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 $p_i=c\lambda^i/i!\ (i\geq 0)$, 求 $P(X>2)$.

解. 根据概率的完备性有

$$1=\sum_{i=0}^{\infty}p_i=c\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}=c\cdot e^{\lambda}$$

从而得到 $c=e^{-\lambda}$, 进一步得到

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

□

例3.3. 从 $\{1,2,\ldots,10\}$ 中不放回随机任意取 5 个数, 令随机变量 X 表示所取 5 个数中的最大值, 求 X 的分布列.

解. 由题意可知 X 的取值为 $5,6,7,8,9,10$, 且

$$P(X=k)=\binom{k-1}{4}\Big/\binom{10}{5}\quad(k=5,6,\ldots,10).$$

由此可得 X 的分布列表格为

X	5	6	7	8	9	10
P	1/252	5/252	15/252	35/252	70/252	126/252

□

3.2 离散型随机变量的期望和方差

随机变量的取值具有一定的随机性, 我们希望研究随机变量的一些不变量, 用以刻画随机变量的特征, 最常见的特征是期望与方差.

3.2.1 期望

定义3.1. 设离散型随机变量 X 的分布列为 $P(X=x_k)=p_k\ (k=1,2,\cdots)$, 若级数 $\sum_{k=1}^{\infty}p_kx_k$ 绝对收敛, 称级数和为随机变量 X 的期望 *(expectation)*, 记为 $E(X)$, 即

$$E(X)=\sum_{k=1}^{\infty}p_kx_k,$$

又被称为均值 *(mean)* 或加权平均 *(weighted average)*.

期望 $E(X)$ 反映随机变量 X 的平均值, 有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值 x_i 根据概率 p_i 加权所得. 级数的绝对收敛保证了级数和不随级数各项次序的改变而改变, 期望 $E(X)$ 反映 X 可能值的平均值, 不会随次序改变而改变. 根据随机变量随机变量 X 的分布列可直接计算其期望.

例3.4. *随机掷一枚骰子, 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 $P(X = i) = 1/6$ $(i \in [6])$. 因此随机变量 X 的期望为

$$E(X) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5.$$

□

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, 用 X 表示有球盒子的最小号码, 求 $E(X)$.*

解. 先给出 X 的分布列

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

进一步可得

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

□

例3.6. *有 n 把钥匙只有一把能打开门, 随机选取一把试开门, 若打不开则除去, 求打开门次数的平均数.*

解. 设随机变量 X 表示试开门的次数, 其分布列为

$$P(X = k) = \frac{\binom{n-1}{k-1}}{\binom{n}{k-1}} \cdot \frac{1}{n - k + 1} = \frac{1}{n},$$

进一步可得打开门次数的平均数

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

□

根据期望的定义有如下性质:

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

性质3.3. *对随机变量 X 和常数 $a, b \in \mathbb{R}$, 有 $E(aX + b) = aE(X) + b$.*

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

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

□

对随机变量函数的期望, 有如下定理:

定理3.1. 设 X 为离散型随机变量, 以及 $g:\mathbb{R}\to\mathbb{R}$ 是连续函数, 若 X 的分布列为 $P(X=x_k)=p_k$ $(k=1,2,\ldots)$, 且 $\sum_{k\geq 1}g(x_k)p_k$ 绝对收敛, 则有

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

此定理意义在于计算随机变量 $Y=g(X)$ 的期望时, 不需计算 Y 的分布列, 而只需利用 X 的分布列即可完成 $E[Y]$ 的计算.

Proof. 证明的思想是利用绝对收敛保证无穷级数任意重排后的级数仍收敛于原无穷级数的和. 根据题意有 X 的分布列为 $P(X=x_k)=p_k$ 以及随机变量函数 $Y=g(X)$ 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	$\cdots$

注意 y_i 可能等于 y_j $(i\neq j)$, 因此 $P(Y=y_j)=p_j$ 不是随机变量 Y 的分布列. 为构造 Y 的分布列, 我们将 $x_1,x_2,\ldots,x_n,\ldots$ 进行重新分组,

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

其中 $y_i'\neq y_j'$ $(i\neq j)$. 由此可得随机变量 Y 的分布列为

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

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级数重排后其和不变. □

推论3.1. 设 X 为离散型随机变量, 以及函数 $g_i:\mathbb{R}\to\mathbb{R}$ 是连续函数 $(i\in[n])$, 且 $E(g_i(X))$ 存在. 对任意常数$c_1,c_2,\ldots,c_n$, 有

$$E\left(c_1g_1(X)+c_2g_2(X)+\cdots+c_ng_n(X)\right)=\sum_{i=1}^{n}c_iE(g_n(X)).$$

Proof. 根据定理 3.1有

$$\begin{aligned}
&E\left(c_1g_1(X)+c_2g_2(X)+\cdots+c_ng_n(X)\right)\\
=&\sum_{k=1}^{\infty}p_k(c_1g_1(x_k)+c_2g_2(x_k)+\cdots+c_ng_n(x_k))\\
=&\sum_{i=1}^{n}c_i\sum_{k=1}^{\infty}p_kg_i(x_k)=\sum_{i=1}^{n}c_iE(g_i(X)).
\end{aligned}$$

□

基于此推理有 $E(X^2+X+\sin X+4)=E(X^2)+E(X)+E(\sin X)+4$. 给定随机变量函数 $Y=g(X)$, 下面探讨 $E(g(X))$ 和 $g(E(X))$ 之间的关系:

定义3.2. *设函数 $g\colon[a,b]\to\mathbb{R}$, 若对任意 $x_1,x_2\in[a,b]$ 和 $\lambda\in[0,1]$, 有 $g(\lambda x_1+(1-\lambda)x_2)\leq\lambda g(x_1)+(1-\lambda)g(x_2)$ 成立, 称函数 $g(x)$ 是定义在 $[a,b]$ 的凸函数;*

若对任意 $x_1,x_2\in[a,b]$ 和 $\lambda\in[0,1]$, 有 $g(\lambda x_1+(1-\lambda)x_2)\geq\lambda g(x_1)+(1-\lambda)g(x_2)$ 成立, 称函数 $g(x)$ 是定义在 $[a,b]$ 的凹函数;

下面介绍著名的 Jensen 不等式, 经常应用于人工智能的各种推导证明中.

定理3.2. *设 X 为 $[a,b]$ 的离散型随机变量, 若 $g\colon[a,b]\to\mathbb{R}$ 是连续的凸函数, 则有*

$$g(E(X))\leq E(g(X));$$

若 $g\colon[a,b]\to\mathbb{R}$ 是连续的凹函数, 则有

$$g(E(X))\geq E(g(X)).$$

Proof. 这里仅给出有限样本空间和凸函数的证明. 不妨假设随机变量 X 的取值为 $x_1,x_2,\cdots,x_n$, 其分布列为 $P(X=x_k)=p_k\geq0$, 根据概率性质有 $\sum_k p_k=1$. 需要证明

$$g(E(X))=g(p_1x_1+p_2x_2+\cdots+p_nx_n)\leq p_1g(x_1)+p_2g(x_2)+\cdots+p_ng(x_n)=E[g(X)]. \tag{2}$$

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 $(m\geq3)$, 下面证明当 $n=m$ 亦成立. 首先有

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

这里利用 $g(p_1x_1+(1-p_1)x_1')\leq p_1g(x_1)+(1-p_1)g(x_1')$, 其中 $x_1'=(x_2p_2+\cdots+x_mp_m)/(1-p_1)$. 容易发现 $p_i/(1-p_1)\geq0$ 且 $\sum_{i=2}^m p_i/(1-p_1)=1$, 根据归纳假设有

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

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

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

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 $p_k=P(X=x_k)$ $(k\geq 0)$, 若期望 $E(X)=\sum_k x_k p_k$ 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 $Var(X)$ 或 $D(X)$, 即*

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

称 $\sqrt{Var(X)}$ 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, 且可在线 (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 $Var(X)=0$.*

性质3.5. 对随机变量 X 和常数 $a, b \in \mathbb{R}$, 有

$$Var(aX+b) = a^2 Var(X).$$

Proof. 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

$$Var(X) = E(X-E(X))^2 \leq E(X-a)^2.$$

Proof. 我们有

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

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

$$Var[X] \leq (b-E(X))(E(X)-a) \leq (b-a)^2/4.$$

Proof. 对任意随机变量 $X \in [a, b]$, 有 $(b-X)(X-a) \geq 0$, 两边同时对随机变量取期望, 整理可得

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

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 $(b-E(X))(E(X)-a) \leq (b-a)^2/4$. □

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.