 个结果
0个T 1	1个T 4	2个T 6	3个T 4	4个T 1						$2^4 = 16$ 个结果
0个T 1	1个T 5	2个T 10	3个T 10	4个T 5	5个T 1					$2^5 = 32$ 个结果
0个T 1	1个T 6	2个T 15	3个T 20	4个T 15	5个T 6	6个T 1				$2^6 = 64$ 个结果

使用帕斯卡三角形显然比把H和T的所有可能组合一一写出来轻松得多。例如，要确定在总共6次抛硬币中恰好出现2次反面的概率，我们只需要看三角形中 $n = 6$ 那一行、 $r = 2$ 那一项，就能看到恰好有2个反面的结果共有15个，而总结果数为 $2^6 = 64$ 。于是我们立刻就可以把这些数代入计数公式，得到

$$P(6 \text{ 次抛掷中有 } 2 \text{ 次反面}) = \frac{\text{在6次抛掷中得到2次反面的方式数}}{\text{6次抛掷的结果总数}} = \frac{15}{64} \approx 23.4\%.$$

帕斯卡三角形中的项记作 $C_{n,r}$ 。一般来说，我们可以利用这种“计数组”的技巧，得到一个关于 n 次抛硬币中恰好有 r 次反面的概率公式：

$$P(n \text{ 次抛掷中有 } r \text{ 次反面}) = \frac{\text{在 } n \text{ 次抛掷中得到 } r \text{ 次反面的方式数}}{n \text{ 次抛掷的结果总数}} = \frac{C_{n,r}}{2^n}.$$

当然，出现正面的概率与出现反面的概率相同，因此上面的公式也具有对称性，它同样可以用来计算 n 次抛掷中恰好出现 r 次正面的概率。

4 补充：生日悖论

令 A 为事件“两个人在房间里生日相同”。在这种情形下，概率 $P(A)$ 会随着 n （房间里的人数）而发生剧烈变化。比如，当 $n = 2$ 时，我们会认为这两个人生日相同的概率极低。然而，当 $n > 365$ 时，房间里必定会有两个人生日相同，因为一年只有365个可能的生日。由此可见，当 n 很小时， $P(A)$ 也很小；而当 n 很大时， $P(A)$ 最终会等于1。于是我们可以提出这样一个问题：

当 n 取什么值时， $P(A)$ 会超过50%？

换句话说，房间里至少要有多少人，才会使得“至少有两个人同一天生日”这件事更有可能发生而不是不发生？事实上，答案相当令人惊讶：只需要23个人！

这个令人惊讶的事实被称为“生日悖论”。它并不是我们在课程前面见过的“理发师悖论”那种意义上的悖论。理发师悖论中，我们得到的是无限循环的真假判断，以至于逻辑本身的功能失效。相比之下，生日悖论是一种“真悖论”（veridical paradox）的例子，也就是说，它只是一个数学结果，惊人到让人感觉不对劲。但是，正如我们接下来要展示的，生日悖论其实只是由概率的一些基本应用推导出来的。

4.1 化解生日悖论

为了化解生日悖论，我们将把问题倒过来看，并重新表述。我们不直接计算 $P(A)$ ，而是去计算 $P(A^c)$ ，即事件 A 的补事件的概率。回忆一下，补事件就是概率论里与“非”对应的概念。因此，这里补事件 A^c 的意思是：“房间里的每个人生日都不同。”

还记得，补事件的概率就是 $1 - P(A)$ 。因此，如果 $P(A^c) < 50\%$ ，那么 $P(A) > 50\%$ ，因为

$$P(A) + P(A^c) = 1.$$

我们现在假设每个生日的出现都是等可能的，因此任意一个生日的概率都是 $1/365$ 。下面我们先对较小的 n 来计算 $P(A^c)$ ，看看整个过程是如何运作的。

当 $n = 2$ 时，我们要计算的是两个人生日不同的概率。显然，第一个人的生日可以任意选择。但是，为了让两个人的生日不同，第二个人的生日必须避开第一个人的那一天。与第一个人的生日不同的日子共有364天，因此（如果把它看作一个随机事件），第二个人生日不同的概率为

$$P(A^c) = \frac{\text{与第1个人生日不同的日期数}}{\text{一年中的总天数}} = \frac{364}{365} \approx 0.997 = 99.7\%.$$

现在假设房间里再多加一个人，也就是 $n = 3$ 。我们刚刚已经确定，两个人生日不同的概率是 $\frac{364}{365}$ 。但是我们还需要第三个人的生日既不同于第1个人，也不同于第2个人。随机选择下一个人的生日是一个条件事件，因为它取决于前面已经选出的结果。为了计算这种“并且”类型的概率，我们需要把相应的概率相乘：

$$\begin{aligned} P(A^c) &= \frac{364}{365} \times \frac{\text{与第1、2个人生日都不同的日期数}}{\text{一年中的总天数}} \\ &= \frac{364}{365} \cdot \frac{363}{365} \approx 0.992 = 99.2\%. \end{aligned}$$

我们可以对更大的 n 不断重复这个模式。每当把 n 的值增加1，我们就会发现，可供第 $(n+1)^{st}$ 个人选择的生日正在不断“被用掉”。这意味着，在每一步之后，我们都必须乘以越来越小的数。例如：

$$\text{当 } n = 4 \text{ 时: } P(A^c) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \approx 0.983 = 98.3\%$$

$$\text{当 } n = 5 \text{ 时: } P(A^c) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \left(\frac{361}{365}\right) \approx 0.973 = 97.3\%$$

$$\text{当 } n = 6 \text{ 时: } P(A^c) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \left(\frac{361}{365}\right) \left(\frac{360}{365}\right) \approx 0.960 = 96.0\%$$

对于一般的 n ，我们得到公式：

$$P(A^c) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \cdots \left(\frac{365 - (n - 1)}{365}\right).$$

要化解这个悖论，我们只需找到最小的 n ，使得概率 $P(A^c)$ 下降到50% 以下，因为这就意味着 $P(A)$ 会上升到 50% 以上。如果我们把这个公式代入越来越大的 n ，就会看到当 $n = 23$ 时：

$$P(A^c) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \cdots \left(\frac{342}{365}\right) \approx 0.490 < 0.5.$$

因此，当 $n = 23$ 时， $P(A) > 50\%$ 。