概率论第二次作业

211240042 林凡琪

Problem 1 (Warm-up problems)

- [**Function of random variable (I)**] Let X be a random variable and $g:\mathbb{R}\to\mathbb{R}$ be a continuous and strictly increasing function. Show that $Y=g(X)$ is a random variable.

  解答:

  随机变量是一个函数，它将样本空间中的每个结果映射到实数轴上的某个值。由于 X 是一个随机变量，它将样本空间中的每个结果映射到实数轴上的某个值。由于 g 是连续且严格递增的，它将每个实数映射到另一个实数。因此，$Y=g(X)$ 将样本空间中的每个结果映射到另一个实数。因此，Y 也是一个随机变量。

- [**Function of random variable (II)**] Let X be a random variable with distribution function $\max(0,\min(1,x))$. Let F be a distribution function which is continuous and strictly increasing. Show that $Y=F^{-1}(X)$ be a random variable with distribution function F.

  解答:

  由于X是随机变量，它将样本空间中的每个结果映射到介于0到1之间的某个实数。F是一个连续且严格递增的分布函数，它将介于0到1之间的实数映射到另一个介于0到1之间的实数。由此可知，$F^{-1}(X)$就是将样本空间的每个结果映射到另一个介于0到1之间的实数。因此，$Y=F^{-1}(X)$是一个分布函数为F的随机变量。

- [**Marginal distribution**] Let (X_1,X_2) be a random vector satisfying $\mathbf{Pr}[(X_1,X_2)=(0,0)]=\mathbf{Pr}[(X_1,X_2)=(1,0)]=\mathbf{Pr}[(X_1,X_2)=(0,1)]=\frac{1}{3}$. Find out the marginal distribution of X_1.

  解答:

  $$\begin{aligned}\mathbf{Pr}[X_1=0]&=\mathbf{Pr}[(X_1,X_2)=(0,0)]+\mathbf{Pr}[(X_1,X_2)=(0,1)]\\&=\frac{1}{3}+\frac{1}{3}\\&=\frac{2}{3}\end{aligned}$$

  和

  $$\begin{aligned}\mathbf{Pr}[X_1=1]&=\mathbf{Pr}[(X_1,X_2)=(1,0)]\\&=\frac{1}{3}.\end{aligned}$$

  因此，X_1 的边缘分布是:

  $$\mathbf{Pr}[X_1=0]=\frac{2}{3},$$
  $$\mathbf{Pr}[X_1=1]=\frac{1}{3}.$$

- [**Independence**] Show that discrete random variables X and Y are independent if and only if $p_{X,Y}(x,y)$ can be written as $g(x)h(y)$ for some function g,h, where $p_{X,Y}$ is the joint mass function of (X,Y).

  解答:

  (根据独立随机变量的定义，$F(x,y)=F(x)F(y)$，其中$F(x,y)$是X、Y 的联合分布函数，$F(x)$、$F(y)$ 分别是X、Y的分布函数。)

  1. 必要性:

两个随机变量X、Y 独立，则满足$P(X \cap Y) = P(X)P(Y)$

即$p_{X,Y}(x,y) = P(X = x \& Y = y) = P(X)P(Y)$，其中$P(X)$、$P(Y)$ 是相互独立的函数。

2. 充分性：

根据联合分布函数的定义，$0 \leq p(x,y) \leq 1 \wedge \sum\sum_{(x,y)} p(x,y) = 1$ ，可得
$0 \leq g(x)h(y) \leq 1 \wedge \sum\sum_{(x,y)} g(x)h(y) = 1$

$p_{X,Y}(x,y) = g(x)h(y)$ 即代表

$$\begin{aligned} p_{X,Y}(x,y) &= P(X = x \& Y = y) \\ &= p_X(x)p_Y(y) \\ &= \sum_y P(X|Y)P(Y) \sum_y P(Y|X)P(X) \\ &= g(x)h(y) \end{aligned}$$

即，$\sum_y P(X|Y)P(Y)$ 和事件Y没有关系，$\sum_y P(Y|X)P(X)$ 和事件X没有关系。

所以可得

$$\begin{aligned} p_{X,Y}(x,y) &= P(X = x \& Y = y) \\ &= p_X(x)p_Y(y) \\ &= \sum_y P(X|Y)P(Y) \sum_y P(Y|X)P(X) \\ &= g(x)h(y) \\ &= P(X)P(Y) \end{aligned}$$

得证

- [**Entropy of discrete random variable**] Let X be a discrete random variable with range of values $[N] = \{1, 2, \ldots, N\}$ and probability mass function p. Define $H(X) = -\sum_{n \geq 1} p(n) \log p(n)$ with convention $0 \log 0 = 0$. Prove that $H(X) \leq \log N$ using Jensen's inequality.

解答：

Jensen不等式表明，如果f是一个凸函数，X是一个随机变量，则$E[f(X)] \geq f(E[X])$。

令$f(x) = -x \log x$。由于f是凸函数，我们有Jensen不等式：

$$f(\frac{x_1 + x_2 + \ldots + x_n}{n}) \geq \frac{f(x_1) + f(x_2) + \ldots + f(x_n)}{n}$$
$$(\text{当且仅当} x_1 = x_2 = \ldots = x_n \text{时取等})$$

$$\begin{aligned} H(X) &= -\sum_{n=1}^{N} p(n) \log p(n) \\ &= \sum_{n=1}^{N} f(p(n)) \\ &\leq N f(\sum_{n=1}^{N} p(n)/N) \\ &= N f(1/N) \\ &= \log N. \end{aligned}$$

(其中已知$\sum_{n=1}^{N} p(n) = 1$)

所以

$$H(X) \leq \log N$$

- [**Law of total expectation**] Let $X \sim \mathrm{Geom}(p)$ for some parameter $p \in (0,1)$. Calculate $\mathbf{E}[X]$ using the law of total expectation.

让 $X \sim \mathrm{Geom}(p)$，其中 $p \in (0,1)$。几何分布给出了第一次成功需要 k 个独立试验的概率。X 的概率质量函数为 $P(X=k) = (1-p)^{k-1}p$，其中 $k = 1, 2, \ldots$。

使用全期望定理计算 $\mathbf{E}[X]$：

$$\begin{aligned}\mathbf{E}[X] &= \sum_{k=1}^{\infty} kP(X=k) \\ &= \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\ &= p\sum_{k=1}^{\infty} k(1-p)^{k-1}\end{aligned}$$

对 $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ 求导：

$$\begin{aligned}\frac{d}{dx}\left(\sum_{k=0}^{\infty} x^k\right) &= \frac{d}{dx}\left(\frac{1}{1-x}\right) \\ \sum_{k=0}^{\infty} kx^{k-1} &= \frac{1}{(1-x)^2}\end{aligned}$$

将两边都乘以 x 并再次求导得到：

$$\begin{aligned}\frac{d}{dx}\left(x\sum_{k=0}^{\infty} kx^{k-1}\right) &= \frac{d}{dx}\left(\frac{x}{(1-x)^2}\right) \\ \sum_{k=0}^{\infty} kx^{k} &= \frac{x+1}{(1-x)^2}\end{aligned}$$

代入 $x = 1-p$，得到：

$$\begin{aligned}\mathbf{E}[X] &= p\sum_{k=1}^{\infty} k(1-p)^{k-1} \\ &= p\sum_{k=0}^{\infty} (k+1)(1-p)^{k} \\ &= p\frac{(1-p)+1}{(p-1)^2} \\ &= \frac{1}{p}.\end{aligned}$$

因此，$\mathbf{E}[X] = \frac{1}{p}$。

- [**Random number of random variables**] Let $\{X_n\}_{n\geq 1}$ be identically distributed random variable and N be a random variable taking values in the non-negative integers and independent of the X_n for all $n \geq 1$. Prove that $\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbf{E}[N]\mathbf{E}[X_1]$.

  解答：

  设 $S_k = X_1 + X_2 + \cdots + X_k$ 是前 k 个随机变量的和。

  根据Wald 恒等式，如果 $\{X_n\}_{n\geq 1}$ 是独立同分布的随机变量，其有限均值为 μ。

  由期望的线性可知，

  $$\begin{aligned}\mathbf{E}[S_N] &= \mathbf{E}[X_1 + X_2 + \cdots + X_N] \\ &= \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_N] \\ &= N\mu,\end{aligned}$$

  N 是取非负整数的随机变量，与 X_n 独立，则

  $$\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbf{E}[S_N] = \mathbf{E}[N]\mu = \mathbf{E}[N]\mathbf{E}[X_1]$$

Problem 2 (Distribution of random variable)

- [**Cumulative distribution function (CDF)**] Let X be a random variable with cumulative distribution function F.

  1. Show that $Y = aX + b$ is a random variable where a and b are real constants, and express the CDF of Y by F. (Hint: Try expressing the event $Y = aX + b \leq y$ by countably many set operations on the events defined on X.)
  2. Let G be the CDF of random variable $Z : \Omega \to \mathbb{R}$ and $0 \leq \lambda \leq 1$, show that
     - $\lambda F + (1 - \lambda)G$is a CDF function.
     - The product FG is a CDF function, and if Z and X are independent, then FG is the CDF of $max\{X, Z\}$.

  解答：

  1. 要证明 $Y = aX + b$ 是一个随机变量，需要证明对于任意 $y \in \mathbb{R}$，事件 $\{Y \leq y\}$ 是可测的。

     当$a > 0$

     有

     $$\begin{aligned}\{Y \leq y\} &= \{aX + b \leq y\} \\ &= \left\{X \leq \frac{y-b}{a}\right\}.\end{aligned}$$

     由于 F 是 X 的累积分布函数，有

     $$\begin{aligned}F_Y(y) &= P(Y \leq y) \\ &= P(aX + b \leq y) \\ &= P(X \leq (y-b)/a) \\ &= F_X((y-b)/a).\end{aligned}$$

     因此，当$a > 0$时，Y 的累积分布函数为 $F_Y(y) = F_X((y-b)/a)$。

     当$a = 0$，$Y = b$是个常量，不再是随机变量。

     当$a < 0$

     有

     $$\begin{aligned}\{Y \leq y\} &= \{aX + b \leq y\} \\ &= \left\{X \geq \frac{y-b}{a}\right\}.\end{aligned}$$

     由于 F 是 X 的累积分布函数，有

     $$\begin{aligned}F_Y(y) &= P(Y \leq y) \\ &= P(aX + b \leq y) \\ &= P(X \geq (y-b)/a) \\ &= 1 - F_X((y-b)/a).\end{aligned}$$

     因此，当$a < 0$时，Y 的累积分布函数为 $F_Y(y) = 1 - F_X((y-b)/a)$。

  2.
     - 因为 F 和 G 都是非降的右连续函数，所以 $\lambda F + (1 - \lambda)G$ 也是一个非降的右连续函数。

       其次，有

       $$\begin{aligned}\lim_{x \to -\infty}(\lambda F(x) + (1-\lambda)G(x)) &= \lim_{x \to -\infty} \lambda F(x) + \lim_{x \to -\infty}(1-\lambda)G(x) \\ &= 0 + 0 \\ &= 0,\end{aligned}$$

$$\begin{aligned}\lim_{x\to+\infty}(\lambda F(x)+(1-\lambda)G(x)) &= \lim_{x\to+\infty}\lambda F(x)+\lim_{x\to+\infty}(1-\lambda)G(x)\\ &=\lambda+(1-\lambda)\\ &=1。\end{aligned}$$

- 首先，由于 F 和 G 都是非降的右连续函数，因此 FG 也是一个非降的右连续函数。

  其次，有

$$\begin{aligned}\lim_{x\to-\infty}FG(x) &= \lim_{x\to-\infty}F(x)G(x)\\ &=0\cdot G(-\infty)\\ &=0,\end{aligned}$$

$$\begin{aligned}\lim_{x\to+\infty}FG(x) &= \lim_{x\to+\infty}F(x)G(x)\\ &=1\cdot G(+\infty)\\ &=1。\end{aligned}$$

  因此，FG 是一个累积分布函数。

  如果 Z 和 X 是独立的，那么它们的联合分布函数为 $F(x)G(z)$。因此，

$$P(\max(X,Z)\le x)=P(X\le x,Z\le x)=F(x)G(x),$$

  所以 FG 是 $\max\{X,Z\}$ 的累积分布函数。

- [**Probability mass function (PMF)**] We toss n coins, and each one shows heads with probability p, independently of each of the others. Each coin which shows head is tossed again. (If the coin shows tail, it won't be tossed again.) Let X be the number of heads resulting from the **second** round of tosses, and Y be the number of heads resulting from **all** tosses, which includes the first and (possible) second round of each toss.

  1. Find the PMF of X and Y.
  2. Find $E[X]$ and $E[Y]$.
  3. Let p_X be the PMF of X, show that $p_X(k-1)p_X(k+1)\le[p_X(k)]^2$ for $1\le k\le n-1$.

  解答：

  1. 设Z是第一轮投掷中出现正面的硬币数：

     $P(Z=k)=\binom{n}{k}p^k(1-p)^{n-k}$

     那么在第二轮投掷中：

$$\begin{aligned}P(X=j) &= \sum_{k=j}^{n}[P(Z=k)\binom{k}{j}p^j(1-p)^{k-j}]\\ &=\sum_{k=j}^{n}[\binom{n}{k}p^k(1-p)^{n-k}\binom{k}{j}p^k(1-p)^{k-j}]\\ &\quad 其中\, j=0,1,\dots,n\end{aligned}$$

     再来看Y的PMF：

     假设在第一轮有k枚硬币正面朝上，因此，在给定 $X=k$ 的条件下，Y 的条件 PMF 为

     $P(Y=j\mid X=k)=\binom{k}{j-k}p^{j-k}(1-p)^{k-(j-k)}$

     其中 $j=k,k+1,\dots,n+k$。

     根据全概率公式，

     $P(Y=j)=\sum_{k=0}^{n}P(Y=j\mid X=k)P(X=k)$

     其中 $j=0,1,\dots,2n$

     将 X 和给定 $X=k$ 的条件下的 Y 的 PMF 的表达式代入上式，得到

     $P(Y=j)=\sum_{k=0}^{n}\binom{k}{j-k}p^{j-k}(1-p)^{k-(j-k)}\binom{n}{k}p^k(1-p)^{n-k}$

其中 $j = 0, 1, \ldots, 2n$

接下来，我们来看 Y 的 PMF。

2. 在第一轮投掷中，每枚硬币以概率 p 出现正面，因此在第一轮投掷中，期望出现正面的硬币数为 np。在第二轮投掷中，每枚出现正面的硬币以概率 p 再次出现正面，因此在第二轮投掷中，期望出现正面的硬币数为

   $E(X) = np^2$

   因此，在所有投掷中，期望出现正面的硬币数为：

   $E(Y) = np + np^2$

3. 设 Z 为第一轮投掷中正面朝上的硬币数量。$P(Z = k) = \binom{n}{k} p^k (1-p)^{n-k}$，其中 $k = 0, 1, \ldots, n$。

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

   其中 $j = 0, 1, \ldots, n$。

   这就是 X 的 PMF，记作 p_X.

   需要证明 $p_X(k-1)p_X(k+1) \leq [p_X(k)]^2$, 其中 $1 \leq k \leq n-1$.

   考虑比值 $\dfrac{p_X(k-1)p_X(k+1)}{[p_X(k)]^2}$，得到

   $$\frac{p_X(k-1)p_X(k+1)}{[p_X(k)]^2} = \frac{\left[\sum_{i=0}^{n} \binom{i}{k-1} p^{k-1}(1-p)^{i-(k-1)} \binom{n}{i} p^i (1-p)^{n-i}\right] \left[\sum_{j=0}^{n} \binom{j}{k+1} p^{k+1}(1-p)^{j-(k+1)} \binom{n}{j} p^j (1-p)^{n-j}\right]}{\left[\sum_{l=0}^{n} \binom{l}{k} p^k (1-p)^{l-k} \binom{n}{l} p^l (1-p)^{n-l}\right]^2}.$$

   用此恒等式来简化：

   $\binom{i}{k-1}\binom{n}{i} = \binom{n}{k-1}\binom{n-k+1}{i-k+1}$,

   对所有 $0 \leq k-1 \leq i \leq n$ 的整数 i, k, 和 n成立

   可得

   $$\frac{\left[\sum_{i=0}^{n} \binom{n-k+1}{i-k+1} \binom{n}{k-1} (p(1-p))^{k-1}(1-p)^{(n-k+2)(i-k+2)}\right] \left[\sum_{j=0}^{n} \binom{n-k-1}{j-k-2} \binom{n}{k+2} (p(1-p))^{k+2}(1-p)^{(n-k)(j-k)}\right]}{\left[\sum_{l=0}^{n} \binom{n-k}{l-k} \binom{n}{k} (p(1-p))^{k}(1-p)^{(n-k+1)(l-k+1)}\right]^2}.$$

   根据 Cauchy-Schwarz 不等式，对任意两个实数序列 $a_1, \ldots, a_n$ 和 $b_1, \ldots, b_n$，都有

   $\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$.

   令 $a_i = \sqrt{\binom{n-k+1}{i-k+1}\binom{n}{k-1}(p(1-p))^{k-1}(1-p)^{(n-k+2)(i-k+2)}}$ 和

   $b_i = \sqrt{\binom{n-k-1}{j-k-2}\binom{n}{k+2}(p(1-p))^{k+2}(1-p)^{(n-k)(j-k)}}$，则

   $$\frac{\left[\sum_{i=0}^{n} \binom{n-k+1}{i-k+1} \binom{n}{k-1} (p(1-p))^{k-1}(1-p)^{(n-k+2)(i-k+2)}\right] \left[\sum_{j=0}^{n} \binom{n-k-1}{j-k-2} \binom{n}{k+2} (p(1-p))^{k+2}(1-p)^{(n-k)(j-k)}\right]}{\left[\sum_{l=0}^{n} \binom{n-k}{l-k} \binom{n}{k} (p(1-p))^{k}(1-p)^{(n-k+1)(l-k+1)}\right]^2}$$

   $$= \frac{\left(\sum_{i=0}^{n} a_i b_i\right)^2}{\left(\sum_{l=0}^{n} a_l^2\right)\left(\sum_{l=0}^{n} b_l^2\right)} \leq 1.$$

   因此，$p_X(k-1)p_X(k+1) \leq [p_X(k)]^2$ 对于所有 $1 \leq k \leq n-1$ 成立。

Problem 3 (Discrete random variable)

- [**Geometric distribution (I)**] Every package of some intrinsically dull commodity includes a small and exciting plastic object. There are c different types of object, and each package is equally likely to contain any given type. You buy one package each day.

  1. Find the expected number of days which elapse between the acquisitions of the j-th new type of object and the $(j+1)$-th new type.

2. Find the expected number of days which elapse before you have a full set of objects.

解答：

1. 设X_j是获得第j种和第j+1种物品中间经过的天数。

   获得j种后，还剩c-j种没有获得。每天获得新品种的可能性是$\frac{c-j}{c}$，也就是说，不能获得新品种的概率为$1-\frac{c-j}{c}=\frac{j}{c}$.

   所以X_j符合概率常量为$\frac{c-j}{c}$的几何分布

   所以

$$E[X_j]=\frac{1}{\frac{c-j}{c}}=\frac{c}{c-j}$$
$$(j=0,1,\ldots,c-1)$$

2. 设Y是收集完整套物品所需要的期望时间。

   $Y=X_0+X_1+\ldots+X_{c-1}$

   由期望的线性可知：

$$\begin{aligned}E[Y]&=E[X_0+X_1+\ldots+X_{c-1}]\\&=E[X_0]+E[X_1]+\ldots+E[Xc-1]\\&=c\sum_{i=0}^{c-1}\frac{1}{c-i}\\&=c\sum_{i=1}^{c}\frac{1}{i}\\&=cH_c\end{aligned}$$

- [**Geometric distribution (II)**] Prove that geometry distribution is the only discrete memoryless distribution with range values $\mathbb{N}_+$.

  解答：

  设 X 是一个取值范围为 $\mathbb{N}_+$ 的离散无记忆随机变量。这意味着对于所有 $n,m\in\mathbb{N}_+$，我们有 $P(X>n+m|X>m)=P(X>n)$。

  设 $p=P(X=1)$。那么对于任意 $n\in\mathbb{N}_+$有

$$\begin{aligned}P(X>n)&=P(X>n|X>n-1)\\&=P(X>1)\\&=1-p.\end{aligned}$$

  因此，对于任意 $n\in\mathbb{N}_+$有

$$\begin{aligned}P(X=n)&=P(X>n-1)-P(X>n)\\&=(1-p)-(1-p)\\&=p(1-p)^{n-1}.\end{aligned}$$

  这是参数为 p 的几何分布的概率质量函数。因此，几何分布是唯一一个取值范围为 $\mathbb{N}_+$ 的离散无记忆分布。

- [**Binomial distribution**] Let $n_1,n_2\in\mathbb{N}_+$ and $0\leq p\leq 1$ be parameters, and $X\sim\mathrm{Bin}(n_1,p),Y\sim\mathrm{Bin}(n_2,p)$ be independent random variables. Prove that $X+Y\sim\mathrm{Bin}(n_1+n_2,p)$.

  解答：

  设 $n_1,n_2\in\mathbb{N}_+$ 和 $0\leq p\leq 1$ 为参数，且设 $X\sim\mathrm{Bin}(n_1,p)$ 和 $Y\sim\mathrm{Bin}(n_2,p)$ 为独立随机变量。我们要证明 $X+Y\sim\mathrm{Bin}(n_1+n_2,p)$。

  参数为 n 和 p 的二项分布的概率质量函数为

  $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$

其中 $k = 0, 1, \ldots, n$。因此，对于任意 $k = 0, 1, \ldots, n_1 + n_2$，

$$\begin{aligned} P(X+Y=k) &= P(Y=k-X) \\ &= \sum_{i=0}^{n_1} P(Y=k-i)P(X=i) \\ &= \sum_{i=0}^{n_1} \binom{n_2}{k-i} p^{k-i}(1-p)^{n_2-(k-i)} \binom{n_1}{i} p^i (1-p)^{n_1-i} \\ &= p^k (1-p)^{n_1+n_2-k} \sum_{i=0}^{n_1} \binom{n_2}{k-i}\binom{n_1}{i}. \end{aligned}$$

根据范德蒙德恒等式，

$\sum_{i=0}^{n_1} \binom{n_2}{k-i}\binom{n_1}{i} = \binom{n_1+n_2}{k}$.

因此，

$P(X+Y=k) = p^k(1-p)^{n_1+n_2-k}\binom{n_1+n_2}{k}$

其中 $k = 0, 1, \ldots, n_1 + n_2$。这是参数为 $n_1 + n_2$ 和 p 的二项分布的概率质量函数。因此，$X + Y \sim \mathrm{Bin}(n_1 + n_2, p)$。

- [**Negative binomial distribution**] Let X follows the negative binomial distribution with parameter $r \in \mathbb{N}_+$ and $p \in (0,1)$. Calculate $\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.

  解答：

  设 X 服从参数为 $r \in \mathbb{N}_+$ 和 $p \in (0,1)$ 的负二项分布。

  $\mathbf{E}[X] = \frac{r(1-p)}{p}$。

  所以

  $\mathbf{Var}[X] = \mathbf{E}[X^2] - \left(\frac{r(1-p)}{p}\right)^2$.

$$\begin{aligned} E\left(X^2\right) =& \sum_{x=r}^{\infty} x^2 f(x) \\ =& \sum_{x=r}^{\infty} x^2 \binom{x-1}{r-1} p^r (1-p)^{x-r} \\ =& \frac{r}{p} \sum_{x=r}^{\infty} x \binom{x}{r} p^{r+1}(1-p)^{x-r} \\ =& \frac{r}{p}\left[\sum_{x=r}^{\infty}(x+1)\binom{x}{r}p^{r+1}(1-p)^{x-r}\right. \\ & \left. -\sum_{x=r}^{\infty}\binom{x}{r}p^{r+1}(1-p)^{x-r}\right] \\ =& \frac{r}{p}\left(\frac{r+1}{p}-1\right) \\ =& \frac{r(r-p+1)}{p^2} \end{aligned}$$

这意味着参数为 r 和 p 的负二项分布的方差为

$\mathrm{Var}(X) = E\left(X^2\right) - [E(X)]^2 = \frac{r(1-p)}{p^2}$

- **[Hypergeometric distribution]**An urn contains N balls, b of which are blue and $r = N - b$ of which are red. A random sample of n balls is drawn **without replacement** (无放回) from the urn. Let B the number of blue balls in this sample. Show that if N, b and r approach $+\infty$ in such a way that $b/N \to p$ and $r/N \to 1-p$, then $\mathbf{Pr}(B=k) \to \binom{n}{k}p^k(1-p)^{n-k}$ for

$0 \leq k \leq n$.

解答:

超几何分布的极限定理。

当 N, b 和 r 趋近于 $+\infty$ 时，超几何分布会趋近于二项分布。

首先，定义超几何分布的PMF为 $\mathbf{Pr}(B=k) = \frac{\binom{b}{k}\binom{r}{n-k}}{\binom{N}{n}}$。

可得

$$\mathbf{Pr}(B=k) = \frac{\binom{b}{k}\binom{r}{n-k}}{\binom{N}{n}} = \frac{\frac{b!}{k!(b-k)!}\frac{r!}{(n-k)!(r-n+k)!}}{\frac{N!}{n!(N-n)!}}$$

接下来，我们可以将阶乘展开为 $\Gamma(x+1) = x!$ 的形式，并使用斯特林公式来近似阶乘：

当然。我们可以继续化简上面的表达式，得到：

$$\begin{aligned}
\mathbf{Pr}(B=k) &= \frac{\sqrt{\frac{r}{b}}\left(\frac{b}{N}\right)^k\left(\frac{r}{N}\right)^{n-k}}{\sqrt{\frac{(N-n)(n-k)}{nk(N-k)}}} \\
&= \frac{\sqrt{\frac{(N-b)n}{bk(N-n)}}\left(\frac{b}{N}\right)^k\left(\frac{N-b}{N}\right)^{n-k}}{\sqrt{\frac{(N-n)(n-k)}{nk(N-k)}}} \\
&= \frac{\sqrt{\frac{(N-b)n}{bk(N-n)}}}{\sqrt{\frac{(N-n)(n-k)}{nk(N-k)}}}\left(\frac{b}{N}\right)^k\left(1-\frac{b}{N}\right)^{n-k}
\end{aligned}$$

根据 $b/N \to p$ 和 $r/N \to 1-p$

得到，$\left(\frac{b}{N}\right)^k \to p^k$ 和 $\left(1-\frac{b}{N}\right)^{n-k} \to (1-p)^{n-k}$。

前面的系数写成：

$$\begin{aligned}
\frac{\sqrt{\frac{(N-b)n}{bk(N-n)}}}{\sqrt{\frac{(N-n)(n-k)}{nk(N-k)}}} &= \sqrt{\frac{n(N-n)}{(N-k)k}} \times \sqrt{\frac{(N-b)k}{b(N-k)}} \\
&= \binom{n}{k} \times \sqrt{\frac{(N-b)k}{b(N-k)}}
\end{aligned}$$

由于 $b/N \to p$，所以 $(N-b)/N \to 1-p$。

因此，当 N, b 和 r 趋近于 $+\infty$ 时，$\sqrt{\frac{(N-b)k}{b(N-k)}} \to 1$。

综上，当 N, b 和 r 趋近于 $+\infty$ 时，$\mathbf{Pr}(B=k) \to \binom{n}{k}p^k(1-p)^{n-k}$。

- [**Poisson distribution**] In your pocket is a random number N of coins, where N has the Poisson distribution with parameter λ. You toss each coin once, with heads showing with probability p each time. Let X be the (random) number of heads outcomes and Y be the (also random) number of tails.

  1. Find the joint mass function of (X, Y).
  2. Find PMF of the marginal distribution of X in (X, Y). Are X and Y independent?

  解答:

  1. 因为 $X+Y=N$，所以可以通过计算 X 和 N 的联合分布来得到 X 和 Y 的联合分布。

首先，计算 X 在给定 $N = n$ 的条件下的条件分布。

由于每枚硬币都是独立抛掷的，所以在给定硬币总数为 n 的情况下，正面朝上的硬币数量 X 服从二项分布，即

$$\mathbf{Pr}(X = k \mid N = n) = \binom{n}{k} p^k (1-p)^{n-k}$$

使用全概率公式计算 X 和 N 的联合分布：

$$\begin{aligned}\mathbf{Pr}(X = k, N = n) &= \mathbf{Pr}(X = k \mid N = n)\mathbf{Pr}(N = n) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \frac{(p\lambda)^k ((1-p)\lambda)^{n-k}}{k!(n-k)!} e^{-\lambda}\end{aligned}$$

由于 $X + Y = N$，所以我们可以得到 X 和 Y 的联合分布为

$$\mathbf{Pr}(X = k, Y = n-k) = \frac{(p\lambda)^k ((1-p)\lambda)^{n-k}}{k!(n-k)!} e^{-\lambda}$$

2. X 的边缘分布：

$$\begin{aligned}\mathbf{Pr}(X = k) &= \sum_{y=0}^{\infty} \mathbf{Pr}(X = k, Y = y) \\ &= \sum_{y=0}^{\infty} \frac{(p\lambda)^k ((1-p)\lambda)^y}{k!y!} e^{-\lambda} \\ &= \frac{(p\lambda)^k}{k!} e^{-\lambda} \sum_{y=0}^{\infty} \frac{((1-p)\lambda)^y}{y!} \\ &= \frac{(p\lambda)^k}{k!} e^{-\lambda} e^{(1-p)\lambda} \\ &= \frac{(p\lambda)^k}{k!} e^{-p\lambda}\end{aligned}$$

X 服从参数为 $p\lambda$ 的泊松分布。

由于 $Y = N - X$，所以 Y 的边缘分布为

$$\mathbf{Pr}(Y = k) = \frac{((1-p)\lambda)^k}{k!} e^{-(1-p)\lambda}$$

但是

$$\mathbf{Pr}(X = k, Y = n-k) \neq \mathbf{Pr}(X = k)\mathbf{Pr}(Y = n-k)$$

所以，X 和 Y 不是独立的。

- [**Conditional distribution (I)**] Let X and Y be independent $Bin(n, p)$ random variables, and let $Z = X + Y$. Show that the conditional distribution of X given $X + Y = n$ is the hypergeometric distribution.

  解答：

  由于 X 和 Y 是独立的二项分布随机变量，所以它们的和 $Z = X + Y$ 服从二项分布，即 $Z \sim Bin(2n, p)$。

  根据条件概率的定义，我们有

  $$\mathbf{Pr}(X = k \mid X + Y = n) = \frac{\mathbf{Pr}(X = k, X + Y = n)}{\mathbf{Pr}(X + Y = n)}$$

  由于 $X + Y = Z$，所以

  $$\mathbf{Pr}(X = k \mid X + Y = n) = \frac{\mathbf{Pr}(X = k, Z = n)}{\mathbf{Pr}(Z = n)}$$

  由于 X 和 Y 是独立的，所以

$$\begin{aligned}\mathbf{Pr}(X=k \mid X+Y=n) &= \frac{\mathbf{Pr}(X=k)\mathbf{Pr}(Y=n-k)}{\mathbf{Pr}(Z=n)} \\ &= \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{n-k}p^{n-k}(1-p)^k}{\binom{2n}{n}p^n(1-p)^n} \\ &= \frac{\binom{n}{k}\binom{n}{n-k}}{\binom{2n}{n}}\end{aligned}$$

上式表明，在给定 $X+Y=n$ 的条件下，X 服从超几何分布。

Promblem 4 (Linearity of Expectation)

- [**Inversion**] Given a sequence of n elements $a_1, a_2, \ldots, a_n$, an inversion is a pair of integer (i, j) such that $1 \leq i < j \leq n$ and $a_i > a_j$. For instance, in the sequence $a = [1, 2, 5, 1, 3]$ there are three inversions: $(2, 4), (3, 4), (3, 5)$. Suppose we choose a sequence ρ from $[q]^n$ uniformly at random, where n and q are given integers with $q \geq 1$. Find the expected number of inversions in ρ.

  解答：

  设 X 表示序列 ρ 中逆序对的数量。

  使用指示器随机变量来计算 X 的期望。

  对于每一对整数 (i,j)，其中 $1 \leq i < j \leq n$，定义指示器随机变量 $X_{i,j}$ 为

  $$X_{i,j} = \begin{cases} 1 & \text{if } \rho_i > \rho_j \\ 0 & \text{otherwise} \end{cases}$$

  则有

  $$X = \sum_{1 \leq i < j \leq n} X_{i,j}$$

  由于 ρ 是从 $[q]^n$ 中均匀随机选取的，所以对于任意一对整数 (i,j)，ρ_i 和 ρ_j 是独立且等概率地取自集合 $[q]$。

  由对称性，

  $$\mathbf{Pr}(\rho_i > \rho_j) = \mathbf{Pr}(\rho_i < \rho_j)$$

  另外，$\rho_i = \rho_j$ 的概率为 $\frac{1}{q}$。

  具体来说，设 A 表示事件 $\{\rho_i = \rho_j\}$，则

  $$\mathbf{Pr}(A) = \sum_{k=1}^{q} \mathbf{Pr}(\rho_i = k, \rho_j = k) = \sum_{k=1}^{q} \mathbf{Pr}(\rho_i = k)\mathbf{Pr}(\rho_j = k) = q \times \frac{1}{q} \times \frac{1}{q} = \frac{1}{q}$$

  因此，$\rho_i = \rho_j$ 的概率为 $\frac{1}{q}$。

  根据

  $$\mathbf{Pr}(\rho_i > \rho_j) + \mathbf{Pr}(\rho_i < \rho_j) + \mathbf{Pr}(\rho_i = \rho_j) = 1$$

  结合上面两个式子，

  $$2\mathbf{Pr}(\rho_i > \rho_j) + \frac{1}{q} = 1$$

  解得

  $$\mathbf{Pr}(\rho_i > \rho_j) = \frac{q-1}{2q}$$

  因此，

  $$\mathbf{E}[X_{i,j}] = \frac{q-1}{2q}$$

由期望的线性性质，

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{1\le i<j\le n} X_{i,j}\right] = \sum_{1\le i<j\le n} \mathbf{E}[X_{i,j}] = \frac{n(n-1)(q-1)}{4q}$$

因此，ρ 中逆序对的期望数量为 $\frac{n(n-1)(q-1)}{4q}$。

- [**Number of cycles**] At a banquet, there are n people who shake hands according to the following process: In each round, two idle hands are randomly selected and shaken (**these two hands are no longer idle**). After n rounds, there will be no idle hands left, and the n people will form several cycles. For example, when $n = 3$, the following situation may occur: the left and right hands of the first person are held together, the left hand of the second person and the right hand of the third person are held together, and the right hand of the second person and the left hand of the third person are held together. In this case, three people form two cycles. How many cycles are expected to be formed after n rounds?

  在一场宴会上，有 n 个人按照以下过程握手：在每一轮中，随机选择两只空闲的手握手（这两只手不再空闲）。经过 n 轮后，将没有空闲的手，n 个人将形成若干个环。例如，当 $n = 3$ 时，可能出现以下情况：第一个人的左右手握在一起，第二个人的左手和第三个人的右手握在一起，第二个人的右手和第三个人的左手握在一起。在这种情况下，三个人形成了两个环。经过 n 轮后，期望形成多少个环?

  解答：

  设 E_m 为 m 次握手后期望形成的环数。

  当 $m = 1$ 时，只握一次手，只有当自己握住自己的手的时候才能成环， $E_1 = 1/2n - 1$。

  当 $m > 1$ 时，前面已经握了m-1次，所以还剩$2n - 1 - 2 * (m - 1) = 2n - 2m + 1$ 只手可以握，其中只有一只手是握住之后能成环的。

  得到递推公式：$E_m = E_{m-1} + \frac{1}{2n-2m+1}$

  由此可以推出：$E_n = \sum_{i=1}^{n} \frac{1}{2n-2i+1}$

  因此，在 n 轮握手后，期望形成的环数为 $\sum_{i=1}^{n} \frac{1}{2n-2i+1}$。

- [**Connected vertices**] Consider the perfect binary tree of depth n. Suppose that you delete each edge of the tree independently with probability $1/2$. Let X be the number of nodes which are still connected to the root after the deletion of the edges. Compute the expectation of X.

  设 X_i 为二进制指示器随机变量，表示深度为 n 的完美二叉树中第 i 个节点是否与根节点相连。则 $X = \sum_{i=1}^{2^n-1} X_i$。

  由于每条边都以概率 $\frac{1}{2}$ 被删除，因此第 i 个节点与根节点相连的概率为 $\frac{1}{2^d}$，其中 d 是从根节点到第 i 个节点的路径上的边数。

  所以 $E[X_i] = \frac{1}{2^d}$

  由期望的线性性质：

  $E[X] = E\left[\sum_{i=1}^{2^n-1} X_i\right] = \sum_{i=1}^{2^n-1} E[X_i] = \sum_{d=0}^{n-1} \sum_{i=1}^{2^d} \frac{1}{2^d} = \sum_{d=0}^{n-1} 1 = n$

  因此，在删除边后，与根节点仍然相连的节点数的期望值为 n。

Problem 5 (Probability meets graph theory)

In this part we will work on undirected simple graphs. For any $G = (V, E)$, a **clique** (*团，完全子图*) is a subset of vertices $C \subset V$, such that every two distinct vertices of C are adjacent, and an **independent set** (*独立集*) is a subset of vertices $I \subset V$ such that no two distinct vertices of I are adjacent.

- [**Erdős–Rényi random graph**] Consider $G \sim G(n,p)$ where $G(n,p)$is the Erdős–Rényi random graph model.

  1. Let $p \in (0,1)$. A "triangle" in a graph is a clique of size 3. Find the expected number of triangles in G. (Hint: use indicators and the linearity of expectation.)
  2. Let $p \in (0,1)$. For any $2 \leq q \leq n$, let the random variable N_q be the number of q-cliques. Here, a q-clique is a clique of size q. Find $\mathbf{E}[N_q]$.
  3. Let $p = 1/2$. For an undirected graph G, define $\alpha(G) = \max\{|S| : S \text{ is an independent set}\}$. Show that when $n \to \infty$, $\mathbf{Pr}[\alpha(G) \geq 3\log_2 n + 1] \to 0$. Also, please interpret this result in the context of social networks, in which the vertices represent people, and the edges represent friendship.

  解答:

  1. 首先计算图中三角形的总数。

     对于每个顶点组合 (u,v,w)，定义一个指示器随机变量 $X_{u,v,w}$，其值为1当且仅当 (u,v,w) 形成一个三角形。

     由于 $G \sim G(n,p)$，所以每个三角形出现的概率是 p^3。因此，$E[X_{u,v,w}] = p^3$。

     由于期望的线性，图中三角形的总数的期望值为所有指示器随机变量的期望值之和。也就是说，$E[\sum_{u,v,w} X_{u,v,w}] = \sum_{u,v,w} E[X_{u,v,w}] = \binom{n}{3}p^3$。

     因此，图中三角形的期望数量为 $\binom{n}{3}p^3$。

  2. 计算q边形的总数.

     对于每个顶点组合$(x_1, x_2, \ldots, x_q)$定义一个指示器随机变量$X_{(x_1,x_2,\ldots,x_q)}$，当且仅当$(x_1, x_2, \ldots, x_q)$形成q边形的时候其值为一。

     每个q边形出现的概率是p^q，因此$E[X_{(x_1,x_2,\ldots,x_q)}] = p^q$

     由于期望的线性，图中三角形的总数的期望值为所有指示器随机变量的期望值之和。也就是说，

     $$E[N_q] = E[\sum_{(x_1,x_2,\ldots,x_q)} X_{(x_1,x_2,\ldots,x_q)}] = \sum_{(x_1,x_2,\ldots,x_q)} E[X_{(x_1,x_2,\ldots,x_q)}] = \binom{n}{q}p^q$$

  3. 设 $X = \alpha(G)$ 。要证明 $\mathbf{Pr}[X \geq 3\log_2 n + 1] \to 0$ 当 $n \to \infty$。

     由于 $p = 1/2$，并且在大小为 k 的集合中有 $\binom{k}{2}$ 对顶点，所以这些对中全都不相邻的概率为 $(1/2)^{\binom{k}{2}}$。

     然后需要找到给定的 k 个顶点形成独立集的概率的上界。

     从具有 n 个顶点的图中选择一组大小为 k 的顶点有 $\binom{n}{k}$ 种方法。因此，根据并集界，存在大小至少为 k 的独立集的概率最多为

     $$\binom{n}{k}(1/2)^{\binom{k}{2}}.$$

     将 $k = 3\log_2 n + 1$ 带入

     $$\mathbf{Pr}[X \geq 3\log_2 n + 1] \leq \binom{n}{3\log_2 n + 1}(1/2)^{\binom{3\log_2 n+1}{2}}.$$

     使用斯特林近似，

     $$\mathbf{Pr}[X \geq 3\log_2 n + 1] \leq \frac{n^{3\log_2 n+1}}{(3\log_2 n/e)^{3\log_2 n+1}}(1/2)^{\binom{3\log_2 n+1}{2}}.$$

     进一步简化

$$\mathbf{Pr}[X \geq 3\log_2 n + 1] \leq (en/3\log_2 n)^{3\log_2 n+1}(1/2)^{\binom{3\log_2 n+1}{2}}.$$

取当 $n \to \infty$ 时的极限

$$\lim_{n\to\infty}(en/3\log_2 n)^{3\log_2 n+1}(1/2)^{\binom{3\log_2 n+1}{2}} = 0.$$

即

$$\lim_{n\to\infty}\mathbf{Pr}[X \geq 3\log_2 n + 1] = 0,$$

在social networks种，这个结果代表随着网络中人数的增加，找到一个彼此不是朋友的大型群体变得越来越不可能，一大群里大概率有人相互认识。

- [**Random social networks**] Let $G = (V, E)$ be a **fixed** undirected graph without isolating vertex. Let d_v be the degree of vertex v. Let Y be a uniformly chosen vertex, and Z a uniformly chosen neighbor of Y.

  1. Show that $\mathbf{E}[d_Z] \geq \mathbf{E}[d_Y]$.
  2. Interpret this inequality in the context of social networks, in which the vertices represent people, and the edges represent friendship.

  解答：

  1. 首先求d_Y的期望值。

  $$\mathbf{E}[d_Y] = \frac{1}{n}\sum_{v\in V} d_v,$$

  其中 $n = |V|$ 是图中顶点的数量。

  然后求 d_Z 的期望值。由于 Z 是 Y 的一个均匀选择的邻居，

  $$\mathbf{E}[d_Z] = \sum_{u\in V}\mathbf{Pr}[Z = u]d_u.$$

  要求 $Z = u$ 的概率，就需要求 Y 是 u 的邻居的概率。

  $$\mathbf{Pr}[Z = u] = \frac{u\text{的邻居数量}}{n} = \frac{d_u}{n}.$$

  代入 $\mathbf{E}[d_Z]$ 的表达式中，得到

  $$\mathbf{E}[d_Z] = \sum_{u\in V}\frac{d_u}{n}d_u = \frac{1}{n}\sum_{u\in V} d_u^2.$$

  比较

  $$\mathbf{E}[d_Z] - \mathbf{E}[d_Y] = \frac{\sum_{u\in V} d_u^2}{n} - \frac{\sum_{u\in V} d_u}{n}.$$

  要证明这个量是非负的，只需证明

  $$\sum_{u\in V} d_u^2 \geq \left(\sum_{u\in V} d_u\right)^2.$$

  这个不等式可以由柯西-施瓦茨不等式得出。因此，得到

  $$\mathbf{E}[d_Z] \geq \mathbf{E}[d_Y],$$

  2. 在social networks中，往往“你的朋友比你拥有更多的朋友”。这个现象被称为“友谊悖论”。

- [**Turán's Theorem**] Let $G = (V, E)$ be a **fixed** undirected graph, and write d_v for the degree of the vertex v. Use probablistic method to prove that $\alpha(G) \geq \sum_{v\in V}\frac{1}{d_v+1}$1. (Hint: Consider the following random procedure for generating an independent set I from a graph with vertex set V: First, generate a random permutation of the vertices, denoted as $v_1, v_2, \ldots, v_n$. Then, construct the independent set I as follows: For each vertex $v_i \in V$, add v_i to I if and only if none of its predecessors in the permutation, i.e., $v_1, \ldots, v_{i-1}$, are neighbors of v_i.)

解答：

考虑以下随机过程：

首先，生成顶点的随机排列，记为 $v_1, v_2, \ldots, v_n$。

然后，构造独立集 I：对于每个顶点 $v_i \in V$，当且仅当其在排列中的前面的点，即 $v_1, \ldots, v_{i-1}$，都不是 v_i 的邻居时，将 v_i 添加到 I 中。

对于顶点 v 被添加到独立集中，它的邻居都不能在排列中出现在它之前。由于在排列中有 $d_v + 1$ 个可能的位置用于顶点 v（包括它自己），并且只有其中一个位置允许顶点 v 被添加到独立集中（即当它在其邻居中首先出现时），我们有

$$\mathbf{Pr}[v \text{ 被添加到 } I] = \frac{1}{d_v + 1}.$$

设 X_v 为顶点 v 被添加到独立集 I 的事件的指示随机变量，当且仅当顶点v 被添加到独立集 I 时其值取1

$$\mathbf{E}[X_v] = \mathbf{Pr}[v \text{ 被添加到 } I] = \frac{1}{d_v + 1}.$$

设 $X = \sum_{v \in V} X_v$ 为此随机过程生成的独立集的大小

$$\mathbf{E}[X] = \sum_{v \in V} \mathbf{E}[X_v] = \sum_{v \in V} \frac{1}{d_v + 1}.$$

由于这对于任何固定的无向图 $G = (V, E)$ 都成立，因此存在一个大小至少为

$$\sum_{v \in V} \frac{1}{d_v + 1}.$$

的独立集。

综上得证：

$$\alpha(G) \geq \sum_{v \in V} \frac{1}{d_v + 1},$$

Prmblem 6 (1D random walk)

Let $p \in (0, 1)$ be a constant, and$\{X_n\}_{n \geq 1}$ be independent Bernoulli trials with successful probability p. Define $S_n = 2 \sum_{i=1}^{n} X_i - n$ and $S_0 = 0$.

设 $p \in (0, 1)$ 为一个常数，$\{X_n\}_{n \geq 1}$ 为成功概率为 p 的独立伯努利试验。定义 $S_n = 2 \sum_{i=1}^{n} X_i - n$ 和 $S_0 = 0$。序列 $S_0, S_1, \ldots, S_n$ 的范围 R_n 定义为序列所取的不同值的数量。

- [**Range of random walk**] The range $R_n$of $S_0, S_1, \ldots, S_n$ is defined as the number of distinct values taken by the sequence. Show that $\mathbf{Pr}(R_n = R_{n-1} + 1) = \mathbf{Pr}(\forall 1 \leq i \leq n, S_i \neq 0)$ as $n \to \infty$, and deduce that $n^{-1}\mathbf{E}[R_n] \to \mathbf{Pr}(\forall i \geq 1, S_i \neq 0)$. Hence show that $n^{-1}\mathbf{E}[R_n] \to |2p - 1|$ as $n \to \infty$.

  解答：

  1. 首先， $R_n = R_{n-1} + 1$ 当且仅当 $S_n \neq S_k (k = 1, 2, \ldots, n-1)$

     即 $2\sum_{i=1}^{n} X_i - n \neq 2\sum_{j=1}^{k} X_j - k$

     可得$2\sum_{i=1+k}^{n} X_i \neq n - k$

     令$x = n - k$

     对于$S_x \neq 0$这件事来说，代表着$2\sum_{i=1}^{x} X_i \neq x$

即

$$2\sum_{i=1}^{n-k} X_i = 2\sum_{i=1+k}^{n} X_i \neq n-k$$

所以，对于任意k和任意x来说，$S_n \neq S_k$和$S_x \neq 0$的本质是一样的，所以发生概率是一样的。

因而可以知道$\mathbf{Pr}\left(R_n = R_{n-1}+1\right) = \mathbf{Pr}\left(\forall 1 \leq i \leq n, S_i \neq 0\right)$

2. $$\mathbf{E}\left(R_n\right) = \sum_{i=1}^{n} \Pr(R_i = R_{i-1}+1)$$

由于 $\Pr(R_n = R_{n-1}+1) = \Pr(\forall 1 \leq i \leq n, S_i \leq 0)$，($n \to \infty$)

所以

$$n^{-1}\mathbf{E}\left(R_n\right) = \frac{\sum_{i=1}^{n} \Pr(R_i = R_{i-1}+1)}{n} = \frac{\sum_{j=1}^{n} \Pr(\forall 1 \leq i \leq j, S_i \leq 0)}{n} \to \Pr(\forall i \geq 1, S_i \neq 0)$$

当 $n \to \infty$

3. 因为当走了奇数步时不可能回到原点，所以可得

$$\Pr(S_{2n} = 0) = \binom{2n}{n} p^n (1-p)^n$$
$$\Pr(S_{2n-1} = 0) = 0$$

因此,

$$\Pr(\forall i \geq 1, S_i \neq 0) = \prod_{i=1}^{\infty} \left[1 - \binom{2i}{i} p^i (1-p)^i\right].$$

接下来证明$\prod_{i=1}^{\infty} \left[1 - \binom{2i}{i} p^i (1-p)^i\right] = |2p-1|$

当$p = 0$或$p = 1$，等式显然成立。

所以只需要证明$0 < p < 1$的情况。

使用斯特林公式近似可得

$$\binom{2i}{i} = \frac{(2i)!}{i!^2} \frac{\sqrt{4\pi i}(2i/e)^{2i}}{2\pi i (i/e)^{2i}} = \frac{2^{2i}}{\sqrt{\pi i}}$$

所以

$$\begin{aligned}
\prod_{i=1}^{\infty} \left[1 - \binom{2i}{i} p^i (1-p)^i\right] &\approx \prod_{i=1}^{\infty} \left[1 - \frac{2^{2i}}{\sqrt{\pi i}} p^i (1-p)^i\right] \\
&= \prod_{i=1}^{\infty} \frac{\sqrt{\pi i} - 2^{2i} p^i (1-p)^i}{\sqrt{\pi i}} \\
&= \frac{1}{\sqrt{\pi}} \prod_{i=1}^{\infty} \frac{\sqrt{i} - 2p\sqrt{1-p}}{\sqrt{i}} \\
&= \frac{1}{\sqrt{\pi}} \frac{\sin(\pi)(2p-1)}{2(2p-1)} \\
&= |2p-1|
\end{aligned}$$

因为 $n^{-1}\mathbf{E}[R_n] \to \mathbf{Pr}(\forall i \geq 1, S_i \neq 0)$，所以

$$n^{-1}\mathbf{E}[R_n] = \mathbf{Pr}(\forall i \geq 1, S_i \neq 0) = |2p-1| (\text{当 } n \to \infty)$$

得证

- [**Symmetric 1D random walk (IV)**] Suppose $p = \frac{1}{2}$. Let N_n be the number of points that have been visited by S exactly once up to n, that is the size of set $\{0 \leq i \leq n \mid \forall 0 \leq j \leq n \text{ and } j \neq i, S_i \neq S_j\}$. Prove that $\mathbf{E}[N_n] = 2$ for all $n \geq 1$.

解答：

数学归纳法：

首先，当 $n=1$ 时，$N_1=2$，因为 $S_0=0$ 和 $S_1=\pm 1$ 都只被访问了一次。

现在假设对于某个 $k\geq 1$，$\mathbf{E}[N_k]=2$。我们来证明 $\mathbf{E}[N_{k+1}]=2$。

注意到 $S_{k+1}=S_k\pm 1$，所以有两种情况：

- 如果 S_k 在之前只被访问过一次，那么 $N_{k+1}=N_k+1$。
- 如果 S_k 在之前被访问过两次或更多次，那么 $N_{k+1}=N_k-1$。

由于 X_{k+1} 是一个伯努利试验，所以这两种情况的概率都是 $\frac{1}{2}$。因此，

$$\begin{aligned}\mathbf{E}[N_{k+1}]&=\frac{1}{2}(\mathbf{E}[N_k+1]+\mathbf{E}[N_k-1])\\&=\frac{1}{2}(2+2)\\&=2\end{aligned}$$

所以，根据数学归纳法，对于所有的 $n\geq 1$，$\mathbf{E}[N_n]=2$。

- [**Symmetric 1D random walk (V)**] Suppose $p=\frac{1}{2}$. Prove that $\mathbf{E}[|S_n|]=\Theta(\sqrt{n})$.

解答

根据随机游走的性质，可知

$\mathbf{E}[S_n]=0$ 并且$\mathrm{Var}(S_n)=n$

$|S_n|$ 是一个非负随机变量，所以可以用切比雪夫不等式来估计它的期望值。

对于任意正实数 t

$$\mathbf{P}(|S_n|\geq t)\leq\frac{\mathbf{E}[S_n^2]}{t^2}=\frac{n}{t^2}$$

注意到当 $t\geq\sqrt{n}$ 时，$\frac{n}{t^2}\leq 1$。因此，

$$\mathbf{E}[|S_n|]=\int_0^\infty\mathbf{P}(|S_n|\geq t)dt=\int_0^{\sqrt{n}}\mathbf{P}(|S_n|\geq t)dt+\int_{\sqrt{n}}^\infty\mathbf{P}(|S_n|\geq t)dt$$

第二项使用切比雪夫不等式来估计：

$$\int_{\sqrt{n}}^\infty\mathbf{P}(|S_n|\geq t)dt\leq\int_{\sqrt{n}}^\infty\frac{n}{t^2}dt=n[-t^{-1}]_{\sqrt{n}}^\infty=n[0-(-\sqrt{n})]=n\sqrt{n}$$

第一项直接估计：

$$\int_0^{\sqrt{n}}\mathbf{P}(|S_n|\geq t)dt\leq\int_0^{\sqrt{n}}1dt=[t]_0^{\sqrt{n}}=[\sqrt{n}-0]=\sqrt{n}$$

因此，

$$\mathbf{E}[|S_n|]=\Theta(\sqrt{n})$$

另一方面，由于 $|S_n|$ 是一个非负随机变量，我们有

$$\mathbf{E}[|S_n|]=\int_0^\infty\mathbf{P}(|S_n|>t)dt$$

注意到当 $t<c\sqrt{n}$ 时，$\frac{n}{t^2}>c^{-2}$。因此，

$$\mathbf{E}[|S_n|]=\Theta(\sqrt{n})+c^{-2}\int_{c\sqrt{n}}^\infty dt=\Theta(\sqrt{n})+c^{-2}[t]_{c\sqrt{n}}^\infty=\Theta(\sqrt{n})+c^{-2}[\infty-c\sqrt{n}]=\Theta(\sqrt{n})+c^{-1}\sqrt{n}$$

由于上式对于任意正实数 c 都成立，所以我们可以取极限得到

$$\mathbf{E}[|S_n|] = \Theta(\sqrt{n})$$

致谢

和我一起讨论的陈子元、胡嘉欣、王祉天和徐研同学。