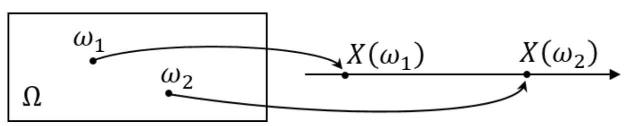
第 3 章　离散型随机变量

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 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 $E(X)$.

解 首先给出随机变量 X 的分布列

$$P(X=1)=\frac{\binom{3}{1}3^2+\binom{3}{2}3+1}{4^3}=\frac{37}{64},\quad P(X=2)=\frac{\binom{3}{1}2^2+\binom{3}{2}2+1}{4^3}=\frac{19}{64},$$
$$P(X=3)=\frac{\binom{3}{1}+\binom{3}{2}+1}{4^3}=\frac{7}{64},\qquad P(X=4)=\frac{1}{64}.$$

由此可得期望

$$E(X)=\frac{37}{64}+2\cdot\frac{19}{64}+3\cdot\frac{7}{64}+4\cdot\frac{1}{64}=\frac{25}{16}.$$

例 3.6 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

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

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 $E\left(\sum_{i=1}^n c_ig_i(X)\right)=\sum_{i=1}^n c_iE(g_i(X))$ 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, 在很多实际应用中仅仅知道随机变量的平均值还是不够的, 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, 其分布列分别为

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

容易得到 $E(X) = E(Y) = E(Z) = 0$, 即三个随机变量的期望一样. 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, 此时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

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

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