4.1. Random variables.

Definition 4.1. A random variable X is an extended real-valued (measurable) function defined on S or equivalently $\{s \in S | X(s) \le a\}$ is an event (measurable set) for all $a \in \mathbb{R}$. For simplicity, write $\{X \in A\}$ for $\{s \in S | X(s) \in A\}$.

Example 4.1 (Coupon collecting problem). Suppose that N balls indexed by $1, ..., N$ are addressed in an urn. Each time, a ball is withdrawn, its number is recorded and then the ball is replaced. Assume that balls are withdrawn uniformly and independently. What is the probability that all balls are withdrawn at least once in n withdrawals?

For convenience, let T denote the first time that all balls are withdrawn at least once. Then, $\{T \le n\}$ is the desired event , while

$$\{T > n\} = \bigcup_{j=1}^{N} A_j,$$

where A_j denotes the event that ball j is never selected in the first n draws. Immediately, one has

$$P(T > n) = \sum_{k=1}^{N} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le N} P(A_{i_1} \cdots A_{i_k}) = \sum_{k=1}^{N} (-1)^{k+1} \binom{N}{k} \left(\frac{N-k}{N} \right)^n .$$

Remark 4.1. Here, we briefly write $P(X \in A)$ for $P(\{X \in A\})$.

Example 4.2. Consider the experiment of flipping a coin infinitely. The sample space consists of infinite sequences of letters H (head) and T (tail). Let X be the time when the first head appears. For instance, $X(TTHTTH \cdots) = 3$ and $X(TTTTHHTT \cdots) = 5$. Then, for $k \ge 1$,

$$\{X = k\} = \{x_1 x_2 ... | x_i = T, \forall 1 \le i < k, x_k = H\}, \quad \{X = \infty\} = \{TTTT \cdots\}.$$

If the heads come up with probability p and P is the probability on S, then

$$P(X = k) = P(\{(x_1, x_2, ...) | x_i = T, \forall 1 \le i < k, x_k = H\}) = p(1-p)^{k-1} \quad \forall k \ge 1.$$

Note that $\sum_{k=1}^{\infty} P(X = k) = 1$, which implies $P(X = \infty) = 1 - P(X < \infty) = 0$.

Remark 4.2. As S is uncountable, how can one define P?

4.2. Discrete random variables.

Definition 4.2. A random variable X is **discrete** if X takes values on a finite or countable set, i.e. $X(S)$ is finite or countable.

Definition 4.3. Let X be a discrete random variable.

(1) The **probability mass function** (or briefly **p.m.f.**) of X is defined to be

$$p(a) = P(X = a) \quad \forall a \in \mathbb{R}.$$

(2) The **(cumulative) distribution function** (or briefly **c.d.f.**) is defined to be

$$F(a) = P(X \le a) \quad \forall a \in \mathbb{R}.$$

Remark 4.3. If $X(S) = \{x_1, x_2, ...\}$ and p and F are p.m.f. and c.d.f. of X, then

$$p(a) = 0, \quad \forall a \notin \{x_1, x_2, ...\}, \quad p(x_i) = P(X = x_i), \quad F(a) = \sum_{i: x_i \le a} p(x_i).$$

Example 4.3. Let $S = \{a, b, c, d\}$ and P be a probability on S defined by

$$P(a) = P(b) = 1/8, \quad P(c) = 1/4, \quad P(d) = 1/2.$$

Consider the random variable X defined by

$$X(a) = 1, \quad X(c) = 2, \quad X(d) = 3, \quad X(b) = 5.$$

If p, F is the p.m.f. and c.d.f. of X, then

$$p(1) = p(5) = 1/8, \quad p(2) = 1/4, \quad p(3) = 1/2$$

and

$$F(a) = \begin{cases} 0 & \text{if } a < 1 \\ 1/8 & \text{if } 1 \le a < 2 \\ 3/8 & \text{if } 2 \le a < 3 \\ 7/8 & \text{if } 3 \le a < 5 \\ 1 & \text{if } a \ge 5 \end{cases}.$$

Lemma 4.1. *Let X be a discrete random variable on S with $X(S) \subset \mathbb{R}$. Then, the distribution function F of X satisfies*

(1) $F(a) \to 1$ *as* $a \to \infty$.
(2) $F(a) \to 0$ *as* $a \to -\infty$.
(3) F *is non-decreasing and right-continuous.*

Proof. For $a \le b$, it is clear that $\{X \le a\} \subset \{X \le b\}$. This implies

$$F(a) = P(X \le a) \le P(X \le b) = F(b).$$

That is, F is non-decreasing. For the right-continuity of F, let $a \in \mathbb{R}$ and $n > 0$. It is an easy exercise to show that $\{X \le a\} = \bigcap_{n=1}^{\infty} \{X \le a + 1/n\}$. This implies

$$F(a) = P(X \le a) = \lim_{n\to\infty} P(X \le a + 1/n) = \lim_{n\to\infty} F(a + 1/n).$$

Since F is non-decreasing, the above identity is equivalent to

$$\lim_{h>0, h\to 0} F(a + h) = F(a).$$

For (1), since X is real-valued,

$$\bigcup_{n=1}^{\infty} \{X \le n\} = S, \quad \bigcap_{n=1}^{\infty} \{X \le -n\} = \emptyset.$$

Again, by the continuity of the probability P,

$$1 = P(S) = \lim_{n\to\infty} P(X \le n) = \lim_{n\to\infty} F(n), \quad 0 = P(\emptyset) = \lim_{n\to\infty} P(X \le -n) = \lim_{n\to-\infty} F(n)$$

or equivalently,

$$\lim_{a\to\infty} F(a) = 1, \quad \lim_{a\to-\infty} F(a) = 0.$$

□

4.3. **Expectation and variance.**

Definition 4.4. Let X be a discrete random variable with probability mass function p. The **expectation** (or **expected value** or **mean**) of X is defined by

$$E(X) = \sum_{x:p(x)>0} xp(x).$$

If the right side in the above equation converges absolutely, i.e. $\sum_{x:p(x)>0} |x|p(x) < \infty$, then $E(X)$ is well-defined. Otherwise, we say that $E(X)$ does not exist.

Remark 4.4. Note that the expectation of a discrete random variable depends only on its probability mass function or, more generally, on its cumulative distribution function.

Remark 4.5. For the case that X is non-negative but $\sum_{x:p(x)>0} xp(x)$ diverges, we write $E(X) = \infty$.

Example 4.4. Let F be an event and X be a random variable defined by

$$X(s) = \begin{cases} 1 & \text{if } s \in F \\ 0 & \text{if } s \in F^c \end{cases}.$$

Then, the probability mass function p of X is given by $p(1) = P(F)$, $p(0) = P(F^c)$ and

$$E(X) = 1 \cdot p(1) + 0 \cdot p(0) = P(F).$$

Example 4.5. Recall the experiment of flipping a coin independently for infinitely many times. Suppose that the coin lands on heads with probability $p \in (0,1)$ and lands on tails with probability $1 - p$. Let X be the random variable of the time when the first head appears. Since $P(TTTT\cdots) = 0$, we may reset the sample space to be

$$S = \{x_1x_2...|x_i \in \{H,T\}\} \setminus \{TTTT\cdots\}$$

and redefine P accordingly. Then, X turns out a discrete random variable with probability mass function

$$p(k) = P(X = k) = p(1-p)^{k-1} \quad \forall k \geq 1.$$

This implies

$$E(X) = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = -p\sum_{k=1}^{\infty} \frac{d(1-p)^k}{dp} = -p \cdot \frac{d}{dp}\left(\sum_{k=1}^{\infty}(1-p)^k\right) = \frac{1}{p}.$$

Proposition 4.2. *Let X be a real-valued discrete random variable with probability mass function p and g be a real-valued function defined on $\mathbb{R}$. Then, $g(X)$ is a discrete random variable and*

$$E(g(X)) = \sum_{x:p(x)>0} g(x)p(x), \tag{4.1}$$

provided the sum in the right side converges absolutely.

Proof. Set $X(S) = \{x_i | i = 1, 2, ...\}$ and $Y = g(X)$. Since $Y(S) = g(X(S)) = \{g(x_i)|i = 1, 2, ...\}$ is a countable set, Y is a discrete random variable. To see the expectation of Y, assume that the right side of (4.1) exists. Let $Y(S) = \{y_i|i = 1, 2, ...\}$ and p_Y be the probability mass function of Y. Then, $\{Y = y\} = \bigcup_{i:g(x_i)=y}\{X = x_i\}$, where the right side is a disjoint union, and

$$p_Y(y) = P(Y = y) = \sum_{i:g(x_i)=y} p(x_i).$$

As a result, we obtain

$$EY = \sum_{j=1}^{\infty} y_j p_Y(y_j) = \sum_{j=1}^{\infty} \sum_{i:g(x_i)=y_j} y_j p(x_i) = \sum_{j=1}^{\infty} \sum_{i:g(x_i)=y_j} g(x_i)p(x_i) = \sum_{i=1}^{\infty} g(x_i)p(x_i)$$

□

Example 4.6. Recall the random variable X of the time when the first head appears in the experiment of independent flips of coins. It has been shown that the probability mass function p of X is given by $p(k) = p(1-p)^{k-1}$ for $k \ge 1$ and $E(X) = 1/p$. Observe that

$$\begin{aligned} E(X(X+1)) &= \sum_{k=0}^{\infty} k(k+1)p(1-p)^{k-1} = p \sum_{k=0}^{\infty} \frac{d^2(1-p)^{k+1}}{dp^2} \\ &= p \cdot \frac{d^2}{dp^2} \left(\sum_{k=0}^{\infty} (1-p)^{k+1} \right) = \frac{2}{p^2}. \end{aligned}$$

Consequently, we obtain

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

Definition 4.5. Let X be a discrete random variable of which expectation exists. The **variance** of X is defined to be $\mathrm{Var}(X) = E[(X - E(X))^2]$ and the **standard deviation** of X is defined to be $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$.

Theorem 4.3. *Let X be a discrete random variable and a, b be real numbers. Assume that $E(X)$ exists.*

(1) $E(aX + b) = aE(X) + b$.
(2) $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
(3) $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$.

In particular, $E(X^2) \ge [E(X)]^2$.

Proof. We first prove (3). Set $\mu = E(X)$. By the definition of variance,

$$\begin{aligned} \mathrm{Var}(X) &= \sum_{x:p(x)>0} (x-\mu)^2 p(x) = \sum_{x:p(x)>0} x^2 p(x) - 2\mu \sum_{x:p(x)>0} xp(x) + \mu^2 \sum_{x:p(x)>0} p(x) \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2. \end{aligned}$$

(1) can be proved in a similar way, where the details are omitted. For (2), one may use (1) to derive

$$[E(aX+b)]^2 = a^2[E(X)]^2 + 2abE(X) + b^2$$

and

$$E[(aX+b)^2] = E(a^2X^2 + 2abX + b^2) = a^2E(X^2) + 2abEX + b^2.$$

The desired identity follows immediately from (3). □

Example 4.7. Recall the random variable X of the time when the first head appears in the experiment of independent flips of coins. We have shown that $E(X) = 1/p$ and $E(X^2) = 2/p^2 - 1/p$. Then, the variance is given by $\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = 1/p^2 - 1/p$.

4.4. **The Bernoulli and binomial random variables.**

Definition 4.6. Let $p \in [0,1]$ and n be a positive integer. A random variable X is

(1) a **Bernoulli random variable** with success probability p (or briefly $X \sim$Bernoulli(p)) if X take values on $\{0,1\}$ and $P(X=1)=p$.
(2) a **binomial random variable** with parameter (n,p) (or briefly $X \sim B(n,p)$) if X takes values on $\{0,1,...,n\}$ and $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$.

Example 4.8. Let $X_1,...,X_n$ be independent Bernoulli(p) random variables and $X = X_1 + \cdots + X_n$. Then, X is a binomial random variable with parameter (n,p). Given $1 \le k \le n$ and $1 \le i_1 < \cdots < i_k \le n$, let $\{j_1,...,j_{n-k}\} = \{1,...,n\} \setminus \{i_1,...,i_k\}$. Note that

$$P(X_{i_1} = \cdots = X_{i_k} = 1,\, X_{j_1} = \cdots = X_{j_{n-k}} = 0) = p^k(1-p)^{n-k}.$$

This implies that

$$\begin{aligned} P(X=k) &= \sum_{1\le i_1<\cdots<i_k\le n} P(X_{i_1} = \cdots = X_{i_k} = 1,\, X_{j_1} = \cdots = X_{j_{n-k}} = 0) \\ &= \sum_{1\le i_1<\cdots<i_k\le n} p^k(1-p)^{n-k} = \binom{n}{k} p^k(1-p)^{n-k}. \end{aligned}$$

One may conclude from the above discussion that a binomial random variable with parameter (n,p) has the same distribution function as the number of successes in a list of n independent Bernoulli random variables of which success probabilities are p.

Remark 4.6. Given a binomial random variable X with parameters (n,p), what is the maximum of its probability mass function p? Note that, for $k \ge 0$,

$$\frac{p(k)}{p(k+1)} = \frac{\binom{n}{k}p^k(1-p)^{n-k}}{\binom{n}{k+1}p^{k+1}(1-p)^{n-k-1}} = \frac{(k+1)(1-p)}{(n-k)p}.$$

This implies $p(k) < p(k+1)$ if and only if $(k+1) \le (n+1)p$. Thus, p achieves its maximum at $\lfloor (n+1)p \rfloor$.

Example 4.9. A communication system consists of n components, where each component functions with probability p independently. The system is effective if at least half of the components function. We first compare the efficiency of systems with $2n$ and $(2n+1)$ components. Let $E_{n,p}$ be the probability that a system of n components is effective. Note that

$$E_{2n}(p) = \sum_{k=n}^{2n} \binom{2n}{k} p^k(1-p)^{2n-k}, \quad E_{2n+1}(p) = \sum_{k=n+1}^{2n+1} \binom{2n+1}{k} p^k(1-p)^{2n+1-k}.$$

Using the identity $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$, we may write

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

and similarly

$$\begin{aligned}
E_{2n+2}(p) &= \sum_{k=n+1}^{2n+1} \binom{2n+1}{k} p^k (1-p)^{2n+2-k} + \sum_{k=n+1}^{2n+2} \binom{2n+1}{k-1} p^k (1-p)^{2n+2-k} \\
&= \sum_{k=n+1}^{2n+1} \binom{2n+1}{k} p^k (1-p)^{2n+2-k} + \sum_{k=n}^{2n+1} \binom{2n+1}{k} p^{k+1} (1-p)^{2n+1-k} \\
&= E_{2n+1}(p) + \binom{2n+1}{n} p^{n+1}(1-p)^{n+1}.
\end{aligned}$$

Consequently, for any two systems with consecutive numbers of components, the one with even number of component works better. For systems with odd numbers of components, the above identities give

$$E_{2n+1}(p) - E_{2n-1}(p) = -\binom{2n}{n} p^n (1-p)^{n+1} + \binom{2n-1}{n-1} p^n (1-p)^n = \binom{2n-1}{n-1} p^n (1-p)^n (2p-1).$$

Obviously, a system of $2n+1$ components works better than a system of $2n-1$ components if and only if $p > 1/2$. Similarly, for systems with even number of components, one has

$$E_{2n+2} - E_{2n} = \frac{(2n+1)p - (n+1)}{n+1} \binom{2n}{n} p^n (1-p)^{n+1}.$$

This implies that a system with $2n+2$ components works better than a system with $2n$ components if and only if $p > (n+1)/(2n+1)$, where $(n+1)/(2n+1) > 1/2$.

Theorem 4.4 (Expectations of sums of random variables)**.** *Let $X_1, ..., X_n$ be discrete random variables and $X = X_1 + \cdots + X_n$. Then, $E(X) = E(X_1) + \cdots + E(X_n)$.*

Proof. It suffices to prove this theorem for $n = 2$. In this case, we let $\{x_i\}$ and $\{y_i\}$ be the values X_1 takes and the values X_2 takes. Clearly, $X_1 + X_2$ takes values on $\{x_i + y_j\}$ and this implies that $X_1 + X_2$ is a discrete random variable. Let p be the p.m.f. of $X_1 + X_2$ and write $\{z_k\} = \{x_i + y_j\}$. Then, for $k \geq 1$,

$$p(z_k) = P(X_1 + X_2 = z_k) = \sum_{(i,j):x_i+y_j=z_k} P(X_1 = x_i, X_2 = y_j).$$

As a result, we have

$$\begin{aligned}
E(X_1 + X_2) &= \sum_k z_k p(z_k) = \sum_k z_k \sum_{(i,j):x_i+y_j=z_k} P(X_1 = x_i, X_2 = y_j) \\
&= \sum_k \sum_{(i,j):x_i+y_j=z_k} (x_i + y_j) P(X_1 = x_i, X_2 = y_j) \\
&= \sum_{i,j} (x_i + y_j) P(X_1 = x_i, X_2 = y_j) \\
&= \sum_{i,j} x_i P(X_1 = x_i, X_2 = y_j) + \sum_{i,j} y_j P(X_1 = x_i, X_2 = y_j) \\
&= \sum_i x_i P(X_1 = x_i) + \sum_j y_j P(X_2 = y_j) = E(X_1) + E(X_2).
\end{aligned}$$

□

Proposition 4.5. *Let X be a binomial random variable with parameters (n, p). Then, $E(X) = np$ and $\mathrm{Var}(X) = np(1-p)$.*

Proof. Note that X and $Y = X_1 + \cdots + X_n$, where $X_1, ..., X_n$ are independent Bernoulli(p) random variables, have the same distribution function and hence the same expectation and variance. By the linearity of expectation, one has

$$E(Y) = \sum_{k=1}^{n} E(X_k) = \sum_{k=1}^{n} P(X_k = 1) = np.$$

To see the variance, note that for $i \neq j$,

$$E(X_i X_j) = P(X_i = 1, X_j = 1) = P(X_i = 1)P(X_j = 1) = p^2,$$

and $E(X_i^2) = E(X_i) = p$. This implies

$$E(Y^2) = E\left(\sum_{i,j=1}^{n} X_i X_j\right) = \sum_{i,j=1}^{n} E(X_i X_j) = \sum_{1 \le i \neq j \le n} p^2 + \sum_{1 \le i = j \le n} p = n(n-1)p^2 + np.$$

Hence, we have

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

□

4.5. **Poisson random variables.**

Definition 4.7. A Poisson random variable X with parameter $\lambda > 0$ is a discrete random variable taking values on $\{0, 1, 2, ...\}$ satisfying

$$P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}, \quad \forall i \ge 0.$$

Example 4.10. Fix $\lambda > 0$. For $n \ge 1$, let X_n be a binomial random variable with parameter (n, p_n). Note that, for $0 \le i \le n$,

$$\begin{aligned} P(X_n = i) &= \binom{n}{i} p_n^i (1-p_n)^{n-i} = \frac{n(n-1)\cdots(n-i+1)}{i!} \times p_n^i (1-p_n)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \times \frac{(np_n)^i}{i!} \times \left(1 - \frac{np_n}{n}\right)^n \left(1 - \frac{np_n}{n}\right)^{-i}. \end{aligned}$$

This implies that, when $np_n \to \lambda$,

$$\lim_{n \to \infty} P(X_n = i) = e^{-\lambda}\frac{\lambda^i}{i!}.$$

Remark 4.7. Many practical examples are related to Poisson random variables. For instance, the number of misprints on a page, the number of wrong phone calls made in a day, the number of customers entering the postoffice on a given day, the number of users connecting to a sever and more.

Proposition 4.6. *Let X be a Poisson random variable with parameter λ. Then, $E(X) = \mathrm{Var}(X) = \lambda$.*

Proof. First, one may compute

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

and

$$E(X(X-1)) = \sum_{i=0}^{\infty} i(i-1) \cdot e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\lambda^2 \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} = \lambda^2.$$

The latter implies $\mathrm{Var}(X) = E(X(X-1)) + E(X) - [E(X)]^2 = \lambda$. □

Example 4.11. Consider a group of n persons whose birthdays are independent and equally likely any day of the 365 days in a year. Let E be the event that no two persons have the same birthday. By the independence, we have

$$P(E) = \frac{\binom{365}{n} n!}{365^n} = \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}.$$

Practically, the computation of the above ratio is not clear. We consider another approach using the Poisson approximation of binomial distributions. First, for $i \neq j$, let $E_{i,j}$ be the event that the ith and the jth persons have the same birthday. Let $X_{i,j}$ be the random variable taking value on 1 if $E_{i,j}$ happens and on 0 otherwise. Obviously, $P(X_{i,j} = 1) = 1/365$ and

$$E = \left\{ \sum_{(i,j):i<j} X_{i,j} = 0 \right\}.$$

Each $X_{i,j}$ is a Bernoulli random variable with parameter $1/365$ and there are $\binom{n}{2}$ random variables. By the Poisson approximation with $\lambda = \binom{n}{2}/365$, we have

$$P(E) = P\left(\sum_{(i,j):i<j} X_{i,j} = 0 \right) \approx e^{-\lambda} = \exp\left\{ -\frac{n(n-1)}{730} \right\}.$$

Concerning the smallest n such that $P(E) \leq 1/2$, one may try the following computation,

$$\exp\left\{ -\frac{n(n-1)}{730} \right\} \leq \frac{1}{2} \quad \Leftrightarrow \quad n(n-1) \geq 730 \ln 2 \approx 505,$$

which implies $n \geq 23$. In fact, this conclusion is correct when one computes $P(E) \leq 1/2$ precisely.

Remark 4.8. Note that, in the previous example, those events, $E_{i,j}$ with $1 \leq i < j \leq n$, are pairwise independent but in fact not independent. For instance, $P(E_{1,2}) = 1/365$ but $P(E_{1,2}|E_{1,3}E_{2,3}) = 1$. The reason why one may still apply the Poisson approximation due to the fact that if any selected pairs (i,j) have no common entries i, j, then they are independent. This means that the dependence among $E_{i,j}$ with $1 \leq i < j \leq n$ is weak and then the Poisson paradigm applies.

Poisson paradigm Let $E_1, ..., E_n$ be events with probabilities $p_1, ..., p_n$. Assume that all the p_i are "small" and the events are independent or "weakly dependent". Then, the probability that exactly i events of $E_1, ..., E_n$ occurs is "approximately" $P(X = i)$, where X is a Poisson random variable with parameter $\lambda = p_1 + \cdots + p_n$.

4.6. **Other discrete random variables.**

Definition 4.8. A random variable X taking values on $\{1, 2, ...\}$ is called a **geometric random variable** with success of probability p if

$$P(X = n) = p(1-p)^{n-1}, \quad \forall n \geq 1.$$

Remark 4.9. The geometric random variable can be interpreted as the time of the first success in an infinite sequence of independent Bernoulli(p) trials.

Proposition 4.7. *Let X be a geometric random variable with success probability $p \in (0,1)$. Then, X is memoryless, that is, $P(X > n+m|X > m) = P(X > n)$. Conversely, if a discrete random variable Y taking values on $\{1, 2, ...\}$ possesses the memoryless property, then Y is geometric with success probability $P(Y = 1)$.*

Proof. For the first part, it is equivalent to show that $P(X > n+m) = P(X > n)P(X > m)$. This is clear since one can compute

$$P(X > n) = \sum_{k=n+1}^{\infty} P(X = k) = \sum_{k=n+1}^{\infty} p(1-p)^{k-1} = (1-p)^n.$$

For the second part, set $f(n) = P(Y > n)$. Then, the memoryless property says $f(n+m) = f(n)f(m)$. This implies

$$f(n) = [f(1)]^n = [P(Y > 1)]^n = [1 - P(Y = 1)]^n.$$

Thus, for $k \geq 1$,

$$P(Y = k) = P(Y > k-1) - P(Y > k) = f(k-1) - f(k) = P(Y = 1)[1 - P(Y = 1)]^{k-1}.$$

□

Proposition 4.8. *Let X be a geometric random variable with success probability $p \in (0,1)$. Then, $E(X) = 1/p$ and* $\mathrm{Var}(X) = 1/p^2 - 1/p$.

See Examples 4.5 and 4.6 for a proof of the above proposition.

Definition 4.9. Let r be a positive integer and $p \in (0,1)$. A random variable X taking values on $\{r, r+1, ...\}$ is called a **negative binomial random variable** with parameters (r,p) if

$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \quad \forall n \geq r.$$

Example 4.12. Consider a sequence of independent Bernoulli(p) trials, $X_1, X_2, ...$ Let r be a positive integer and X be the random variable denoting the time when the rth success happens. Then,

$$X = \inf\{n \geq 1 | X_1 + X_2 + \cdots + X_n = r\}.$$

As $P(X = \infty) = 0$ (Why?), one may suitably reform the sample space S so that X is real-valued and thus $X(S) = \{r, r+1, r+2, ...\}$. Note that, for $n \geq r$,

$$\{X = n\} = \{X_n = 1\} \cap \{\text{there are exactly } r-1 \text{ successes in the first } n-1 \text{ trials}\}.$$

By letting $1 \leq i_1 < i_2 < \cdots < i_{r-1} \leq n-1$ and setting $\{j_1, j_2, ..., j_{n-r}\} = \{1, ..., n\} \setminus \{i_1, ..., i_{r-1}, n\}$, we may rewrite

$$\{X = n\} = \bigcup_{1 \leq i_1 < \cdots < i_{r-1} \leq n-1} \{X_{i_1} = \cdots = X_{i_{r-1}} = X_n = 1, X_{j_1} = \cdots = X_{j_{n-r}} = 0\}.$$

This implies

$$P(X = n) = \sum_{1 \leq i_1 < \cdots < i_{r-1} \leq n-1} p^r (1-p)^{n-r} = \binom{n-1}{r-1} p^r (1-p)^{n-r},$$

which proves that X is a negative binomial random variable with parameters (r,p).

Remark 4.10. For the case of $r = 1$, the negative binomial random variable turns out a geometric random variable.

Remark 4.11. Let $r \geq 1$ and $X_1, ..., X_r$ be independent geometric random variables with parameter p. Define $X = X_1 + \cdots + X_r$. Since $P(X_i \geq 1) = 1$ for all $i = 1, 2, ..., r$,

$P(X \geq r) = 1$. For $n \geq r$,

$$\begin{aligned} P(X = n) &= \sum_{\substack{k_1 \geq 1, ..., k_r \geq 1: \\ k_1 + \cdots + k_r = n}} P(X_1 = k_1, ..., X_r = k_r) = \sum_{\substack{k_1 \geq 1, ..., k_r \geq 1: \\ k_1 + \cdots + k_r = n}} P(X_1 = k_1) \cdots P(X_r = k_r) \\ &= \sum_{\substack{k_1 \geq 1, ..., k_r \geq 1: \\ k_1 + \cdots + k_r = n}} p(1-p)^{k_1 - 1} \cdots p(1-p)^{k_r - 1} = \sum_{\substack{k_1 - 1 \geq 0, ..., k_r - 1 \geq 0: \\ (k_1 - 1) + \cdots + (k_r - 1) = n - r}} p^r (1-p)^{n-r} \\ &= \binom{(n-r) + (r-1)}{r-1} p^r (1-p)^{n-r} = \binom{n-1}{r-1} p^r (1-p)^{n-r}. \end{aligned}$$

This proves that X is negative binomial with parameters (r, p).

Proposition 4.9. *Let X be a negative binomial random variable with parameters (r, p). Then, $E(X) = r/p$ and $\mathrm{Var}(X) = r(1-p)/p^2$.*

Proof. Observe that if $X_1, ..., X_r$ be independent geometric random variables with parameter p, then $X = X_1 + \cdots + X_r$ is negative binomial with parameters (r, p). Recall that $E(X_i) = 1/p$ and $\mathrm{Var}(X_i) = (1-p)/p^2$. This implies $E(X_i^2) = (2-p)/p^2$. By the linearity of expectation, we have

$$E(X) = E(X_1 + \cdots + X_r) = E(X_1) + \cdots E(X_r) = r/p.$$

To see the variance, note that

$$E(X^2) = \sum_{i,j=1}^{r} E(X_i X_j).$$

For $i \neq j$, one has

$$E(X_i X_j) = \sum_{k,l=1}^{\infty} kl P(X_i = k, X_j = l) = \sum_{k,l=1}^{\infty} k P(X_i = k) \cdot l P(X_j = l) = E(X_i) E(X_j) = \frac{1}{p^2}.$$

Thus, $E(X^2) = (r^2 - r)/p^2 + r(2-p)/p^2$. As a result, we obtain

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

□

Definition 4.10. A random variable X is a **hypergeometric random variable** with parameter (n, N, m) if $X(S) = \{0, 1, ..., n\}$ and

$$P(X = i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}, \quad \forall 0 \leq i \leq n,$$

where $\binom{n}{k} := 0$ for $k > n$.

An interpretation on the hypergeometric random variable is as follows. Consider an urn with N balls, where m balls are black and $N - m$ balls are white. An experiment is performed by randomly selecting n balls from the urn. Let X denote the number of black balls selected. Then, X is a hypergeometric random variable with parameters (n, N, m).

Proposition 4.10. *Let X be a hypergeometric random variable with parameters (n, N, m). Assume that $N > 1$ and $n > 1$. Then,*

$$E(X) = \frac{nm}{N}, \quad \mathrm{Var}(X) = \frac{nm}{N}\left[\frac{(n-1)(m-1)}{N-1} + 1 - \frac{nm}{N}\right].$$

Proof. For the expectation, we have

$$E(X) = \sum_{i=1}^{n} i \binom{m}{i}\binom{N-m}{n-i}\Big/\binom{N}{n} = m\sum_{i=1}^{n}\binom{m-1}{i-1}\binom{(N-1)-(m-1)}{(n-1)-(i-1)}\Big/\binom{N}{n}$$
$$= m\binom{N-1}{n-1}\Big/\binom{N}{n} = \frac{mn}{N},$$

where the second to the last equality is a result of the fact $\sum_{i=0}^{k}\binom{m}{i}\binom{n-m}{k-i} = \binom{n}{k}$. For the variance, a similar computation as before implies

$$E(X(X-1)) = \sum_{i=2}^{n} i(i-1)\binom{m}{i}\binom{N-m}{n-i}\Big/\binom{N}{n}$$
$$= m(m-1)\sum_{i=2}^{n}\binom{m-2}{i-2}\binom{(N-2)-(m-2)}{(n-2)-(i-2)}\Big/\binom{N}{n}$$
$$= m(m-1)\binom{N-2}{n-2}\Big/\binom{N}{n} = \frac{m(m-1)n(n-1)}{N(N-1)}.$$

Hence, we have

$$\mathrm{Var}(X) = E(X(X-1)) + E(X) - [E(X)]^2 = \frac{nm}{N}\left[\frac{(n-1)(m-1)}{N-1} + 1 - \frac{nm}{N}\right].$$

□

Remark 4.12. Note that the selection of n balls at a time is equivalent to choose n balls one at a time without replacement. It is natural to compare this model with the experiment of "selecting n balls once at a time with replacement". Let Y be the number of black balls selected in the second experiment. Then, Y is exactly a binomial random variable with parameters $(n, m/N)$. This implies $E(Y) = nm/N$ and $\mathrm{Var}(Y) = (nm/N)(1-m/N)$. Let X be the hypergeometric random variable with parameters (n, N, m). By Proposition 4.10, we have

$$E(X) = \frac{nm}{N}, \quad \mathrm{Var}(X) = \frac{nm}{N}\left[\frac{(n-1)(m-1)}{N-1} + 1 - \frac{nm}{N}\right].$$

Then, $E(X) = E(Y)$. When $m, N \to \infty$ with $m/N \to p \in (0,1)$, one can see that

$$\mathrm{Var}(X) - \mathrm{Var}(Y) \to 0$$

and, for fixed i,

$$P(X=i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} = \frac{m!(N-m)!n!(N-n)!}{i!(m-i)!(n-i)!(N-m-n+i)!N!}$$
$$= \binom{n}{i}\frac{m}{N}\cdot\frac{m-1}{N-1}\cdots\frac{m-i+1}{N-i+1}\cdot\frac{N-m}{N-i}\cdot\frac{N-m-1}{N-i-1}\cdots\frac{(N-m)-(n-i)+1}{(N-i)-(n-i)+1}$$
$$\to \binom{n}{i}p^i(1-p)^{n-i}.$$

This builds up a relationship between the hypergeometric random variable and the binomial random variable.