Chapter 1

Axioms of Probability

1.1 Sample Space and Events

Definition. The set of all possible outcomes of an experiment is called the *sample space* of the experiment and is denoted by Ω.

Example. If the experiment consists of tossing two dice, then the sample space is

$$\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}.$$

Definition. Let Ω be a sample space of an experiment. A family Σ of subsets of Ω is called a *σ-algebra* on Ω if the following conditions hold.

(a) $\Omega \in \Sigma$.

(b) For all $E \in \Sigma$, $\Omega \setminus E \in \Sigma$.

(c) If $E_1, E_2, \ldots$ is a sequence of elements in Σ, then

$$\bigcup_{i=1}^{\infty} E_i \in \Sigma.$$

Definition. If Σ is a σ-algebra on Ω, then (Ω, Σ) is called a *measurable space*, and a subset of Ω that belongs to Σ is called an *event*.

Theorem 1.1. Let I be an index set such that for each $i \in I$, Σ_i is a σ-algebra on Ω. Then

$$\Sigma^* = \bigcap_{i \in I} \Sigma_i$$

is also a σ-algebra on Ω.

Proof.

(a) Since $\Omega \in \Sigma_i$ for each $i \in I$, it follows that $\Omega \in \Sigma$.

(b) We have

$$\begin{aligned} E \in \Sigma \quad &\Rightarrow \quad E \in \Sigma_i \text{ for each } i \in I \\ &\Rightarrow \quad \Omega \setminus E \in \Sigma_i \text{ for each } i \in I \\ &\Rightarrow \quad \Omega \setminus E \in \Sigma. \end{aligned}$$

(c) We have

$$\begin{aligned}
E_1, E_2, \dots \in \Sigma \quad &\Rightarrow \quad E_1, \dots, E_2 \in \Sigma_i \text{ for each } i \in I \\
&\Rightarrow \quad \bigcup_{j=1}^{\infty} E_j \in \Sigma_i \text{ for each } i \in I \\
&\Rightarrow \quad \bigcup_{j=1}^{\infty} E_j \in \Sigma. \qquad \square
\end{aligned}$$

Definition. Let Φ be a family of subsets of Ω. Then the *σ-algebra generated by* Φ, denoted by $\sigma(\Phi)$, is the intersection of all σ-algebras that contains Φ.

Example. Let Φ be the collection of all open intervals on $\mathbb{R}$. Then the σ-algebra generated by Φ is called the *Borel algebra* of $\mathbb{R}$, denoted by $\mathcal{B}$.

1.2 Axioms of Probability

Definition. Two events E and F are *mutually exclusive* if $E \cap F = \varnothing$.

Definition. Let (Ω, Σ) be a measurable space. A function $P : \Sigma \to \mathbb{R}$ is called a *probability function* and (Ω, Σ, P) is a *probability space* if the following conditions hold.

(a) For all $E \in \Sigma$, $P(E) \geq 0$.

(b) $P(\Omega) = 1$.

(c) If $E_1, E_2, \ldots$ is a sequence of events that are pairwise mutually exclusive, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

Theorem 1.2. Let (Ω, Σ, P) be a probability space. Let $E, F \in \Sigma$. Then $P(F \setminus E) = P(F) - P(E \cap F)$.

Proof. Since $E \cap F$ and $F \setminus E$ are mutually exclusive, we have

$$P(F) = P((E \cap F) \cup (F \setminus E)) = P(E \cap F) + P(F \setminus E).$$

Thus, $P(F \setminus E) = P(F) - P(E \cap F)$. $\square$

Corollary. $P(\Omega \setminus E) = 1 - P(E)$ holds for any event E, implying $P(\varnothing) = 0$.

Corollary. If $E \subseteq F$, then $P(E) \leq P(F)$ because $P(F) - P(E) = P(F \setminus E) \geq 0$.

Theorem 1.3 (Inclusive-exclusive Principle)**.** Let (Ω, Σ, P) be a probability space. If $E_1, \ldots, E_n \in \Sigma$, then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \leq i_1 < \cdots < i_r \leq n} P(E_{i_1} \cap \cdots \cap E_{i_r}).$$

Proof. The proof is by induction on n. The theorem holds for $n = 0$ and $n = 1$ trivially. For $n = 2$, since $E_1 \cap E_2$ and $E_1 \setminus E_2$ are mutually exclusive, we have

$$P(E_1) = P((E_1 \cap E_2) \cup (E_1 \setminus E_2)) = P(E_1 \cap E_2) + P(E_1 \setminus E_2).$$

Thus, since $E_1 \setminus E_2$ and E_2 are mutually exclusive, we have

$$\begin{aligned} P(E_1 \cup E_2) &= P((E_1 \setminus E_2) \cup E_2) \\ &= P(E_1 \setminus E_2) + P(E_2) \\ &= P(E_1) - P(E_1 \cap E_2) + P(E_2). \end{aligned}$$

Now suppose that the theorem holds for some $n \geq 2$, and we prove that the theorem is true for $n + 1$. Since $E_1 \cup \cdots \cup E_n$ and E_{n+1} are mutually exclusive, we have

$$P(E_1 \cup \cdots \cup E_n \cup E_{n+1}) = P(E_1 \cup \cdots \cup E_n) + P(E_{n+1}) - P((E_1 \cup \cdots \cup E_n) \cap E_{n+1}),$$

where the first term can be written as

$$P(E_1 \cup \cdots \cup E_n) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le n} P(E_{i_1} \cap \cdots \cap E_{i_r}),$$

and the last term can be written as

$$\begin{aligned}
&P((E_1 \cup \cdots \cup E_n) \cap E_{n+1}) \\
&= P((E_1 \cap E_{n+1}) \cup \cdots \cup (E_k \cap E_{n+1})) \\
&= \sum_{s=1}^{n} (-1)^{s+1} \sum_{1 \le i_1 \cdots \le i_s \le n} P((E_{i_1} \cap E_{n+1}) \cap \cdots \cap (E_{i_s} \cap E_{n+1})) \\
&= \sum_{s=1}^{n} (-1)^{s+1} \sum_{1 \le i_1 \cdots \le i_s \le n} P(E_{i_1} \cap \cdots \cap E_{i_s} \cap E_{n+1}) \\
&= -\sum_{r=2}^{n+1} (-1)^{r+1} \sum_{\substack{1 \le i_1 \cdots \le i_{r-1} \le n \\ i_r = n+1}} P(E_{i_1} \cap \cdots \cap E_{i_{r-1}} \cap E_{i_r}).
\end{aligned}$$

Now we consider r, which is the number of sets in each intersection. The second term is actually the case with $r = 1$, and the last term consists of the cases with $r \ge 2$. Thus,

$$P(E_{n+1}) - P((E_1 \cup \cdots E_n) \cap E_{n+1}) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{\substack{1 \le i_1 \cdots \le i_{r-1} \le n \\ i_r = n+1}} P(E_{i_1} \cap \cdots \cap E_{i_r}).$$

Furthermore, note that the first term consists of the case where E_{n+1} does not appear in the intersection, while the difference above consists of the case where E_{n+1} appears in the intersection. Thus, by summing up all terms, we have

$$P(E_1 \cup \cdots \cup E_n \cup E_{n+1}) = \sum_{r=1}^{n+1} (-1)^{r+1} \sum_{1 \le i_1 \le \cdots \le i_r \le n+1} P(E_{i_1} \cap \cdots \cap E_{i_r}),$$

which completes the proof. $\square$

Example. For any three events E_1, E_2, E_3, we have $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_2 \cap E_3) - P(E_3 \cap E_1) + P(E_1 \cap E_2 \cap E_3)$.

1.3 Sample Spaces with Equally Likely Outcomes

Theorem 1.4. Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite sample space and let P be a probability function such that $P(\{\omega_i\}) = P(\{\omega_j\})$ for $i, j \in \{1, \ldots, n\}$. Then for each event $E \subseteq \Omega$ with $|E| = k$, we have

$$P(E) = \frac{k}{n}.$$

Proof. Let p denote the probability of each elementary event $\{\omega_i\}$ for all $i \in \{1, \ldots, n\}$. Then we have

$$1 = P(\Omega) = P\left(\bigcup_{i=1}^{n} \{\omega_i\}\right) = \sum_{i=1}^{n} P(\{\omega_i\}) = np.$$

Thus,

$$p = \frac{1}{n}.$$

Let $E = \{\omega_{i_1}, \ldots, \omega_{i_k}\}$. Then

$$P(E) = P\left(\bigcup_{r=1}^{k} \{\omega_{i_r}\}\right) = \sum_{r=1}^{k} P(\{\omega_{i_r}\}) = \frac{k}{n}. \qquad \square$$

Chapter 2

Conditional Probability and Independence

2.1 Conditional Probability

Definition. Let (Ω, Σ, P) be a probability space. If E is an event with $P(E) > 0$, then define

$$P(F \mid E) = \frac{P(E \cap F)}{P(E)}$$

for any event F.

Theorem 2.1. Let (Ω, Σ, P) be a probability space. If E is an event with $P(E) > 0$, then the function $P_E : \Sigma \to \mathbb{R}$ is a probability function if

$$P_E(F) = P(F \mid E)$$

for any event F.

Proof. For events E and F,

$$P_E(F) = \frac{P(E \cap F)}{P(E)} \geq 0.$$

Moreover,

$$P_E(\Omega) = \frac{P(E \cap \Omega)}{P(E)} = \frac{P(E)}{P(E)} = 1.$$

If $F_1, F_2, \ldots$ is a sequence of events that are piecewise mutually exclusive, then

$$P_E\left(\bigcup_{i=1}^{\infty} F_i\right) = \frac{P\Big(E \cap \bigcup_{i=1}^{\infty} F_i\Big)}{P(E)} = \frac{P\Big(\bigcup_{i=1}^{\infty}(E \cap F_i)\Big)}{P(E)} = \sum_{i=1}^{\infty} \frac{P(E \cap F_i)}{P(E)} = \sum_{i=1}^{\infty} P_E(F_i).$$

Thus, P_E is a probability function. $\square$

2.2 Bayes' Formula

Definition. A *partition* of Ω is a family of nonempty events such that each element in Ω is in exactly one of these events.

Theorem 2.2. Let $E_1, \ldots, E_n$ form a partition of Ω such that $P(E_i) > 0$ for each $i \in \{1, \ldots, n\}$. Then for any event F,

$$P(F) = \sum_{i=1}^{n} P(F \mid E_i)P(E_i).$$

Proof. Since

$$F = F \cap \Omega = F \cap \bigcup_{i=1}^{n} E_i = \bigcup_{i=1}^{n} (F \cap E_i),$$

it follows that

$$P(F) = P\left(\bigcup_{i=1}^{n} (F \cap E_i)\right) = \sum_{i=1}^{n} P(F \cap E_i) = \sum_{i=1}^{n} P(F \mid E_i)P(E_i). \qquad \square$$

Theorem 2.3 (Bayes' Formula)**.** Let $E_1, \ldots, E_n$ form a partition of Ω such that $P(E_j) > 0$ for each $j \in \{1, \ldots, n\}$. Then for any event F with $P(F) > 0$, for any $i \in \{1, \ldots, n\}$, we have

$$P(E_i \mid F) = \frac{P(F \mid E_i)P(E_i)}{\sum_{j=1}^{n} P(F \mid E_j)P(E_j)}.$$

Proof. By Theorem 2.2, we have

$$P(E_i \mid F) = \frac{P(E_i \cap F)}{P(F)} = \frac{P(F \mid E_i)P(E_i)}{\sum_{j=1}^{n} P(F \mid E_j)P(E_j)}. \qquad \square$$

2.3 Independence

Definition. Let $E_1, \ldots, E_n$ be events in a probability space (Ω, Σ, P).

- $E_1, \ldots, E_n$ are *independent* if

  $$P\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} P(E_i)$$

  holds for any nonempty subset I of $\{1, \ldots, n\}$.

- $E_1, \ldots, E_n$ are *dependent* if they are not independent.

Definition. Let $E_1, \ldots, E_n$ and F be events in a probability space (Ω, Σ, P), where $P(F) > 0$. Then $E_1, \ldots, E_n$ are *independent given event F* if

$$P\left(\bigcap_{i \in I} E_i \,\middle|\, F\right) = \prod_{i \in I} P(E_i \mid F)$$

holds for any nonempty subset I of $\{1, \ldots, n\}$.

Chapter 3

Discrete Random Variables

3.1 Discrete Random Variables

Definition. Let $X : \Omega \to \mathbb{R}$ be a function in a probability space (Ω, Σ, P). Then X is called a *random variable* if for each $S \in \mathcal{B}$, its preimage

$$X^{-1}(S) = \{\omega \in \Omega : X(\omega) \in S\}$$

is an element of Σ.

- A random variable is *discrete* if its range is countable.
- A random variable is *continuous* if its range is uncountable.

Remark. Since each $S \in \mathcal{B}$ is mapped to a event $X^{-1}(S) \in \Sigma$, we will abuse the notation, using conditions related to a ramdom variable to denote events. For example,

$$P(-1 \leq X \leq 1) = P(\{\omega \in \Omega : -1 \leq X(\omega) \leq 1\}).$$

Definition. Let X be a random variable in a probability space (Ω, Σ, P). The *probability mass function* $p_X : \mathbb{R} \to \mathbb{R}$ of X is defined as

$$p_X(x) = P(X = x)$$

for each $x \in \mathbb{R}$.

3.2 Expectation and Variance

Definition. Let X be a discrete random variable in a probability space (Ω, Σ, P). Then the *expectation* of X, denoted by $E[X]$, is defined as follows.

- If X is nonnegative, i.e, $X(\omega) \geq 0$ for each $\omega \in \Omega$, then
$$E[X] = \sum_{x \in X(\Omega)} x \cdot p_X(x).$$
- Otherwise, we define $E[X] = E[X^+] - E[X^-]$, where $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$.

Theorem 3.1. Let X and Y be discrete random variables in a probability space (Ω, Σ, P). If both $E[X]$ and $E[Y]$ exist, then the following statements are true.

(a) $E[aX] = aE[X]$ for $a \in \mathbb{R}$.

(b) $E[X + Y] = E[X] + E[Y]$.

Proof.

(a) First, suppose that $a \geq 0$. If X is nonnegative, then so is aX. Thus, we have
$$E[aX] = \sum_{x \in X(\Omega)} ax \cdot p_X(x) = aE[X].$$
If X is not nonnegative, by the fact that $(aX)^+ = aX^+$ and $(aX)^- = aX^-$, we have
$$E[aX] = E[aX^+] - E[aX^-] = aE[X^+] - aE[X^-] = aE[X].$$
since X^+ and X^- are nonnegative. Thus the statement holds for $a \geq 0$.
For the case that $a < 0$, note that since $(-X)^+ = X^-$ and $(-X)^- = X^+$, it follows that
$$E[-X] = E[X^-] - E[X^+] = -E[X].$$
Thus, we have
$$E[aX] = E[(-a)(-X)] = -aE[-X] = aE[X].$$

(b) To be completed. □

Definition. Let X be a discrete random variable in a probability space (Ω, Σ, P). Then the *variance* of X is defined as
$$\mathrm{Var}(X) = E[(X - E[X])^2].$$

Theorem 3.2. Let X be a discrete random variable. Then $\mathrm{Var}(X) = E[X^2] - (E[X])^2$.

Proof. It is proved by
$$\begin{aligned}
\mathrm{Var}(X) &= E[(X - E[X])]^2 \\
&= E[X^2 - 2X \cdot E[X] + (E[X])^2] \\
&= E[X^2] - 2E[X] \cdot E[X] + (E[X])^2 \\
&= E[X^2] - (E[X])^2.
\end{aligned}$$
□

3.3 Bernoulli and Binomial Random Variables

Definition. Let $0 \leq p \leq 1$. A random variable X is called a *Bernoulli random variable* with parameter p if $p_X(1) = p$ and $p_X(0) = 1 - p$.

Theorem 3.3. Let X be a Bernoulli random variable with parameter p.

(a) $E[X] = p$.

(b) $\mathrm{Var}(X) = p(1-p)$.

Proof. Since $p(0) + p(1) = 1$, we have $p(x) = 0$ for $x \notin \{0, 1\}$.

(a) We have
$$E[X] = \sum_{x\,:\,p_X(x)>0} x \cdot p_X(X) = 1 \cdot p + 0 \cdot (1-p) = p.$$

(b) By (a), we have
$$\begin{aligned}\mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\ &= (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 \\ &= p - p^2 \\ &= p(1-p).\end{aligned}$$
□

Definition. Let n be a nonnegative integer and $0 \leq p \leq 1$. A random variable X is called a *binomial random variable* with parameter (n, p), if
$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$
for each $x \in \{0, \ldots, n\}$.

Theorem 3.4. Let X be a binomial random variable with parameter (n, p).

(a) $E[X] = np$.

(b) $\mathrm{Var}(X) = np(1-p)$.

Proof. We have $p(x) = 0$ for $x \notin \{0, \ldots, n\}$ because
$$\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1.$$
Also, note that
$$\begin{aligned}E[X^k] &= \sum_{x=0}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^{n} x^{k-1} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} \\ &= np \sum_{y=0}^{n-1} (y+1)^{k-1} \binom{n-1}{y} p^y (1-p)^{(n-1)-y} \qquad (\star)\end{aligned}$$
holds for positive integer k.

(a) By $(\star)$, the expectation of X is given by

$$E[X] = np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{(n-1)-y} = np.$$

(b) By $(\star)$, we have

$$E[X^2] = np \sum_{y=0}^{n-1} (y+1) \binom{n-1}{y} p^y (1-p)^{(n-1)-y} = np((n-1)p+1).$$

Thus, the variance of X is given by

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\ &= np((n-1)p+1) - (np)^2 \\ &= np(1-p). \end{aligned}$$

□

3.4 Poisson Random Variable

Theorem 3.5. Let $\lambda > 0$. For integer $n \geq \lambda$, let X_n be a binomial random variable with parameter $(n, \lambda/n)$. Then for nonnegative integer x, we have

$$\lim_{n\to\infty} p_{X_n}(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}.$$

Proof. For $n \geq \lambda$, we have

$$\begin{aligned}
p_{X_n}(x) &= \frac{n!}{(n-x)! \cdot x!} \cdot \left(\frac{\lambda}{n}\right)^x \left(\frac{n-\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \cdot \frac{n!}{(n-x)! \cdot (n-\lambda)^x} \cdot \left(\frac{n-\lambda}{n}\right)^n \\
&= \frac{\lambda^x}{x!} \cdot \prod_{i=1}^{x} \frac{n+1-i}{n-\lambda} \cdot \left(1 - \frac{\lambda}{n}\right)^n.
\end{aligned}$$

Thus,

$$\lim_{n\to\infty} p_{X_n}(x) = \frac{\lambda^x}{x!} \cdot \prod_{i=1}^{x} \left(\lim_{n\to\infty} \frac{n+1-i}{n-\lambda}\right) \cdot \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = \frac{\lambda^x}{x!} \cdot e^{-\lambda}. \qquad \square$$

Definition. Let $\lambda > 0$. A random variable X is called a *Poisson random variable* with parameter λ, if

$$p_X(x) = e^{-x} \cdot \frac{\lambda^x}{x!}$$

holds for any nonnegative integer x.