5 离散分布

Discrete Distributions

理想化的离散随机变量概率模型

> 究其本质，概率论无非是将生活常识简化成数学运算。
>
> ***The theory of probabilities is at bottom nothing but common sense reduced to calculation.***
>
> —— 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

- `matplotlib.pyplot.barh()` 绘制水平直方图
- `matplotlib.pyplot.stem()` 绘制火柴梗图
- `mpmath.pi` mpmath 库中的圆周率
- `numpy.bincount()` 统计列表中元素出现的个数
- `scipy.stats.bernoulli()` 伯努利分布
- `scipy.stats.binom()` 二项分布
- `scipy.stats.geom()` 几何分布
- `scipy.stats.hypergeom()` 超几何分布
- `scipy.stats.multinomial()` 多项分布
- `scipy.stats.poisson()` 泊松分布
- `scipy.stats.randint()` 离散均匀分布
- `seaborn.heatmap()` 产生热图

5.1 概率分布：高度理想化的数学模型

本书前文介绍的事件概率描述一次试验中某一个特定样本发生的可能性。想要了解某个随机变量在样本空间中不同样本的概率或概率密度，我们就需要**概率分布** (probability distribution)。

概率分布描述随机变量取值的概率规律，常用的概率分布都是高度理想化的数学模型。

我们知道随机变量分为离散和连续两种，因此概率分布也分为两类——**离散分布** (discrete distribution)、**连续分布** (continuous distribution)。

图 1 给出几种在数据科学、机器学习领域常用的概率分布。图 1 中，用火柴梗图描绘的是离散随机变量的 PMF，曲线描绘的是连续随机变量的 PDF。

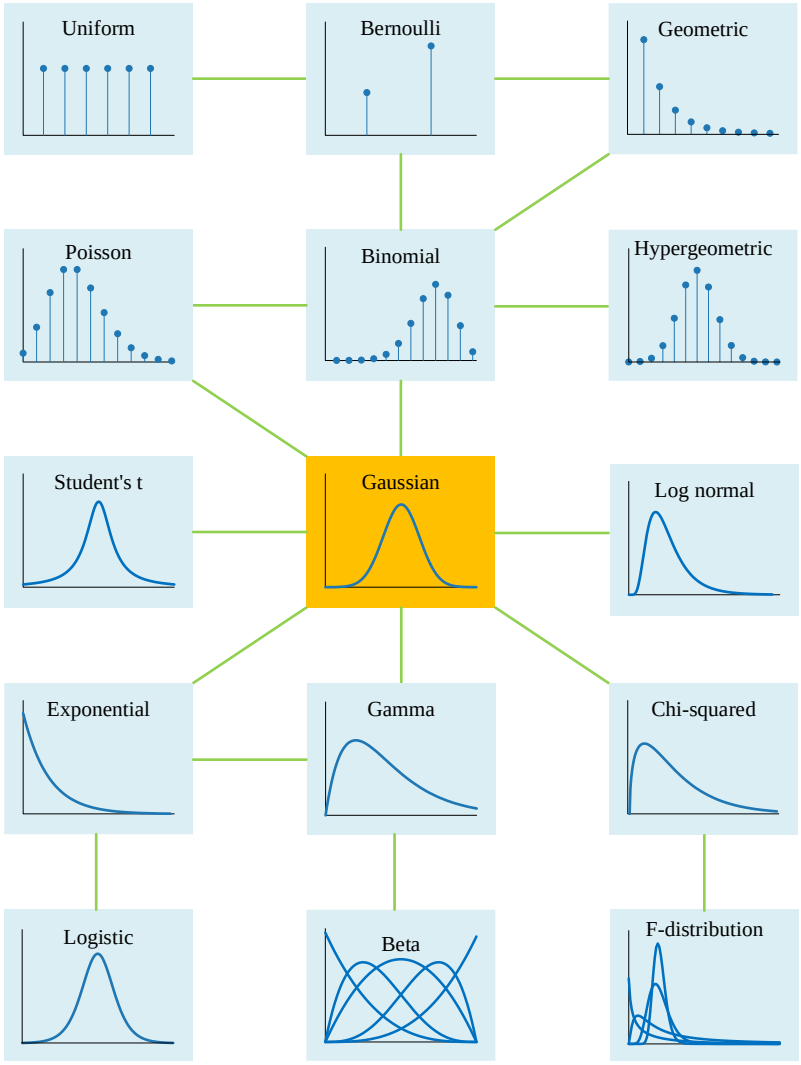


图 1. 常见的几种概率分布，给出多种分布样式

建议大家在学习概率分布时，首先考虑变量是离散还是连续，然后熟悉分布形状以及决定形状的参数，并掌握概率分布的应用场景。

再次强调，离散分布对应的是概率质量函数 PMF，其本质是概率。连续分布对应的是概率密度函数 PDF。概率密度函数积分、二重积分，有时甚至多重积分后，才得到概率值。

本章介绍常见离散分布，本书第 7 章讲解连续分布。建议大家把本章和第 7 章当成是“手册”来看待，以浏览的方式来学习，不需要死记硬背各种概率分布函数。大家后续在应用时，如果遇到某个特定概率分布时，可以回来查“手册”。

5.2 离散均匀分布：不分厚薄

离散均匀分布 (discrete uniform distribution) 应该是最简单的离散概率分布。离散型均匀分布分配给离散随机变量所有结果相等的权重。比如，离散随机变量 X 等概率地取得 [a, b] 区间内所有整数，取得每一个整数对应的概率为：

$$p_X(x)=\frac{1}{b-a+1},\quad x=a,a+1,\cdots b-1,b \tag{1}$$

注意，a、b 为正整数。上述概率质量函数 $p_X(x)$ 显然满足如下等式：

$$\sum_x p_X(x)=1 \tag{2}$$

期望值、方差

满足 (1) 这个离散均匀分布的 X 的期望值为：

$$\mathrm{E}(X)=\frac{a+b}{2} \tag{3}$$

X 的方差为：

$$\mathrm{var}(X)=\frac{(b-a+2)(b-a)}{12} \tag{4}$$

抛骰子试验

定义抛一枚色子结果为离散随机变量 X，假设获得六个不同点数为等概率，则 X 服从离散均匀分布。X 的概率质量函数为：

$$p_X(x)=1/6,\quad x=1,2,3,4,5,6 \tag{5}$$

X 的概率质量函数图像如图 2 所示。请大家自行计算 X 的期望值和方差。

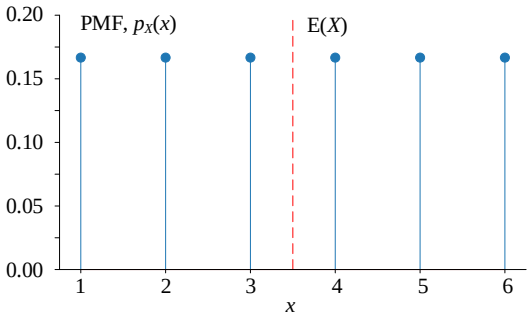


图 2. 离散均匀分布

Bk5_Ch05_01.py 代码文件绘制图 2。

圆周率

我们来看一个《数学要素》第 1 章提过的例子。图 3 所示为圆周率小数点后 1024 位数字的热图。

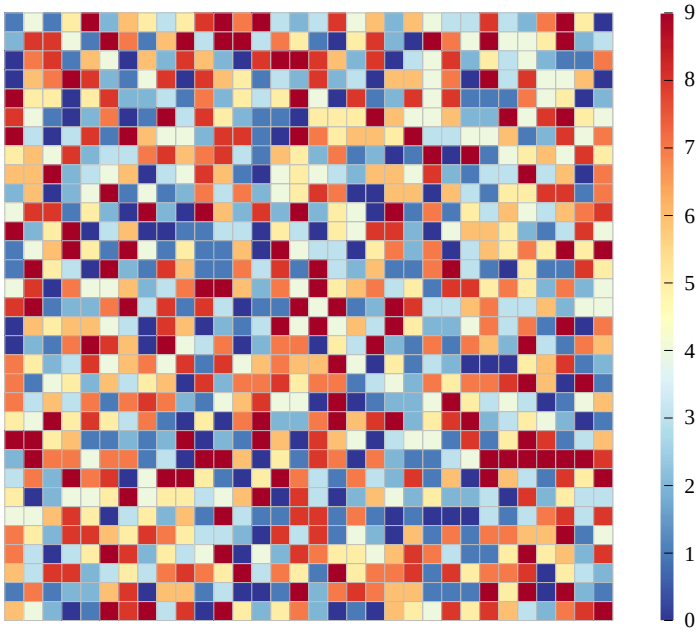


图 3. 圆周率小数点后 1024 位热图，图片来自《数学要素》第 1 章

热图中的数字看似没有任何规律。但是经过分析发现，随着数字数量越大，0 ~ 9 这些数字看上去服从离散均匀分布。图 4 所示为圆周率小数点后 100 位、1,000 位、10,000 位、100,000 位、1000,000 位 0 ~ 9 这些数字分布。

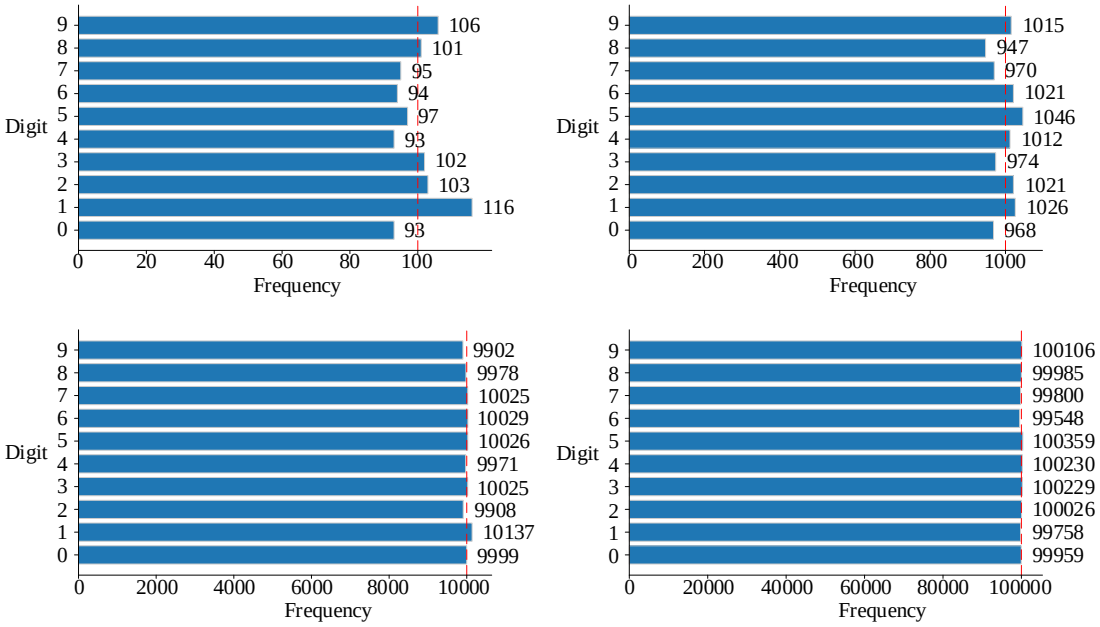


图 4. 圆周率小数点后数字的分布，100 位、1,000 位、10,000 位、100,000 位、1000,000 位

代码 Bk5_Ch05_02.py 绘制图 3 和图 4。

5.3 伯努利分布：非黑即白

在重复独立试验中，如果每次试验结果离散变量 X 仅有两个可能结果，比如 0、1，这种离散分布叫做**伯努利分布** (Bernoulli distribution)，对应的概率质量函数为：

$$p_X(x)=\begin{cases} p & x=1 \\ 1-p & x=0 \end{cases} \tag{6}$$

其中，p 满足 $0<p<1$。

(6) 还可以写成：

$$p_X(x)=p^x(1-p)^{1-x} \quad x\in\{0,1\} \tag{7}$$

请大家将 x = 0、1 分别代入上式检验 PMF 结果。

(6) 对应的概率质量函数显然满足归一化条件：

$$\sum_x p_X(x) = p + (1-p) = 1 \tag{8}$$

满足 (6) 中伯努利分布随机变量 X 的期望和方差分别为：

$$\begin{aligned} \mathrm{E}(X) &= p \\ \mathrm{var}(X) &= p(1-p) \end{aligned} \tag{9}$$

抛硬币

本书前文介绍的抛一枚硬币的试验就是常见的伯努利分布。如果硬币质地均匀，获得正面 (X = 1)、反面 (X = 0) 的概率均为 0.5，则 X 的概率质量函数为：

$$p_X(x) = \begin{cases} 0.5 & x = 1 \\ 0.5 & x = 0 \end{cases} \tag{10}$$

如果硬币质地不均匀，假设获得正面的概率为 0.6，则对应获得背面的概率为 1 – 0.6 = 0.4。则 X 的概率质量函数为：

$$p_X(x) = \begin{cases} 0.6 & x = 1 \\ 0.4 & x = 0 \end{cases} \tag{11}$$

请大家把 (10) 和 (11) 写成 (7) 这种形式。

Python 中伯努利分布函数常用 scipy.stats.bernoulli()。

抽样试验

再次强调，伯努利分布是离散分布，只有两种对立的可能结果，即结果样本空间只有 2 个元素。伯努利分布的参数只有 p。

从抽样试验角度，伯努利试验还可以看成是只有两个结果的放回抽样试验。比如，如图 5 所示，10 只动物中有 6 只兔子、4 只鸡。每次放回抽取一只动物，取到兔子的概率为 0.6，取到鸡的概率为 0.4。

图 5. 从抽样试验角度看伯努利试验

5.4 二项分布：杨辉三角

二项分布 (binomial distribution)，也叫二项式分布，建立在伯努利分布之上。

举个例子，一枚硬币抛 n 次，每次抛掷结果服从伯努利分布，即正面出现的概率为 p，反面出现的概率为 $1-p$，而且各次抛掷相互独立。进行 n 次独立的试验，令 X 为获得正面次数，X 对应的概率质量函数：

$$p_X(x) = C_n^x p^x (1-p)^{n-x}, \quad x = 0,1,...,n \tag{12}$$

(12) 所示二项式概率质量函数 $p_X(x)$ 满足归一化：

$$\begin{aligned} \sum_x p_X(x) &= C_n^0 p^0 (1-p)^n + C_n^1 p^1 (1-p)^{n-1} + \cdots + C_n^n p^n (1-p)^0 \\ &= \left(p + (1-p)\right)^n = 1 \end{aligned} \tag{13}$$

如果 X 服从 (12) 中给出的二项分布，X 的期望和方差分别为：

$$\begin{aligned} \mathrm{E}(X) &= n \cdot p \\ \mathrm{var}(X) &= n \cdot p(1-p) \end{aligned} \tag{14}$$

质地均匀硬币

为了方便大家理解二项分布，我们假定硬币质地均匀，即 $p = 0.5$。

先从 $n = 1$ 说起，也就是说试验中抛 1 枚均匀硬币。令 X 为正面为朝上的次数，X 的概率质量函数 PMF 为：

$$p_X(x) = \begin{cases} 1/2 & x = 0 \\ 1/2 & x = 1 \end{cases} \tag{15}$$

这本质上是伯努利分布。

当 $n = 2$，即抛 2 枚均匀硬币，X 的概率质量函数为：

$$p_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \end{cases} \tag{16}$$

抛 3 枚均匀硬币，X 的概率质量函数为：

$$p_X(x) = \begin{cases} C_3^0 \cdot (1/2)^3 = 1/8 & x = 0 \\ C_3^1 \cdot (1/2)^3 = 3/8 & x = 1 \\ C_3^2 \cdot (1/2)^3 = 3/8 & x = 2 \\ C_3^3 \cdot (1/2)^3 = 1/8 & x = 3 \end{cases} \tag{17}$$

试验中，抛 n 枚均匀硬币，令 X 为正面为朝上的次数，X 的概率质量函数为：

$$p_X(x)=\begin{cases} C_n^0\cdot(1/2)^n & x=0 \\ C_n^1\cdot(1/2)^n & x=1 \\ \cdots & \cdots \\ C_n^n\cdot(1/2)^n & x=n \end{cases} \tag{18}$$

图 6 所示为 $p = 0.5$ 时，n 取不同值，二项分布的概率质量函数分布。随着 n 不断增大，大家仿佛看到了“高斯分布”。请大家特别注意，高斯分布对应连续随机变量，而二项分布对应离散随机变量。

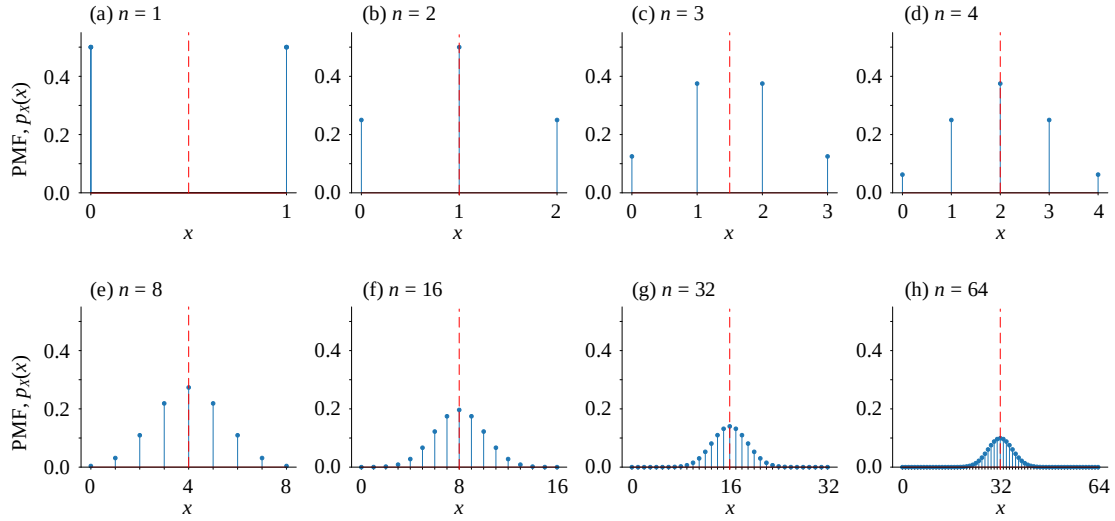


图 6. 二项分布，$p = 0.5$

质地不均匀硬币

如果硬币不均匀，假设正面朝上的概率为 $p = 0.8$。试验中，抛 n 枚均匀硬币，令 X 为正面为朝上的次数，X 的概率质量函数为：

$$p_X(x)=\begin{cases} C_n^0\cdot 0.8^0(1-0.8)^n & x=0 \\ C_n^1\cdot 0.8^1(1-0.8)^{n-1} & x=1 \\ \cdots & \cdots \\ C_n^n\cdot 0.8^n(1-0.8)^0 & x=n \end{cases} \tag{19}$$

图 7 所示为 $p = 0.8$ 时，n 取不同值，二项分布的概率质量函数分布。

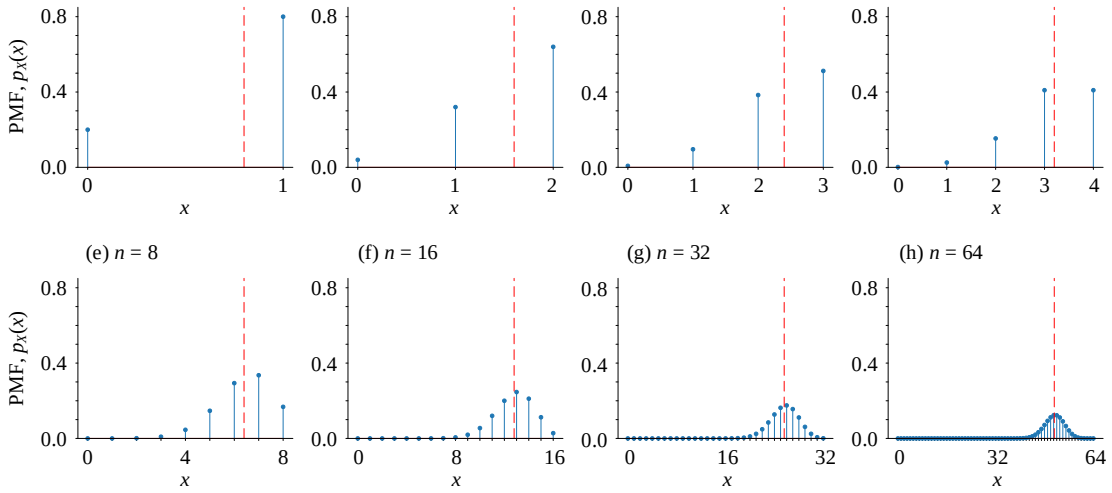


图 7. 二项分布，$p = 0.8$

显然，二项分布概率质量函数形状是由 n、p 两个参数确定下来的。容易发现，当 $p = 1/2$ 时，PMF 关于 $x = n/2$ 对称。当 $p > 1/2$ 时，PMF 图像偏向 n；当 $p < 1/2$ 时，PMF 图像偏向 0。随着 n 不断增大，分布的偏度逐渐变小，而且形状上不断近似高斯分布。

必须再次强调的是，二项分布对应离散随机变量，而高斯分布对应连续随机变量。二项分布 $p_X(x)$ 为概率质量函数，而高斯分布 $f_X(x)$ 为概率密度函数。

有放回 vs 不放回

总结来说，二项分布是 n 个独立进行的伯努利试验。二项分布 PMF 有两个参数——n、p。

从抽样试验角度，二项分布强调“独立”，每次抽取后再放回，这样总体本身不发生变化。还是利用鸡兔做例子，每次抽取时，取得兔子的概率为 0.6，取得鸡的概率为 0.4。计算 $n = 10$ 次有放回抽取中有 5 只兔子的概率，用的就是二项分布。

若是不放回抽取，即每次抽取之后不放回，则总体随之变化，分别取得鸡、兔的概率不断变化。二项分布则无法处理无放回抽样，我们需要用到超几何分布。超几何分布是本章后续要介绍的分布类型。

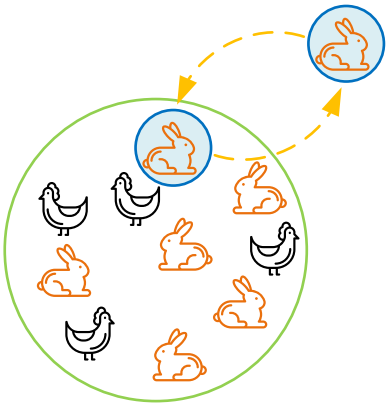
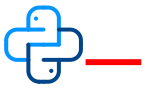


图 8. 从抽样试验角度看二项分布



代码 Bk5_Ch05_03.py 绘制图 6 和图 7。

5.4 多项分布：二项分布推广

多项分布 (multinomial distribution)，也叫多项式分布，是二项式分布的推广。多项分布的概率质量函数为：

$$p_{X_1,\ldots,X_K}(x_1,\ldots,x_K;n,p_1,\ldots,p_K)\begin{cases}\dfrac{n!}{(x_1!)\times(x_2!)\cdots\times(x_K!)}\times p_1^{x_1}\times\cdots\times p_K^{x_K} & \text{when } \sum_{i=1}^{K}x_i=n\\ 0 & \text{otherwise}\end{cases} \tag{20}$$

其中 x_i (i = 1, 2, …, K) 为非负整数，且 $\sum_{i=1}^{k}p_i=1$。

注意，为了避免混淆，本书用 "|" 引出条件概率中的条件，用分号 ";" 引出概率分布的参数。

举个例子

假设一个农场有大量动物，其中 60%为兔子，10%为猪，30%为鸡。如果随机抓取 8 只动物，其中有 2 只兔子、3 只猪、3 只鸡的概率为多少？

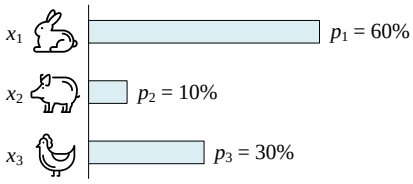


图 9. 农场兔、猪、鸡的比例

计算这个概率就用到了多项分布。当 K = 3 且 n = 8 时，多项式分布的概率质量函数为：

$$f(x_1,x_2,x_3;p_1,p_2,p_3)=\begin{cases}\dfrac{8!}{(x_1!)\times(x_2!)\times(x_K!)}\times p_1^{x_1}\times p_1^{x_2}\times p_3^{x_3} & \text{when } x_1+x_2+x_3=8\\ 0 & \text{otherwise}\end{cases} \tag{21}$$

其中，x_1、x_2、x_3 均为非负整数。

将 x_1 = 2，x_2 = 3，x_3 = 3，p_1 = 0.6，p_2 = 0.1，p_3 = 0.3 代入上式得到：

$$f\left(\underset{x_1}{2},\underset{x_2}{3},\underset{x_3}{3};\underset{p_1}{0.6},\underset{p_2}{0.1},\underset{p_3}{0.3}\right)=\frac{8!}{(2!)\times(3!)\times(3!)}\times 0.6^2\times 0.1^3\times 0.3^3\approx 0.0054 \tag{22}$$

散点图、热图、火柴梗图

下面，我们分别用三维散点图、二维散点图、热图、火柴梗图可视化多项分布。

给定参数，$n = 8$，$p_1 = 0.6$，$p_2 = 0.1$，$p_3 = 0.3$，多项分布的三维散点图如图 10 (a) 所示。图中每一个散点代表一个 (x_1, x_2, x_3) 组合，注意这三个均为非负整数。由于 $x_1 + x_2 + x_3 = 8$，所以 (x_1, x_2, x_3) 散点均在一个平面上。散点的颜色代表概率质量 PMF 值大小。

将这些散点投影在 x_1x_2 平面上，便得到图 10 (b)。这说明只要给定 x_1 和 x_2，根据 $x_3 = 8 - (x_1 + x_2)$，x_3 便确定下来。

图 11 所示为上述多项分布的 PMF 热图和散点图。

图 12、图 13 和图 14、图 15 可视化另外两组参数的多项分布，请大家自行比较分析。

多项分布常常和 Beta 分布 (第 7 章) 一起出现在**贝叶斯推断** (Bayesian inference) 中，这是本书第 21、22 章要介绍的内容。

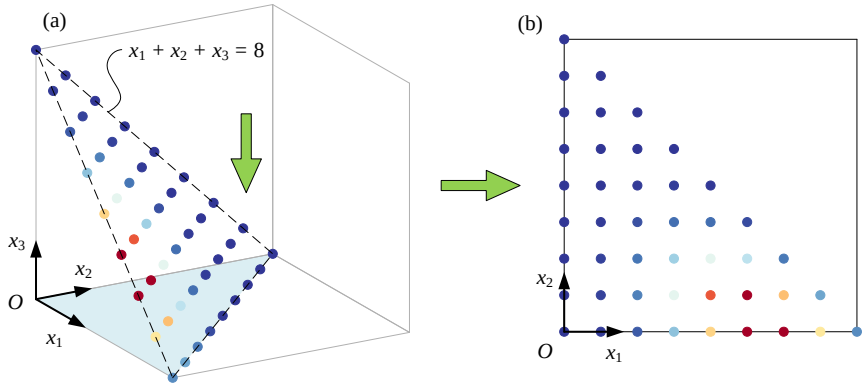


图 10. 多项分布 PMF 三维和平面散点图，$n = 8$，$p_1 = 0.6$，$p_2 = 0.1$，$p_3 = 0.3$

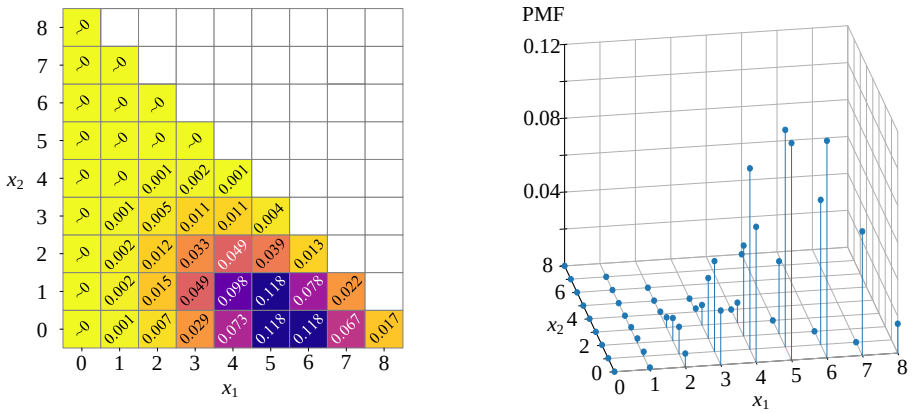


图 11. 多项分布 PMF 热图和火柴梗图，$n = 8$，$p_1 = 0.6$，$p_2 = 0.1$，$p_3 = 0.3$

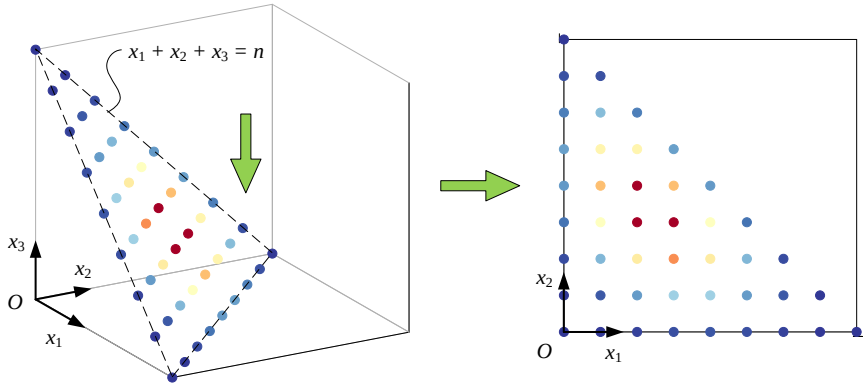


图 12. 多项分布 PMF 三维和平面散点图，$n = 8$，$p_1 = 0.3$，$p_2 = 0.4$，$p_3 = 0.3$

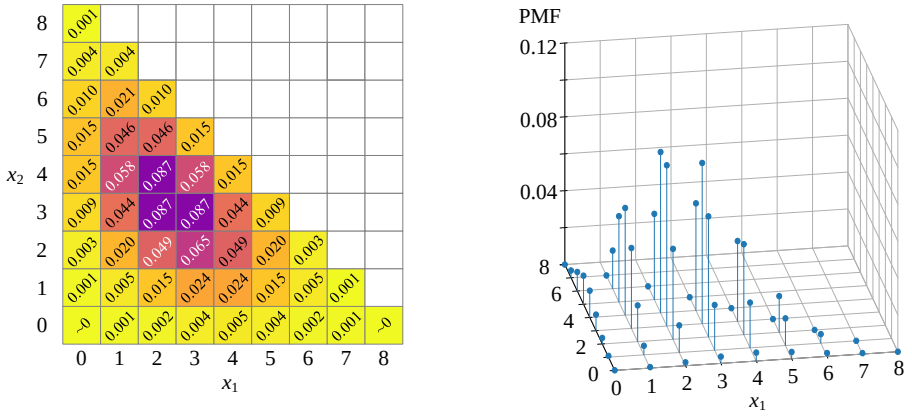


图 13. 多项分布 PMF 热图和火柴梗图，$n = 8$，$p_1 = 0.3$，$p_2 = 0.4$，$p_3 = 0.3$

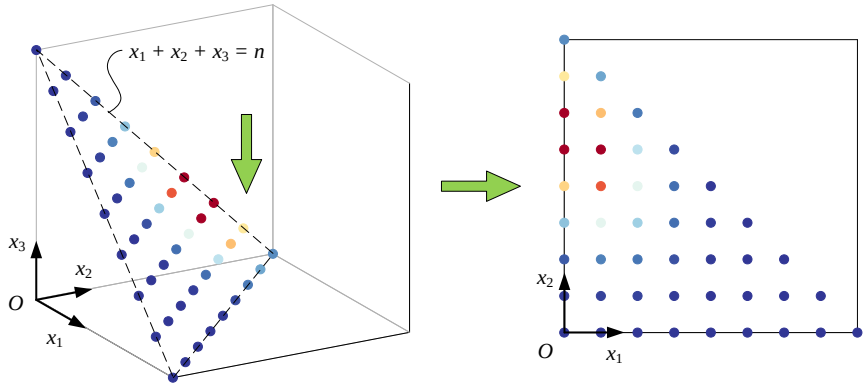


图 14. 多项分布 PMF 三维和平面散点图，$n = 8$，$p_1 = 0.1$，$p_2 = 0.6$，$p_3 = 0.3$

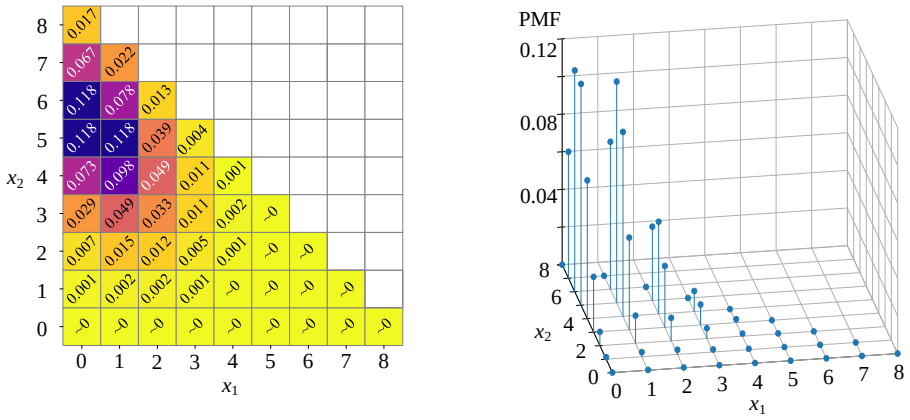


图 15. 多项分布 PMF 热图和火柴梗图，$n = 8$，$p_1 = 0.1$，$p_2 = 0.6$，$p_3 = 0.3$

Bk5_Ch05_04.py 绘制本节图像。

5.5 泊松分布：建模随机事件的发生次数

如果二项分布的试验次数 n 非常大，事件每次发生的概率 p 非常小，并且它们的乘积 np 存在有限的极限 λ，则这个二项分布趋近于另一种分布——**泊松分布** (Poisson distribution)。泊松分布的概率质量函数为：

$$p_X(x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad x = 0,1,2,... \tag{23}$$

图 16 所示为泊松分布概率质量函数随 λ 变化。

满足 (23) 泊松随机变量的期望和方差都是 λ：

$$\mathrm{E}(X) = \mathrm{var}(X) = \lambda \tag{24}$$

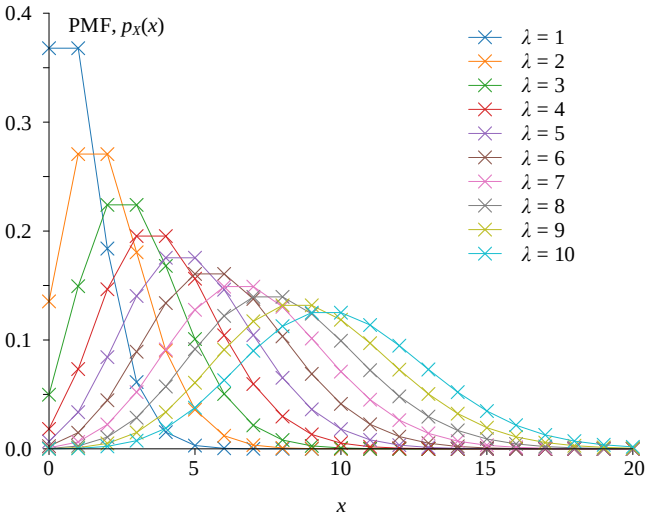


图 16. 泊松分布概率质量函数随 λ 变化

泊松分布是单参数离散概率分布，我们一般用泊松分布描述在给定的时间段、距离、面积等范围内随机事件发生的概率。应用泊松分布的例子包括每小时走入商店的人数，以及网络上每分钟的数据包丢包数等等。

代码 Bk5_Ch05_05.py 绘制图 16。

5.6 几何分布：滴水穿石

几何分布 (geometric distribution) 也是一个单参数概率分布，几何分布模拟一系列独立伯努利试验中一次成功之前的失败次数。其中，每次试验要么成功要么失败，并且任何单独试验的成功概率是恒定的。

比如，抛 n 次硬币 (伯努利试验)，前 $x-1$ 次均为反面，在第 x 次为正面。

在连续抛硬币的试验中，每次抛掷正面出现的概率为 p，反面出现的概率为 $1-p$，每次抛掷相互独立。令 X 为连续抛掷一枚硬币，直到第一次出现正面所需要的次数。X 的概率质量函数为：

$$p_X(x) = (1-p)^{x-1} p, \quad x = 1, 2, \ldots \tag{25}$$

满足 (25) 几何分布的离散随机变量 X 的期望和方差分别为：

$$\begin{aligned} \mathrm{E}(X) &= \frac{1}{p} \\ \mathrm{var}(X) &= \frac{1-p}{p^2} \end{aligned} \tag{26}$$

图 17 所示为当 $p = 0.5$ 时，几何分布概率质量函数 PMF 和 CDF。

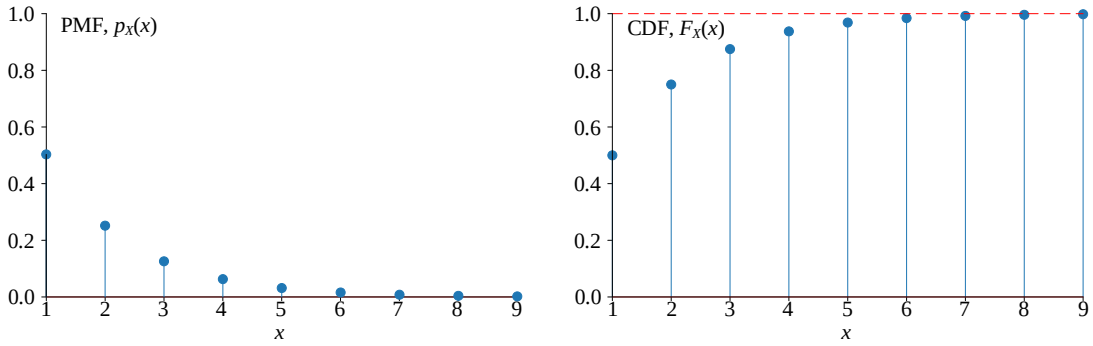


图 17. 几何分布概率质量函数 PMF 和 CDF，$p = 0.5$

图 18 所示为几何分布概率质量函数 PMF 随 p 变化。

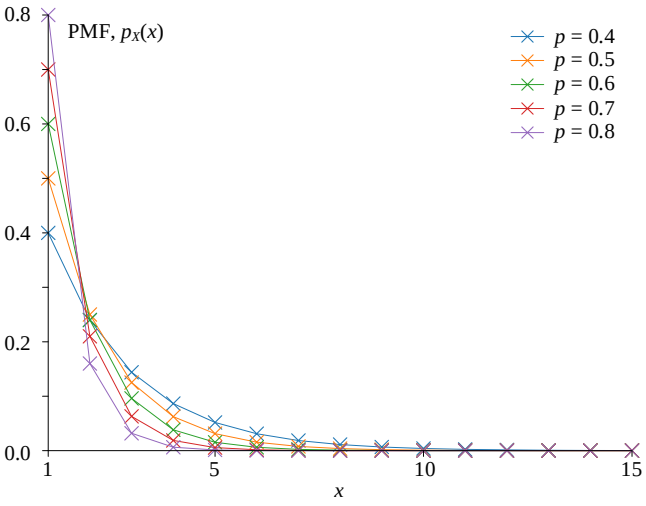


图 18. 几何分布概率质量函数 PMF 随 p 变化

代码 Bk5_Ch05_06.py 绘制图 17 和图 18。

5.7 超几何分布：不放回

我们在介绍二项分布时，特别强调二项分布在抽样时放回。如果抽样时不放回，我们便得到**超几何分布** (hypergeometric distribution)。

举个例子，假如某个农场总共有 N 个动物，其中 K 只兔子。从 N 只动物不放回抽取 n 个动物，其中有 x 只兔子的概率为：

$$p_X(x) = \frac{C_K^x C_{N-K}^{n-x}}{C_N^n}, \quad \max(0, n+K-N) \le x \le \min(K, n) \tag{27}$$

这个分布就是超几何分布。

比如，如图 19 所示，有 50 (N) 只动物，其中有 15 (K) 只兔子 (30%)。从 50 (N) 只动物中不放回地抽取 20 (n) 只动物，其中有 x 只兔子对应的概率为：

$$p_X(x) = \frac{C_{15}^x C_{35}^{20-x}}{C_{50}^{20}} \tag{28}$$

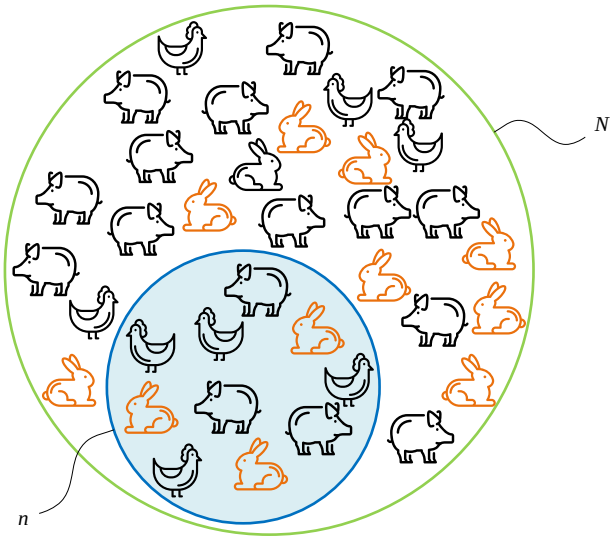


图 19. 超几何分布原理

上式中概率质量函数 $p_X(x)$ 对应的图像如图 20 所示。

总结来说，超几何分布的核心是“不放回”。超几何分布 PMF 输入有四个，N、K 描述整体，n、x 描述采样。

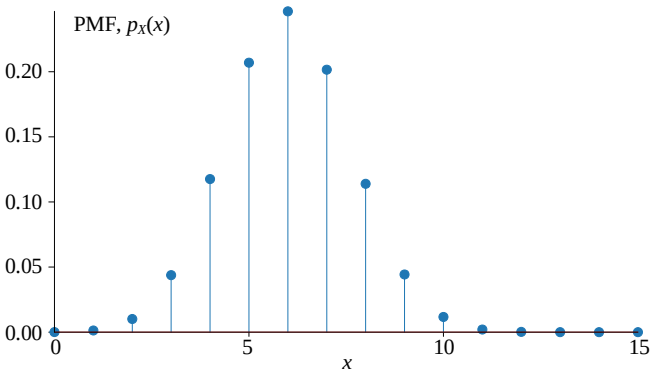


图 20. 超几何分布概率质量函数，$N = 50$，$K = 15$，$n = 20$

代码 Bk5_Ch05_07.py 绘制图 20。

二项分布 vs 超几何分布

注意，如果总体数量 N 很大，抽取数量 n 很小，不管抽样时是否放回，都可以用二项分布近似。

举个例子，兔子占整体的比例确定为 $p = 0.3$ (30%)，而动物总体数量分别为 $N = 100$、200、400、800 条件下，放回抽取 (二项分布)、不放回抽取 (超几何分布) $n = 20$ 只动物，兔子数量 x 对应概率分布如图 21 所示。

观察这四幅子图，我们发现当 N 不断增大，二项分布和超几何分布的 PMF 曲线逐渐靠近。

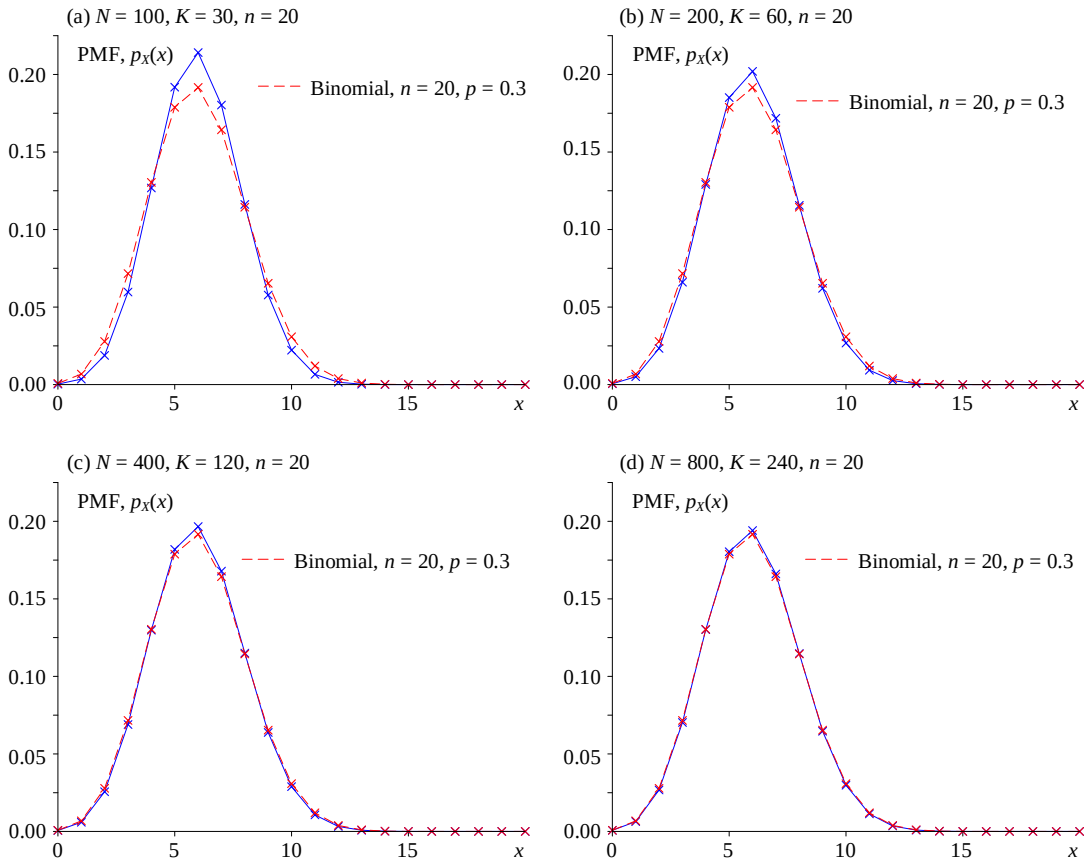


图 21. 超几何分布 PMF 和二项分布 PMF 关系

代码 Bk5_Ch05_08.py 绘制图 21。

各种分布之间的联系，请大家参考：

http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

http://www.math.wm.edu/~leemis/2008amstat.pdf