这里对 n 采用归纳法证明, 当 $n=2$ 时由凸函数的定义直接可证. 不妨假设 $n=m-1$ 时成立 $(m\geqslant 3)$, 下面证明当 $n=m$ 亦成立. 首先有

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

这里利用 $g(p_1x_1+(1-p_1)x_1')\leqslant p_1g(x_1)+(1-p_1)g(x_1')$, 其中 $x_1'=(x_2p_2+\cdots+x_mp_m)/(1-p_1)$. 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

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

代入即可完成证明.

对任意离散型随机变量 X, 根据 Jensen 不等式有

$$(E(X))^2\leqslant E(X^2) \qquad 和 \qquad e^{E(X)}\leqslant E(e^X).$$

3.2.2 方差

数学期望反映了 X 取值的平均值, 对三个随机变量 X, Y 和 Z, 其分布列分别为

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

尽管随机变量的均值相同 $EX=EY=EZ=0$, 但这三个随机变量与期望的偏离程度有很大的差异, 本节研究随机变量 X 与期望 $E(X)$ 的偏离程度, 即方差.

定义 3.3 离散性随机变量 X 的分布列为 $p_k=P(X=x_k)$ $(k\geqslant 0)$, 若期望 $E(X)=\sum_k x_kp_k$ 存在, 以及 $E(X-E(X))^2=\sum_k p_k(x_k-E(X))^2$ 存在, 称 $E(X-E(X))^2$ 为随机变量 X 的方差 (variance), 记为 $\mathrm{Var}(X)$ 或 $D(X)$, 即

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

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

根据方差的绝对收敛性可知方差不会随随机变量取值的顺序改变而改变, 进而根据期望的性质有

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

由此给出方差的另一种等价性定义

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

尽管方差的两种定义完全等价, 但在实际应用中可能带来不同的计算量.

例 3.7 设随机变量 X 的分布列为 $P(X = x_i) = 1/n$ $(i \in [n])$, 试问: 计算随机变量 X 的方差需要遍历 $x_1, x_2, \cdots, x_n$ 几遍.

解 若利用定义 $\mathrm{Var}(X) = E(X - E(X))^2$, 则需要遍历数据 $x_1, x_2, \cdots, x_n$ 两遍, 第一遍计算期望 $E(X)$, 第二遍计算方差 $\mathrm{Var}(X)$. 若利用定义 $\mathrm{Var}(X) = E(X^2) - (E(X))^2$, 则只需要遍历 $x_1, x_2, \cdots, x_n$ 一遍, 可在线 (online) 计算方差, 不需存储数据.

下面给出方差的性质:

性质 3.4 若随机变量 $X \equiv c$, 则 $\mathrm{Var}(X) = 0$.

性质 3.5 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

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

一般情况下方差不具有线性性, 即 $\mathrm{Var}[f(X) + g(X)] \neq \mathrm{Var}[f(X)] + \mathrm{Var}[g(X)]$.

性质 3.6 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 我们有

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

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

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

进一步对二次函数 $f(t) = -t^2 + (a+b)t - ab$ 求最大值, 可得 $(b-E(X))(E(X)-a) \leqslant (b-a)^2/4$.

3.3 常用的离散型随机变量

本节介绍几种常用的离散型随机变量, 并研究其数字特征.

3.3.1 离散型均匀分布

定义 3.4 设随机变量 X 的取值为 $x_1, x_2, \cdots, x_n$, 且 $P(X = x_i) = 1/n$, 称 X 服从离散型均匀分布.

由定义可知

$$\begin{aligned} E(X) &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \mathrm{Var}(X) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2. \end{aligned}$$

下面来看一个离散型均匀分布的例子.

例 3.8 (德国坦克问题) 假设德国生产了 N 辆坦克, 编号分别为 $1, 2, \cdots, N$, 盟军战斗中随机击毁了 k 辆, 被随机击毁坦克编号分别为 $x_1, x_2, \ldots, x_k$, 如何估计 N 的大小.

解 对题意进行分析, 坦克被随机击毁可看作坦克被击毁的服从离散型均匀分布, 即 $P[X = i] = 1/N$ $(i \in [N])$, 因此有 $E(X) = (1+N)/2$, 可用被击毁坦克编号的平均值去近似, 即

$$E(X) = \frac{1+N}{2} \approx \frac{1}{k}\sum_{i=1}^{k} x_i \qquad \Longrightarrow \qquad N \approx \frac{2}{k}\sum_{i=1}^{k} x_i - 1.$$

若 k 较大时, 上述表达式对 N 有一个较好的估计; 但 k 较小时, 往往更关注于被毁坦克的最大编号 m.

方法二 问题转化为从 $\{1,2,\cdots,N\}$ 中以不放回随机抽取 k 个数, 观察到 k 个数中最大数为 m, 如何利用 m 和 k 估计 N. 假设随机变量 X 表示抽取 k 个数中的最大数, 其分布列为

$$Pr[X=i]=\frac{\binom{i-1}{k-1}}{\binom{N}{k}} \quad (k\leqslant i\leqslant N).$$

由此得到

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

为计算期望 $E(X)$, 考虑从 $N+1$ 个元素中随机取 $k+1$ 个元素, 可等价于按所抽取 $k+1$ 个元素中最大元进行分类, 即

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

代入 $E(X)$ 可得

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

由于仅做了一次观察, 可以将一次观察的最大值 m 看作为 $E[X]$ 的近似, 即

$$m\approx E(X)=\frac{k(N+1)}{k+1} \quad\Longrightarrow\quad N\approx m(1+k^{-1})-1,$$

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 $N=212\times(1+1/6)-1\approx 246$. 针对二战德国坦克数量的实际估计情况可参见下表, 统计估计比情报估计准确得多, 接近德国的实际产量.

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

3.3.2 0-1分布

定义 3.5 随机变量 X 的取值为 $\{0,1\}$, 其分布列 $P(X=1)=p$, $P(X=0)=1-p$, 称 X 服从参数为 p 的 $0-1$ 分布, 又称两点分布, 或 Bernoulli 分布, 记 $X\sim\mathrm{Ber}(p)$.

由定义可知

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

$0-1$分布是很多概率模型的基础.

3.3.3 二项分布

Bernoulli 试验只有两个结果: A 和 $\bar{A}$, 设 $P(A)=p$ ($p\in[0,1]$), 因此 $P(\bar{A})=1-p$. 将 Bernoulli 试验独立重复地进行 n 次, 称 n 重 Bernoulli 试验, 是一种非常重要的概率模型, 具有广泛的应用.

定义 3.6 用随机变量 X 表示 n 重 Bernoulli 试验试验中事件 A 发生的次数, 则 X 的取值为 $0,1,\cdots,n$, 其分布列为

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

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

我们称之为二项分布是因为与二项展开式 $(a+b)^n=\sum_{k=1}^n\binom{n}{k}a^kb^{n-k}$ 相似, 容易验证

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

对二项分布, 我们有

性质 3.7 对随机变量 $X\sim B(n,p)$ 有

$$E(X)=np\qquad 和\qquad \operatorname{Var}(X)=np(1-p).$$

证明 由定义可知

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

为计算 $E(X)$, 对二项展开式 $(1+x)^n=\sum_{k=0}^n\binom{n}{k}x^k$ 两边求导数有

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

将 $x=p/(1-p)$ 带入可得

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

对于方差, 首先计算

$$E(X^2)\ =\ \sum_{k=0}^n k^2\binom{n}{k}p^k(1-p)^{n-k}$$

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

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

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

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

例 3.9 有 5 个选择题, 每个选择题有 4 种答案, 只有一种正确, 求一学生随机猜对 4 个选择题的概率?

解 将每一个选择题看作一次 Bernoulli 试验, 事件 A 表示猜正确, 则有 $P(A) = 1/4$. 整个问题等价于 5 重 Bernoulli 试验, 用 X 表示学生猜对题的个数, 则 $X \sim B(5, 1/4)$, 从而得到

$$
P(X=4) = \binom{5}{4} \cdot \frac{1}{4^4} \cdot \frac{3}{4} = \frac{15}{4^5}.
$$

3.3.4 几何分布

定义 3.7 在多重 Bernoulli 试验, 设事件 A 发生的概率为 p. 用随机变量 X 表示事件 A 首次发生时的试验次数, 则 X 的取值为 $1, 2, \cdots$, 其分布列为

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

称 X 服从参数为 p 的几何分布, 记为 $X \sim G(p)$.

首先可知 $P(X=k) = (1-p)^{k-1}p \geqslant 0$ 以及

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

从而验证了几何分布构成一个分布列. 对几何分布, 我们有