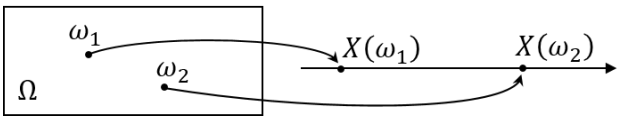
第 3 章　离散型随机变量

在很多随机现象中, 随机试验的结果可能与某些数值直接相关, 例如, 抛一枚骰子的点数分别为 $1, 2, \cdots, 6$; 国家一年内出生的婴儿数; 一批出厂的产品中包含的废品数; 等等. 有些看起来与数值无关的随机现象, 也可以通过数值来描述, 例如在抛硬币试验中, 每次试验结果为正面或反面朝上, 与数值没无关, 我们可以用 1 表示 '正面朝上', 用 0 表示 '反面朝上', 通过数值进行描述. 针对更一般的随机事件 A, 也可与数值产生联系, 如

$$X = \begin{cases} 1 & \text{如果事件 } A \text{ 发生,} \\ 0 & \text{如果事件 } A \text{ 不发生.} \end{cases}$$

针对随机现象中的每种结果 ω (即每个基本事件), 都能与实数值 $X(\omega)$ 建立某种数值对应关系, 并且随着基本事件 ω 的不同而 $X(\omega)$ 的取值也不同, 称这样的函数 $X = X(\omega)$ 为随机变量, 如下图所示:



定义 3.1　设 Ω 是一个样本空间, 如果对每个基本事件 $\omega \in \Omega$, 都对应于一个实数 $X(\omega)$, 称这样的单射实值函数 $X(\omega) : \Omega \to \mathbb{R}$ 为 **随机变量** (random variable), 一般简写为 X.

随机变量 X 的取值随试验结果的不同而不同, 具有一定的随机性; 由于各试验结果的出现具有一定的概率, X 的取值具有统计规律性, 因此随机变量与普通函数存在着本质的不同. 通过随机变量来描述随机现象或随机事件, 使得我们可以利用各种数学分析工具, 通过对随机变量的研究来分析随机现象. 可以用 $\{X \leqslant -\infty\}$ 表示不可能事件, 以及 $\{X \leqslant +\infty\}$ 表示必然事件. 一般用大写字母 X, Y, Z 表示随机变量. 下面给出一些随机变量的例子:

- 抛一枚骰子, 用随机变量 X 表示出现的点数, 则随机变量 $X \in \{1, 2, 3, 4, 5, 6\}$. 出现的点数不超过 4 的事件可表示为 $\{X \leqslant 4\}$; 出现偶数点的事件可表示为 $\{X = 2, 4, 6\}$.
- 用随机变量 X 表示一盏电灯的寿命, 其取值为 $[0, +\infty)$, 电灯寿命不超过 500 小时的事件可表示为 $\{X \leqslant 500\}$.

根据取值的类型, 可将随机变量分为离散型随机变量和非离散型随机变量. 若随机变量 X 的取值是有限的、或无限可列的, 则称 X 为 **离散型随机变量**; 若随机变量 X 的取值是无限不可列的, 则称 X 为 **非离散型随机变量**. 本章主要研究离散型随机变量.

3.1 离散型随机变量及分布列

离散型随机变量的取值是有限或无限可列的, 要完全刻画它的概率属性, 需要首先了解它所有可能的取值, 以及这些取值发生的概率.

定义 3.2 设随机变量 X 所有可能的取值为 $x_1, x_2, \cdots, x_k, \cdots$, 事件 $\{X = x_k\}$ 的概率为

$$p_k = P(X = x_k), \quad k = 1, 2, \cdots,$$

称之为随机变量 X 的 **概率分布列** 或 **概率分布**, 简称 **分布列**.

概率分布列能一目了然的看出随机变量的取值以及相应的概率, 也可以通过下面的表格给出:

X	x_1	x_2	$\cdots$	x_n	$\cdots$
P	p_1	p_2	$\cdots$	p_n	$\cdots$

根据概率的非负性和完备性可知分布列应具有如下性质:

性质 3.1 随机变量 X 的分布列 $p_k = P(X = x_k)$ 满足 $p_k \geqslant 0$ 和 $\sum_k p_k = 1$.

反之, 任何满足上面两条性质的数列 $\{p_k\}$, 都可以作为一个随机变量的分布列.

例 3.1 设随机变量 X 的分布列 $P(X = k) = c/4^k$ $(k = 0, 1, 2, \cdots)$, 求概率 $P(X = 1)$.

解 根据概率的完备性有

$$1 = \sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{c}{4^k} = \frac{4}{3}c,$$

求解得到 $c = 3/4$, 进一步有 $P(X = 1) = 3/16$.

例 3.2 给定常数 $\lambda > 0$, 随机变量 X 的分布列 $p(X = i) = c\lambda^i/i!$ $(i \geqslant 0)$, 求 $P(X > 2)$.

解 根据概率的完备性有

$$1 = \sum_{i=0}^{\infty} P(X = i) = c \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = c \cdot e^{\lambda}$$

从而得到 $c = e^{-\lambda}$, 进一步得到

$$P(X > 2) = 1 - P(X \leqslant 2) = 1 - p_0 - p_1 - p_2 = 1 - e^{-\lambda}(1 + \lambda + \lambda^2/2).$$

3.2 离散型随机变量的期望

针对一个具体的问题, 完全求解出概率分布列可能不是一件容易的事; 很多时候也不需要知道精确的概率分布列, 而是要掌握它的整体特征. 例如, 在统计某个地区的工资水平时, 我们可能更关心该地区工资的平均水平、贫富差距等特征, 而不是每个人的具体工资. 这些刻画随机变量某些方面特征的数值称为 **随机变量的数字特征**.

数字特征在概率统计中起着重要的作用, 它从宏观的角度刻画了随机变量某些基本特性, 有助于对随机变量的总体理解. 针对一些常用的随机变量, 我们可能只需要知道他们的一些数字特征, 就可以完全确定其概率分布. 常用的数字特征包括随机变量的期望、方差、相关系数和矩等, 本节介绍离散型随机变量的期望和性质, 其它数字特征在后续章节中介绍.

定义 3.3 设离散型随机变量 X 的分布列为 $p_k = P(X = x_k)$ $(k \geqslant 1)$. 若级数

$$\sum_{k=1}^{\infty} p_k x_k$$

绝对收敛, 则称该级数和为随机变量 X 的 **期望** (expectation), 又称为 **均值** (mean), 记为 $E(X)$, 即

$$E(X) = \sum_{k=1}^{\infty} p_k x_k \ .$$

期望 $E(X)$ 反映随机变量 X 的平均值, 由随机变量的分布列决定, 是常量而不是变量, 其本质是随机变量的取值 x_i 根据概率 p_i 加权所得. 级数的绝对收敛确保了期望 $E(X)$ 的唯一性, 即级数和不会随级数各项次序的改变而改变. 除非特别说明, 我们通常都直接利用定义计算期望, 不需考虑其绝对收敛性.

例 3.3 随意掷一枚骰子, X 表示观察到的点数, 求 $E[X]$.

解 随机变量 X 的取值为 $1, 2, \cdots, 6$, 且每点等可能发生, 其分布列为 $P(X = k) = 1/6$, $k \in [6]$. 因此随机变量 X 的期望为 $E(X) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 7/2$.

我们来看一个期望不存在的例子:

例 3.4 设随机变量 X 的分布列为 $P\left(X = (-2)^k/k\right) = 1/2^k$, $k = 1, 2, \cdots$, 求期望 $E(X)$.

解 尽管根据定义有

$$E(X) = \sum_{k=1}^{+\infty} P\left(X = \frac{(-2)^k}{k}\right) \frac{(-2)^k}{k} = \sum_{k=1}^{+\infty} \frac{(-1)^k}{k} = -\ln 2.$$

但是

$$\sum_{k=1}^{+\infty} P\left(X = \frac{(-2)^k}{k}\right) \left|\frac{(-2)^k}{k}\right| = \sum_{k=1}^{+\infty} \frac{1}{k} \to +\infty.$$

该级数并非绝对收敛, 其级数和可能随着求和顺序的改变而改变, 级数和并非唯一的数值, 故该随机变量的期望 $E(X)$ 不存在.

例 3.5 若 n 把钥匙中只有一把能开门, 现随机选取一把钥匙开门, 若打不开门则去掉该钥匙, 再随机选取剩下的钥匙进行尝试, 求打开门需要尝试的平均次数.

解 设随机变量 X 表示尝试开门的次数, 其分布列为

$$P(X=k)=\frac{\binom{n-1}{k-1}}{\binom{n}{k-1}}\cdot\frac{1}{n-k+1}=\frac{1}{n}\qquad k\in[n]\ .$$

因此打开门需要尝试的平均次数

$$E(X)=\sum_{k=1}^{n}\frac{k}{n}=\frac{(1+n)n}{2n}=\frac{n+1}{2}\ .$$

下面介绍期望的一些性质, 除非特别说明, 这些性质不仅对离散型随机变量成立, 对其它任何类型的随机变量都成立. 为了证明的可读性, 在证明过程中仅考虑离散型随机变量.

性质 3.2 设 $c\in\mathbb{R}$ 是常数, 若随机变量 $X\equiv c$, 则 $E(X)=c$.

性质 3.3 若随机变量 X 的取值非负, 即 $X\geqslant 0$, 则 $E(X)\geqslant 0$.

性质 3.4 对随机变量 X 和常数 $a,b\in\mathbb{R}$, 有 $E(aX+b)=aE(X)+b$.

证明 设随机变量 X 的分布列为 $p_k=P(X=x_k)$, 则随机变量 $Y=aX+b$ 的分布列为 $p_k=P(Y=ax_k+b)$, 进而有

$$E[aX+b]=\sum_{k\geqslant 1}(ax_k+b)p_k=a\sum_{k\geqslant 1}x_kp_k+b\sum_{k\geqslant 1}p_k=aE[X]+b.$$

性质 3.5 若离散型随机变量 X 所有可能的取值为非负整数 $\{0,1,2,\cdots\}$, 则

$$E(X)=\sum_{i=1}^{+\infty}P(X\geqslant i)\ .$$

证明 根据期望的定义有

$$E[X]=\sum_{j=1}^{+\infty}jP(X=j)=\sum_{j=1}^{+\infty}\sum_{i=1}^{j}P(X=j)=\sum_{i=1}^{+\infty}\sum_{j=i}^{+\infty}P(X=j)=\sum_{i=1}^{+\infty}P(X\geqslant i)\ ,$$

由此完成证明.

针对随机变量的函数的期望, 有如下定理:

定理 3.1 设离散型随机变量 X 的分布列为 $p_k = P(X = x_k)$ $(k \geqslant 1)$. 对任意的实值函数 $g : \mathbb{R} \to \mathbb{R}$, 若级数 $\sum_{k\geqslant 1} g(x_k)p_k$ 绝对收敛, 则有

$$E[g(X)] = \sum_{k=1}^{\infty} g(x_k)p_k.$$

证明 证明的核心思想是利用无穷级数的绝对收敛确保任意重排后的级数和等于原级数和. 根据题意有 X 的分布列为 $p_k = P(X = x_k)$ 和随机变量函数 $Y = g(X)$ 有

X	x_1	x_2	$\cdots$	x_n	$\cdots$
P	p_1	p_2	$\cdots$	p_n	$\cdots$
Y	y_1	y_2	$\cdots$	y_n	$\cdots$

其中 $y_i = g(x_i)$. 上面的表格给出了随机变量 X 的分布列, 但并没给出随机变量 Y 的分布列, 因为可能存在 $y_i = g(x_i) = y_j = g(x_j)$. 为了得到随机变量 Y 的分布列, 需要将 $x_1, x_2, \ldots, x_n, \ldots$ 进行重新排列分组为

$$\underbrace{x_{1,1}, x_{1,2}, \cdots, x_{1,k_1}}_{y_1'=g(x_{1,j})\ (j\in[k_1])},\quad \underbrace{x_{2,1}, x_{2,2}, \cdots, x_{2,k_2}}_{y_2'=g(x_{2,j})\ (j\in[k_2])},\quad \cdots,\quad \underbrace{x_{n,1}, x_{n,2}, \cdots, x_{n,k_n}}_{y_n'=g(x_{n,j})\ (j\in[k_n])},\quad \cdots$$

满足当 $i \neq j$ 时有 $y_i' \neq y_j'$ 成立. 由此可得随机变量 Y 的分布列为

$$P(Y = y_i') = \sum_{j=1}^{k_i} p_{i,j} = \sum_{k\geqslant 1, y_i'=g(x_k)} p_k\ ,$$

进一步得到随机变量 Y 的期望为

$$E[Y] = \sum_{i=1}^{\infty} y_i' P[Y = y_i'] = \sum_{i=1}^{\infty} y_i' \sum_{j=1}^{k_i} p_{i,j} = \sum_{i=1}^{\infty}\sum_{j=1}^{k_i} g(x_{i,j})p_{i,j} = \sum_{k=1}^{\infty} g(x_k)p_k\ ,$$

最后一个等式成立是因为绝对收敛的无穷级数在重排前与重排后其级数和不变.

基于上述定理, 我们可以直接计算随机变量 $Y = g(X)$ 的期望, 而不需要知道 Y 的分布列, 即通过 X 的分布列计算期望 $E[Y]$. 此外基于该定理有

推论 3.1 设 X 是离散型随机变量, 以及 $g_i : \mathbb{R} \to \mathbb{R}$ 是实值函数 $(i \in [n])$. 若期望 $E(g_i(X))$ 存在, 则对任意常数 $c_1, c_2, \ldots, c_n$ 有 $E\left(\sum_{i=1}^n c_i g_i(X)\right) = \sum_{i=1}^n c_i E(g_i(X))$ 成立.

基于此推论很容易得到

$$E(X^4 + \sin(X) + 4) = E(X^4) + E(\sin(X)) + 4.$$

最后探讨当函数 $g(x)$ 满足什么样的性质时, 期望 $E(g(X))$ 和 $g(E(X))$ 之间都存在一定的大小比较关系. 相关的知识在实际应用和科研中具有重要意义, 因为即使不知道随机变量的具体概率分布, 仍可以对期望进行一定的估计或推理. 看一个例子: 设离散型随机变量 X 的分布列为 $P(X=1)=P(X=2)=P(X=0)=1/3$, 很容易发现

$$(E(X))^2 \leqslant E(X^2) \quad \text{和} \quad \sqrt{E(X)} \geqslant E(\sqrt{X}).$$

针对更一般的情况, 考虑两类函数:

定义 3.4 设函数 $g\colon [a,b]\to\mathbb{R}$, 对任意 $x_1,x_2\in[a,b]$ 和 $\lambda\in[0,1]$,

- 若 $g(\lambda x_1+(1-\lambda)x_2)\leqslant \lambda g(x_1)+(1-\lambda)g(x_2)$, 则称函数 $g(x)$ 是定义在 $[a,b]$ 上的 **凸函数**;
- 若 $g(\lambda x_1+(1-\lambda)x_2)\geqslant \lambda g(x_1)+(1-\lambda)g(x_2)$, 则称函数 $g(x)$ 是定义在 $[a,b]$ 上的 **凹函数**.

凸函数和凹函数具有很多良好的数学性质, 例如凸函数的一阶导数单调性、二阶导数小于或等于零, 大家可以参考一些数学分析或优化书籍. 下面介绍著名的 **琴生不等式** (Jensen's inequality).

定理 3.2 设随机变量 $X\in[a,b]$ 和实值函数 $g\colon [a,b]\to\mathbb{R}$,

- 若 $g(x)$ 在 $[a,b]$ 上是凸函数, 则有 $g(E(X))\leqslant E(g(X))$;
- 若 $g(x)$ 在 $[a,b]$ 上是凹函数, 则有 $g(E(X))\geqslant E(g(X))$.

定理 3.2 中的不等式对所有的随机变量都成立. 即使在不知道随机变量 X 的概率分布情况下, 根据该定理可知

$$(E(X))^2\leqslant E(X^2),\qquad \sqrt{E(X)}\geqslant E(\sqrt{X})\qquad \text{和}\qquad e^{E(X)}\leqslant E(e^X)\ .$$

证明 这里仅给出离散型随机变量具有有限个取值和凸函数的证明. 设随机变量 X 的取值为 $x_1,x_2,\cdots,x_n$, 以及它的分布列为 $p_k=P(X=x_k)>0$, 易知 $\sum_k p_k=1$. 我们需要证明的不等式为

$$g(p_1x_1+p_2x_2+\cdots+p_nx_n)\leqslant p_1g(x_1)+p_2g(x_2)+\cdots+p_ng(x_n). \tag{3.1}$$

针对上式采用归纳法证明, 当 $n=2$ 时利用凸函数的定义结论显然成立. 不妨假设当 $n=m-1$ 时 (3.1) 成立, 下面证明当 $n=m$ 时 (3.1) 亦成立. 首先有

$$\begin{aligned} g(p_1x_1+p_2x_2+\cdots+p_mx_m) &= g\left(p_1x_1+(1-p_1)\left[\frac{p_2}{1-p_1}x_2+\cdots+\frac{p_m}{1-p_1}x_m\right]\right)\\ &\leqslant p_1g(x_1)+(1-p_1)g\left(\frac{p_2}{1-p_1}x_2+\cdots+\frac{p_m}{1-p_1}x_m\right)\ , \end{aligned}$$

这里将凸函数的定义应用到两个点 x_1 和 $x_1'=(x_2p_2+\cdots+x_mp_m)/(1-p_1)$. 容易发现 $p_i/(1-p_1)\geqslant 0$

且 $\sum_{i=2}^{m} p_i/(1-p_1)=1$, 根据归纳假设有

$$g\left(\frac{p_2}{1-p_1}x_2+\cdots+\frac{p_m}{1-p_1}x_m\right)\leqslant\frac{p_2}{1-p_1}g(x_2)+\cdots+\frac{p_m}{1-p_1}g(x_m),$$

由此可完成证明.

3.3 离散型随机变量的方差

数学期望反映了随机变量的平均值, 在很多实际应用中我们不仅仅要知道随机变量的平均值, 还需要进一步了解随机变量的取值与期望之间的偏离程度. 例如, 考虑三个随机变量 X, Y 和 Z, 它们的分布列分别为

$$P(X=0)=1;\quad P(Y=1)=P(Y=-1)=1/2;\quad P(Z=2)=1/5, P(Z=-1/2)=4/5.$$

容易得到 $E(X)=E(Y)=E(Z)=0$, 即三个随机变量的期望相同. 然而很显然这三个随机变量存在着明显的差异, 如何刻画它们的不同之处, 可以考虑三个随机变量的取值与期望的偏离程度, 即本节所研究随机变量的方差.

定义 3.5 设离散随机变量 X 的分布列为 $p_k=P(X=x_k)>0$, 若期望 $E(X)$ 和 $E(X-E(X))^2$ 存在, 则称 $E(X-E(X))^2$ 为随机变量 X 的 **方差** (variance), 记为 $\mathrm{Var}(X)$, 即

$$\mathrm{Var}(X)=E(X-E(X))^2=\sum_k p_k(x_k-E(X))^2=\sum_k p_k\left(x_k-\sum_k x_k p_k\right)^2. \tag{3.2}$$

称 $\sqrt{\mathrm{Var}(X)}$ 为 **标准差** (standard deviation), 记为 $\sigma(X)$.

结合方差的定义和期望的性质有

$$\begin{aligned}\mathrm{Var}(X) &= E(X-E(X))^2=E(X^2-2XE(X)+E^2(X))\\ &= E(X^2)-2E(X)E(X)+(E(X))^2=E(X^2)-(E(X))^2,\end{aligned}$$

由此给出方差的另一种定义

$$\mathrm{Var}(X)=E(X^2)-(E(X))^2. \tag{3.3}$$

尽管方差的两种定义等价, 然而在实际应用中却存在着不同的用处, (3.3) 给出了方差的物理含义, 而 (3.3) 更有利于方差的计算, 例如,

例 3.6 设随机变量 X 的分布列为 $P(X=x_k)=1/n$ ($k\in[n]$), 这里 $x_1,x_2,\cdots,x_n$ 的数值各不相同, 需遍历数据几次才能计算出方差 $\mathrm{Var}(X)$.

解 若采用 $\mathrm{Var}(X)=E(X-E(X))^2$, 则需要遍历数据 $x_1,x_2,\cdots,x_n$ 两次, 第一次计算期望 $E(X)$, 第二次计算方差 $\mathrm{Var}(X)$.

若采用 $\mathrm{Var}(X) = E(X^2) - (E(X))^2$, 则只需要遍历数据 $x_1, x_2, \cdots, x_n$ 一次, 在遍历数据的过程中计算 $E(X^2)$ 和 $(E(X))^2$, 从而计算方差. 在此过程中也不需要将全部数据 $x_1, x_2, \cdots, x_n$ 存在内存中, 可以一个个轮流存取数据.

下面给出方差的性质:

性质 3.6 设 $c \in \mathbb{R}$ 是常数, 若随机变量 $X \equiv c$, 则 $\mathrm{Var}(X) = 0$.

性质 3.7 对随机变量 X 和常数 $a, b \in \mathbb{R}$, 有

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

证明 根据期望的性质有 $E(aX + b) = aE(X) + b$, 代入可得

$$\mathrm{Var}(aX + b) = E(aX + b - E(aX + b))^2 = a^2 E(X - E(X))^2 = a^2 \mathrm{Var}(X).$$

值得注意的是, 方差通常不具有线性性, 即 $\mathrm{Var}(f(X) + g(X)) \neq \mathrm{Var}(f(X)) + \mathrm{Var}(g(X))$.

性质 3.8 对随机变量 X 和常数 $a \in \mathbb{R}$, 有

$$\mathrm{Var}(X) = E(X - E(X))^2 \leqslant E(X - a)^2.$$

证明 根据期望的性质有

$$\begin{aligned} E(X - c)^2 &= E(X - E(X) + E(X) - c)^2 \\ &= E(X - E(X))^2 + E[(X - E(X))(E(X) - c)] + (E(X) - c)^2 \\ &= E(X - E(X))^2 + (E(X) - c)^2 \\ &\geqslant E(X - E(X))^2, \end{aligned}$$

从而完成证明.

定理 3.3 (Bhatia-Davis不等式) 对随机变量 $X \in [a, b]$, 有

$$\mathrm{Var}[X] \leqslant (b - E(X))(E(X) - a) \leqslant (b - a)^2/4.$$

证明 对任意随机变量 $X \in [a, b]$, 有

$$(b - X)(X - a) \geqslant 0,$$

两边同时对随机变量取期望, 整理可得

$$E(X^2) \leqslant (a + b)E(X) - ab.$$

根据方差的定义有

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 = -(E(X))^2 + (a+b)E(X) - ab = (b - E(X))(E(X) - a).$$

利用二次函数 $f(t) = (b-t)(t-a) = -t^2 + (a+b)t - ab$ 的最大值可得

$$(b - E(X))(E(X) - a) \leqslant (b-a)^2/4.$$

3.4 常用离散型随机变量

本节介绍几种常用的离散型随机变量, 并研究它们的性质.

3.4.1 0-1分布

0-1分布是概率统计中最经典、最简单的分布, 是很多概率模型的基础.

定义 3.6 设随机变量 X 的分布列 $P(X=1) = p$, $P(X=0) = 1-p$, 等价于

$$P(X=k) = p^k(1-p)^{1-k} \qquad k = 0, 1,$$

则称随机变量 X 服从 **参数为 p 的 0-1 分布**, 又称 **两点分布**, 或 **伯努利分布** (Bernoulli distribution), 记 $X \sim \mathrm{Ber}(p)$.

0-1 分布也可以通过表格表示为

X	0	1
P	$1-p$	p

根据定义容易得到

引理 3.1 若随机变量 $X \sim \mathrm{Ber}(p)$, 则有 $E(X) = p$ 和 $\mathrm{Var}(X) = p(1-p)$.

由此可知 0-1 分布也可由它的数学期望唯一确定.

若一次试验只考虑事件 A 发生或不发生两种情况, 称这样的试验为 **伯努利试验**, 可以通过 0-1 分布来描述伯努利试验:

$$X = \begin{cases} 1 & \text{若事件 } A \text{ 发生,} \\ 0 & \text{否则.} \end{cases}$$

此时容易得到 $E[X] = P(A)$, 即随机变量 X 的期望等于事件 A 发生的概率.

3.4.2 二项分布

伯努利试验只考虑事件 A 发生或不发生两种结果, 不妨设事件 A 发生的概率 $P(A) = p \in (0,1)$. 将一个伯努利试验独立重复地进行 n 次, 称这一系列重复的独立试验为 n **重伯努利试验**. 它是一种非常重要的概率模型, 衍生出很多的概率分布.

在 n 重伯努利试验中, 我们关心随机事件 A 发生了多少次, 用随机变量 X 表示, 其所有可能的取值为 $0,1,2,\cdots,n$. 随机事件 $\{X=k\}$ 表示在 n 重伯努利试验中事件 A 发生了 k 次, 到底是哪 k 次发生的, 共有 $\binom{n}{k}$ 种不同的情况. 针对一种具体的情况, 不妨设前 k 次事件 A 发生, 后 $n-k$ 次事件 A 不发生, 此种情况发生的概率为

$$\underbrace{p\times p\times\cdots\times p}_{k\text{ 个}}\times\underbrace{(1-p)\times(1-p)\times\cdots\times(1-p)}_{n-k\text{ 个}}=p^k(1-p)^{n-p}.$$

由此可知在 n 重伯努利试验中事件 A 发生了 k 次的概率为 $P(X=k)=\binom{n}{k}p^k(1-p)^{n-p}$.

定义 3.7 若随机变量 X 的分布列为

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}\qquad k=0,1,2,\ldots,n, \tag{3.4}$$

则称随机变量 X 服从 **参数为 n 和 p 的二项分布** (binomial distribution), 记 $X\sim B(n,p)$.

容易发现 (3.4) 中 $P(X=k)$ 是二项式 $(1-p+xp)^n$ 展开式中 x^k 项的系数, 所以该分布被称为二项分布. 进一步可检验

$$\sum_{k=0}^{n}P(X=k)=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}=(p+1-p)^n=1\ .$$

若 $n=1$, 则二项分布退化为 0-1 分布, 即 $B(1,p)=\mathrm{Ber}(p)$. 关于二项分布的数字特征有

引理 3.2 若随机变量 $X\sim B(n,p)$, 则有 $E(X)=np$ 和 $\mathrm{Var}(X)=np(1-p)$.

若知道二项分布的期望和方差, 可反解出参数 n 和 p, 因此二项分布可由它的期望和方差唯一确定.

证明 根据定义有

$$E(X)=\sum_{k=0}^{n}P(X=k)k=\sum_{k=1}^{n}k\binom{n}{k}p^k(1-p)^{n-k}=(1-p)^n\sum_{k=1}^{n}\binom{n}{k}k\left(\frac{p}{1-p}\right)^k\ .$$

对二项展开式 $(1+x)^n=\sum_{k=0}^{n}\binom{n}{k}x^k$ 两边同时求导后乘 x 可得

$$nx(1+x)^{n-1}=\sum_{k=1}^{n}\binom{n}{k}kx^k\ ,$$

将 $x=p/(1-p)$ 带入上式可得

$$E(X)=(1-p)^n\sum_{k=0}^{n}\binom{n}{k}k\left(\frac{p}{1-p}\right)^k=(1-p)^n\frac{np}{1-p}\frac{1}{(1-p)^{n-1}}=np\ .$$

对于方差, 首先计算

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=2}^{n} k(k-1) \binom{n}{k} p^k (1-p)^{n-k} + np \\ &= (1-p)^n \sum_{k=2}^{n} k(k-1) \binom{n}{k} \left(\frac{p}{1-p}\right)^k + np \ . \end{aligned}$$

对二项展开式 $(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$ 两边同时求导两次后乘 x^2 可得

$$n(n-1)x^2(1+x)^{n-2} = \sum_{k=2}^{n} \binom{n}{k} k(k-1)x^k \ ,$$

将 $x = p/(1-p)$ 带入上式有

$$E(X^2) = n(n-1)p^2 + np = n^2p^2 + np(1-p) \ ,$$

从而得到 $\mathrm{Var}(X) = E[X^2] - (E[X])^2 = np(1-p)$.

下面给出几个二项分布的概率分布示意图:

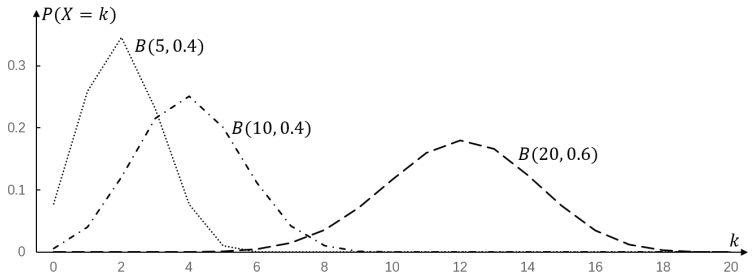


图 3.1 二项分布的概率分布示意图

根据上述分布图可知: 若随机变量 $X \sim B(n,p)$, 则有 $P(X=k)$ 从一开始单调递增, 然后一直单调递减, 一般在期望 np 附近的整数点取得最大值. 也可严格证明: 当 $k \in [0, np+p]$ 时 $P(X=k)$ 单调递增, 当 $k \in [np+p, n]$ 时 $P(X=k)$ 单调递减.

例 3.7 假设有两个箱子, 每个箱子里放了 n 个球, 现在任意选取一个箱子拿走其中一球 (不放回), 重复这一过程, 求一个箱子中的球拿光而另一个箱子还剩下 r 个球的概率.

解 两个箱子分别被表示为第一个箱子和第二个箱子, 考虑的伯努利试验: 在箱子选取过程中是否选取第一个箱子? 用事件 A 表示选取第一个箱子, 根据题意有 $P(A) = 1/2$. 因此可以共发现进行了 $2n-r$ 重伯努利试验, 用 X 表示事件 A 发生的次数, 于是有

$$X \sim B(2n-r, 1/2).$$

最后所求概率为

$$\begin{aligned} & P(X=n)+P(X=n-r) \\ = & \binom{2n-r}{n}(1/2)^n(1/2)^{n-r}+\binom{2n-r}{n-r}(1/2)^{n-r}(1/2)^n=\binom{2n-r}{n}\Big/2^{2n-r-1}\ , \end{aligned}$$

由此完成证明.

例 3.8 一个系统由 n 个独立的元件组成, 每个元件能正常工作的概率为 p, 若该系统中至少有一半的元件能正常工作则整个系统有效, 在什么情况下 5 个元件的系统比 3 个元件的系统更有效?

解 用 X 表示 n 个元件能正常工作的元件数, 则有 $X\sim B(n,p)$. 由此可知包含有 5 个元件的系统有效的概率为

$$\binom{5}{3}p^3(1-p)^2+\binom{5}{4}p^4(1-p)+\binom{5}{5}p^5=p^3(6p^2-15p+10)\ ,$$

而包含有 3 个元件的系统有效的概率为

$$\binom{3}{2}p^2(1-p)+\binom{3}{3}p^3=p^2(3-2p)\ .$$

当 $p^3(6p^2-15p+10)>p^2(3-2p)$ 时, 即当 $3(p-1)^2(2p-1)>0$ 时 5 个元件的系统比 3 个元件的系统更有效, 此时 $p>1/2$.

3.4.3 泊松分布

泊松分布是概率论中另一种重要的分布, 用于描述大量试验中稀有事件出现次数的概率模型. 例如, 一个月内网站的访问量, 一个小时内公共汽车站来到的乘客数, 书中一页出现错误的语法数, 一天中银行办理业务的顾客数, 一年内中国发生的地震次数等.

定义 3.8 给定常数 $\lambda>0$, 若随机变量 X 的分布列为

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}\qquad k=0,1,2,\cdots\ ,$$

则称随机变量 X 服从 **参数为 λ 的泊松分布** (Possion distribution), 记为 $X\sim P(\lambda)$.

容易验证 $P(X=k)=\lambda^k e^{-\lambda}/k!\geqslant 0$, 并根据指数的泰勒展式 $e^x=\sum_{k=0}^{\infty}x^k/k!$ 有

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=e^{-\lambda}\cdot e^{\lambda}=1.$$

关于泊松分布的数字特征有:

引理 3.3 若随机变量 $X\sim P(\lambda)$, 则有 $E(X)=\lambda$ 和 $\mathrm{Var}(X)=\lambda$.

因此泊松分布可由期望或方差唯一确定.

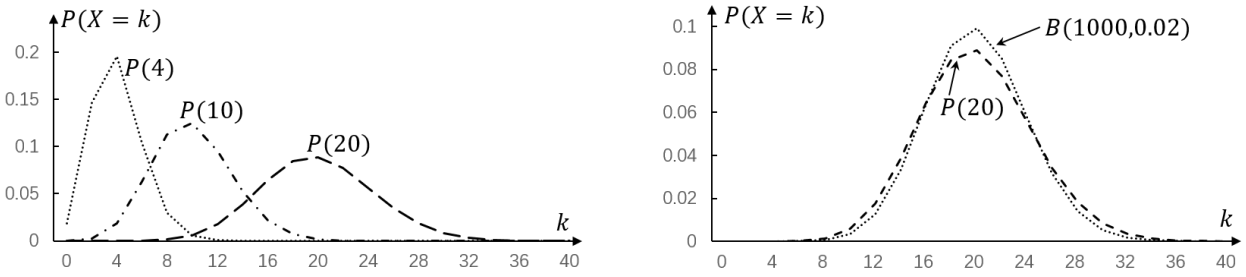
证明 根据指数的泰勒展开式有 $e^x = \sum_{k=0}^{\infty} x^k/k!$ 有

$$E(X) = \sum_{k=0}^{\infty} k \cdot P(X=k) = \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

对于随机变量的方差, 首先计算

$$E[X^2] = \sum_{k=0}^{\infty} k^2 P(X=k) = \sum_{k=1}^{\infty} k(k-1)\frac{\lambda^k}{k!} e^{-\lambda} + \lambda = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 + \lambda \ .$$

从而得到 $\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \lambda$.



a) 泊松分布示意图　　**b)**二项分布$B(1000,0.02)$与泊松分布$P(20)$的逼近性

图 3.2 泊松分布示意图、以及泊松分布与二项分布的逼近图

从图 3.2(a) 中可以观察发现: 若随机变量 $X \sim P(\lambda)$, 则有 $P(X=k)$ 从一开始单调递增, 然而一致单调递减, 在期望 λ 附近取得最大值. 其次, 泊松分布与二项分布的分布图之间有一定的相似性, 如图 3.2(b) 所示, 下面的定理给出了二者之间的近似关系:

定理 3.4 (泊松定理) 设 $\lambda > 0$ 任意给定的常数, n 是一个正整数, 若 $np_n = \lambda$, 则对任意给定的非负整数 k, 有

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

证明 由 $p_n = \lambda/n$, 有

$$\begin{aligned}
\binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!}(1-\frac{1}{n})\cdots(1-\frac{k-1}{n}) \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!}(1-\frac{1}{n})\cdots(1-\frac{k-1}{n}) \left(1-\frac{\lambda}{n}\right)^{\frac{n}{\lambda}\frac{n-k}{n}\lambda}
\end{aligned}$$

当 $n \to \infty$ 时有 $(1-\frac{\lambda}{n})^{\frac{n}{\lambda}} \to e^{-1}$ 以及 $\frac{n-k}{n}\lambda \to \lambda$, 从而完成证明.

泊松分布的应用: 若随机变量 $X \sim B(n,p)$, 当 n 比较大而 p 比较小时, 令 $\lambda = np$, 有

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}.$$

即利用泊松分布近似计算二项分布. 针对彩票中奖、火山爆发、洪水泛滥、意外事故等小概率事件, 当试验的次数较多时, 可以将 n 重伯努利试验中小概率事件发生的次数近似服从泊松分布.

例 3.9 设有 80 台同类型设备独立工作, 每台发生故障的概率为 0.01, 一台设备发生故障时只能由一人处理, 考虑两种方案: I) 由四人维护, 每人单独负责 20 台; II) 由三人共同维护 80 台. 哪种方案更为合理?

解 首先讨论方案 I), 用事件 A_i 表示第 i 人负责的设备发生故障不能及时维修, 用 X_i 为第 i 人负责的 20 台设备同一时刻发生故障的台数, 则有 $X \sim B(20, 0.01)$, 根据泊松定理有近似有 $X \sim P(0.2)$, 进一步有

$$P(A_i) = P(X_i \geqslant 2) = 1 - P(X=0) - P(X=1) \approx 1 - \sum_{k=0}^{2} \frac{(0.2)^k}{k!} e^{-0.2} \approx 0.0175.$$

因四人独立维修, 有设备发生故障时而不能及时的概率

$$P(A_1 \cup A_2 \cup A_3 \cup A_4) \geqslant P(A_1) \approx 0.0175.$$

对方案 II): 设随机变量 Y 为 80 台设备中同一时刻发生故障的台数, 则 $Y \sim B(80, 0.01)$, 根据泊松定理有近似有 $Y \sim P(0.8)$, 则有设备发生故障不能及时维修的概率为

$$P(Y \geqslant 4) = 1 - \sum_{k=0}^{3} P(Y=k) \approx 1 - \sum_{k=0}^{3} \frac{(0.8)^k}{k!} e^{-0.8} \approx 0.0091.$$

由此比较可知方案 II) 更优.

例 3.10 一个公共汽车站有很多路公交车, 若一个时间段内到站的乘客数 $X \sim P(\lambda)$ $(\lambda > 0)$, 所有到站的乘客是相互独立的、且选择 D1 路公交车的概率为 p $(p > 0)$, 求乘坐 D1 路公交车的乘客数 Y 的分布.

解 设一个时间段内到站的乘客数为 k, 该事件发生的概率

$$P(X=k) = \lambda^k e^{-\lambda}/k!\ .$$

根据题意可知到达公交站的 k 个人中乘坐 D1 的人数服从参数为 k 和 p 的二项分布 $B(k,p)$, 即

$$P(Y=i|X=k) = \binom{k}{i} p^i (1-p)^{k-i}\ .$$

根据全概率公式和指数函数 e^x 的泰勒展开式有

$$\begin{aligned} P(Y=i) &= \sum_{k=i}^{+\infty} P(X=k)P(Y=i|X=k) = p^i e^{-\lambda} \sum_{k=i}^{+\infty} \binom{k}{i} \frac{\lambda^k}{k!}(1-p)^{k-i} \\ &= \frac{(p\lambda)^i e^{-\lambda}}{i!} \sum_{k=i}^{+\infty} \frac{((1-p)\lambda)^{k-i}}{(k-i)!} = \frac{(p\lambda)^i e^{-\lambda}}{i!} \sum_{k=0}^{+\infty} \frac{((1-p)\lambda)^k}{(k)!} \\ &= \frac{(p\lambda)^i e^{-\lambda}}{i!} e^{(1-p)\lambda} = \frac{(p\lambda)^i e^{-p\lambda}}{i!} \ , \end{aligned}$$

由此可知乘坐 D1 路公交车的乘客数 $Y \sim P(p\lambda)$.

3.4.4 几何分布

在多重 Bernoulli 试验中, 设事件 A 发生的概率为 p. 用随机变量 X 表示事件 A 首次发生需要的试验次数, 事件 $\{X=k\}$ 发生当且仅当事件 A 在前 $k-1$ 次不发生而第 k 次发生, 根据多重 Bernoulli 试验的独立性可知概率 $P(X=k) = (1-p)^{k-1}p$.

定义 3.9 设 $p \in (0,1)$ 是一个常数, 若随机变量 X 的分布列为

$$P(X=k) = (1-p)^{k-1}p \qquad (k \geqslant 1) \ , \tag{3.5}$$

称 X 服从 **参数为 p 的几何分布** (geometric distribution), 记 $X \sim G(p)$.

容易得到 $P(X=k) \geqslant 0$ 以及

$$\sum_{k=1}^{\infty} P(X=k) = p\sum_{k=1}^{\infty}(1-p)^{k-1} = p \times \frac{1}{1-(1-p)} = 1 \ ,$$

从而验证了 (3.5) 构成概率分布列. 几何分布有一个重要的性质: **无记忆性** (memoryless property).

定理 3.5 设随机变量 $X \sim G(p)$, 对任意正整数 m, n, 有

$$P(X > m+n | X > m) = P(X > n).$$

证明 根据几何分布的定义, 对任何正整数 k 有

$$P(X>k) = \sum_{i=k+1}^{\infty} p(1-p)^{i-1} = p\sum_{i=k+1}^{\infty}(1-p)^{i-1} = p\frac{(1-p)^k}{1-(1-p)} = (1-p)^k \ .$$

根据条件概率的定义有

$$P(X>m+n|X>m) = \frac{P(X>m+n)}{P(X>m)} = \frac{(1-p)^{m+n}}{(1-p)^m} = (1-p)^n = P(X>n) \ ,$$

这里利用事件 $\{X > m+n\} \cap \{X > m\} = \{X > m+n\}$, 从而完成证明.

几何分布无记忆性的直观解释: 假设以前经历了 m 次失败, 从当前起至成功的次数与 m 无关. 例如, 一人赌博时前面总输, 觉得下一次应该会赢了, 然而无记忆性告诉大家: 下一次是否会赢与前面输了多少次没有任何关系.

关于几何分布的数字特征, 我们有

引理 3.4 若随机变量 $X \sim G(p)$ $(0 < p < 1)$, 则有

$$E(X) = \frac{1}{p} \qquad \text{和} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2} \ .$$

证明 根据几何分布的定义有

$$P(X \geqslant i) = \sum_{k=i}^{+\infty} P(X = k) = p \sum_{k=i}^{+\infty} (1-p)^{k-1} = (1-p)^{i-1} \ .$$

对于非负整数的随机变量 X 有

$$E(X) = \sum_{i=1}^{+\infty} P(X \geqslant i) = \sum_{i=1}^{+\infty} (1-p)^{i-1} = 1/p \ .$$

对于随机变量 X 的方差, 首先计算

$$E(X^2) = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = p \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-1} + 1/p \ .$$

对级数展开式 $(1-x)^{-1} = \sum_{k=0}^{\infty} x^k$ 两边先求二阶导后乘 x 有

$$\sum_{k=2}^{\infty} k(k-1)x^{k-1} = \frac{2x}{(1-x)^3} \ .$$

令 $x = 1-p$ 代入可得 $E(X^2) = (2-p)/p^2$. 最后有 $\mathrm{Var}(X) = E(X^2) - (EX)^2 = (1-p)/p^2$.

例 3.11 在古代非常重视生男孩但生存资源有限, 于是规定: 每个家庭可生一个男孩, 如果没男孩则可以继续生育直至有一个男孩; 若已有一个男孩, 则不再生育. 不妨假设每个家庭生男孩的概率为 $p = 1/2$, 问题: 1) 一个家庭恰好有 n 个小孩的概率; 2) 一个家庭至少有 n 个小孩的概率; 3) 男女比例是否会失衡?

解 用随机变量 X 表示一个家庭的小孩个数, 其取值为 $\{1, 2, \cdots\}$, 根据题意可知 X 服从参数为 $p = 1/2$ 的几何分布, 因此一个家庭恰好有 n 个小孩的概率为

$$P(X = n) = p(1-p)^{n-1} = 1/2^n \ .$$

一个家庭至少有 n 个小孩的概率为

$$P(X \geqslant n) = \sum_{k=n}^{+\infty} P(X = k) = 1/2^{n-1} \ .$$

至于男女比例是否会失衡, 考虑一个家庭平均的孩子个数为 $E[X] = 1/p = 2$, 由此可知在平均的情形下, 一个家庭的小孩男女比例 $1:1$, 因此不会造成男女失衡.

几何分布考虑在多重试验中事件 A 首次发生时所进行的试验次数, 可以进一步考虑事件 A 第 r 次发生时所进行的试验次数. 设随机事件 A 发生的概率为 $p \in (0,1)$, 用 X 表示事件 A 第 r 次成功时发生的试验次数, 则 X 取值 $r, r+1, r+2, \cdots$, 其分布列为

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \qquad (k = r, r+1, r+2, \cdots) \ ,$$

称随机变量 X 服从 **参数为 r 和 p 的负二项分布** 或 **帕斯卡分布**. 可以验证上述概率构成一个分布列, 以及随机变量 X 的期望 $E(X) = r/p$ 和方差 $\mathrm{Var}(X) = r(1-p)/p^2$. 相关证明将作为练习题.

3.5 案例分析

3.5.1 德国坦克问题

在二战期间, 同盟国一直在努力确定德国坦克的生产数量, 有助于对德国战力的评估. 这个问题可描述为: 德国生产了 n 辆坦克, 编号分别为 $1, 2, \cdots, n$. 盟军在战斗中任意击毁了 k 辆坦克, 被击毁的坦克编号分别为 $x_1, x_2, \ldots, x_k$, 能否通过被击毁的坦克编号来估计 n 的大小, 即估计德国生产了多少辆坦克.

在没有其它信息的情况下, 不妨假设被随机击毁的坦克是等可能事件, 即第 i 辆坦克被击毁的概率为 $1/n$. 可以将问题看作从集合 $\{1, 2, \cdots, n\}$ 中不放回随机抽取 k 个数, 用 X 表示抽到的 k 个数中的最大数. 则 X 的取值为 $\{k, k+1, \cdots, n\}$ 以及概率

$$P(X = i) = \binom{i-1}{k-1} \Big/ \binom{n}{k} \qquad (i = k, k+1, \cdots, n) \ .$$

于是得到

$$E(X) = \binom{n}{k}^{-1} \sum_{i=k}^{n} \binom{i-1}{k-1} i \ .$$

针对上面的求和表达式, 可以考虑从 $n+1$ 个元素中选取 $k+1$ 个元素, 共有 $\binom{n+1}{k+1}$ 种不同的方法. 将这些不同的方法分情况讨论, 按照选取的 $k+1$ 个元素中最大元素 $i = k+1, k+2, \cdots, n+1$ 进行分类; 若最大元素为 i, 则有 $\binom{i-1}{k}$ 种不同的方法. 于是有

$$\binom{n+1}{k+1} = \sum_{i=k+1}^{n+1} \binom{i-1}{k} = \sum_{i=k}^{n} \binom{i}{k} = \sum_{i=k}^{n} \frac{i}{k} \binom{i-1}{k-1} \ ,$$

代入期望 $E(X)$ 可得

$$E(X)=k\binom{n}{k}^{-1}\sum_{i=k}^{n}\binom{i-1}{k-1}\frac{i}{k}=k\binom{n+1}{k+1}\Big/\binom{n}{k}=\frac{k(n+1)}{k+1}\ .$$

由于仅做了一次观察, 将观察中 k 个数的最大值近似期望 $E[X]$, 即 $E(X)\approx\max(x_1,x_2,\cdots,x_n)$, 由此估计

$$n\approx\max(x_1,x_2,\cdots,x_n)\left(1+\frac{1}{k}\right)-1,$$

从而完成 n 的估计.

例如, 如果观察到被击毁坦克编号分别为 $17,68,94,127,135,212$, 根据上面的推到可估计出

$$n\approx 212\times(1+1/6)-1=246.$$

针对德国坦克数量的实际估计情况见下表, 可以发现利用上述所提的统计估计方法接近德国的实际产量, 比英国的情报估计准确得多.

时间	统计估计	英国情报估计	德国实际产量
1940-06	169	1000	122
1941-06	244	1550	271
1942-08	327	1550	342

3.5.2 集卡活动

很多小朋友喜欢各种集卡活动, 如奥特曼卡和叶罗丽卡等. 事实上很多成年人也对集卡游戏并不陌生, 例如 80 年代的葫芦娃洋画、或 90 年代的小虎队旋风卡等. 问题可以描述为: 市场上有 n 种不同类型的卡片, 假设一个小朋友每次都能以等可能概率、独立地收集一张卡片, 问一个小朋友在平均情况下至少要收集多少张卡才能收集齐 n 种不同类型的卡片.

这里先补充一个需要用到的引理, 后面将给出详细的证明:

引理 3.5 对任意的随机变量 $X_1,X_2,\cdots,X_n$ 有

$$E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)\ .$$

用 X 表示收集齐 n 种不同类型的卡片所需要的收集次数, 用 X_k 表示收集齐第 $k-1$ 种和第 k 种不同类型卡片之间所需要的收集次数 ($k\in[n]$), 于是有 $X=X_1+X_2+\cdots+X_n$. 我们的问题是计算期望 $E(X)$.

很容易发现随机变量 X_k 服从参数为 p_k 的几何分布. 当已经收集到 $k-1$ 种不同类型的卡片时, 再获得一张新卡的概率

$$p_k=1-(k-1)/n\ .$$

根据几何分布的性质有 $E[X_k] = 1/p_k = n/(n-k+1)$. 利用引理 3.5 有

$$E(X) = E\left(\sum_{k=1} X_k\right) = \sum_{k=1} E(X_k) = \sum_{k=1}^{n} \frac{n}{n-k+1} = n\sum_{k=1}^{n} \frac{1}{k} = nH(n) \ ,$$

这里 $H(n)$ 表示参数为 n 的调和数, 即 $H(n) = \sum_{k=1}^{n} 1/k$. 关于调和数有

引理 3.6 调和数 $H(n) \in [\ln(n+1), 1+\ln(n)]$.

证明 因为函数 $1/x$ 在 $x \in (0, +\infty)$ 单调递减, 有

$$\ln(n+1) = \int_{x=1}^{n+1} \frac{1}{x} dx \leqslant \sum_{k=1}^{n} \frac{1}{k} = 1 + \sum_{k=2}^{n} \frac{1}{k} \leqslant 1 + \int_{x=1}^{n} \frac{1}{x} dx = 1 + \ln(n).$$

最后得到 $n\ln(n+1) \leqslant E(X) \leqslant n + n\ln n$.

3.5.3 随机二叉树叶子结点的高度

在机器学习中, 随机树和随机森林是一类经典的分类或回归算法, 随机树叶子结点的高度估计对学习算法性能的分析具有重要作用. 本节考虑完全随机的二叉树中一个叶子结点的平均高度. 随机二叉树的构造过程非常简单: 首先给定二叉树的根结点, 然后在每一轮的迭代过程中执行以下两步操作:

- 在当前所有的叶子结点中随机选择一个叶子结点作为划分结点;
- 被选中的叶子结点变成一个内部结点, 生长出左、右两个叶子结点.

重复上述过程 n 步, 最后得到具有 n 个叶子结点的随机二叉树. 在这一构造过程中, 最关键的一步是随机选择的叶子结点作为划分结点. 随机二叉树构造的示意图如下所示:

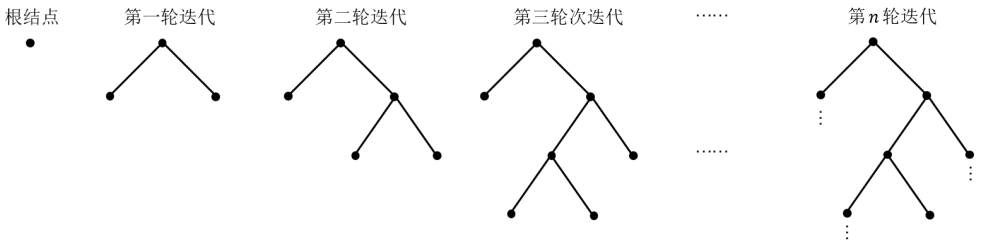


图 3.3 随机二叉树构造的示意图

一个叶子结点的高度是从根节点到该叶子结点的路径中边的条数. 求解的问题: 在最后生成的随机二叉树中, 求任意一个叶结点的平均高度.

用随机变量 X 表示任意给定的一个叶结点的高度, 并用随机变量 X_i 表示在第 i 轮迭代过程中该叶子的祖先结点是否恰好被选中作为划分结点, 而在第 i 轮迭代过程中恰好有 i 个叶结点, 则有

$$X_i = \mathrm{Ber}(1/i) \qquad 且 \qquad X = X_1 + X_2 + \cdots + X_n.$$

根据期望的性质和引理 3.6 有

$$E[X] = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} 1/i = H(n) \in [\ln(n+1), 1+\ln(n)] \ .$$

由此可知一个叶子结点的平均高度为 $\Theta(\ln n)$.

习题

3.1 从 $\{1,2,...,10\}$ 中有放回地任取 5 个数, 令 X 表示五个数中的最大值, 求 X 的分布列, 并求在无放回地情况下的分布列.

3.2 将一枚骰子任意投掷三次, 用 X 表示三次中得到最小点的点数, 求 X 的分布列及期望.

3.3 有 4 个盒子编号分别为 $1,2,3,4$. 将 3 个不同的球随机放入 4 个盒子中, 同一盒子内的球无顺序关系, 用 X 表示有球盒子的最小编号, 求 $E(X)$.

3.4 设离散型随机变量 $X \in [a,b]$ 的取值有有限中可能, 实值函数 $g\colon [a,b] \to \mathbb{R}$ 是凹函数, 证明 $g(E(X)) \geqslant E(g(X))$.

3.5 若随机变量 $X \sim B(n,p)$, 证明当 $k \in [0, np+p]$ 时 $P(X=k)$ 单调递增, 当 $k \in [np+p, n]$ 时 $P(X=k)$ 单调递减.

3.6 设随机变量 X 服从参数为 λ 的泊松分布, 且 $P(X=1) = P(X=2)$, 求 $P(X \geqslant 4)$.

3.7 设随机变量 X 的取值为 $r, r+1, \cdots$ 以及事件 $\{X=k\}$ 的概率为

$$P(X=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \qquad p \in (0,1), \quad k = r, r+1, r+2, \cdots,$$

检验上面的概率构成一个分布列.

3.8 设随机变量 X 的分布列为

$$P(X=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \qquad p \in (0,1), \quad k = r, r+1, r+2, \cdots,$$

证明: 随机变量 X 的期望 $E(X) = r/p$ 和方差 $\mathrm{Var}(X) = r(1-p)/p^2$

3.9 现需要 100 个符合规格的元件, 从市场上购买该元件的废品率为 0.01, 现准备在市场上买 $100+x$ 个元件, 要使得其中至少有 100 个符合规格元件的概率大于 0.95, 求 x 的最小值?

3.10 设随机变量的分布列为 $P(X = (-5)^k/k) = 4/5^k$ $(k = 1, 2, \cdots)$, 证明 X 的期望不存在.

3.11 一个箱子中有一个白球和一个红球, 若从箱子中随机摸到一个白球则再放入一个白球, 若摸到一个红球则结束这个游戏. 证明: 游戏结束时的摸球次数的期望不存在.