Ch 3 离散型随机变量

回顾前一次课

随机变量　　　　　　　　　　离散型随机变量

分布列：$p_k = P(X = x_k)$　　　分布列性质

期望: $E(X) = \sum_k p_k x_k$ 反映随机变量的平均值

对随机变量X和常数$a, b \in R$, 有$E(aX + b) = aE(X) + b$

函数的期望：$E[g(X)] = \sum_{k \geq 1} g(x_k) p_k$

凸函数、Jensen不等式

方差(variance)

数学期望反映了随机变量的平均值

假设三个随机变量X,Y和Z的分布列分别为

$P(X=0)=1$ 期望：$E(X)=0$

$P(Y=1)=P(Y=-1)=1/2$ 期望：$E(Y)=0$

$P(Z=2)=1/5, P(Z=-1/2)=4/5$ 期望：$E(Z)=0$

随机变量期望一样，但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

方差的定义

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$E(X) = \sum_{k\geq 0} x_k p_k$ 存在, 以及

$$\boldsymbol{E\left(X - E(X)\right)^2 = \sum_{k\geq 0} p_k\left(x_k - E(X)\right)^2}$$

存在, 称$E\left(X - E(X)\right)^2$为随机变量X的方差, 记$\mathrm{Var}\,(X)$ 或$D(X)$

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

根据定义可知：**方差不会随随机变量取值的顺序改变而改变**

方差的另一种定义

方差的等价定义：$\mathbf{Var}(\boldsymbol{X}) = \boldsymbol{E}\left(\boldsymbol{X}^2\right) - [\boldsymbol{E}(\boldsymbol{X})]^2$

尽管两种定义完全等价, 在实际应用中可能带来不同的计算量

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

方差的性质

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

对随机变量X和常数$a, b \in\ R$, 有

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

一般情况下方差不具有线性性, 即

$$\mathrm{Var}\big(f(X) + g(X)\big) \neq\ \mathrm{Var}\big(f(X)\big) + \mathrm{Var}\big(g(X)\big)$$

方差的性质

对随机变量X和常数$a \in R$, 有

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

Bhatia-Davis不等式

对随机变量$X \in [a, b]$, 有

$$Var(X) \le \big(b - E(X)\big)(E(X) - a) \le (b - a)^2/4$$

Ch 3.3 常用的离散型随机变量

离散均匀分布

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 **均匀分布**

期望：$E(X) = \sum_{i=1}^{n} \frac{x_i}{n}$

方差: $E(X) = \sum_{i=1}^{n} \frac{x_i^2}{n} - \left(\sum_{i=1}^{n} \frac{x_i}{n}\right)^2$

二战德国坦克数量问题

假设德国生产N辆坦克, 编号为$1,2,\cdots,N$, 盟军战斗中随机击毁 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

二战德国坦克数量问题

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

$$N = 212 \times \left(1 + \frac{1}{6}\right) - 1 \approx 246$$

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

0/1分布

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**, 或 **Bernoulli 分布**, 记 $\boldsymbol{X \sim \mathrm{Ber}(p)}$

若 $X \sim \mathrm{Ber}(p)$，则

$$E(X) = p \qquad \mathrm{Var}(X) = p(1 - p)$$

二项分布

Bernoulli试验有两个结果: A和$\bar{A}$, 设 $P(A) = p$ $(p \in [0,1])$

将Bernoulli试验独立重复地进行n次, 称 **n重Bernoulli试验**

一种非常重要的概率模型, 具有广泛的应用

用随机变量X表示n重Bernoulli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

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

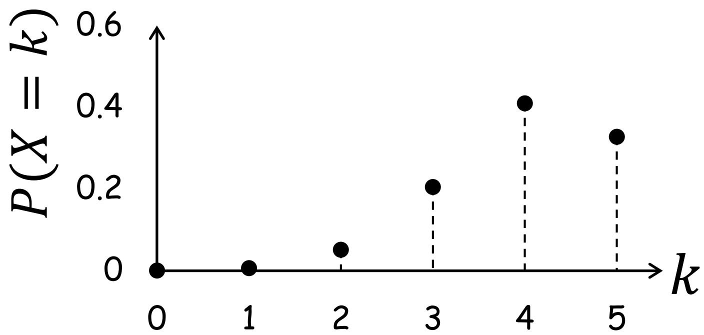
称随机变量X服从参数为n和p的**二项分布 (binomial distribution)**, 记 $\boldsymbol{X \sim B(n,p)}$

性质

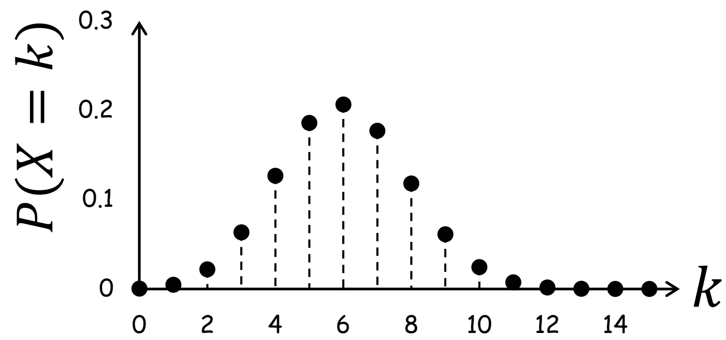
对随机变量 $X \sim B(n,p)$ 有

$$E(X) = np \qquad \qquad \mathrm{Var}(X) = np(1-p)$$

二项分布图



$B(5, 0.8)$



$B(15, 0.4)$

例子

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?