2. AXIOMATIC PROBABILITY

2.1 The axioms

The formulation for classical probability in which all outcomes or points in the sample space are equally likely is too restrictive to develop a useful theory of probability. In this section we give a general definition. For the moment we will assume that the sample space of outcomes Ω is either a finite or countable set.

A **probability distribution** $\mathbb{P}$ on Ω is a real-valued function defined on subsets (events) of Ω satisfying the following three conditions:

1. $0 \leqslant \mathbb{P}(A) \leqslant 1$, for all $A \subseteq \Omega$.
2. $\mathbb{P}(\Omega) = 1$.
3. For a finite or infinite collection of disjoint events $\{A_i\}_i$, that is $A_i \cap A_j = \emptyset$ for $i \neq j$, we have
$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}\left(A_i\right).$$

First observe the following consequence of Axiom 3.

4. $\mathbb{P}\left(\emptyset\right) = 0$.

This follows because if we take $A_i = \emptyset$, for each i in Axiom 3 then we get a contradiction unless $\mathbb{P}\left(\emptyset\right) = 0$. Note that Axiom 3 here implies the property 3. that we had in Chapter 1 if we take $A_i = \emptyset$ for $i > 2$, $A_1 = A$ and $A_2 = B$. This of course extends to n disjoint events $A_1, \ldots, A_n$ for which $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$. An important consequence of this axiom is a continuity property of probabilities in that for events $B_1 \subseteq B_2 \subseteq B_3 \subseteq \cdots$ we have
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \lim_{i\to\infty} \mathbb{P}\left(B_i\right). \tag{2.1}$$

This follows by taking $A_1 = B_1$ and for $i > 1$, $A_i = B_i \cap B_{i-1}^c$, then the $\{A_i\}$ are disjoint with $\bigcup_{j=1}^{i} A_j = B_i$ and $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ so that $\mathbb{P}\left(B_i\right) = \sum_{j=1}^{i} \mathbb{P}\left(A_j\right)$, and
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(A_i\right) = \lim_{i\to\infty} \sum_{j=1}^{i} \mathbb{P}\left(A_j\right) = \lim_{i\to\infty} \mathbb{P}\left(B_i\right).$$

As in Chapter 1 we may deduce that

5. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
6. If $A \subseteq B$ then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.
7. $\mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Using (2.1) and Property 5. we may see a corresponding continuity property for probabilities on decreasing events, in that for events $B_1 \supseteq B_2 \supset B_3 \supseteq \cdots$, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = \lim_{i \to \infty} \mathbb{P}\left(B_i\right).$$

Boole's Inequality We may deduce from Property 7, that for any events $A_1, \ldots, A_n$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} \mathbb{P}\left(A_i\right).$$

The case $n = 2$ follows from Property 7 (and Axiom 1) immediately and the case for general n by induction. We may use (2.1) to extend Boole's Inequality to the case of countably many events $A_1, A_2, \ldots$ so that

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

2.2 Inclusion-exclusion formula

The inclusion-exclusion formula gives a method of calculating that at least one of a number of events occurs.

Theorem 2.2 (Inclusion-exclusion) *For any events $A_1, A_2, \ldots, A_n$,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i} \mathbb{P}\left(A_i\right) - \sum_{i_1 < i_2} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) + \sum_{i_1 < i_2 < i_3} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap A_{i_3}\right) - \cdots + (-1)^{n-1} \mathbb{P}\left(\cap_{i=1}^{n} A_i\right).$$

Proof. The proof is by induction on n. The case $n = 2$ is just Property 7. Assume the result holds for any n events and consider $A_1, \ldots, A_{n+1}$. Now see from Property 7. again that

$$\mathbb{P}\left(\bigcup_1^{n+1} A_i\right) = \mathbb{P}\left(\bigcup_1^{n} A_i\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left(\left(\bigcup_1^{n} A_i\right) \cap A_{n+1}\right)$$
$$= \mathbb{P}\left(\bigcup_1^{n} A_i\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left(\bigcup_1^{n} (A_i \cap A_{n+1})\right), \tag{2.3}$$

and then apply the inductive hypothesis to the first and the third terms to complete the induction and obtain the conclusion. □

Example 2.4 What is the probability that in a round of Bridge at least one player holds exactly two aces and two kings? Let Ω be the sample space of all possible (unordered) hands for 4 players, then the number of points in Ω is $N = \binom{52}{13\ 13\ 13\ 13}$. For each player i, $i = 1, 2, 3, 4$, let A_i be the event that player i holds exactly two aces and two kings. Then

$$\mathbb{P}(A_i) = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9}\binom{39}{13\ 13\ 13}}{\binom{52}{13\ 13\ 13\ 13}} = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9}}{\binom{52}{13}};$$

think of choosing the aces for player i, then the kings and then 9 other cards and finally distributing the remaining 39 cards among the other three players. If $i \neq j$, then

$$\mathbb{P}(A_i \cap A_j) = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9\ 9\ 26}\binom{26}{13}}{\binom{52}{13\ 13\ 13\ 13}} = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9\ 9\ 26}}{\binom{52}{13\ 13\ 26}};$$

as before, think of picking the aces and kings for player i with the remaining aces and kings going to player j, and then picking 9 other cards for each of i and j. If i, j, k are unequal then $A_i \cap A_j \cap A_k = \emptyset$, so that by inclusion exclusion we have the required probability is

$$\mathbb{P}\left(\bigcup_{i=1}^{4} A_i\right) = 4 \times \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9}}{\binom{52}{13}} - \binom{4}{2} \times \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9\ 9\ 26}}{\binom{52}{13\ 13\ 26}}.$$ □

Example 2.5 Suppose that n students leave their n coats outside a lecture room and when they leave they pick up their coats at random. What is the probability that at least one student has his own coat? Let Ω consist of all permutations of $1, \ldots, n$, so that if $(i_1, \ldots, i_n) \in \Omega$ then i_j is the index of the coat got by student j. Denote by A_i the event that student i gets his own coat. Then for $i_1 < i_2 < \cdots < i_r$, the probability that all of the students $i_1, \ldots, i_r$ get their own coats is

$$\mathbb{P}\left(\bigcap_{k=1}^{r} A_{i_k}\right) = \frac{(n-r)!}{n!},$$

since the numerator is the number of ways that the remaining $n-r$ coats may be permuted among the remaining $n-r$ students. Then by inclusion-exclusion we have

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{r=1}^{n}\left[(-1)^{r-1} \sum_{i_1<\cdots<i_r} \mathbb{P}\left(\bigcap_{k=1}^{r} A_{i_k}\right)\right] \\
&= \sum_{r=1}^{n}(-1)^{r-1}\binom{n}{r}\frac{(n-r)!}{n!} = \sum_{r=1}^{n}(-1)^{r-1}\frac{1}{r!},
\end{aligned}$$

so that for large n the probability is approximately equal to

$$1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots = 1 - e^{-1} \approx 0{\cdot}632\,. \qquad \square$$

Corollary 2.6 (Bonferroni's Inequalities) *For any events $A_1, A_2, \ldots, A_n$ and for any r, $1 \leqslant r \leqslant n$,*

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \ \overset{\leqslant}{\underset{\geqslant}{\mathit{or}}}\ & \sum_{i} \mathbb{P}(A_i) - \sum_{i_1<i_2} \mathbb{P}(A_{i_1} \cap A_{i_2}) \\
&+ \sum_{i_1<i_2<i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{r-1} \sum_{i_1<\cdots<i_r} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_r}),
\end{aligned}$$

according as r is odd or even. That is, if the sum in the inclusion-exclusion formula is truncated after r terms it overestimates or underestimates the probability of the union of the n events according as r is odd or even.

Proof. The proof proceeds by induction on n. Assume it is true for n, then for $n+1$ events it is true when $r = n+1$, by the inclusion-exclusion formula, and for $r \leqslant n$, apply the inductive hypothesis to probability of the two unions of n events in relation (2.3) to get the result. $\square$

2.3 Conditional probability

Suppose that $B \subseteq \Omega$ is an event with $\mathbb{P}(B) > 0$. For any event $A \subseteq \Omega$, the **conditional probability of A given B** is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)};$$

it is the probability that the event A occurs, that is that the outcome of the 'experiment' is in A, if it is known that the outcome is in the event B, i.e., that B occurs.

Example 2.7 What is the probability that a Bridge hand contains the ace of hearts given that it contains exactly 5 hearts? Let A be the event that the hand contains the ace of hearts and B the event that the hand contains exactly 5 hearts. Then

$$\mathbb{P}(B) = \frac{\binom{13}{5}\binom{39}{8}}{\binom{52}{13}} \quad \text{and} \quad \mathbb{P}(A \cap B) = \frac{\binom{12}{4}\binom{39}{8}}{\binom{52}{13}},$$

whence $\mathbb{P}(A \mid B) = \binom{12}{4} \Big/ \binom{13}{5} = \frac{5}{13}$. $\square$

The first thing to observe is that $\mathbb{P}(\,\cdot \mid B)$ is a probability distribution on the sample space B, because it satisfies the axioms for a probability distribution as follows:

1. For $C \subseteq B$, $\mathbb{P}(C \mid B) = \mathbb{P}(C)/\mathbb{P}(B)$, so that $0 \leqslant \mathbb{P}(C \mid B) \leqslant 1$.
2. $\mathbb{P}(B \mid B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$.
3. For disjoint events $C_1, C_2, \ldots$ in B,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i \,\middle|\, B\right) &= \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i\right)}{\mathbb{P}(B)} \\ &= \frac{\sum_{i=1}^{\infty} \mathbb{P}(C_i)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(C_i \mid B).\end{aligned}$$

Multiplication rule The next thing to notice is the multiplication rule that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B),$$

so that the probability of two events occurring can be broken up into calculating successive probabilities–firstly the probability that B occurs and then given that B has occurred the probability that A occurs. This is one of two central procedures for calculating probabilities (the second is the Law of Total Probability introduced below).

Example 2.8 Suppose that two students are selected, without replacement, from a class of 5 women and 13 men. What is the probability that the first student selected is a man and the second is a woman? Let B be the event that the first is a man and A the event that the second is a woman, then $\mathbb{P}(A \cap B) = \mathbb{P}(\, A \mid B)\mathbb{P}(B) = \frac{5}{17} \times \frac{13}{18}$. □

More generally we can write down the multiplication rule for events $A_1, \ldots, A_n$,

$$\mathbb{P}\left(A_1 \cap A_2 \cdots \cap A_n\right) = \mathbb{P}(A_1)\mathbb{P}(\, A_2 \mid A_1)\mathbb{P}(\, A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(\, A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Example 2.9 In drawing three cards without replacement from a pack what is the probability of three successive aces? Here A_i would be the event that an ace is obtained on draw i, $i = 1, 2, 3$, then

$$\begin{aligned} \mathbb{P}\left(A_1 \cap A_2 \cap A_3\right) &= \mathbb{P}(A_1)\mathbb{P}(\, A_2 \mid A_1)\mathbb{P}(\, A_3 \mid A_1 \cap A_2) \\ &= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}. \end{aligned}$$ □

Law of Total Probability A collection $\{B_i\}_{i=1}^{\infty}$ of disjoint events for which $\bigcup_{i=1}^{\infty} B_i = \Omega$ is said to be a **partition** of the sample space Ω. For any partition of the sample space, $\{B_i\}$, and for any event A, we may write

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}\left(A \cap B_i\right) = \sum_{i} \mathbb{P}\left(\, A \mid B_i\right) \mathbb{P}(B_i), \tag{2.10}$$

where the second summation extends only over those events B_i in the partition for which $\mathbb{P}(B_i) > 0$. The identity in (2.10) is known as the Law of Total Probability. It follows immediately from Axiom 3 and the multiplication rule, because the event A may be represented as $A = \bigcup_{i=1}^{\infty} (A \cap B_i)$.

Example 2.11 An urn contains b black balls and r red balls from which two balls are drawn without replacement. What is the probability that the second ball drawn is black? Let A represent the event that the second ball is black and B the event that the first ball is black. Then B and B^c form a partition of the sample space (think of the other events in the partition as being $\emptyset$). Then

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\left(A \mid B\right) \mathbb{P}(B) + \mathbb{P}\left(A \mid B^c\right) \mathbb{P}(B^c) \\ &= \left(\frac{b-1}{b+r-1}\right)\left(\frac{b}{b+r}\right) + \left(\frac{b}{b+r-1}\right)\left(\frac{r}{b+r}\right) = \frac{b}{b+r}\,.\end{aligned}$$

This probability may be seen to be the same as the probability that the second ball is black when sampling with replacement. □

Example 2.12 A survey of US voters shows the following figures for proportions of voters registered with the main parties and the proportions of voters registered for each of the parties who declare an intention to vote for Dubya.

	Registered	Proportion for Bush
Democratic	45%	10%
Republican	35%	60%
Not affiliated	20%	40%

Suppose that a voter is chosen at random, what is the probability (s)he is a Bush voter? Here the partition of the sample space is D (Democrat), R (Republican) and NA (No affiliation), and if B is the event the voter is a Bush supporter then

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \mid D)\mathbb{P}(D) + \mathbb{P}(B \mid R)\mathbb{P}(R) + \mathbb{P}(B \mid NA)\mathbb{P}(NA) \\ &= 0{\cdot}1 \times 0{\cdot}45 + 0{\cdot}6 \times 0{\cdot}35 + 0{\cdot}4 \times 0{\cdot}2 = 0{\cdot}335.\end{aligned}$$ □

Bayes' Theorem For any events A and B, for which $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}\left(B \mid A\right) = \frac{\mathbb{P}\left(A \mid B\right) \mathbb{P}(B)}{\mathbb{P}(A)}. \tag{2.13}$$

Example 2.14 In Example 2.12, given that the voter chosen is a Bush supporter, what is the probability that (s)he is a Republican?

$$\mathbb{P}(R \mid B) = \frac{\mathbb{P}\left(B \mid R\right) \mathbb{P}(R)}{\mathbb{P}(B)} = 0{\cdot}6 \times 0{\cdot}35/0{\cdot}335 \approx 0{\cdot}63.$$ □

Combining the Law of Total Probability with the statement in (2.13), gives the general statement of Bayes' Theorem:

Theorem 2.15 (Bayes' Theorem) *Suppose that $\{B_i\}_i$ is a partition of the sample space and that A is an event for which $\mathbb{P}(A) > 0$. Then for any event, B_i, in the partition with $\mathbb{P}(B_i) > 0$, we have*

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)},$$

where the summation in the denominator extends over all j for which $\mathbb{P}(B_j) > 0$.

Example 2.16 Consider a diagnostic test for some disease for which the outcome of the test is either positive, $+$, or negative, $-$, and which is 99% accurate, so that if D represents the event that the patient has the disease then

$$\mathbb{P}(+ \mid D) = 0{\cdot}99 = \mathbb{P}(- \mid D^c)\,.$$

Suppose that 0·1% of patients have the disease. A patient is chosen at random and tests positive, what is the probability that (s)he has the disease? Then from Bayes' Theorem

$$\begin{aligned}\mathbb{P}(D \mid +) &= \frac{\mathbb{P}(+ \mid D)\,\mathbb{P}(D)}{\mathbb{P}(+ \mid D)\,\mathbb{P}(D) + \mathbb{P}(+ \mid D^c)\,\mathbb{P}(D^c)} \\ &= \frac{0{\cdot}99 \times 0{\cdot}001}{0{\cdot}99 \times 0{\cdot}001 + 0{\cdot}01 \times 0{\cdot}999} \approx 0{\cdot}09\,.\end{aligned}$$

At first sight this conclusion, that the probability that someone testing positive has a somewhat low probability of having the disease, may be counter intuitive. It arises because the probability of a 'false positive', $\mathbb{P}(+ \mid D^c)$, is large compared to the probability of the disease in the population, $\mathbb{P}(D)$. Notice that

$$\mathbb{P}(D \mid +) = \frac{1}{1 + (\mathbb{P}(+ \mid D^c)\,\mathbb{P}(D^c)/\mathbb{P}(+ \mid D)\,\mathbb{P}(D))};$$

typically $\mathbb{P}(D^c)$ and $\mathbb{P}(+ \mid D)$ will be close to 1 so that $\mathbb{P}(D \mid +)$ will be large or small according as the ratio $\mathbb{P}(+ \mid D^c)/\mathbb{P}(D)$ is small or large. This becomes less mysterious when you consider a population of 1,000 patients with just one person suffering from the disease. Of the 999 not suffering from the disease there will be about 10 who will test positive so there will be about 11 in the population who will test positive; given that the person selected has tested positive then there is about a 1 in 11 chance that the person is the one suffering from the disease. A similar calculation to the above shows that $\mathbb{P}(D \mid -) \approx 0{\cdot}000001$. □

Example 2.17 *Simpson's paradox* One example of conditional probability that appears counter-intuitive when first seen is the following situation which can arise frequently. Consider one individual chosen at random from 50 men and 50 women applicants to a particular College. Figures on the 100 applicants are given in the following table indicating whether they were educated at a state school or at an independent school and whether they were admitted or rejected.

All applicants	Admitted	Rejected	% Admitted
State	25	25	50%
Independent	28	22	56%

Note that overall the probability that an applicant is admitted is 0·53, but conditional on the candidate being from an independent school the probability is 0·56 while conditional on being from a state school the probability is lower at 0·50. Suppose that when we break down the figures for men and women we have the following figures.

Men only	Admitted	Rejected	% Admitted
State	15	22	41%
Independent	5	8	38%

Women only	Admitted	Rejected	% Admitted
State	10	3	77%
Independent	23	14	62%

It may now be seen that now for both men and women the conditional probability of being admitted is higher for state school applicants, at 0·41 and 0·77, respectively. This may seem to be a surprising result in that while the aggregate data suggests that independent school applicants have a better chance of being admitted, the individual tables for both men and women show that in both cases state school applicants have a higher acceptance rate. A result of this type in tables of this sort (called **contingency tables** in statistics) is known as Simpson's paradox. Strictly, it is not a paradox in that it has an easy explanation. Note that overall women have a much higher acceptance rate (66%) than men (40%) whereas the proportion of men from state schools is 74% with 26% from independent schools while those proportions are reversed for women. This situation is known in statistics as **confounding**; it occurs when figures for two separate and different populations are aggregated to give misleading conclusions. It may be difficult or impossible to determine whether data such as that presented in the first table arise from two disparate populations and so confounding is present.

The example shows that if A, B, C are three events it is possible to have the three inequalities

$$\mathbb{P}(A \mid B \cap C) > \mathbb{P}(A \mid B \cap C^c), \quad \mathbb{P}(A \mid B^c \cap C) > \mathbb{P}(A \mid B^c \cap C^c), \quad \text{and} \tag{2.18}$$

$$\mathbb{P}(A \mid C^c) > \mathbb{P}(A \mid C), \tag{2.19}$$

holding simultaneously. In this example, A would be the event 'being admitted', B the event 'being a man' (with B^c being a woman) and C being 'state school' (with C^c being independent school). One situation where the three inequalities in (2.18) and (2.19) cannot hold simultaneously is when

$$\mathbb{P}(B \mid C) = \mathbb{P}(B \mid C^c). \tag{2.20}$$

To see this, suppose that (2.18) and (2.20) hold, then we see that

$$\mathbb{P}(A \cap B \mid C) > \mathbb{P}(A \cap B \mid C^c) \quad \text{and} \quad \mathbb{P}(A \cap B^c \mid C) > \mathbb{P}(A \cap B^c \mid C^c)$$

and then adding and observing that, for example,

$$\mathbb{P}(A \cap B \mid C) + \mathbb{P}(A \cap B^c \mid C) = \mathbb{P}(A \mid C),$$

we obtain $\mathbb{P}(A \mid C) > \mathbb{P}(A \mid C^c)$, which contradicts (2.19). In this example it would not be possible to engineer that (2.20) holds, but consider the situation where a clinical trial of a new drug for some illness is being conducted to test its effectiveness against a standard drug. Then A would be the event that a patient recovers from the illness and C would be the event that the patient receives the new drug (C^c the event that (s)he receives the standard drug); again B or B^c would correspond to the patient being a man or woman, respectively. Then to avoid the sort of situation described in this example, (2.20) would require that the trial be designed so that the relative proportions of men and women that receive the new drug are the same as the relative proportions that receive the standard drug. $\square$

2.4 Independence

We say that two events A and B are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Example 2.21 Roll a die twice and let A be the event that a 1 is obtained on the first roll and B be the event that a 5 is obtained on the second roll. The sample space $\Omega = \{(i,j) : 1 \leqslant i \leqslant 6, 1 \leqslant j \leqslant 6\}$ has 36 points, all of which have the same probability $\frac{1}{36}$. Then $A \cap B$ is just the outcome $(1,5)$ and $\mathbb{P}(A) = \frac{1}{6} = \mathbb{P}(B)$ so we can conclude that A and B are independent. $\square$

Notice that if $\mathbb{P}(B) > 0$ then A and B are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Furthermore, if A and B are independent then

(i) A and B^c are independent;

(ii) A^c and B^c are independent; and

(iii) A^c and B are independent.

For (i),

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c),$$

and (ii) and (iii) follow from (i).

We need to generalize the notion of independence to more than two events. Events $A_1, A_2, \ldots$ are independent if for all choices of $i_1 < i_2 < \cdots < i_r$, we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_r}). \tag{2.22}$$

The definition in (2.22) implies that if we take any two of the events A_i and A_j (i distinct from j) then they are independent; so the events are said to be **pairwise independent**. However, it should be noted that **pairwise independence does not imply independence** as the next example shows.

Example 2.23 Suppose that a fair coin is tossed twice (by a fair coin we mean the 4 possible outcomes HH, HT, TH and TT are equally likely and equal to $\frac{1}{4}$, so that a Head or a Tail on either toss has probability $\frac{1}{2}$). Let A_1 be the event that a head is obtained on the first toss, A_2 the event that there is a head on the second toss and A_3 the event that exactly 1 head is obtained. Then $\mathbb{P}(A_i) = \frac{1}{2}$ for $i = 1, 2, 3$. It may be seen that the events are pairwise independent since, for example,

$$\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1 \cap A_2^c) = \frac{1}{4} = \mathbb{P}(A_1)\mathbb{P}(A_3),$$

and similarly for A_2 and A_3 (and A_1 and A_2). But

$$\mathbb{P}\left(A_1 \cap A_2 \cap A_3\right) = 0 \neq \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{1}{8},$$

so the three events are not independent. □

We need to see how to model the notion of independent experiments where the outcome of one experiment does not influence the outcome of the other and see how it relates to this definition of independence. Suppose that Ω_1 and Ω_2 are the sample spaces for two experiments with probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively. Let $\Omega = \Omega_1 \times \Omega_2$ be the sample space corresponding to both experiments being performed, so that $\Omega = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$. We can define the probability distribution $\mathbb{P}$ on Ω by specifying $\mathbb{P}$ for any point (singleton) of Ω (see next section), by letting $\mathbb{P}\left(\{(\omega_1, \omega_2)\}\right) = \mathbb{P}_1(\{\omega_1\})\mathbb{P}_2(\{\omega_2\})$. If now $A_i \subseteq \Omega_i$, $i = 1, 2$, and both experiments are performed we can identify the event A_1 with the event $A_1 \times \Omega_2$ in Ω and the event A_2 with the event $\Omega_1 \times A_2$ in Ω, and the intersection $A_1 \cap A_2$ with $A_1 \times A_2$, so that

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2) &= \sum_{\omega_1 \in A_1} \sum_{\omega_2 \in A_2} \mathbb{P}\left(\{(\omega_1, \omega_2)\}\right) = \sum_{\omega_1 \in A_1} \sum_{\omega_2 \in A_2} \mathbb{P}_1\left(\{\omega_1\}\right) \mathbb{P}_2(\{\omega_2\}) \\ &= \sum_{\omega_1 \in A_1} \mathbb{P}_1\left(\{\omega_1\}\right) \sum_{\omega_2 \in A_2} \mathbb{P}_2(\{\omega_2\}) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2);\end{aligned}$$

but we then have $\mathbb{P}(A_1 \cap \Omega_2) = \mathbb{P}_1(A_1)$, and $\mathbb{P}(\Omega_1 \cap A_2) = \mathbb{P}_2(A_2)$, whence we can interpret this as $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$ which ties in with the definition of independence given above. This extends to n independent experiments in the obvious way.

2.5 Distributions

We are considering a finite or countable sample space $\Omega = \{\omega_i\}_i$, and for each point $\omega_i \in \Omega$ let $p_i = \mathbb{P}(\{\omega_i\})$ be the probability of the event consisting of the single point ω_i. Then the sequence $\{p_i\}_i$ satisfy

$$p_i \geqslant 0, \quad \text{for all} \quad i, \quad \text{and} \quad \sum_i p_i = 1, \tag{2.24}$$

because $\Omega = \bigcup_i \{\omega_i\}$. There is a one-to-one correspondence between sequences $\{p_i\}_i$ satisfying (2.24) and probability distributions as defined in this chapter through the relation $\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i$. Because of this correspondence the term **probability distribution** is also applied to sequences satisfying (2.24). We consider a number of particular distributions:

Example 2.25 *Bernoulli distribution* Consider a sample space with just two points $\Omega = \{H, T\}$, where the two points may be thought of as $H =$ "heads", and $T =$ "tails", so we are modelling a coin toss where we will take p as the probability of heads and $1-p$ as the probability of tails. Then $p = \mathbb{P}(H)$ and $1-p = \mathbb{P}(T)$, with $0 \leqslant p \leqslant 1$. □

Example 2.26 *Binomial distribution* This models the number of heads obtained in n successive tosses of the coin in the previous example. Then $\Omega = \{0, 1, 2, \ldots, n\}$ and the probability of k heads is

$$p_k = \mathbb{P}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leqslant k \leqslant n. \tag{2.27}$$

Notice that this defines a probability distribution since, by the Binomial Theorem

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1.$$

To see why this corresponds to the probability of k heads in n tosses of the coin–let $\Omega' = \{(i_1, \ldots, i_n) : i_j = H \text{ or } T\}$ be the sample space where an outcome records whether a head or a tail is obtained on each toss. Then, if the sequence $\underline{i} = (i_1, \ldots, i_n)$ represents an outcome in Ω', let $N(\underline{i})$ represent the number of indices j with $i_j = H$; that is, $N(\underline{i})$ is the number of heads in $\underline{i}$. By independence the probability of the outcome $\underline{i}$ is

$$\mathbb{P}(\{\underline{i}\}) = p^{N(\underline{i})}(1-p)^{n-N(\underline{i})}.$$

Now the number of sequences $\underline{i} \in \Omega'$ for which $N(\underline{i}) = k$, that is the number of sequences for which there are exactly k heads, is $\binom{n}{k}$ (think of choosing the k positions for the heads from the n possible positions) and hence we get (2.27). $\square$

Example 2.28 *Poisson distribution* This distribution is often used to model the number of occurrences of some event in a specified period of time, such as the number of accidents on a particular stretch of road, for example, or the number of customers who enter a particular shop. Here the probability space is $\Omega = \{0, 1, 2, \ldots\}$, the non-negative integers, and the probability of the point k is

$$p_k = \mathbb{P}(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

for some fixed $\lambda > 0$. Check that

$$\sum_{k=0}^{\infty} p_k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1. \qquad \square$$

Suppose that we consider customers entering a shop during an hour-long period, $(0, 1]$. Think of dividing the period into n segments, $((i-1)/n, i/n]$, for $i = 1, \ldots, n$, and suppose that 1 customer enters the shop in each segment with probability p, $0 < p < 1$. Then the probability that k customers enter in the hour is the binomial probability

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

Now suppose that $n \to \infty$ and $p \to 0$ in such a way that $np \to \lambda$, then we have, for each fixed k,

$$\lim_{n\to\infty}\left[\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\right] = \lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{n-k}\frac{(np)^k}{k!}\frac{n!}{n^k(n-k)!}\right] = e^{-\lambda}\frac{\lambda^k}{k!},$$

because

$$\lim_{n\to\infty}\left[\frac{n!}{n^k(n-k)!}\right]=\lim_{n\to\infty}\left[1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right]=1.$$

This has proved the following result.

Theorem 2.29 (Poisson approximation to the binomial) *Suppose that $n\to\infty$ and $p\to 0$ so that $np\to\lambda$, then*

$$\binom{n}{k}p^k\,(1-p)^{n-k}\to e^{-\lambda}\frac{\lambda^k}{k!},\quad \textit{for}\quad k=0,1,2,\ldots. \qquad \square$$

Example 2.30 *Geometric distribution* This is a distribution which models the number of tosses of a coin required to obtain the first occurrence of a head, when p, $0<p<1$, is the probability of a head on each toss; for the probability space $\Omega=\{1,2,\ldots\}$ we have

$$p_k=\mathbb{P}\,(k)=p\,(1-p)^{k-1}\quad \text{for}\quad k=1,2,\ldots.$$

Check that $\sum_1^\infty p_k=p/\left(1-(1-p)\right)=1$. Note that the term geometric distribution is also applied to the distribution on the probability space $\Omega=\{0,1,2\ldots\}$ with $p_k=p(1-p)^k$, for $k\geqslant 0$; this would be modelling the number of tails before first obtaining a head, but no confusion should arise between the slightly different usage of the term. $\square$

Example 2.31 *Hypergeometric distribution* Consider an urn with n_1 red balls and n_2 black balls of which n are drawn without replacement, $n\leqslant n_1+n_2$. The probability that there are exactly k red balls drawn is

$$p_k=\frac{\binom{n_1}{k}\binom{n_2}{n-k}}{\binom{n_1+n_2}{n}},\quad \text{for}\quad \max(0,n-n_2)\leqslant k\leqslant\min(n,n_1).$$

For example, the probability that there are exactly 5 hearts in a bridge hand is

$$\frac{\binom{13}{5}\binom{39}{8}}{\binom{52}{13}}. \qquad \square$$

9 January 2007