SMSTC (2007/08)

Probability

www.smstc.ac.uk

Contents

SMSTC (2007/08)
Probability

Lecture 2: Conditioning and independence

Stan Zachary, Heriot-Watt University[a]

www.smstc.ac.uk

Contents

2.1 Conditional probability

2.1.1 Introduction

As always we consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, which we think of as a mathematical model for some physical experiment.

Recall that for any event F, say, the probability $\mathbf{P}(F)$ measures our state of knowledge about the likelihood of the event F, and that this will change as our general state of knowledge changes.

Now let E be any event such that $\mathbf{P}(E) > 0$. Suppose we learn that the event E has in fact occurred (*but know nothing else about the outcome of the experiment*). Given this information, we can calculate a revised probability for every event F in the sample space Ω.

Definition 2.1. The *conditional probability* $\mathbf{P}(F \mid E)$ of the event F, *given* that the event E has occurred is

$$\mathbf{P}(F \mid E) = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(E)}. \tag{2.1}$$

[a] s.zachary@hw.ac.uk

The probability measure $\mathbf{P}_E$, defined by $\mathbf{P}_E(F) = \mathbf{P}(F \mid E)$ for all $F \in \mathcal{F}$ (we could just say the probability measure $\mathbf{P}(\cdot \mid E)$), represents our complete revision of all probabilities once we know the event E to have occurred. Note that it is easy to check that $\mathbf{P}_E$ *is* a probability measure, i.e. that the axiomatic properties $P1$–$P3$ of Lecture 1 are satisfied whenever $\mathbf{P}$ is replaced by $\mathbf{P}_E$.

Example 2.1. A fair coin is tossed three times. Suppose we learn that the event A that the third toss is a head has in fact occurred (and we know nothing else). What is the revised probability of the event B that exactly two heads are obtained. We require

$$\mathbf{P}(B \mid A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

This is greater than the unconditional probability of the same event, which is 3/8.

Example 2.2. A couple decide to have 2 children. Each is equally likely to be a boy or a girl, independently of the other, so that a suitable probability model is given by

$$\begin{array}{rcccc} \text{outcome}: & bb & bg & gb & gg \\ \text{probability}: & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{array}$$

Let A be the event that the first child is a girl, B be the event that at least one child is a girl, and C be the event that both children are girls. Find:

- the (conditional) probability that both children are girls, given that the first child is a girl; this is
$$\mathbf{P}(C \mid A) = \frac{\mathbf{P}(A \cap C)}{\mathbf{P}(A)} = \frac{\mathbf{P}(C)}{\mathbf{P}(A)} = \frac{1}{2};$$
- the (conditional) probability that both children are girls, given that at least one child is a girl; this is
$$\mathbf{P}(C \mid B) = \frac{\mathbf{P}(B \cap C)}{\mathbf{P}(B)} = \frac{\mathbf{P}(C)}{\mathbf{P}(B)} = \frac{1}{3}.$$

  Does this last answer seem surprising?

2.1.2 The chain rule.

The *chain rule* (also called the *multiplication rule*) states that, for any sequence of n events $E_1, E_2, \dots, E_n$,

$$\mathbf{P}(E_1 \cap E_2 \cap \cdots \cap E_n) = \mathbf{P}(E_1)\mathbf{P}(E_2 \mid E_1) \dots \mathbf{P}(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}) \qquad (2.2)$$

In the case $n = 2$, the expression (2.2) reduces to

$$\mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1)\mathbf{P}(E_2 \mid E_1).$$

which, for $\mathbf{P}(E_1) > 0$, is immediate from the definition (2.1) of conditional probability, and, for $\mathbf{P}(E_1) = 0$, is trivial (since then $\mathbf{P}(E_1 \cap E_2) = 0$ also).

The use of the chain rule is that conditional probabilities are often more easily obtained than unconditional probabilities. Indeed probability models are often most naturally specified via sequences of initially unconditional, and then conditional, probabilities.

Example 2.3. An urn contains 6 numbered balls. Two balls are chosen at random, in succession, and without replacement. What is the probability of the event that both balls chosen have an even number?

Let A be the event that the first ball chosen has an even number and let B be the event that the second ball chosen has an even number. Then, in a reasonable probability model, $\mathbf{P}(A) = 1/2$ and $\mathbf{P}(B \mid A) = 2/5$ (since, if the first ball chosen has an even number, then of the 5 remaining balls only 2 now have an even number). The required probability is now

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B \mid A) = \frac{1}{2} \times \frac{2}{5} = \frac{1}{5}.$$

This result may also be obtained by noting that, again in any reasonable probability model, each of the $6 = 30$ possible ordered sequences (n_1, n_2) of two numbered balls is equally likely to be obtained, and for $3 \times 2 = 6$ of these sequences both numbers will be even. Therefore the probability of the event that both balls chosen have an even number should again be $6/30 = 1/5$.

2.1.3 The partition rule.

Let $E_1, \dots, E_n$ be a *partition* of the sample space Ω, i.e. a collection of *disjoint* events whose union is *all* of Ω. Thus exactly one of the events $E_1, E_2, \dots, E_n$ always occurs.

Then, for any event F,

$$\begin{aligned} \mathbf{P}(F) &= \sum_{1=1}^{n} \mathbf{P}(E_i \cap F) && \text{(by the addition rule)} \\ &= \sum_{i=1}^{n} \mathbf{P}(E_i)\mathbf{P}(F \mid E_i) && \text{(by the chain rule).} \qquad (2.3) \end{aligned}$$

The result (2.3) is known as the *partition rule* or, more obscurely, *the law of total probability.* The result is illustrated in the Venn diagram of Figure 2.1.

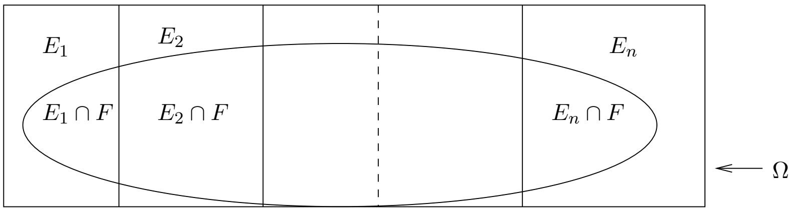

Figure 2.1: Partition rule: the rectangles correspond to the partition $E_1, \dots, E_n$ while the ellipse corresponds to the event F

Example 2.4. The probability that a randomly chosen individual will develop a particular disease is 10^{-4}. There is a test for the disease which is not entirely reliable. An individual who has the disease will test positive with probability 0.95 (and negative with probability 0.05.) An individual who does not have the disease will test positive with probability 0.1 (and negative with probability 0.9.) What is the probability that a randomly chosen individual tests positive.

Let A_1 be the event that the individual has the disease, and let A_2 be the event that (s)he does not. Note that A_1 and A_2 form a partition of Ω, and that $\mathbf{P}(A_1) = 0.0001$, $\mathbf{P}(A_2) = 0.9999$. Let B be the event that the individual tests positive. Then we require

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \mathbf{P}(A_2)\mathbf{P}(B \mid A_2) \\ &= 0.0001 \times 0.95 + 0.9999 \times 0.10 \\ &= 0.000095 + 0.099990 \\ &= 0.100085. \end{aligned}$$

Note that nearly all the positive results come from those who do not have the disease.

2.1.4 Bayes' Theorem

Again let $E_1, \dots, E_n$ be a *partition* of the sample space Ω. Suppose that the probabilities of $E_1, \dots, E_n$ are known (and sum to 1). Suppose also that some event F is observed to have occurred. *Given* this event, what are the *conditional* revised probabilities of $E_1, \dots, E_n$?

Bayes' Theorem states that, for each i,

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

Note that these revised probabilities $\mathbf{P}(E_i \mid F)$ also sum to 1.

The proof is almost immediate from the *chain rule* and *partition rule*, and is left as an exercise.

Example 2.4 (again). Suppose that, in the earlier example, the test gives a positive result. What is the probability that the individual has the disease. By Bayes' Theorem, this is

$$\begin{aligned}\mathbf{P}(A_1 \mid B) &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \mathbf{P}(A_2)\mathbf{P}(B \mid A_2)} \\ &= \frac{0.0001 \times 0.95}{0.0001 \times 0.95 + 0.9999 \times 0.10} \\ &= 0.00095.\end{aligned}$$

Thus, given a positive result, it is about 10 times as likely as before that the individual has the disease. However, the probability is still very small.

2.2 Independence

In this section we discuss the concept of probabilistic *independence*—of *events*, *random variables*, etc. However, it turns out that an *understanding* of *independence* is most naturally expressed in terms of *σ-algebras*, and we shall try to make this clear. We start with the usual (simple) concept of *independence* of *events.*

We continue to consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$

2.2.1 Independence of events

Typically events are *independent* if they depend on *physically independent* experiments or situations.

Definition 2.2. *Two* events E and F are *independent* if and only if

$$\mathbf{P}(E \cap F) = \mathbf{P}(E)\mathbf{P}(F).$$

If we compare this with the earlier *chain rule* of Section 2.1 (valid for any two events E and F) we see that, provided $\mathbf{P}(E) > 0$, the events E and F are *independent* if and only if $\mathbf{P}(F \mid E) = \mathbf{P}(F)$. This is in agreement with our intuitive understanding of *independence*, which is that *information* about whether or not E has occurred conveys *no information* about whether or not F has occurred (and conversely).

It is important to note that that if E and F are *independent*, then so also are E and F^c, and also E^c and F, and also E^c and F^c: for example, we have

$$\begin{aligned}&\mathbf{P}(E \cap F^c)\mathbf{P}(E) - \mathbf{P}(E \cap F) \\ &\qquad = \mathbf{P}(E) - \mathbf{P}(E)\mathbf{P}(F) \qquad \text{(independence of } E \text{ and } F\text{)} \\ &\qquad = \mathbf{P}(E)(1 - \mathbf{P}(F)) \\ &\qquad = \mathbf{P}(E)\mathbf{P}(F^c).\end{aligned}$$

Example 2.5. A fair die is rolled twice, in such a way that all 36 paired outcomes are equally likely. For $k = 2, \dots, 12$, let A_k be the event that the total shown on the two rolls is k. Let B be the event that the first roll shows 6. Show that the events A_k and B are independent only for $k = 7$, and give the intuitive understanding of this result.

We now extend the definition of independence to any countable number of events.

Definition 2.3. The events $E_1, E_2, E_3, \dots$ are *independent* if and only if, for every finite subset $E_{i_1}, \dots, E_{i_k}$, we have

$$\mathbf{P}(E_{i_1} \cap \dots \cap E_{i_k}) = \mathbf{P}(E_{i_1}) \dots \mathbf{P}(E_{i_k}). \tag{2.4}$$

Again this is in agreement with our intuitive understanding of *independence* of *events*, which is that *information* about whether or not each of any given number of these has occurred conveys *no information* about whether or not any of the remainder have occurred.

Note that it is easy to construct examples in which (2.4) holds for every *pair* of the given *events*, but does *not* hold for larger collections, i.e. *pairwise independence does not imply independence.*

Again *independence* of *events* implies *independence* of their *complements.*

2.2.2 Independence of σ-algebras

Recall that a *sub σ-algebra* of $\mathcal{F}$ is a collection of events within $\mathcal{F}$ which is itself a σ-algebra (i.e. is closed under the taking of complements and of countable unions of events). The simplest example is the *σ-algebra generated* by a single event E, which is $\mathcal{G}(E) = \{E, E^c, \Omega, \emptyset\}$. In general we can associate a *σ-algebra* with some precise state of *information* about the *sample space* Ω. For example, if, for a given collection of events $E_1, E_2, \dots$, we know whether or not each of them has occurred, then we also know this for every event in the *σ-algebra* $\mathcal{G}(E_1, E_2, \dots)$ *generated* by these events (the smallest σ-algebra which contains all of them). (This follows since it is easy to see that the entire collection of events in $\mathcal{F}$ about which we can say whether or not each of them has occurred is a *σ-algebra*—check it has the right closure properties; hence if this collection contains $E_1, E_2, \dots$, it necessarily contains $\mathcal{G}(E_1, E_2, \dots)$.)

The following definition is now natural.

Definition 2.4. The *σ-algebras* $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \dots$ are *independent* if and only if, for all $E_i \in \mathcal{F}_i$, $i \geq 1$,

$$\mathbf{P}\left(\bigcap_{i\geq 1} E_i\right) = \prod_{i\geq 1} \mathbf{P}(E_i). \tag{2.5}$$

By recalling that any of the events E_i in the above definition may be taken to be Ω (so that we do not formally need to consider finite sub-collections in (2.5) (in contrast to (2.4)), we see that we may equivalently say that the *σ-algebras* $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \dots$ are *independent* if and only if, for all $E_i \in \mathcal{F}_i$, $i \geq 1$, the events $E_1, E_2, E_3, \dots$ are *independent.*

By recalling also the earlier result that *independence* of *events* implies *independence* of their *complements*, we conversely have that events $E_1, E_2, E_3, \dots$ are *independent* if and only if the *σ-algebras* $\mathcal{G}(E_1), \mathcal{G}(E_2), \mathcal{G}(E_3), \dots$ which they *generate* are *independent.*

2.2.3 Independence of random variables

For any *random variable* X we may similarly define the *σ-algebra* $\mathcal{G}(X)$ *generated* by X as

$$\mathcal{G}(X) = \{X^{-1}(B),\, B \in \mathcal{B}\}, \tag{2.6}$$

where as usual $\mathcal{B}$ is the *Borel σ-algebra* on $\mathbb{R}$. (Recall from Lecture 1 that $\mathcal{G}(X)$ thus defined *is* a *σ-algebra.*) This is the collection of *events* $X^{-1}(B) = \{\omega \colon X(\omega) \in B\}$ *which may be defined* *in terms of the random variable* X, e.g. the events $X \le x$ or $a \le X \le b$ for any x, a, b. It may again be thought of as corresponding to the *information* about the *sample space* Ω given by knowledge of the value of the *random variable* X, i.e. $\mathcal{G}(X)$ is the set of *events* whose occurrence or otherwise is determined by the value of X.

Similarly, the *σ-algebra generated* by any *collection* of *random variables* is the *smallest* which contains the σ-algebras they individually generate, and has a similar interpretation.

We can now make the following definition of independence of random variables.

Definition 2.5. The (finite or countably infinite collection of) *random variables* $X_1, X_2, \dots$ are *independent* if and only if the corresponding *σ-algebras* $\mathcal{G}(X_1), \mathcal{G}(X_2), \dots$ are *independent.*

In view of the definition of *independence* of *σ-algebras*, this is equivalent to the requirement that, for all *Borel sets* $B_1, B_2, \dots \in \mathcal{B}$,

$$\mathbf{P}\Big(\bigcap_{i\ge 1}\{X_i \in B_i\}\Big) = \prod_{i\ge 1}\mathbf{P}(X_i \in B_i). \tag{2.7}$$

Thus, informally, *random variables* are *independent* if and only if they define *independent events.*

Example 2.6. Consider again the fair die which is rolled twice. The more natural way to model this is to think of it as a sequence of two *independent* experiments, in each of which each of the 6 outcomes is equally likely. Let the random variables N_1 and N_2 be the numbers obtained on the two successive rolls. Then N_1 and N_2 are *independent* random variables. Define also the total $N = N_1 + N_2$. Then, for example,

$$\begin{aligned}\mathbf{P}(N = 7) &= \sum_{i=1}^{6}\mathbf{P}(N_1 = i,\, N_2 = 7 - i)\\ &= \sum_{i=1}^{6}\mathbf{P}(N_1 = i)\mathbf{P}(N_2 = 7 - i) \qquad \text{(independence of } N_1, N_2\text{)}\\ &= \sum_{i=1}^{6}\frac{1}{6}\times\frac{1}{6} = \frac{1}{6}.\end{aligned}$$

Of course in this case, since the die is fair, all *outcomes* of the whole experiment, i.e. all *ordered pairs* of numbers which may be obtained, are equally likely, and so the above probability, and indeed the entire distribution of the random variable N, may be determined simply by *counting.*

2.2.4 Use of independence to construct probability models

The real value of independence is in the *construction* of probability models for complex experiments.

Suppose that we wish to construct a probability model for the *joint* outcome of a sequence of experiments which are *physically independent* in the sense that the outcomes of any of them do not affect the outcomes of any of the others. An example is a sequence of *independent trials* (identical experiments, such as coin tosses or repeated measurements of a physical quantity). It is natural to do this by

- specifying the probabilities associated with each *individual experiment,*
- requiring that events which depend on *separate* experiments are (probabilistically) *independent* of each other.

Example 2.7. A coin is given repeated *independent* tosses, on each of which it lands heads with probability $p \in (0,1)$. Then, in the first n tosses, the probability of obtaining a *given* sequence of heads and tails, in which the total number of heads is k, is $p^k(1-p)^{n-k}$. Thus, also, the probability that, in the first n tosses, exactly k heads are obtained is $\binom{n}{k}p^k(1-p)^{n-k}$.

We need to be sure that the above procedure leads to a *valid* probability model for the *entire sequence* of experiments. We assume the existence of a probability model $(\Omega_i, \mathcal{F}_i, \mathbf{P}_i)$ for each *individual* experiment i. Then the probability model $(\Omega, \mathcal{F}, \mathbf{P})$ for the entire sequence is given by the following *product* construction:

- The *sample space* Ω is the *product* of the individual sample spaces Ω_i, i.e. the set of ω of the form $\omega = (\omega_1, \omega_2, \dots)$ where each $\omega_i \in \Omega_i$. The sample points $\omega \in \Omega$ represent the possible *outcomes* of the entire sequence of experiments;
- We regard each $\mathcal{F}_i$ as a *σ-algebra* on Ω consisting of those *events* determined by the ith experiment. The *σ-algebra* $\mathcal{F}$ on Ω is then the *smallest* σ-algebra necessary to contain all the σ-algebras $\mathcal{F}_i$. The closure properties of the σ-algebra $\mathcal{F}$ then ensure that it contains *all* events which may reasonably be described in terms of the entire sequence of experiments.
- The *probability measure* $\mathbf{P}$ is defined by the following two requirements:
  - for each i, the probability measure $\mathbf{P}$ should assign the same probabilities as $\mathbf{P}_i$ to events in the σ-algebra $\mathcal{F}_i$, i.e. to events which depend only on the ith experiment;
  - *the σ-algebras $\mathcal{F}_i$, $i \geq 1$*, should be *independent* with respect to $\mathbf{P}$; this means that *events*, and similarly *random variables*, determined by different *individual* experiments will be *independent.*

  Standard (if somewhat tedious) arguments in measure theory show that the above two requirements define a *unique probability measure* $\mathbf{P}$ (satisfying the axioms P1–P3 of Lecture 1) on the *σ-algebra* $\mathcal{F}$ for the entire sequence of experiments.

Finally, it is usual to *complete*, if necessary, the model by requiring that $\mathcal{F}$ be extended to the minimum σ-algebra which also includes all subsets of sets of probability measure 0.

2.3 Sequences of Bernoulli trials

2.3.1 Definition

As an illustration of some of the ideas of this lecture we study sequences of *Bernoulli trials*, i.e. sequences of *independent identical trials*, each of which is a *success* with probability p, for some fixed $p \in (0,1)$, and a *failure* with probability $1-p$. Such sequences of trials form one of the fundamental models of probability theory, with many applications. We may conveniently speak of *successes* and *failures* no matter what is being modelled: thus, in the case of repeatedly tossing a coin which lands *heads* with probability p we may regard a *head* as a *success* and a *tail* as a *failure.*

The probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is defined as in Section 2.2.4. We define a sequence of *independent identically distributed* random variables $\xi_1, \xi_2, \dots$ by setting $\xi_i = 1$ if the ith trial is a *success*, and $\xi_i = 0$ if it is a *failure.* Thus

$$\mathbf{P}(\xi = 0) = 1 - p, \qquad \mathbf{P}(\xi = 1) = p.$$

Define also $S_0 = 0$, $S_n = \sum_{i=1}^n \xi_i$ for $n \geq 1$.

(A common alternative is to write $\xi'_i = 1$ or $\xi'_i = -1$ in the event of the ith trial being respectively a *success* or a *failure.* We similarly set $S'_0 = 0$, $S'_n = \sum_{i=1}^n \xi'_i$ for $n \geq 1$. Since, for the same sequence of trials, we have $\xi'_i = 2\xi_i - 1$ for all i, and thus $S'_n = 2S_n - n$ for all n, it follows that the two probability models are equivalent. The sequence of random variables $(S'_n,\, n \geq 0)$ is usually referred to as a *simple random walk*. In the case $p = 1/2$—so that, for each i, $\xi'_i = 1$ and $\xi'_i = -1$ are equally likely—the random walk is further said to be *symmetric*.)

The probability model defined by the sequence $(\xi_i,\, i \geq 1)$ is deceptively simple; the behaviour of the model, in particular of the sequence $(S_n,\, n \geq 0)$ of partial sums, is extremely subtle and frequently counter-intuitive, and it is possible to pose questions of almost arbitrary depth and difficulty about it. We shall only look at a few simple questions.

2.3.2 Behaviour of S_n

We consider first the distribution, for fixed n, of the random variable S_n. Since this is the total number of successes in a sequence of n independent identically distributed trials, we already know that S_n has a *binomial distribution* with parameters n and p, i.e. that

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \dots, n, \tag{2.8}$$

and we write $S_n \sim \mathrm{Bin}(n, k)$.

Weak law of large numbers. Given any $\epsilon > 0$,

$$\mathbf{P}\left(\left| \frac{S_n}{n} - p \right| \leq \epsilon \right) \to 1 \qquad \text{as } n \to \infty. \tag{2.9}$$

This result implies that, for large n, the *proportion* of *successes* S_n/n thus far obtained is likely to be very close to the *probability* p of a single *success*. The result may of course be obtained by careful manipulation of (2.8), but there is a simple direct proof for more general sequences of independent identically distributed random variables, the pleasure of which will be postponed to later.

Central limit theorem. A further question of interest is whether and how the distribution of $S_n/n - p$, or equivalently of $S_n - np$, may be *normalised* so as to *converge* to a non-degenerate *limit* (in some appropriate sense). For given p, once n is sufficiently large the *binomial distribution* $\mathrm{Bin}(n, k)$ is well approximated by the *normal distribution* with the same *mean* and *standard deviation*. More precisely we have the following result (the *de Moivre-Laplace central limit theorem*). For all $z \in \mathbb{R}$,

$$\mathbf{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} < z \right) \to \Phi(z) \qquad \text{as } n \to \infty, \tag{2.10}$$

where Φ, given by

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{z} e^{-u^2/2}\, du,$$

is the distribution function of the *normal distribution* with *mean* 0 and *standard deviation* 1. This result may also be obtained by careful manipulation of (2.8). However, it will again be considered and proved, for very much more general sequences of independent identically distributed random variables, in a later lecture.

The *central limit theorem* implies in particular that *deviations* of S_n from its *mean* np are typically of order $n^{1/2}$. It is a stronger result than the *weak law of large numbers* which is a trivial consequence of it.

A considerably deeper result (of which the proof is yet again postponed) is the following.

Strong law of large numbers.

$$\mathbf{P}\left(\frac{S_n}{n} \to p \text{ as } n \to \infty\right) = 1. \tag{2.11}$$

(we usually say $S_n/n \to p$ *almost surely* (*a.s.*) as $n \to \infty$).

It is important to understand the relationship between the conclusions of the *weak* and *strong laws of large numbers.* For each $\epsilon > 0$ and for each positive integer n, define the events

$$\begin{aligned} A_n^\epsilon &= \left\{ \left| \frac{S_{n'}}{n'} - p \right| \le \epsilon \text{ for all } n' \ge n \right\}, \\ B_n^\epsilon &= \left\{ \left| \frac{S_n}{n} - p \right| \le \epsilon \right\}. \end{aligned}$$

The *strong law of large numbers* (2.11) is equivalent to the assertion that

$$\mathbf{P}\left(\bigcap_{\epsilon>0} \bigcup_{n \ge 1} A_n^\epsilon \right) = 1, \tag{2.12}$$

and since $\bigcup_{n\ge1} A_n^\epsilon$ is decreasing as ϵ decreases, (2.12) is further equivalent to the assertion that

$$\mathbf{P}\left(\bigcup_{n \ge 1} A_n^\epsilon \right) = 1, \qquad \text{for all } \epsilon > 0. \tag{2.13}$$

Since, for each fixed ϵ, A_n^ϵ is increasing as n increases, it finally follows from (2.13) that the *strong law of large numbers* is equivalent to the assertion that

$$\lim_{n\to\infty} \mathbf{P}\left(A_n^\epsilon\right) = 1 \qquad \text{for all } \epsilon > 0. \tag{2.14}$$

Since, for each ϵ and each n, we have that $A_n^\epsilon \subseteq B_n^\epsilon$, it follows from the *strong law of large numbers* that

$$\lim_{n\to\infty} \mathbf{P}\left(B_n^\epsilon\right) = 1 \qquad \text{for all } \epsilon > 0. \tag{2.15}$$

which is just the *weak law of large numbers.* The conclusion of the *strong law* cannot, however, be deduced from that of the *weak law.*

Interpretation of strong law of large numbers. The strong law of large numbers says that, in repeated *independent trials*, in each of which an event occurs with *probability* p, the long term *frequency* of occurrence of that event tends to p—at least with probability 1 (we say *almost surely*). Thus, for example, in repeated *independent* tosses of a coin which lands heads with probability p, the long-term proportion of heads will *always* tend to p, no matter how perverse the results of any initial sequence—however long—of the tosses. This is one interpretation of what it means to say that the coin lands heads with *probability* p.

In practice, if we took a coin we thought to be fair ($p = 1/2$) and then obtained 80 heads in the first 100 tosses of that coin, we would regard this as strong evidence that, after all, $p > 1/2$. We would therefore expect that, in further tosses of the coin, the number of heads would continue to considerably exceed the number of tails. But this is a *statistical* issue, arising because we do not know for certain the *true* value of p. In *probability theory* we reason as if probability models were known precisely. In the real world, such reasoning is an essential component of making *inference* in the presence of uncertainty.

2.3.3 Time to the occurrence of a given pattern

We think of our Bernoulli trials as occurring sequentially at times $n = 1, 2, \dots,$ and ask about (the distribution of) how long we have to wait until we observe some given pattern. Note that a general recipe for answering this question for any pattern is quite sophisticated (but an attractive example of the *martingale* theory of later lectures). We give here some arguments for particular examples.

Time to the first occurrence of '1'. The simplest such question is that of the distribution of the random variable T_1 defined to be the first time n such that $\xi_n = 1$, i.e.

$$T_1 = \min\{n \ge 1 \colon \xi_n = 1\}.$$

Clearly we have that, for all $n \ge 1$,

$$\begin{aligned}\mathbf{P}(T_1 = n) &= \mathbf{P}(\xi_1 = 0, \dots, \xi_{n-1} = 0, \xi_n = 1)\\ &= \mathbf{P}(\xi_1 = 0) \dots \mathbf{P}(\xi_{n-1} = 0)\mathbf{P}(\xi_n = 1) \qquad \text{(independence)}\\ &= (1-p)^{n-1}p,\end{aligned}$$

i.e. T_1 has a *geometric* distribution with parameter p (we write $T \sim \mathrm{Geo}(p)$).

Time to the first occurrence of '11'. Define the random variable

$$T_{11} = \min\{n \ge 2 \colon \xi_{n-1} = 1, \xi_n = 1\}.$$

For $n \ge 1$, define $p_n = \mathbf{P}(T_{11} = n)$. Then $p_1 = 0$ and $p_2 = p^2$. For $n \ge 3$, in order for the event $T_{11} = n$ to have any possibility of occurring, we must either have $\xi_1 = 0$ or else $\xi_1 = 1$, $\xi_2 = 0$; hence, by the *partition rule*,

$$\begin{aligned}p_n &= \mathbf{P}(T_{11} = n)\\ &= \mathbf{P}(\xi_1 = 0)\mathbf{P}(T_{11} = n \,|\, \xi_1 = 0) + \mathbf{P}(\xi_1 = 1, \xi_2 = 0)\mathbf{P}(T_{11} = n \,|\, \xi_1 = 1, \xi_2 = 0)\\ &= \mathbf{P}(\xi_1 = 0)\mathbf{P}(T_{11} = n-1) + \mathbf{P}(\xi_1 = 1, \xi_2 = 0)\mathbf{P}(T_{11} = n-2)\\ &= (1-p)p_{n-1} + p(1-p)p_{n-2},\end{aligned}$$

where the third line above follows since, in the case of the occurrence of either of the events $\xi_1 = 0$ or $\xi_2 = 0$, we effectively restart the process of waiting for the given pattern. Hence we obtain a simple linear recurrence relation, enabling the determination of successive values of p_n. As usual, p_n has an explicit solution in terms of the roots of a quadratic equation, but this is not very enlightening.

Time to the occurrence of the first k '1's. We now consider something slightly different, the time T_k to the occurrence of the first k '1's, without requiring that they should all occur together. For $n \ge k$, in order for T_k to be equal to n, we require that $\xi_n = 1$ and that, of the random variables $\xi_1, \dots, \xi_{n-1}$, exactly $k - 1$ of them should be equal to 1. We thus have

$$\mathbf{P}(T_k = n) = \mathrm{Bin}(n-1, k-1)p^k(1-p)^{n-k}. \tag{2.16}$$

The random variable T_k is said to have a *negative binomial distribution* with *parameters* k and p.

2.4 Exercises

2–1. A letter is placed in a desk with some probability $p < 1$, and is then equally likely to be placed in any one of 8 drawers in the desk. A total of 7 of the drawers are opened, and the letter is not found in any of them. Given this event, what is the probability that the letter is in the 8th drawer? *[Answer:* $p/(8 - 7p)$*.]*

2–2. An urn contains 5 red and 3 green balls. Three balls are chosen at random, in succession, and without replacement.

(a) Show that the probability that all three balls drawn are red is 5/28.
(b) Use the chain and partition rules to show that the probability that the third ball drawn is red is 5/8. Can you give also a quick derivation of this result?
(c) Show that the (conditional) probability that the third ball drawn is red, given that the first ball drawn is red, is 4/7; show also that the (conditional) probability that the first ball drawn is red, given that the third ball drawn is red, is 4/7. Can you draw any further interesting conclusions?

2–3. An urn contains 12 red, 8 green and 10 blue balls. Two balls are chosen at random without replacement (the order of their choice being irrelevant). Given the event that they have different colours, find the conditional probabilities that

(a) neither ball is blue *[answer:* 12/37*]*;
(b) at least one ball is blue *[answer:* 25/37*]*.

2–4. Of three cards, one is red on both sides, one is black on both sides, and one is red on one side and black on the other. A card is chosen at random (each choice being equally likely) and then placed flat so that either side is equally likely to show. Given that a red side shows, what is the (conditional) probability that the other side shows red? *[Answer:* 2/3*.]*

2–5. *[Chung.]* Telegraphic signals *dot* and *dash* are sent in the proportion 3 : 4. Due to conditions causing very erratic transmission a *dot* becomes a *dash* with probability 1/4, whereas a *dash* becomes a *dot* with probability 1/3. If a *dot* is received, what is the (conditional) probability it is sent as a *dot*? *[Answer:* 27/43*.]*

2–6. A tells the truth with probability 3/4, while B tells the truth with probability 4/5. From an urn containing 8 black and 1 white balls, 1 ball is selected at random. Given that both A and B declare the ball to be white, find the (conditional) probability that it *is* white. *[Answer:* 3/5*.]*

2–7. A fair die is rolled twice, the successive rolls being independent. Define the following events:

$$\begin{aligned} A &= \{\text{the first roll shows an odd number}\} \\ B &= \{\text{the second roll shows an even number}\} \\ C &= \{\text{the sum of the two numbers obtained is odd}\} \end{aligned}$$

Show that the events A, B, and C are *pairwise independent* but are not *independent.*

2–8. Show that *disjoint* events A and B are *independent* if and only if either $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 0$.

2–9. A system contains 5 components and is such that, for $1 \leq i \leq 5$, component i works with probability p_i. The components work, or fail to work, independently of each other. Find the probability that the system works under each of the following sets of conditions:

(a) the system works if and only if all five components work *[answer:* $\prod_{i=1}^{5} p_i$*]*;

(b) the system works if and only if at least one component works *[answer:* $1 - \prod_{i=1}^{5}(1 - p_i)$*]*;

(c) the system works if and only if components 1, 2, 3 and 4 all work *or* components 3, 4 and 5 all work *[answer:* $p_3p_4(1 - (1 - p_1p_2)(1 - p_5))$*]*.

Now consider any rule whereby whether or not the system works is a function of whether or not each of the components works, provided only that it satisfies the following natural condition: for each component i, if, for any given combination of states of the remaining components, the entire system works when component i is *not* working, then it also works when component i *is* working. Show that the probability that the entire system works is an increasing function of each p_i.

2–10. A standard pack of 52 playing cards contains 4 aces. A bridge player is dealt 13 cards which may be considered as being randomly chosen from the 52. Show that the probability that (s)he is dealt no aces is 0.01279. Find also the probability that, in 3 successive independent deals, no aces are ever obtained. *[Answer:* 2.093×10^{-6}*.]*

2–11. Five teams play in a football tournament. Every team plays every other team. Each match is equally likely to be won by either team (a draw cannot happen) *independently* of the outcome of the other matches. Find the probability of the (one) event that, at the end of the tournament, each team has won precisely two matches. *[Answer:* $3/128$*]*.

2–12. Six points are chosen *uniformly* at random, and *independently* of each other, from the interval $[0, a]$.

(a) For $0 < b < a$, find the probability that 2 of the points are less than (or equal to) b, while 4 are greater than b. *[Answer:* $\binom{6}{2}\frac{b^2(a-b)^4}{a^6}$*.]*

(b) For $0 < b < c < a$, find the probability that 2 of the points are less than (or equal to) b, 1 is between b and c, while 3 are greater than c. *[Answer:* $\frac{6!}{2!1!3!}\frac{b^2(c-b)(a-c)^3}{a^6}$*.]*

2–13. Consider 4 cities A, B, C, D. The following pairs are connected by roads: A with B, B with D, D with C, C with A, and B with C. Each road is independently blocked by snow with probability p.

(a) Show that the probability that it is possible to travel by road from A to D is given by $(1 - p)^2(1 + 2p + p^2 - 2p^3)$.

(b) Show that, conditional on the event that it is possible to travel from A to D, the probability that the road BC is not blocked is given by $\frac{(1-p)(1+p)^2}{1+2p+p^2-2p^3}$.

2–14. A certain hereditary characteristic has associated with it (as usual) the three *genotypes* (*gene* pairs) AA, Aa, aa. Assume that in a large population these exist in the proportions $u : 2v : w$ where $u > 0$, $v > 0$, $w > 0$ and $u + 2v + w = 1$. Under *random mating*, a randomly chosen individual (for which $\mathbf{P}(AA) = u$, $\mathbf{P}(Aa) = 2v$, $\mathbf{P}(aa) = w$) mates with another such individual whose genotype is independent of the first. The genotype of the offspring is obtained by combining one ‘letter’ (gene) chosen at random from the genotype of one parent—either choice being equally likely—with one ‘letter’ (gene) chosen independently at random from the genotype of the other parent. (Thus, for example, if both parents have genotype Aa, the genotype of the offspring will be AA, Aa, aa with respective probabilities 1/4, /2, 1/4.) Suppose now that the initial population dies out and is replaced by its first generation offspring. Show that the genotypes in this first generation population are in the proportions $p^2 : 2pq : q^2$, where $p = u + v$, $q = v + w$. Deduce *carefully* that these proportions are maintained in all subsequent generations.

2–15. Consider a sequence of *independent* Bernoulli trials in which on each occasion the probability of obtaining a '1' is 1/2. Let T_{01} be the time to the first occurrence of the pattern '01', i.e.

$$T_{01} = \min\{n \geq 2 \colon \xi_{n-1} = 0,\, \xi_n = 1\}.$$

Show that, for all ≥ 1, $\mathbf{P}(T_{01} > n) = (n+1)/2^n$, and deduce that $\mathbf{P}(T_{01} = n) = (n-1)/2^n$.

Show that it is easier to obtain the pattern '01', than to obtain the pattern '11', in the sense that if T_{11} is the random variable defined in Section 2.3.3, then, for all $n \geq 2$,

$$\mathbf{P}(T_{01} \leq n) \geq \mathbf{P}(T_{11} \leq n).$$

[This is tricky.] Understand intuitively why this is so.

2–16. *[Problem for research.]* Consider the simple symmetric random walk $(S'_n,\, n \geq 0)$ defined in Section 2.3.1 (recall that here, for all i, $\mathbf{P}(\xi'_i = 1) = \mathbf{P}(\xi'_i = -1) = 1/2$). Show that, for all integer $n \geq 1$,

$$\mathbf{P}(S'_1 \neq 0, \ldots, S'_{2n} \neq 0) = \mathbf{P}(S'_{2n} = 0).$$