1

Finite trials

In probability theory an experiment, an observation, or a survey is called a *trial*. A trial is called a *finite trial* if it has a finite number of outcomes. In this chapter we will explore elementary concepts in probability theory, limiting ourselves to finite trials. For this purpose, knowledge of elementary set theory is sufficient. However, once the reader understands the fundamental concepts of probability theory through observation of finite trials, he or she can easily proceed to *infinite trials*–the subject of modern probability theory–using the results in modern analysis.

1.1 Probability spaces

In observing the number obtained when throwing a die, we have six outcomes: $1, 2, 3, 4, 5, 6$. Each outcome is called a *sample point*, and the set of all sample points is called the *sample space* for the trial. The sample points and the sample space are considered in any trial. A trial is called a finite trial or an infinite trial according to whether the sample space is a finite set or an infinite set. In this chapter we observe only finite trials, which we call simply trials.

Let T be a trial and Ω its sample space. For any set $A \subset \Omega$ we say that *A occurs* instead of saying that a sample point in A is realized in the trial. In this sense a subset A of Ω is called an *event* A.

The complement A^c of A occurs if and only if A does not occur. Hence A^c is called the *complementary event* of A. The union of A and B, $A \cup B$, occurs if and only if at least one of A and B occurs. Hence $A \cup B$ is called the *sum event* of A and B. The intersection of A and B, $A \cap B$, occurs if and only if both A and B occur. Hence $A \cap B$ is called the *intersection event* of A and B. The difference of A from B, $A \setminus B$ occurs if and only if A occurs and B does not. Hence $A \setminus B$ is called the *difference event* of A and B. The inclusion relation $A \subset B$ means that B occurs whenever A does. In this case the difference $B \setminus A$ is called the *proper difference*, written $B - A$. When we use the notation $B - A$, we understand that $A \subset B$ is implicitly assumed. The events A and B are called *exclusive* if and only if the sets A and B are exclusive; that is, $A \cap B = \emptyset$. In this case $A \cup B$ is called the *direct sum* of A and B, written $A + B$. When we use the notation $A + B$, we understand that $A \cap B = \emptyset$ is assumed implicitly. Similarly for the sum $A_1 \cup A_2 \cup \cdots \cup A_n$ (or $\bigcup_{i=1}^n A_i$) and the direct sum $A_1 + A_2 + \cdots + A_n$ (or $\sum_{i=1}^n A_i$).

Let T be a trial and Ω its sample space. For any set $A \subset \Omega$ we denote by $P(A)$ the *probability* that the event A occurs. By the intuitive meaning of probability it is natural to impose the following on $P(A)$.

(p.1) $P(A) \geqq 0$,

(p.2) $P(A + B) = P(A) + P(B)$, (additivity)

(p.3) $P(\Omega) = 1$.

In general a set function $P(A)$, $A \subset \Omega$, having these properties is called a *probability measure* on Ω, and a set Ω endowed with a probability measure P on Ω is called a *probability space* (Ω, P). Hence (Ω, P) is called the probability space for a trial T, if Ω is the sample space of T and $P(A)$ is the probability of occurrence of $A \subset \Omega$. Conversely, for any given probability space (Ω, P) we can consider a trial T of drawing a sample point from Ω with probability $P(A)$ for every $A \subset \Omega$, so that the probability space for T turns out to be (Ω, P).

Since the probabilistic aspect of a trial is completely determined by its probability space, any trials having the same probability space are identified in probability theory. Suppose that T_1 is the trial of observing the number obtained when throwing a die and T_2 is the trial of observing the number on a card drawn at random from a box with six cards numbered 1, 2, 3, 4, 5, and 6. Then T_1 and T_2 have the same probability space (Ω, P), where

$$\Omega = \{1,2,3,4,5,6\}, \qquad P(A) = \#A/6 \ (\#: \text{the number of points})$$

Hence T_1 and T_2 are identical with each other. In view of this observation, a trial with the probability space (Ω, P) is called simply a trial (Ω, P).

Theorem 1. If P is a probability measure on Ω, then we have the following:

(i) $P(\sum_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$,

(ii) $P(B - A) = P(B) - P(A)$,

(iii) $P(A^c) = 1 - P(A)$,

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,
(v) $P(A) = \Sigma_{\omega \in A} P\{\omega\}$ (Hence P is determined if $P\{\omega\}$ is determined for every $\omega \in \Omega$.)

Proof:
(i) Derive this from (p.2) by induction on n.
(ii) Apply (p.2) to $B = (B - A) + A$. Remember that $A \subset B$ is assumed implicitly when we use the notation $B - A$.
(iii) Set $B = \Omega$ in (ii).
(iv) Setting $C = A \cap B$, $A_1 = A - C$, and $B_1 = B - C$, we have

$$A = A_1 + C, \qquad B = B_1 + C, \qquad A \cup B = A_1 + B_1 + C,$$

which implies (iv) by (i).
(v) Apply (i) to $A = \Sigma_{a \in A}\{a\}$. ■

We have mentioned that an event is represented by a set. We also represent an event by a condition. Let $\alpha(\omega)$ be a condition concerning a generic sample point ω. We say that α occurs if a sample point ω satisfying $\alpha(\omega)$ is realized as an outcome of the trial in consideration. In view of this, a condition $\alpha(\omega)$ is often called an event $\alpha(\omega)$. If we set $A = \{\omega | \alpha(\omega)\}$, then occurrence of α is equivalent to that of A. Hence the probability of occurrence of α, written $P(\alpha)$, is equal to $P\{\omega | \alpha(\omega)\}$.

The negation of α, written $\alpha^{\neg}$, is the complementary event of α, because $\{\omega | \alpha^{\neg}(\omega)\} = \{\omega | \alpha(\omega)\}^c$. The condition "$\alpha$ or β", written $\alpha \vee \beta$, is the sum event of α and β, because

$$\{\omega | \alpha(\omega) \vee \beta(\omega)\} = \{\omega | \alpha(\omega)\} \cup \{\omega | \beta(\omega)\}.$$

Similarly the condition "α and β," written $\alpha \wedge \beta$, is the intersection event of α and β.

Exercise 1. Prove the following *inclusion–exclusion formulas*:
(i) $P(\bigcup_{i=1}^n A_i) = \Sigma_{k=1}^n (-1)^{k-1} \Sigma_{i_1 < i_2 < \cdots < i_k} P(\bigcap_{\kappa=1}^k A_{i_\kappa})$.
(ii) The formula obtained by switching $\cup$ and $\cap$.

[*Hint*: Derive these formulas from Theorem 1(iv), using induction on n.]

1.2 Real random variables and random vectors

Let (Ω, P) be the probability space for a trial T. A real-valued function $X(\omega)$ defined on Ω is called a *real random variable* on (Ω, P), which intuitively means a quantity varying in accordance with the result of T. As we mentioned in the last section, the trial of observing the number obtained when throwing a die is represented by the probability space

$$\Omega = \{1,2,3,4,5,6\}, \qquad P(A) = \#A/6.$$

If we are to get a prize of an amount double the number obtained, the prize is a random variable $X(\omega)$ defined by $X(\omega) = 2\omega$.

The trial of observing the numbers obtained when throwing a die twice is represented by the probability space

$$\Omega = \{(i, j) \mid i, j = 1,2,3,4,5,6\} = \{1,2,3,4,5,6\}^2,$$

$$P(A) = \#A/36.$$

If we denote by X_1, X_2, and X the number obtained in the first throw, the number obtained in the second throw, and the sum of these numbers, respectively, then X_1, X_2, and X are real random variables represented by

$$X_1(i, j) = i, \qquad X_2(i, j) = j, \qquad X(i, j) = i + j.$$

Let $X = X(\omega)$ be a real random variable on (Ω, P). The image $X(\Omega)$, the set of all the values $X(\omega)$ can take, is called the sample space of the random variable X, written Ω^X. For any set $B \subset \Omega^X$ the probability of the event that the value of X lies in B is equal to $P\{\omega \mid X(\omega) \in B\}$, that is, $P(X^{-1}(B))$. Viewing this as a function of a set $B \subset \Omega^X$, we denote it by $P^X(B)$. It is easy to check that P^X satisfies the conditions (p.1), (p.2), and (p.3) in Section 1.1 using the relations

$$X^{-1}(B_1 + B_2) = X^{-1}(B_1) + X^{-1}(B_2), \qquad X^{-1}(\Omega^X) = \Omega.$$

The probability measure P^X is called the *probability distribution* of X. In particular, we have

$$P^X\{b\} = P(X^{-1}(b)), \qquad b \in \Omega^X.$$

Since P^X is a probability measure, we have

$$P^X(B) = \sum_{b \in B} P^X\{b\},$$

which shows that P^X is completely determined by assigning $P^X\{b\}$ for every $b \in \Omega^X$. (Ω^X, P^X) is called the *probability space* of the random variable X.

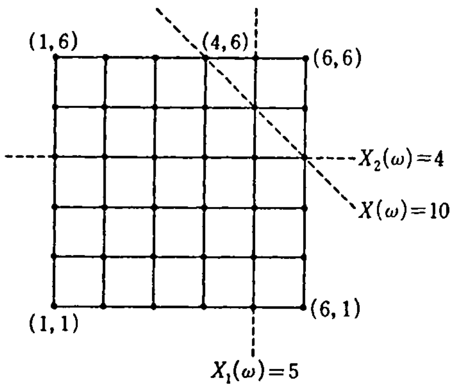

Figure 1.1

Let us determine the probability spaces of the random variables X_1, X_2, and X in the example of throwing a die twice, mentioned above.

$$\Omega^{X_1} = \Omega^{X_2} = \{1,2,3,4,5,6\}, \qquad \Omega^X = \{2,3,4,\ldots,12\},$$

$$P^{X_i}\{k\} = P(X_i^{-1}(k)) = \frac{\# X_i^{-1}(k)}{36} = \frac{6}{36} = \frac{1}{6}, \qquad i = 1,2,$$

$$P^X\{k\} = P(X^{-1}(k)) = \frac{\# X^{-1}(k)}{36} = \begin{cases} (k-1)/36 & (2 \leqq k \leqq 7), \\ (13-k)/36 & (8 \leqq k \leqq 12), \end{cases}$$

as we can see in Figure 1.1.

Let $X_1(\omega)$ and $X_2(\omega)$ be real random variables. Then the pair $X(\omega) = (X_1(\omega), X_2(\omega))$ is a function on Ω with values in $\mathbb{R}^2$, called a (two-dimensional) *vector-valued random variable* or simply a *random vector*. The sample space and the probability law of X are defined as follows:

$$\Omega^X = X(\Omega),$$

$$P^X(B) = P\{\omega | X(\omega) \in B\} = P(X^{-1}(B)), \qquad B \subset \Omega^X.$$

Ω^X is a subset of $\mathbb{R}^2$ and P^X is a probability measure on Ω^X. It is obvious that

$$\Omega^X \subset \Omega^{X_1} \times \Omega^{X_2}.$$

Note that these two sets do not always coincide with each other. $X_1(\omega)$ and $X_2(\omega)$ are *component variables* of $X(\omega)$, and $X(\omega)$ is called the *joint variable* of $X_1(\omega)$ and $X_2(\omega)$. The probability law P^X is often called the *joint probability distribution* of $X_1(\omega)$ and $X_2(\omega)$. For any natural number n we can define n-dimensional random vectors similarly.

The map that carries $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ to its kth component x_k is called the *k-projection*, written π_k. If $X(\omega)$ is the joint variable of $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$, then we have

$$X_k(\omega) = \pi_k(X(\omega)), \qquad \text{i.e., } X_k = \pi_k \circ X,$$

where the small circle $\circ$ denotes composition of maps.

Let $X(\omega)$ be a real random variable. For any real-valued function φ defined on Ω^X, we obtain a real random variable $Y(\omega)$ defined by

$$Y(\omega) = \varphi(X(\omega)) \qquad (\text{i.e., } Y = \varphi \circ X).$$

The probability space of Y is given by

$$\Omega^Y = (\varphi \circ X)(\Omega),$$

$$P^Y(C) = P\big((\varphi \circ X)^{-1}(C)\big).$$

A real random variable $Y(\omega)$ thus obtained from $X(\omega)$ is called a function of $X(\omega)$. Similarly an m-dimensional random vector $Y(\omega)$ is defined to be a function of an n-dimensional random vector $X(\omega)$ if there exists a map φ: $\Omega^X(\subset \mathbb{R}^n) \to \mathbb{R}^m$ such that $Y(\omega) = \varphi(X(\omega))$ for every ω. For example the kth component $X_k(\omega)$ of $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$ is a function of $X(\omega)$, since $X_k(\omega) = \pi_k(X(\omega))$.

Theorem 1.2.1. If $Y(\omega) = \varphi(X(\omega))$, then

$$\Omega^Y = \varphi(\Omega^X), \qquad P^Y(C) = P^X(\varphi^{-1}(C)).$$

Proof: Observing that $Y(\omega) = (\varphi \circ X)(\omega)$ and that

$$\omega \in (\varphi \circ X)^{-1}(C) \Leftrightarrow (\varphi \circ X)(\omega) \in C \Leftrightarrow \varphi(X(\omega)) \in C$$
$$\Leftrightarrow X(\omega) \in \varphi^{-1}(C) \Leftrightarrow \omega \in X^{-1}(\varphi^{-1}(C)),$$

we obtain

$$Y(\omega) = \varphi(X(\omega)) = (\varphi \circ X)(\omega)$$

$$\Omega^Y = (\varphi \circ X)(\Omega) = \varphi(X(\Omega)) = \varphi(\Omega^X),$$

$$P^Y(C) = P\big((\varphi \circ X)^{-1}(C)\big) = P\big(X^{-1}(\varphi^{-1}(C))\big) = P^X(\varphi^{-1}(C)).$$

■

For a real random variable X the *mean value* (or *expectation*) of X, written EX, is defined by

$$EX = \sum_{\omega \in \Omega} X(\omega) P\{\omega\}.$$

For any set $A \subset \Omega$ we denote

$$\sum_{\omega \in A} X(\omega) P\{\omega\}$$

by $E(X, A)$. The mean value of a random vector $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$ is defined by

$$EX = (EX_1, EX_2, \ldots, EX_n) \in \mathbb{R}^n.$$

Theorem 1.2.2. Let X and Y be random vectors. Then we have the following:

(i) (*additivity of mean values*)

$$E(aX + bY) = aEX + bEY \qquad (a, b\text{: constant})$$

(ii) $E(X, \sum_{i=1}^n A_i) = \sum_{i=1}^n E(X, A_i)$

(iii) If $X(\omega) = a$ (a: constant vector) on A, then $E(X, A) = P(A)a$. In particular, if $X(\omega) = a$ on Ω, then $EX = a$.

(iv) $EX = \sum_{x \in \Omega^X} P^X\{x\} x$

(v) If $Y(\omega) = \varphi(X(\omega))$ on Ω, then

$$EY = \sum_{x \in \Omega^X} \varphi(x) P^X\{x\}.$$

Let X and Y be real random variables, then we have the following:

(vi) $X(\omega) \geqq Y(\omega) \Rightarrow EX \geqq EY$, $E(X, A) \geqq E(Y, A)$.

(vii) $X(\omega) \geqq 0$, $A \subset B \Rightarrow E(X, A) \leqq E(X, B)$.

Proof: We will only prove (iv) and (v) here, because the other statements are trivial.

(iv) Setting $A_x = X^{-1}(x)$, we have

$$\Omega = \sum_{x \in \Omega^X} A_x,$$

so

$$E(X) = E(X, \Omega) = \sum_{x \in \Omega^X} E(X, A_x)$$

by (ii). Since $X(\omega) = x$ on A_x, we have

$$E(X, A_x) = P(A_x)x = P^X\{x\}x.$$

Hence we obtain (iv) at once.

(v) In the same way as above, we have

$$E(\varphi(X)) = \sum_x E(\varphi(X), A_x) = \sum_x \varphi(x) P(A_x)$$
$$= \sum_x \varphi(x) P^X\{x\}$$

where x runs over Ω^X. ■

Let A be a subset of Ω. The *indicator* of A, written 1_A, is defined by

$$1_A = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^c. \end{cases}$$

This is a random variable that takes 1 or 0 according to whether A occurs or not. It is obvious that

$$1_\Omega(\omega) = 1 \qquad \text{for every } \omega \in \Omega.$$

$1_\Omega(\omega)$ is a random variable that takes 1 always, whereas 1 is just a fixed number. However, we often denote $1_\Omega(\omega)$ simply by 1 and accordingly $a \cdot 1_\Omega(\omega)$ (a: constant) simply by a unless there is any possibility of confusion. It is obvious that

$$E 1_A = P(A), E(X, A) = E(X 1_A).$$

Theorem 1.2.3. Let $e, e', e_1, e_2, \ldots, e_n$ denote the indicators of $A, A^c, A_1, A_2, \ldots, A_n$, respectively.

(i) $e' = 1 - e$.

(ii) If $A = \bigcap_{i=1}^n A_i$, then

$$e = \prod_{i=1}^n e_i \qquad \text{and} \qquad e = \min(e_1, e_2, \ldots, e_n).$$

(iii) If $A = \bigcup_{i=1}^n A_i$, then

$$e = 1 - \prod_{i=1}^n (1 - e_i) \qquad \text{and} \qquad e = \max(e_1, e_2, \ldots, e_n).$$

Proof: Assertions (i) and (ii) and the second part of (iii) are obvious. To prove the first part of (iii), observe the following:

$$A^c = \bigcap_{i=1}^n A_i^c \qquad \text{by de Morgan's law,}$$

$$1 - e = \prod_{i=1}^n (1 - e_i) \qquad \text{by (i) and (ii).}$$ ■

The *variance* of X, written $V(X)$, is defined as follows:

$$V(X)=E\big((X-EX)^2\big).$$

Setting $\varphi(x)=(x-EX)^2$ in Theorem 1.2.2(v), we have

$$V(X)=\sum_{x\in\Omega^X}(x-EX)^2P^X\{x\}.$$

The *covariance* of X and Y, written $V(X,Y)$, is defined as

$$V(X,Y)=E((X-EX)(Y-EY)).$$

Replacing X by (X,Y), x by (x,y), and $\varphi(x)$ by $(x-EX)(y-EY)$ in Theorem 1.2.2(v), we obtain

$$V(X,Y)=\sum_{(x,y)\in\Omega^{(X,Y)}}(x-EX)(y-EY)P^{(X,Y)}\{(x,y)\}.$$

The following properties are verified easily:

$$V(X,Y)=V(Y,X),$$

$$V(X,a)=0 \qquad (a\text{: constant}),$$

$$V(X,X)=V(X)\geqq 0.$$

Theorem 1.2.4.

(i) $V(aX+b)=a^2V(X), \qquad V(aX+b,cY+d)=acV(X,Y).$

(ii) $V(aX+bY)=a^2V(X)+2abV(X,Y)+b^2V(Y).$

(iii) $V(X)=EX^2-(EX)^2, \qquad V(X,Y)=E(XY)-EXEY.$

(iv) $|V(X,Y)|\leqq\sqrt{V(X)V(Y)}.$

Proof: To prove (i), (ii), and (iii), take the mean values of both sides of the following:

$$((aX+b)-E(aX+b))^2=a^2(X-E(X))^2,$$

$$((aX+b)-E(aX+b))((cY+d)-E(cY+d))$$
$$=ac(X-EX)(Y-EY),$$

$$((aX+bY)-E(aX+bY))^2=a^2(X-EX)^2+2ab(X-EX)$$
$$\times(Y-EY)+b^2(Y-EY)^2,$$

$$(X-EX)^2=X^2-2(EX)X+(EX)^2,$$

$$(X-EX)(Y-EY)=XY-(EX)Y-(EY)X+EXEY.$$

To prove (iv), observe that

$$\begin{aligned} 0 &\leqq V(tX+Y) = E((tX+Y)-E(tX+Y))^2 \\ &= E\left(t^2(X-EX)^2 + 2t(X-EX)(Y-EY) + (Y-EY)^2\right) \\ &= t^2V(X) + 2tV(X,Y) + V(Y). \end{aligned}$$

The last quadratic function of t is nonnegative for every real value of t. Hence

$$V(X,Y)^2 - V(X)V(Y) \leqq 0, \quad \text{i.e.,} \ |V(X,Y)| \leqq \sqrt{V(X)V(Y)}.$$

■

The *standard deviation* of X, $\sigma(X)$ and the *correlation coefficient* of X and Y, $R(X,Y)$ are defined as follows:

$$\sigma(X) = \sqrt{V(X)},$$

$$R(X,Y) = \frac{V(X,Y)}{\sigma(X)\sigma(Y)} \qquad \text{if } \sigma(X)\sigma(Y) > 0.$$

A condition on real random variables $X(\omega)$, $Y(\omega)$, and $Z(\omega)$, for example, $X(\omega)+Y(\omega) \geqq Z(\omega)$, is also regarded as a condition on ω. Hence it is an event, and the probability of occurrence of this event is

$$P\{\omega | X(\omega)+Y(\omega) \geqq Z(\omega)\},$$

which is denoted simply by

$$P\{X(\omega)+Y(\omega) \geqq Z(\omega)\} \qquad \text{or} \qquad P\{X+Y \geqq Z\}.$$

If the probability of occurrence of α equals 1; that is, $P(\alpha)=1$, we say that $\alpha(\omega)$ holds (or occurs) *almost surely* or that $\alpha(\omega)$ holds with probability 1, written $\alpha(\omega)$ a.s. Mean values, variances, covariances, and correlation coefficients are the same for the random variables equal to each other almost surely:

$$X(\omega) = X'(\omega) \text{ a.s.} \Rightarrow EX = EX',\ V(X) = V(X'),\ \sigma(X) = \sigma(X'),$$

$$X(\omega) = X'(\omega) \text{ a.s.},\ Y(\omega) = Y'(\omega) \text{ a.s.} \Rightarrow V(X,Y) = V(X',Y'),$$

$$R(X,Y) = R(X',Y').$$

To prove these properties it suffices to show that

$$X(\omega) = X'(\omega) \text{ a.s.} \Rightarrow EX = EX'.$$

Let A denote the set $\{\omega | X(\omega) \neq X'(\omega)\}$. Then $P(A)=0$, so $P\{\omega\}=0$ for every $\omega \in A$. This implies that $E(X,A) = E(X',A) = 0$. Since $X(\omega) = X'(\omega)$ on A^c, we have $E(X,A^c) = E(X',A^c)$. Hence we have

$$E(X) = E(X,A) + E(X,A^c) = E(X',A) + E(X',A^c) = E(X').$$