第六周　常见随机变量的期望与方差和应用实例

6.1 二项分布与泊松分布的期望与方差

二项分布　$X \sim b(n,p)$，　$P(X=k)=C_n^k p^k q^{n-k}$

$$E(X)=\sum_{k=0}^{n} x_k P(X=x_k)=\sum_{k=0}^{n} k\cdot C_n^k p^k q^{n-k}=\sum_{k=0}^{n} k\cdot\frac{n!}{k!(n-k)!}p^k q^{n-k}$$

$$=\sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k}=np\cdot\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}q^{n-k}$$

$$=np\cdot\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!\left[(n-1)-(k-1)\right]!}p^{k-1}q^{(n-1)-(k-1)}$$

$$=np\cdot\sum_{j=0}^{n-1}\frac{(n-1)!}{j!(n-1-j)!}p^j q^{n-1-j}$$

$$=np\cdot(p+q)^{n-1}=np$$

****

$$E(X^2)=\sum_{k=0}^{n} x_k{}^2 p(x_k)=\sum_{k=0}^{n} k^2\cdot C_n^k p^k q^{n-k}=\sum_{k=1}^{n}\left[k(k-1)+k\right]\cdot\frac{n!}{k!(n-k)!}p^k q^{n-k}$$

$$=\sum_{k=1}^{n} k(k-1)\frac{n!}{k!(n-k)!}p^k q^{n-k}+\sum_{k=1}^{n} k\cdot\frac{n!}{k!(n-k)!}p^k q^{n-k}$$

$$=\sum_{k=2}^{n}\frac{n!}{(k-2)!(n-k)!}p^k q^{n-k}+np$$

$$=n(n-1)p^2\cdot\sum_{k=2}^{n}\frac{(n-2)!}{(k-2)!\left[(n-2)-(k-2)\right]!}p^{k-2}q^{(n-2)-(k-2)}+np$$

$$=n(n-1)p^2\cdot\sum_{j=0}^{n-2}\frac{(n-2)!}{j!(n-2-j)!}p^j q^{n-2-j}+np$$

$$= n(n-1)p^2 \cdot (p+q)^{n-2} + np = n(n-1)p^2 + np = n^2p^2 + np(1-p)$$

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

****

二项分布随机变量 $X \sim b(n,p)$ 期望和方差的另一种理解

考虑 n 个独立的 $0-1$ 随机变量 $X_k \sim b(1,p)$， $k = 1,\cdots,n$，

满足 $P(X_k = 1) = p$， $P(X_k = 0) = 1-p$， 则 $X = X_1 + X_2 + \cdots + X_n \sim b(n,p)$；

对所有 $k = 1,\cdots,n$， $E(X_k) = 1\times p + 0\times(1-p) = p$， $E(X_k^2) = 1\times p + 0\times(1-p) = p$

$$Var(X_k) = E(X_k{}^2) - E(X_k)^2 = p - p^2 = p(1-p),$$

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = np$$

$$Var(X) = Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n) = np(1-p)$$

****

泊松分布 $X \sim P(\lambda)$， $\lambda > 0$， $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$， $k = 0,1,2,\cdots$

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

$$= \lambda\sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} \overset{j=k-1}{=} \lambda\sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda。$$

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

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

****

用二项分布极限的观点理解泊松分布的期望和方差

泊松分布 $X \sim P(\lambda)$，$\lambda > 0$，$P(X=k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$，$k = 0,1,2,\cdots$。

$Y_n \sim b(n,p)$，$np \to \lambda$，则 $Y_n \to X$

则 $E(Y_n) = np$，$Var(Y_n) = np(1-p)$，$n \to \infty$，$p \to 0$，$1-p \to 1$。

$E(X) = \lambda$，$Var(X) = \lambda$。

****

6.2 几何分布的期望与方差

几何分布 $X \sim Ge(p)$，$0 < p < 1$，$P(X=k) = p\cdot(1-p)^{k-1}$，$k = 1,2,\cdots$

$$E(X) = \sum_{k=1}^{\infty} k\cdot P(X=k) = \sum_{k=1}^{\infty} k\cdot p(1-p)^{k-1} = \sum_{k=1}^{\infty}\left[(k-1)+1\right]\cdot p(1-p)^{k-1}$$

$$= \sum_{k=1}^{\infty}(k-1)\cdot p(1-p)^{k-1} + \sum_{k=1}^{\infty} p(1-p)^{k-1} = (1-p)\sum_{k=2}^{\infty}(k-1)\cdot p(1-p)^{k-2} + 1$$

$$= (1-p)\sum_{j=1}^{\infty} j\cdot p(1-p)^{j-1} + 1 = (1-p)E(X) + 1$$

$$E(X) = (1-p)E(X)+1 \quad \Rightarrow \quad E(X) = \frac{1}{p}。$$

$$E(X^2) = \sum_{k=1}^{\infty} k^2 P(X=k) = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = \sum_{k=1}^{\infty}\left[(k-1)^2 + 2(k-1) + 1\right]\cdot p(1-p)^{k-1}$$

$$= \sum_{k=1}^{\infty}\left[(k-1)^2 + 2k - 1\right]\cdot p(1-p)^{k-1}$$

$$=\sum_{k=1}^{\infty}(k-1)^2\cdot p(1-p)^{k-1}+2\sum_{k=1}^{\infty}k\cdot p(1-p)^{k-1}-\sum_{k=1}^{\infty}p(1-p)^{k-1}$$

$$=(1-p)\sum_{j=1}^{\infty}j^2\cdot p(1-p)^{j-1}+2E(X)-1=(1-p)E(X^2)+2E(X)-1$$

$$E(X^2)=(1-p)E(X^2)+2E(X)-1 \quad\Rightarrow\quad E(X^2)=\frac{2}{p^2}-\frac{1}{p}$$

$$Var(X)=E(X^2)-E(X)^2=\frac{1-p}{p^2}$$

备注：

计算泊松分布和几何分布随机变量的期望过程中，做了诸如将k拆分为$(k-1)+1$，以及k^2拆分为$k(k-1)+k$或$(k-1)^2+2(k-1)+1$等等的等价变形处理，这一类的拆分是概率统计计算中常用的处理方法，目的是为了凑出随机变量的分布列求和或期望等的求和式，利用求和式的概率意义和已知的概率结果往往可以简化计算。

****

6.3 均匀、指数和正态分布的期望与方差

均匀分布 $X\sim U(a,b)$，$f(x)=\begin{cases}\dfrac{1}{b-a} & X\in[a,b]\\ 0 & \text{其他}\end{cases}$

$$E(X)=\int_{-\infty}^{+\infty}x\cdot f(x)dx=\int_a^b\frac{x}{b-a}dx=\frac{1}{b-a}\cdot\frac{x^2}{2}\Big|_a^b=\frac{1}{b-a}\frac{b^2-a^2}{2}=\frac{a+b}{2},$$

$$Var(X)=E(X^2)-E(X)^2=\int_a^b\frac{x^2}{b-a}dx-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}。$$

b-a 越大，则随机变量取值越分散，其方差也越大。

****

指数分布 $X \sim Exp(\lambda)$，$f(x)=\begin{cases}0 & ,x\le 0\\ \lambda e^{-\lambda x} & ,x>0\end{cases}$

$$E(X)=\int_{-\infty}^{+\infty} x\cdot f(x)dx=\int_0^{+\infty} x\lambda e^{-\lambda x}dx=\int_0^{+\infty} xd\left(-e^{-\lambda x}\right)$$

$$=-xe^{-\lambda x}\Big|_0^{\infty}-\int_0^{+\infty}-e^{-\lambda x}dx=-xe^{-\lambda x}\Big|_0^{\infty}+\int_0^{+\infty}e^{-\lambda x}dx=\frac{1}{\lambda}\int_0^{+\infty}\lambda e^{-\lambda x}dx=\frac{1}{\lambda}$$

$$E(X^2)=\int_{-\infty}^{+\infty} x^2\cdot f(x)dx=\int_0^{+\infty} x^2\lambda e^{-\lambda x}dx=\int_0^{+\infty} x^2d\left(-e^{-\lambda x}\right)$$

$$=-x^2e^{-\lambda x}\Big|_0^{\infty}+\int_0^{+\infty}e^{-\lambda x}dx^2=\int_0^{+\infty}2xe^{-\lambda x}dx$$

$$=\frac{2}{\lambda}\int_0^{+\infty}x\lambda e^{-\lambda x}dx=\frac{2}{\lambda}E(X)=\frac{2}{\lambda^2}$$

$$Var(X)=E(X^2)-E(X)^2=\frac{2}{\lambda^2}-\left(\frac{1}{\lambda}\right)^2=\frac{1}{\lambda^2}$$

****

标准正态分布 $X\sim N(0,1)$，$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$　　$x\in R$

$$E(X)=\int_{-\infty}^{+\infty}x\cdot\varphi(x)dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x\cdot e^{-\frac{x^2}{2}}dx=0$$

$x\cdot e^{-\frac{x^2}{2}}$ 为奇函数，且 $\int_0^{+\infty}x\cdot e^{-\frac{x^2}{2}}dx$ 有界，所以 $\int_{-\infty}^{+\infty}x\cdot e^{-\frac{x^2}{2}}dx=0$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2\cdot\varphi(x)dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^2\cdot e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}-xde^{-\frac{x^2}{2}}$$

$$=\frac{1}{\sqrt{2\pi}}\left[-xe^{-\frac{x^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}dx\right]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}dx=1,$$

$Var(X)=E(X^2)-E(X)^2=1$。

****

一般正态分布随机变量 $X \sim N(\mu,\sigma^2)$，$\dfrac{X-\mu}{\sigma} \sim N(0,1)$，

$$E\left(\frac{X-\mu}{\sigma}\right)=0 \Rightarrow \frac{1}{\sigma}E(X-\mu)=0 \Rightarrow E(X)=\mu$$

$$Var\left(\frac{X-\mu}{\sigma}\right)=1 \Rightarrow \frac{1}{\sigma^2}Var(X-\mu)=1 \Rightarrow \frac{1}{\sigma^2}Var(X)=1 \Rightarrow Var(X)=\sigma^2$$

所以，正态分布的参数 mu 和 sigma 方具有明确的概率意义，就是正态随机变量的期望和方差。

****

6.4 随机变量数学期望的应用实例

验血问题(blood tesing problem)

二战期间，美国大量的征募年轻人入伍，应征报名入伍的人都需要通过体检。患有某种罕见传染性疾病的人不准入伍，有一种验血方法可以经过一次化验有效地查出血样中是否含有这种传染病的病毒，即使病毒含量非常低的时候，此方法也能够很灵敏地显示阳性结果。

最简单的办法是将每个人的血样检查一遍，如果有 10 万人应征报名入伍，就要做 10 万次化验，需要巨大的工作量。1943 年，一个叫 Robert Dorfman 的年轻学者提出了一种分组验血的策略，可以显著地提高检验效率。他的方法是将每个人的血样提取一部分，以 k 个人一组混合。如果混合血样显示阴性，则这 k 个人只需这一次化验即可确认无病。如果试验结果为阳性，再分别检查该组成员的每份血样，确定患病者。

我们先通过具体的假设场景分析一下该方案的有效性：

假设有 100 个人参加检查，其中有 3 人患有疾病。如果用最简单的办法，将每人的血样化验一遍，需要 100 次检验。如将 100 人分为 10 组，每组 10 人，现将每个人的血样提取一部分，将每组十个人的部分血样合成一份混合血样。则 10 组混合血样中最多有 3 组包含患病血样。也就是说至多需要 10 组混合血样的化验加上 3 组

共 30 份血样的逐一化验，共计 40 次即可确定出所有患病者；而如果 3 个患病血样恰好被分到了同一组，则只需要 20 次化验即可确定出所有患病者。所以对这个这特定的例子，分组策略最多可能节省 80%的工作量，最少也能节省 60%的工作。这一策略，确实可能减少工作量。下面我们建立概率模型对分组验血的效率进行分析。

****

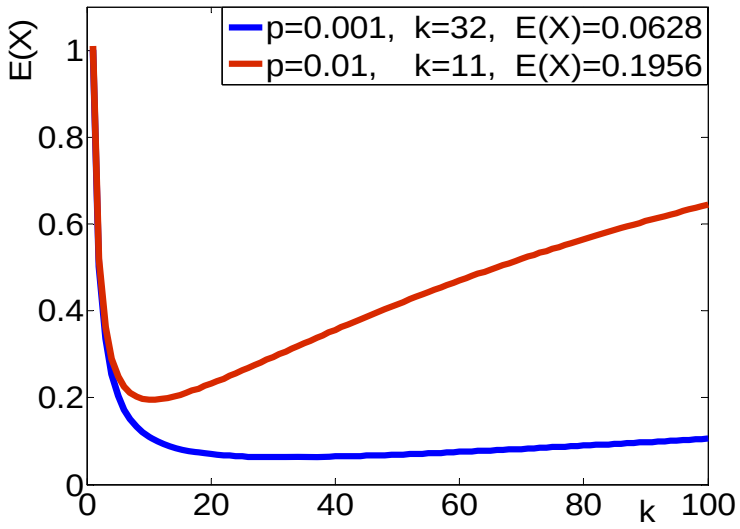
例 6.4.1 验血问题（blood tesing problem）

N 个人验血普查某种疾病，每人血样单独检验共需 N 次。现采用 k 人一组的方式，若结果阴性，则此 k 人均正常；若结果阳性，则将此 k 人逐一检验。假设发病率为 p，则此方法是否可以节省检验次数。

解：设每人血样的化验次数为随机变量 X，则 X 服从分布 $\begin{pmatrix} \frac{1}{k} & 1+\frac{1}{k} \\ (1-p)^k & 1-(1-p)^k \end{pmatrix}$

$$E(X)=\frac{1}{k}(1-p)^k+\left(1+\frac{1}{k}\right)\cdot\left[1-(1-p)^k\right]=1-\left[(1-p)^k-\frac{1}{k}\right]$$

当 $(1-p)^k>\frac{1}{k}$ 时，可在平均意义下节省化验次数。



图中列出了 p=1%和 p=0.1%时，不同分组规模对应的 E（X）值，即平均化验次数。横坐标代表分组规模 k，纵坐标代表平均化验次数 E(X)。

****

例 6.4.2 优惠券收集问题（Coupon Collector's Problem）

假设为促销饼干，商家在每盒饼干内放 1 张优惠券，共有n种不同的优惠券。当收集到全套的优惠券时，可以得到奖品。假定每盒饼干中的优惠券是从n种不同的优惠券中随机选取的，问要想获得奖品，平均需要购买多少盒饼干。

解：令X表示收集到每种优惠券至少一种所需要的购买的饼干的盒数，求$E(X)$。

记X_k表示已经收集到了$k-1$张不同优惠券后，为了一张新的优惠券所购买的饼干数，则有$X=\sum_{k=1}^{n}X_k$。

恰好有$k-1$张不同优惠券后，购买到新的一盒饼干中装有一种新优惠券的概率是$p_k=\frac{n-k+1}{n}$，X_k服从参数为p_k的几何分布，$E(X_k)=\frac{n}{n-k+1}$。

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

****