

20

Fundamentals of Probability

概率入门

从杨辉三角到古典概率模型



这个世界的真正逻辑是概率的推演。

The true logic of this world is the calculus of probabilities.

——詹姆斯·克拉克·麦克斯韦 (James Clerk Maxwell) | 英国数学物理学家 | 1831 ~ 1879



```
◀ ax.invert_xaxis() 调转 x 轴
◀ ax1.spines['right'].set_visible(False) 除去图像右侧黑框线
◀ ax1.spines['top'].set_visible(False) 除去图像上侧黑框线
◀ itertools.combinations() 无放回抽取组合
◀ itertools.combinations_with_replacement() 有放回抽取组合
◀ itertools.permutations() 无放回排列
◀ matplotlib.pyplot.barh() 绘制水平直方图
◀ matplotlib.pyplot.stem() 绘制火柴梗图
◀ numpy.concatenate() 将多个数组进行连接
◀ numpy.stack() 将矩阵叠加
◀ numpy.zeros_like() 用来生成和输入矩阵形状相同的零矩阵
◀ scipy.special.binom() 产生二项式系数
◀ sympy.Poly 将符号代数式转化为多项式
```

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



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

20.1 概率简史：出身赌场

概率是现代人类的自然思维方式。大家在日常交流时，用到“预测”、“估计”、“肯定”、“百分之百的把握”、“或许”、“百分之五十可能性”、“大概”、“可能”、“恐怕”、“绝无可能”等字眼时，思维已经进入概率的范畴。

概率论的目的就是将这些字眼数学化、量化。

意大利学者**吉罗拉莫·卡尔达诺** (Girolamo Cardano, 1501 ~ 1576) 可以说是文艺复兴时期百科全书式人物。他做过执业医生，第一个发表三次代数方程式的一般解法，他还是赌场常胜将军。

卡尔达诺死后才向世人公布自己创作的赌博秘籍《论赌博的游戏》(*Book on Games of Chance*)，这本书首次对概率进行系统介绍。他在书中用投色子游戏讲解等可能事件和其他概率概念。值得一提的是，卡尔达诺的父亲和达芬奇是好友。和达芬奇一样，卡尔达诺也是私生子。

概率论的基本原理是在**帕斯卡** (Blaise Pascal, 1623 ~ 1662) 和**费马** (Pierre de Fermat, 1607 ~ 1665) 的一系列来往书信中搭建起来的。他们在书信中讨论的是著名的赌博奖金分配问题。

举个例子说明赌博奖金分配问题。 A 、 B 两人玩抛硬币游戏，每次抛一枚硬币，硬币朝上 A 得一分，硬币朝下 B 得一分，谁先得到 10 分谁就赢得所有奖金。但是，游戏进行到途中突然中断，此时 A 得分 7 分， B 得分 5 分，两人此时应该如何分配奖金？

在帕斯卡和费马的讨论中，他们提出了枚举法。一些书信中也能看到他们谈到利用杨辉三角和二项式展开求解赌博奖金分配问题。

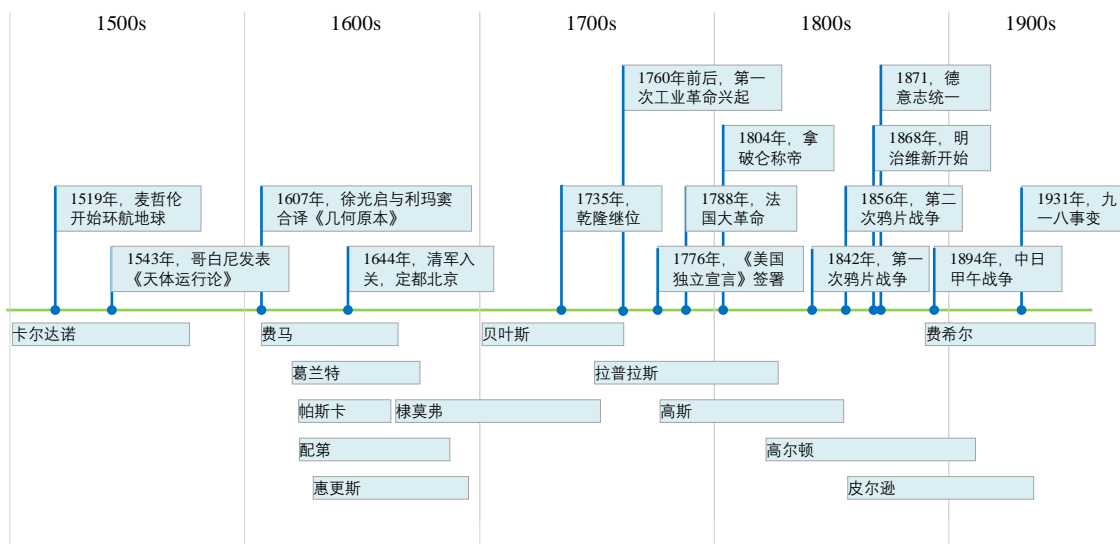


图 1. 概率论、统计学发展时间轴

克里斯蒂安·惠更斯 (Christiaan Huygens, 1629 ~ 1695) 扩展了帕斯卡和费马的理论。惠更斯 1657 年发表了《论赌博中的计算》(*On Reasoning in Games of Chance*)，被很多人认为是概率论诞生的标志。

法国数学家**亚伯拉罕·棣莫弗** (Abraham de Moivre, 1667 ~ 1754) 继续推动概率论的发展，他首先提出正态分布、中心极限定理等。在处理莱布尼兹-牛顿微积分发明权之争，棣莫弗还被选做裁决人之一。

贝叶斯 (Thomas Bayes, 1701 ~ 1761) 在自己的论文《解决机会学说中的问题》(*An Essay Towards Solving a Problem in the Doctrine of Chances*) 中探讨了条件概率，这使得贝叶斯成为贝叶斯学派的开山鼻祖。

在概率领域，**高斯** (Carl Friedrich Gauss, 1777 ~ 1855) 发明最小二乘法。虽然正态分布常被称作高斯分布，但是高斯不是正态分布的第一发明者。

弗朗西斯·高尔顿 (Francis Galton, 1822 ~ 1911) 则提出回归、相关系数等重要统计学概念。有趣的是，高尔顿是查尔斯·达尔文的表弟。

概率论和统计学两门学科相互交融，而且发展历史跨度很大，太多学者起到推动作用。很可惜，限于篇幅本节只能走马观花用几句话概括关键人物的生平。

20.2 二叉树：一生二、二生三

杨辉三角可谓是算数、代数、几何、数列、概率的完美结合体。沿着帕斯卡和费马的思路，本章从杨辉三角入手来和大家探讨概率论的核心思想。

本节首先从一个全新视角解读杨辉三角——**二叉树** (binomial tree)。将本书第 4 章介绍的杨辉三角逆时针旋转 90 度，得到图 2 这个二叉树。图 2 中每个点称作**节点** (node)。

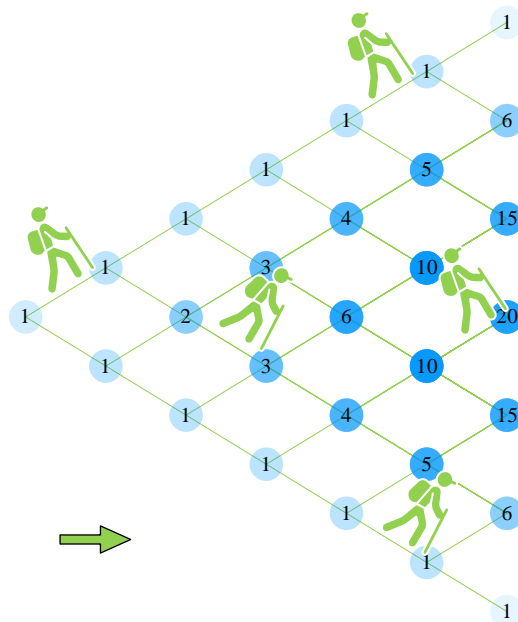


图 2. 杨辉三角逆时针旋转 90 度得到一个二叉树

试想，一名登山者从最左侧初始点出发，沿着二叉树规划的路径向右移动，到达最右侧任意节点结束。途中每个节点处，登山者可以向右上方或右下方走，但是不能往回走。

这样，图 2 中的数字便有了另外一层内涵——登山者到达对应节点的可能路径。

二叉树原理

下面解释一下原理。

如图 3 所示，当 $n = 1$ 时，二叉树叫做**一步二叉树** (one-step binomial tree)。也就是说，登山者从初始点出发，只有两个路径到达两个不同终点。

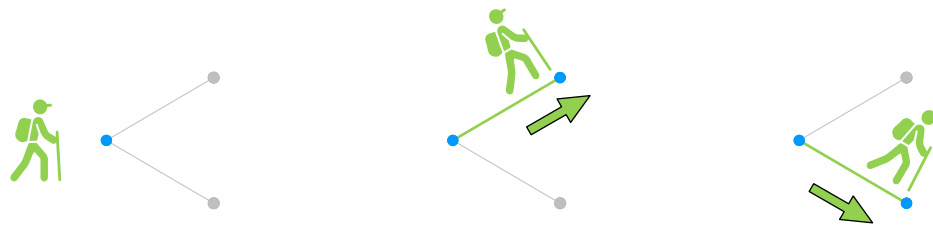


图 3. $n = 1$ ，向上、向下走的路径

如图 4 所示， $n = 2$ 时，二叉树为**两步二叉树** (two-step binomial tree)。从起点到终点，一共有 4 条路径，二项式系数 1、2、1 则相当于到达对应 A、B、C 终点的可能路径数量。

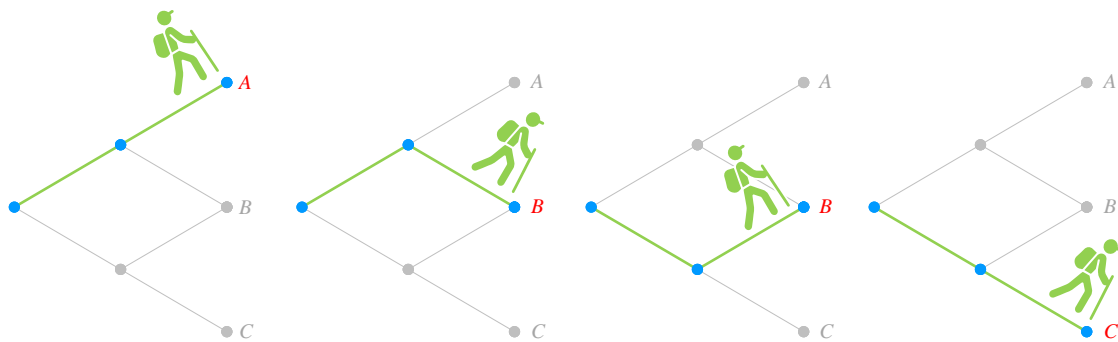


图 4. $n = 2$ ，通向最终节点路径

当二叉树的层数不断增多，到达终点的路径的数量呈现指数增长趋势。

如图 5 (a) 所示， $n = 3$ 时，路径数量为 $8 (= 1 + 3 + 3 + 1 = 2^3)$ 。如图 5 (b) 所示， $n = 4$ 时，路径数量为 $16 (= 1 + 4 + 6 + 4 + 1 = 2^4)$ 。如图 5 (c) 所示， $n = 5$ 时，路径数量为 $32 (= 1 + 5 + 10 + 10 + 5 + 1 = 2^5)$ 。

这个结果也不难理解，二叉树每增加一层，登山者就多一次二选一的机会。从路径数量角度，就是再乘 2。

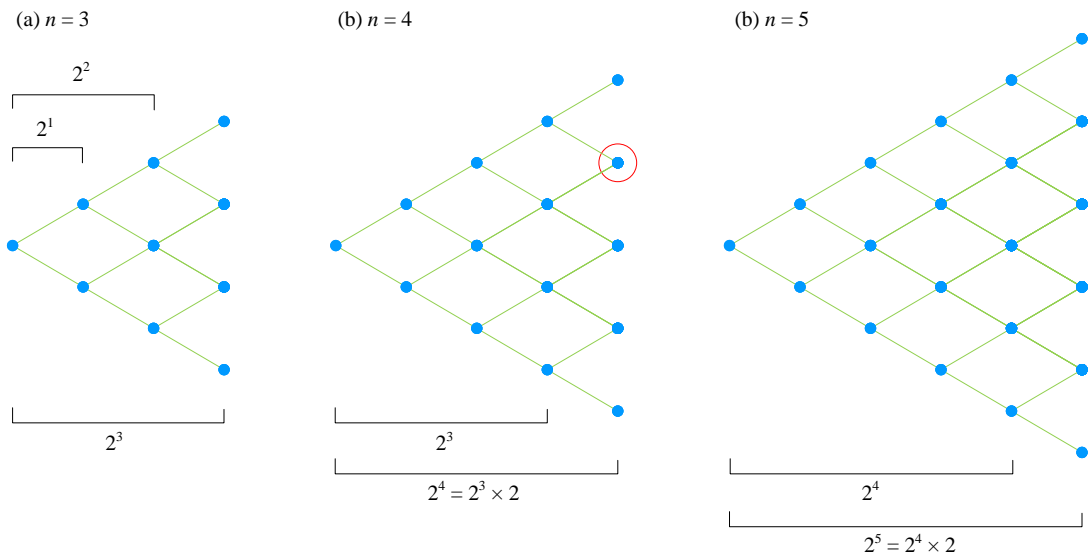


图 5. $n = 3, 4, 5$ ，通向最终节点路径

图 6 所示为 4 条到达图 5 (b) 二叉树画红圈终点节点路径。4 这个结果和组合数有着密切关系。下面我们聊一下如何用组合数解释到达不同终点路径数。

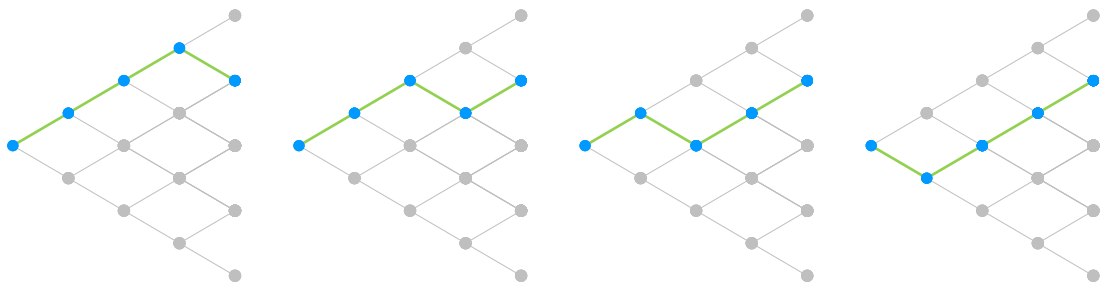


图 6. 四条到达同一终点节点的路径

组合数

利用**水平条形图** (horizontal bar graph) 可视化图 5 二叉树路径数。如图 7 所示， $n = 3$ 时，到达二叉树终点节点的路径分别有 1、3、3、1 条，总共有 8 条路径，写成组合数：

$$C_3^0 + C_3^1 + C_3^2 + C_3^3 = 1 + 3 + 3 + 1 = 8 = 2^3 \quad (1)$$

大家可能会问，组合数在这里扮演的角色是什么？

很容易理解，登山者在图 7 所示二叉树需要做三次“向上走或向下走”的决策。

C_3^0 可以理解为，3 次决策中 0 次向下； C_3^1 可以理解为，3 次决策中 1 次向下； C_3^2 可以理解为，3 次决策中 2 次向下； C_3^3 可以理解为，3 次决策中 3 次向下。

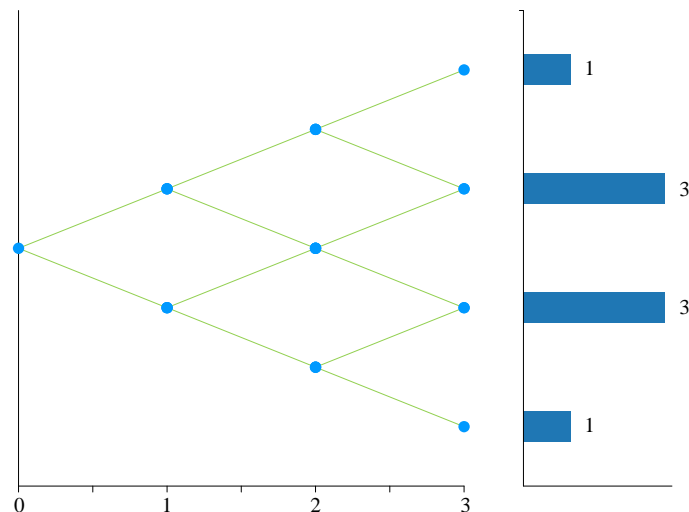


图 7. $n = 3$ ，二叉树路径数分布

如图 8 所示， $n = 4$ 时，到达二叉树终点节点的路径分别有 1、4、6、4、1 条，总共有 16 条路径：

$$C_4^0 + C_4^1 + C_4^2 + C_4^3 + C_4^4 = 1 + 4 + 6 + 4 + 1 = 16 = 2^4 \quad (2)$$

也就是说，这种情况登山者面临 4 次“二选一”的决策。

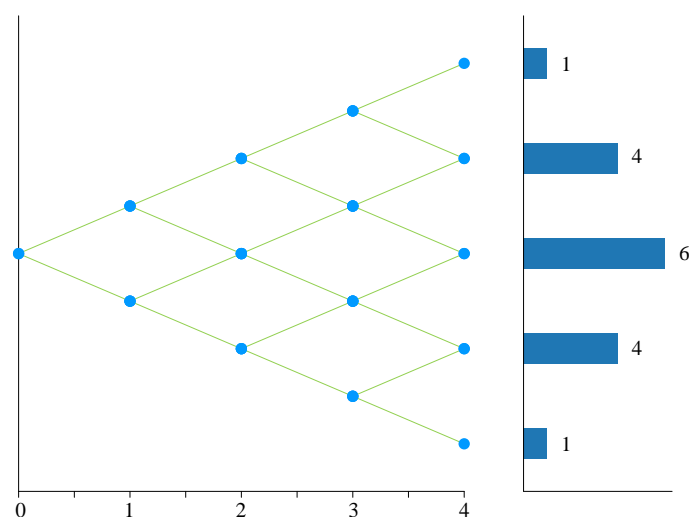


图 8. $n = 4$ ，二叉树路径数分布

如图 9 所示, $n = 5$ 时, 登山者有 5 次“二选一”决策, 到达二叉树终点节点的路径分别有 1、5、10、10、5、1 条, 总共有 32 条路径:

$$C_5^0 + C_5^1 + C_5^2 + C_5^3 + C_5^4 + C_5^5 = 1 + 5 + 10 + 10 + 5 + 1 = 32 = 2^5 \quad (3)$$

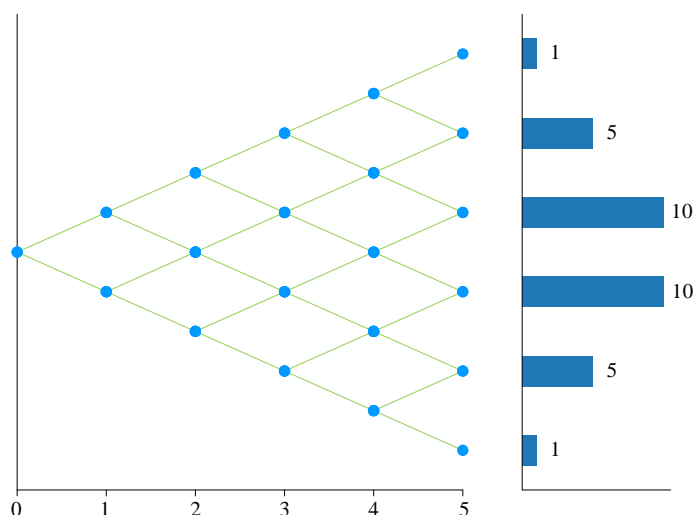


图 9. $n = 5$, 二叉树路径数分布

从概率统计角度, 图 9 右侧的直方图常被称作**频数直方图** (frequency histogram)。频数也称次数, 是对总数据按某种标准进行分组, 统计出各个组内含个体的个数。

杨辉三角和二叉树体现出来的规律像极了老子所言“道生一, 一生二, 二生三, 三生万物。”



代码文件 Bk3_Ch20_1.py 中 Bk3_Ch20_1_A 部分绘制图 7 ~ 图 9。

20.3 抛硬币：正反面概率

确定与随机

在自然界和社会实践活动中, 人类遇到的各种现象可分为两大类: 确定现象, 随机现象。

随机现象的准确定义是: 在一定条件下, 出现的可能结果不止一个, 事前无法确切知道哪一个结果一定会出现, 但大量重复试验中其结果又具有统计规律的现象称为随机现象。

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

一年 24 节气轮替，太阳东升西落，这是确定性现象。某一年是干旱少雨，还是洪涝灾害频发，某一天是否会下雨，什么时候下雨，降水量多大，这些事情的结果都是随机的。

天地不仁，以万物为刍狗——感觉这句就是在说随机性。

但是，随机之中有确定。举个例子，抛一枚硬币，谁也不能准确预测硬币落地时是正面还是反面朝上。但是，大量抛硬币，却发现硬币的正反面平均值有一定的规律。

人类虽然不能百分之百准确预测明年今天的晴雨状况。但是，通过研究大量气象数据，我们可以找到降水周期性规律，并在一定范围内预测降水量。

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。

随机试验

随机试验 (random experiment) 是在相同条件下对某随机现象进行的大量重复观测。随机试验需要满足三个条件：


- a) 可重复，在相同条件下试验可以重复进行；
- b) 结果集合明确，每次试验的可能结果不止一个，并且能事先明确试验的所有可能结果；
- c) 单次试验结果不确定，进行一次试验之前不能确定哪一个结果会出现，但必然出现结果集合中的一个。

给定一个随机试验，所有的结果构成的集合为样本空间 Ω 。样本空间 Ω 中的每一个元素为一个样本点。

概率

概率 (probability) 反映随机事件出现的可能性大小。

给定任意一个事件 A ， $\Pr(A)$ 为事件 A 发生的概率 (the probability of event A occurring)。

 注意本书概率记法， \Pr 为正体。

对于任意事件 A ， A 发生的概率满足：

$$\Pr(A) \geq 0 \quad (4)$$

整个样本空间 Ω 的概率为 1，即，

$$\Pr(\Omega) = 1 \quad (5)$$

空集 \emptyset 不包含任何样本点，也称作不可能事件，因此对应的概率为 0，即：

$$\Pr(\emptyset) = 0 \quad (6)$$

白话说，一定会发生的事情，概率值为 1 (100%); 一定不会发生的事情，概率值为 0 (0%)。不一定会发生的事情，概率值在 0、1 之间。这就是量化“可能性”的基础。

等可能

等可能性是指设一个试验的所有可能发生的结果有 n 个，它们都是随机事件，每次试验有且只有其中的一个结果出现。

如果每个结果出现的机会均等，那么说这 n 个事件的发生是等可能试验的结果。设样本空间 Ω 由 n 个等可能的试验结果构成，事件 A 的概率为：

$$\Pr(A) = \frac{n_A}{n} \quad (7)$$

其中， n_A 为含于事件 A 的试验结果数量。

这种基本事件个数有限且等可能的概率模型，称为古典概率模型。所谓概率模型是对不确定现象的数学描述。

抛硬币

举最简单的例子，抛一枚硬币，1 代表落地结果正面、0 代表反面。抛一枚硬币的可能结果样本空间 Ω 为：

$$\Omega = \{0, 1\} \quad (8)$$

根据生活常识，如果硬币质地均匀。获得正面和反面的概率相同均为 $1/2$ ，即等可能：

$$\Pr(0) = \Pr(1) = \frac{1}{2} \quad (9)$$

连续抛 100 枚硬币，并记录每次硬币正 (1)、反面 (0) 结果。图 10 所示为每一次试验硬币正反面结果以及累计结果平均值变化。可以发现，随着抛硬币的次数不断增多，硬币正反面平均值愈靠近 $1/2$ 。

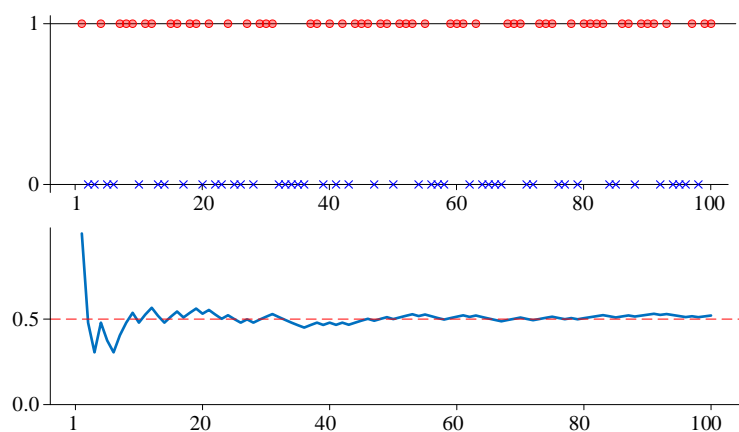


图 10. 抛硬币 100 次试验，硬币正反面结果，以及平均值变化



Bk3_Ch20_2.py 绘制图 10。

20.4 聊聊概率：向上还是向下

本节引入概率，给杨辉三角增加一个新视角。

登山者在二叉树始点或中间节点时，都会面临“向上”或“向下”这种二选一抉择。如果登山者通过抛硬币，决定每一步的行走路径——正面，向右上走；反面，向右下走。

生活经验告诉我们，如果硬币质地均匀，抛硬币时获得正面和反面的可能性相同。这个可能性，就是上一节提到的概率。

对于图 11 (a)，当登山者位于红色点 ●，他通过抛一枚硬币决定向上走和向下走的概率（可能性）相同，均为 0.5 (50%)。

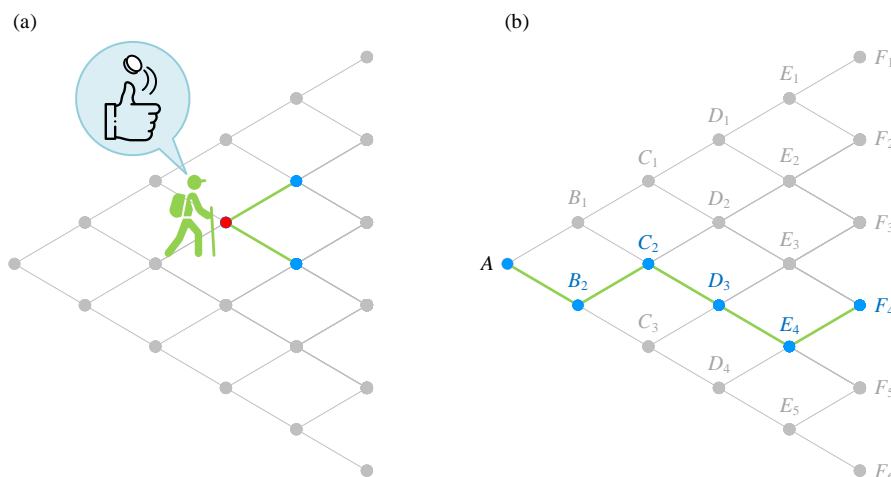


图 11. 二叉树路径与可能性

等可能角度

通过本章前文学习，大家已经清楚图 11 (b) 二叉树一共有 32 条路径。显然，从初始点到某一特定终点节点，登山者采用任意路径的可能性相同。也就是说图 11 (b) 中 $A \rightarrow B_2 \rightarrow C_2 \rightarrow D_3 \rightarrow E_4 \rightarrow F_4$ 这条路径被采纳的概率(可能性)为：

$$\Pr(A \rightarrow B_2 \rightarrow C_2 \rightarrow D_3 \rightarrow E_4 \rightarrow F_4) = \frac{1}{32} = 0.03125 = 3.125\% \quad (10)$$

二选一角度

再换一个角度，登山者在 A 、 B 、 C 、 D 、 E 这 5 个节点都面临二选一的抉择，而选择向上或向下的概率均为 $1/2$ ；因此，登山者选择图 11 (b) 中 $A \rightarrow B_2 \rightarrow C_2 \rightarrow D_3 \rightarrow E_4 \rightarrow F_4$ 路径的概率为：

$$\Pr(A \rightarrow B_2 \rightarrow C_2 \rightarrow D_3 \rightarrow E_4 \rightarrow F_4) = \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.03125 = 3.125\% \quad (11)$$

结果和 (10) 完全一致。

组合数

图 11 (b) 二叉树从起点 A 到终点 ($F_1 \sim F_6$) 一共有 32 条路径，而到达 F_4 点一共有 10 条路径。也就是说从 A 点出发，最终到达 F_4 点的概率为：

$$\Pr(F_4) = \frac{C_5^3}{2^5} = \frac{10}{32} = 0.3125 = 31.25\% \quad (12)$$

同理，我们可以计算得到到达 F_1 、 F_2 、 F_3 、 F_5 、 F_6 这几个终点的概率：

$$\begin{aligned}
\Pr(F_1) &= \frac{C_5^0}{2^5} = \frac{1}{32} = 0.03125 \\
\Pr(F_2) &= \frac{C_5^1}{2^5} = \frac{5}{32} = 0.15625 \\
\Pr(F_3) &= \frac{C_5^2}{2^5} = \frac{10}{32} = 0.3125 \\
\Pr(F_4) &= \frac{C_5^3}{2^5} = \frac{10}{32} = 0.3125 \\
\Pr(F_5) &= \frac{C_5^4}{2^5} = \frac{5}{32} = 0.15625 \\
\Pr(F_6) &= \frac{C_5^5}{2^5} = \frac{1}{32} = 0.03125
\end{aligned} \tag{13}$$

举个例子，从 A 点出发，不管中间走那条路线，到达 F_2 的概率为 15.625%。

这些概率值求和，得到结果为 1；这就是说，按照既定规则，登山者从起点出发，必然到达终点。1 量化了“必然”这一论述：

$$\begin{aligned}
\left(\frac{1}{2} + \frac{1}{2}\right)^5 &= C_5^0 \left(\frac{1}{2}\right)^5 + C_5^1 \left(\frac{1}{2}\right)^5 + C_5^2 \left(\frac{1}{2}\right)^5 + C_5^3 \left(\frac{1}{2}\right)^5 + C_5^4 \left(\frac{1}{2}\right)^5 + C_5^5 \left(\frac{1}{2}\right)^5 \\
&= \frac{1}{32} + \frac{5}{32} + \frac{10}{32} + \frac{10}{32} + \frac{5}{32} + \frac{1}{32} \\
&= 0.03125 + 0.15625 + 0.3125 + 0.3125 + 0.15625 + 0.03125 = 1
\end{aligned} \tag{14}$$

概率直方图

将上述概率值做成水平条形图，放在二叉树路径的右侧，我们得到图 12。

这种直方图被称作**概率直方图** (probability histogram)。大家可能已经发现，图 9 所示的频数直方图结果除以总数 32，就得到图 12 这幅概率直方图。也就是说，频数直方图和概率直方图可以很容易相互转化。

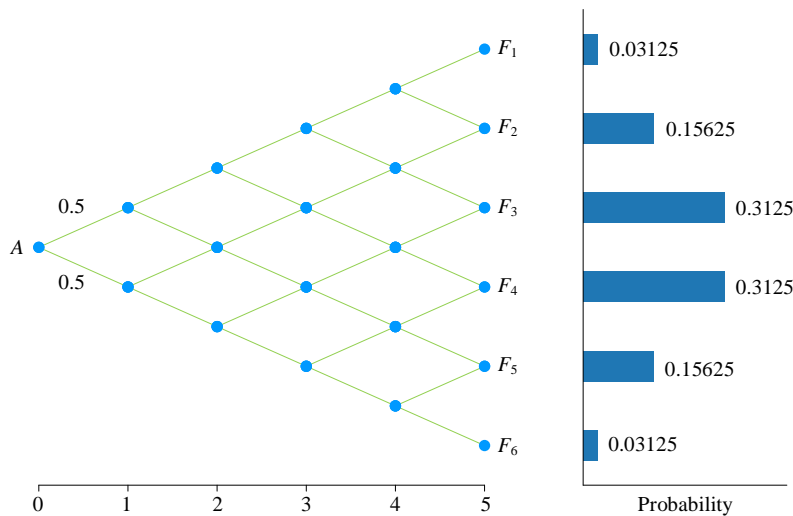


图 12. $n = 5$, 到达二叉树终点节点概率分布, 向上、向下概率均为 0.5

20.5 一枚质地不均匀的硬币

前文假设硬币质地均匀, 即抛一枚硬币获得正面背面朝上的概率相同, 均为 0.5 (50%); 但是, 假设一种情况, 硬币质地不均匀, 抛这枚硬币时, 得到正面的可能性为 60%, 反面的可能性为 40%。

下面计算一下抛这枚硬币决定在图 11 所示二叉树中登山者从起点到达终点的选取不同路径的可能性。

在五次“二选一”的决策中, 向上走的可能性为 0.6, 向下走的可能性为 0.4, 利用组合数容易得到, 到达 F_1 、 F_2 、 F_3 、 F_4 、 F_5 、 F_6 对应的概率分别为:

$$\begin{aligned}\Pr(F_1) &= C_5^0 \times 0.6^5 \times 0.4^0 = 0.07776 \\ \Pr(F_2) &= C_5^1 \times 0.6^4 \times 0.4^1 = 0.2592 \\ \Pr(F_3) &= C_5^2 \times 0.6^3 \times 0.4^2 = 0.3456 \\ \Pr(F_4) &= C_5^3 \times 0.6^2 \times 0.4^3 = 0.2304 \\ \Pr(F_5) &= C_5^4 \times 0.6^1 \times 0.4^4 = 0.0768 \\ \Pr(F_6) &= C_5^5 \times 0.6^0 \times 0.4^5 = 0.01024\end{aligned}\tag{15}$$

到达 F_1 、 F_2 、 F_3 、 F_4 、 F_5 、 F_6 对应的概率之和仍然为 1:

$$\begin{aligned}(0.6 + 0.4)^5 &= \underbrace{C_5^0 \times 0.6^5 \times 0.4^0}_{\Pr(F_1)} + \underbrace{C_5^1 \times 0.6^4 \times 0.4^1}_{\Pr(F_2)} + \underbrace{C_5^2 \times 0.6^3 \times 0.4^2}_{\Pr(F_3)} + \\ &\quad \underbrace{C_5^3 \times 0.6^2 \times 0.4^3}_{\Pr(F_4)} + \underbrace{C_5^4 \times 0.6^1 \times 0.4^4}_{\Pr(F_5)} + \underbrace{C_5^5 \times 0.6^0 \times 0.4^5}_{\Pr(F_6)} \\ &= 0.07776 + 0.2592 + 0.3456 + 0.2304 + 0.0768 + 0.01024 = 1\end{aligned}\tag{16}$$

但是对比图 12 和图 13, 容易发现登山者倾向于“向上走”; 这显然是因为硬币不均匀, 抛硬币得到正面的概率高于反面。而且图 13 右侧的概率直方图不再对称。

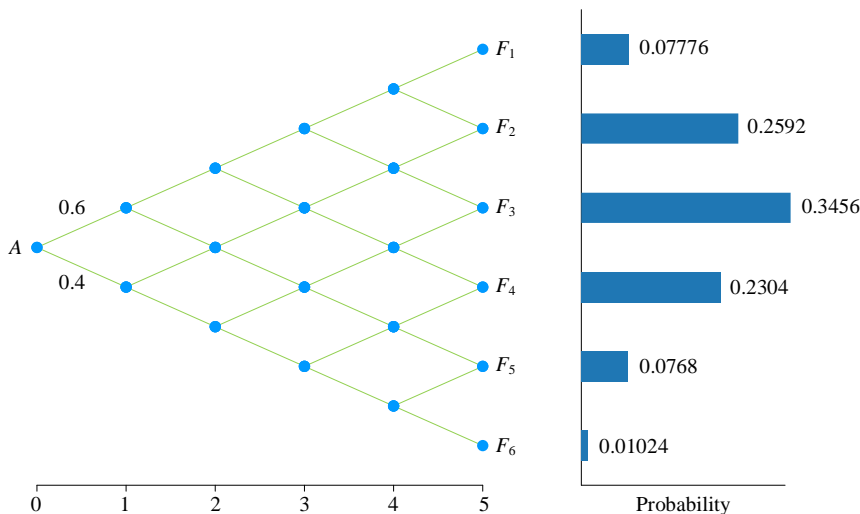


图 13. $n = 5$ ，到达二叉树终点节点概率分布，向上、向下概率分别为 0.6、0.4

如果我们恰好能够找到另外一枚质地不均匀的硬币，抛这枚硬币时，得到正面的可能性为 30%，反面的可能性为 70%。登山者通过抛这枚硬币确定向上走或向下走，如图 14 所示，登山者更倾向于向下走。

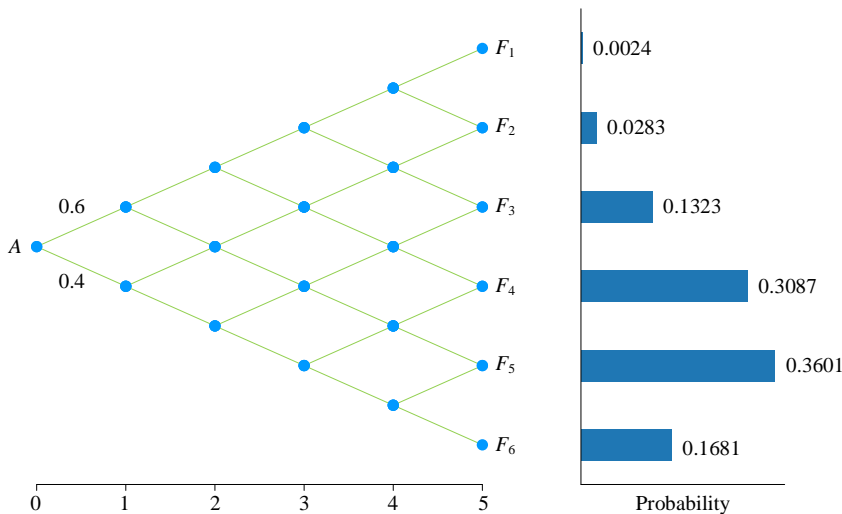


图 14. $n = 5$ ，到达二叉树终点节点概率分布，向上、向下概率分别为 0.3、0.7

这一节的内容，实际上就是我们要在丛书《概率统计》一本中要讲解的**二项式分布** (binomial distribution)。概率是数据科学和机器学习中重要的板块，本系列丛书《概率统计》一本将全面讲解。



代码文件 Bk3_Ch20_1.py 中 Bk3_Ch20_1_B 部分绘制图 12、图 13、图 14。请读者修改代码中 p 值。



在 Bk3_Ch20_1.py 基础上，我们做了一个 App 展示不同概率值对到达终点不同点概率的影响。请参考 Streamlit_Bk3_Ch20_1.py。

20.6 随机中有规律

本节还是用二叉树来探讨随机和确定之间的辩证关系。

在给定的二叉树网格中，登山者在不同节点“随机”确定向上走、向下走，得到的结果就是一种**随机漫步** (random walk)。

图 15 所示为 20 步二叉树网格，根据前文所学，我们知道从起点到终点，这个网格对应 2^{20} (1048576) 条路径。图 15 四幅图给出的是登山者“可能”走的 2、4、8、16 条随机路径。

随着路径数量增多，我们似乎可以预感，到达终点时登山者在中间的可能性会高于两端。

为了验证这一直觉，并相对准确确定登山者到达终点位置的规律，我们不断增加随机路径的数量，并根据终点位置绘制频率直方图。如图 16 所示，50、100、5000 条随机路径条件下，登山者终点位置概率直方图。

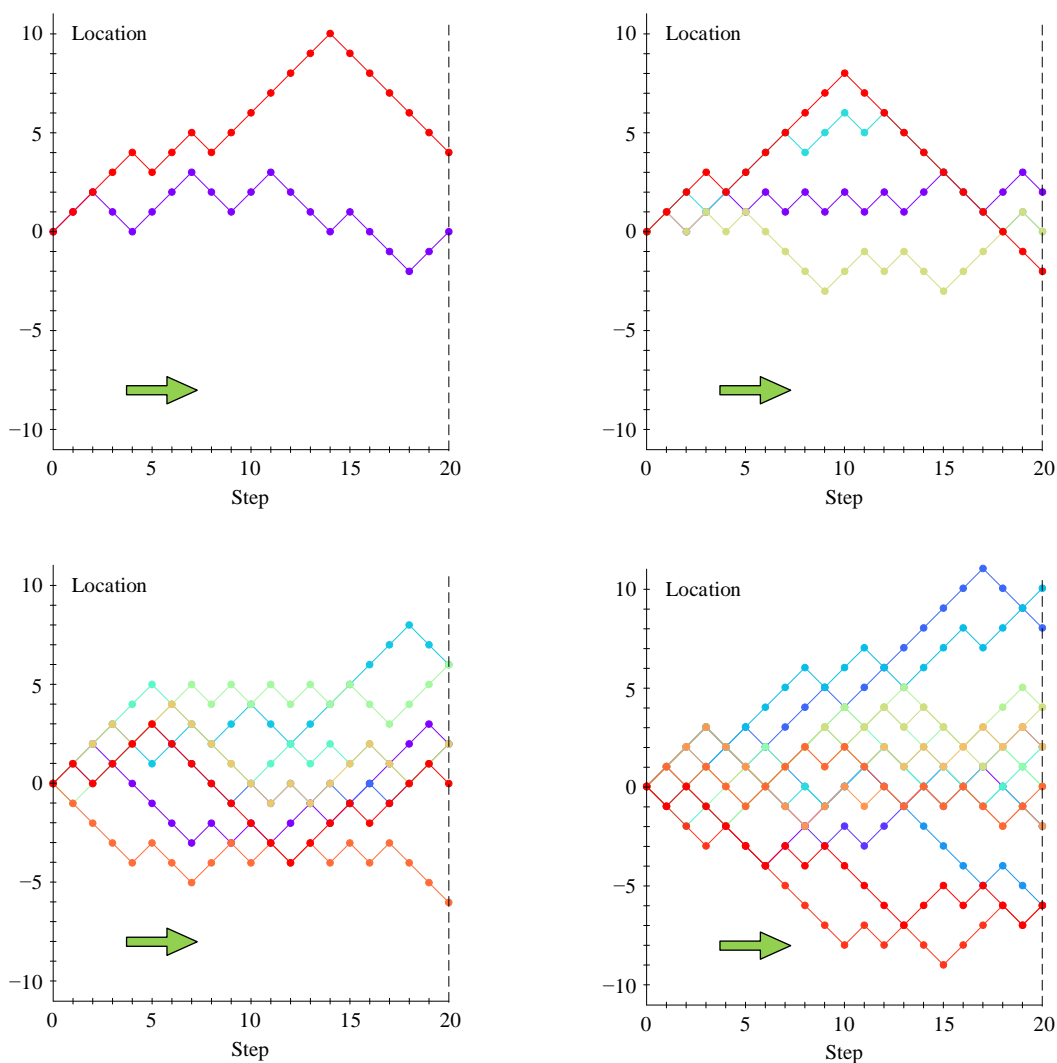


图 15. 随机漫步，2、4、8、16 条路径

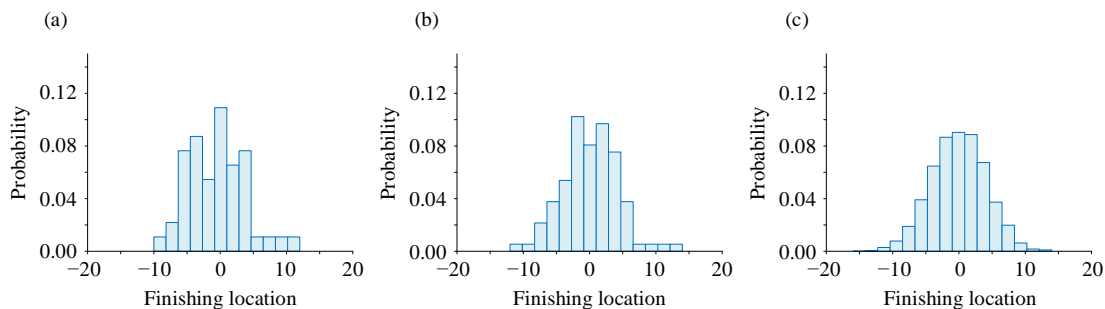


图 16. 随机漫步结束位置概率直方图，50、100、5000 条路径

实际上，二叉树网格限制了登山者向上或向下运动的步幅。更进一步，如果我们放开二叉树网格的限制，让登山者按照某种规律自行决定向上或向下的步幅，就可以得到图 17。

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

单看图中任意一条或几条路径，我们很难抓住任何规律；但是随着随机路径的数量不断增加，运动的规律就不言自明了。

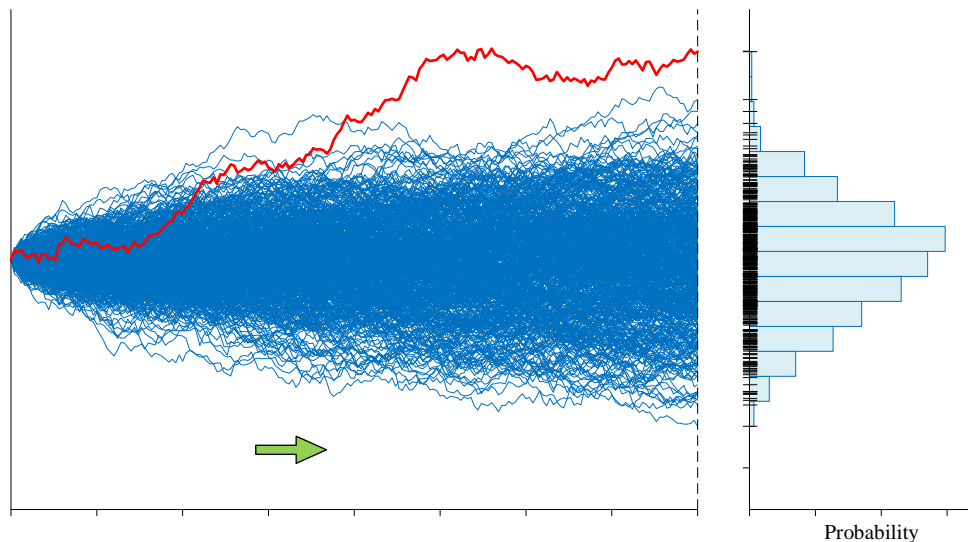


图 17. 不受二叉树网格限制的随机漫步

生活中这种随机中存在规律的情况不胜枚举。

举个例子，图 18 左图所示为一段时间内某只股票日收益率，红线以上为股价上涨，红线以下为股价下跌。单看某几天的股价涨跌很难把握住规律。但是，把一段时间内股价的日收益率数据绘制成直方图，如图 18 右图，我们就可以发现股价涨跌规律的端倪。

➔ 当然，为了得出更有意义的结论，我们还需要掌握更多的概率统计工具。本系列丛书将在《概率统计》和《数据科学》两本书中介绍更多概率统计知识，以及如何将它们应用到数据分析和预测实践中。

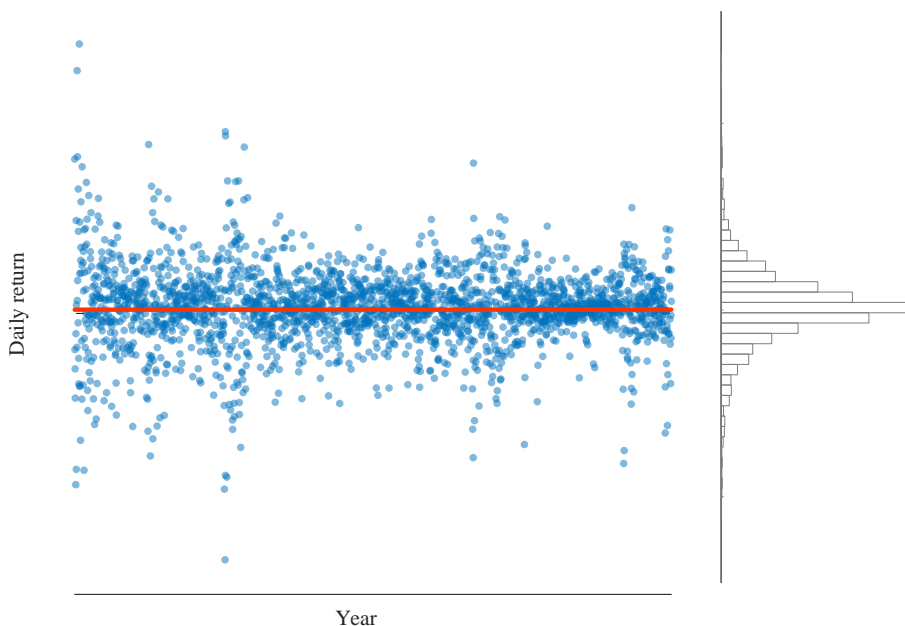


图 18. 股价日收益率和一段时间内的分布情况

高斯分布

观察图 16、图 17、图 18 直方图，似乎某种神秘的规律，或者一条神秘的曲线，呼之欲出。这就是“宇宙终极分布”——**高斯分布** (Gaussian distribution)。

高斯分布之众多概率分布中较为常用的一种。所谓**概率分布** (probability distribution) 描述随机变量取值的概率规律。

下式是高斯分布的**概率密度函数** (probability density function, PDF) 曲线解析式：

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (17)$$

其中， μ 为均值， σ 为标准差；下一章将介绍均值和标准差。

满足 (17) 的高斯分布常记做 $N(\mu, \sigma^2)$ 。连续型随机变量的概率密度函数 PDF 描述随机变量的在某个确定的取值点附近的可能性的函数。

(17) 实际上就是本书第 12 章介绍过的高斯函数通过函数变换得到的解析式。

图 19 所示为三个不同参数的一元高斯分布概率密度函数曲线。高斯分布，形态上极富美感；公式优雅精巧，包含数学中两个重要两个无理数 π 和 e 。高斯分布可以解释自然界很多纷繁复杂的规律；有人说，高斯分布似乎代表着宇宙幕后终极秩序。



本系列丛书《概率统计》一册将深入介绍高斯分布。

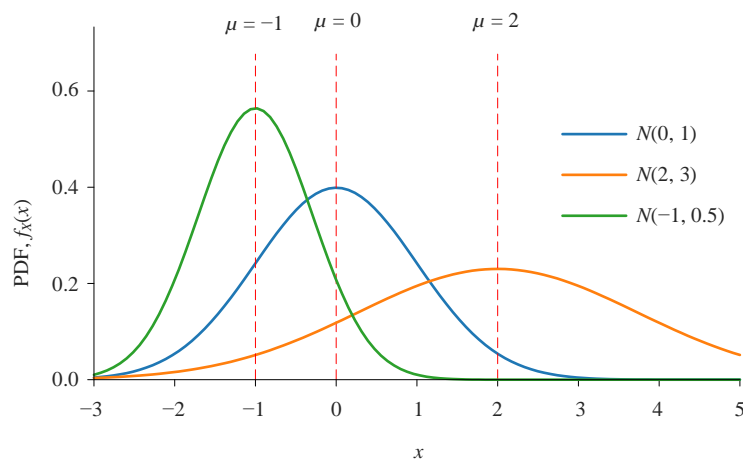
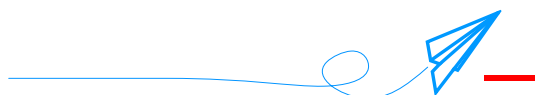


图 19. 三个不同一元高斯分布的概率密度函数曲线



本书前文利用杨辉三角，将算数、代数、几何、数列等数学知识联系起来，本章又将杨辉三角的触角伸到二叉树、概率和随机等概念；这正是丛书的重要目的之一——打破数学板块之间的壁垒，将它们有机联结。

希望大家通过本章的学习能够获得有关概率和随机的直观感受。随着，本系列丛书内容的不断深入，大家不但能够获得解释随机现象的数学工具，还能将它们用在解决数据科学和机器学习具体问题中去。