Chapter 2

Probability Theory for the Coalescent

The aims of this chapter are somewhat more modest than the title would imply. There is a large body of abstract probability theory which one needs in order to do any fundamental mathematical work in coalescent theory. Even Kingman's (1982a) proof of the standard coalescent model is a little too technical for us here. The focus instead is to develop, starting with some very basic notions, some of the less abstract ideas from probability theory, which in turn can be used to frame the coalescent. This culminates, in this chapter, with an introduction to Poisson processes, while Chapter 5 continues with an introduction to the theory of Markov processes. There are two goals. The first is to generate an intuitive understanding of coalescent processes so that, at the very least, the primary literature can be read without too much discomfort. The second goal is to build sufficient expertise in these methods that anyone who masters them can begin to address problems of biological interest using theory. The presentation is informal, and references are made to the phenomena we will encounter in subsequent chapters. Readers are encouraged to have a probability textbook nearby, such as Ross (2000) which contains all of what follows here.

2.1 Fundamentals of Probability Theory

2.1.1 Events, Probabilities, and Random Variables

Any stochastic process we might be interested in modeling will have a sample space. This is the world of all possible outcomes, *i.e.* if we could observe the process in an experiment. In population genetics, the processes we need to model include mutation, recombination, migration, and the randomness of reproduction, and what we observe are samples of genetic data. An everyday example of a stochastic process is the toss of a coin. The sample space of a coin toss is heads or tails; no other outcomes are possible. Each outcome has associated with it a probability, which is simply the chance that that event will happen. The probability of getting tails when a fair coin is tossed is equal to 1/2 or 0.5. Another example is the roll of a die, in which the probability of rolling any particular number, for instance a four, is equal to 1/6 or about 0.167. On average, we expect that one out of every two times we flip a coin it will come up tails, and one out of every six times we roll a die it will show a four.

The probability of an event $\{A\}$ which will be denoted $P\{A\}$, or sometimes just p, must be between zero and one inclusive: $0 \leq p \leq 1$. An event is a subset of the sample space. If $p = 1$ then we are certain that the event in question will occur, and if $p = 0$ we are certain that it

won't. We represent the sample space as a set. For instance the sample space of a dice roll is $\{1, 2, 3, 4, 5, 6\}$. The total probability of all the possible outcomes that make up the sample space must be equal to one. To calculate the probability of observing an event that consists of different possible outcomes, we simply sum the probabilities of the individual outcomes. For example, the chance of rolling either a two or a six with one throw of a die is

$$P\{2 \cup 6\} = P\{2\} + P\{6\} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}. \tag{2.1}$$

The symbol $\cup$ denotes the union of two events and can be read as "or." More complicated experiments, such as when a die is rolled and a coin is tossed, necessitate additional concepts. To calculate the probability of observing more than one event when events can co-occur and they are independent, we simply multiply the probabilities of the events to obtain

$$P\{tails \cap 2\} = P\{tails\} \times P\{2\} = \left(\frac{1}{2}\right)\left(\frac{1}{6}\right) = \frac{1}{12}. \tag{2.2}$$

The definition of independent is that the outcome of one event does not influence the outcome of the other event. The result is that the individual probabilities are multiplied together to get the joint probability, as in equation 2.2. The symbol $\cap$ denotes the intersection of events and can be read as "and." The sample space associated with this experiment contains twelve possible outcomes as there are two ways for the coin and six ways for the die to fall, and $2 \times 6 = 12$.

For events that are non-independent we bring in the notion of conditional probability. Consider the following experiment. We have two dice, one numbered one through five and a regular one numbered one through six. A coin toss will determine which one we roll: heads, the six-side die; tails, the five-sided die. Now the probability of rolling a two is conditional on whether the coin comes up tails or heads. Using $\{A|B\}$ to mean "$\{A\}$ given $\{B\}$," equation 2.2 becomes

$$P\{tails \cap 2\} = P\{tails\} \times P\{2|tails\} = \left(\frac{1}{2}\right)\left(\frac{1}{5}\right) = \frac{1}{10}, \tag{2.3}$$

and similarly we have

$$P\{heads \cap 2\} = P\{heads\} \times P\{2|heads\} = \left(\frac{1}{2}\right)\left(\frac{1}{6}\right) = \frac{1}{12}. \tag{2.4}$$

Conditional probabilities are calculated using the formula

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}, \tag{2.5}$$

which can be seen in equations 2.3 and 2.4. Thus, when events are non-independent, we must use $P\{A \cap B\} = P\{A|B\}P\{B\}$, but for independent events we use $P\{A \cap B\} = P\{A\}P\{B\}$, as in 2.2, because in this case $P\{A|B\} = P\{A\}$.

The fundamental formula 2.5 is nicely illustrated by a graphical representation of the sample space. Figure 2.1 is drawn so that the areas of the different possible outcomes are equal to their probabilities. Thus, the total area of the sample space is 1.0. One half of this is *heads* and the other half is *tails*, the result of the coin toss. Because a six-sided die is rolled when the coin comes up heads, the left half is divided into six pieces of equal area. For simplicity, only the event that a two is rolled is shown. Thus, one sixth of one half of the sample space (see equation 2.3) is contained within the left-hand portion of the oval. By similar logic, the

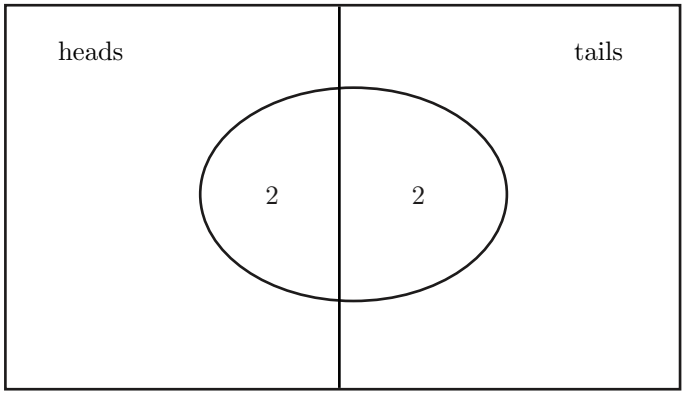

Figure 2.1: The sample space for the experiment discussed in the text.

right-hand portion of the oval occupies one tenth of the entire sample space (equation 2.4). For the sake of illustration, we can use equation 2.5 to compute something we already know, namely that the probability of rolling a two, given tails, is equal to 1/5:

$$P\{2|tails\} = \frac{P\{2 \cap tails\}}{P\{tails\}} = \frac{1/10}{1/2} = \frac{1}{5}.$$

In figure 2.1, this is the proportion of the tails side of the graph that is contained within the oval.

The overall or unconditional probability of rolling a two in this experiment is just the total area of the oval. This is obtained by summing the areas of the left-hand portion and the right-hand portion, *i.e.* $1/12 + 1/10 = 11/60$ which is midway between one fifth and one sixth. More formally, we use

$$P\{A\} = \sum_{i=1}^{k} P\{A \cap B_i\} \tag{2.6}$$

$$= \sum_{i=1}^{k} P\{A|B_i\}P\{B_i\} \tag{2.7}$$

to calculate the overall probability of event $\{A\}$ when the entire sample space is divided up into k non-overlapping events $\{B_i; i = 1, \ldots, k\}$. Equation 2.6 is called the law of total probability.

Finally, note that equation 2.1 is incorrect for two events that are not mutually exclusive. For overlapping events, we instead use

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}. \tag{2.8}$$

For example, in the experiment depicted in figure 2.1, we know that $P\{heads\} = 1/2$ and we just computed $P\{2\} = 11/60$. However, the sum of these is greater than $P\{heads \cup 2\}$ because both events $\{heads\}$ and $\{2\}$ contain the event $\{heads \cap 2\}$. Equation 2.8 corrects for this by substracting off the probability, or area in the figure, that was counted twice.

Random Variables, Distributions, and Moments

A random variable takes different values by chance. The result of a coin toss is a random variable. It takes the value *heads* with probability 1/2 and it takes the value *tails* also with probability 1/2. A random variable encodes the sample space as a set of numbers, and also records the probability of each. For example, if $X = 1$ if the coin shows *heads* and $X = 0$ if the coin shows *tails*, then X is a real-valued random variable and $P\{X = 1\} = P\{X = 0\} = 1/2$. The probabilities of all the possible values of a random variable together are called the probability function of the random variable. Random variables comes in two types: discrete and continuous. A discrete random variable assumes particular values with certain probabilities, as with a coin toss. A typical range for a discrete random variable is the integers, $1, 2, 3, \ldots$ Continuous random variables can vary infinitesimally, so the chance that they take on any particular value is equal to zero. This may seem odd, since then naive reasoning would imply that the total probability is equal to zero, but all it means is that we must consider the probability that a continuous random variable takes on values in a particular interval. A typical range for a continuous random variable is the positive real numbers: $(0, \infty)$, and examples will follow.

First consider discrete random variable X that can assume the values $x_1, x_2, \ldots$ The probability function is given by $P\{X = x_i\}$, or simply $p(x_i)$, and by definition must sum to one over the entire range

$$\sum_{i=1}^{\infty} p(x_i) = 1. \tag{2.9}$$

We picture the probability function of a discrete random variable with a histogram (figure 2.2). The heights of the bars of a histogram are the probability values associated with each possible value of the random variable. If we consider the width of each bar to be equal to one, then the total area of the bars sums to one. The histogram is simply another way to represent the probability function and the sample space, different only in format from depictions like figure 2.1.

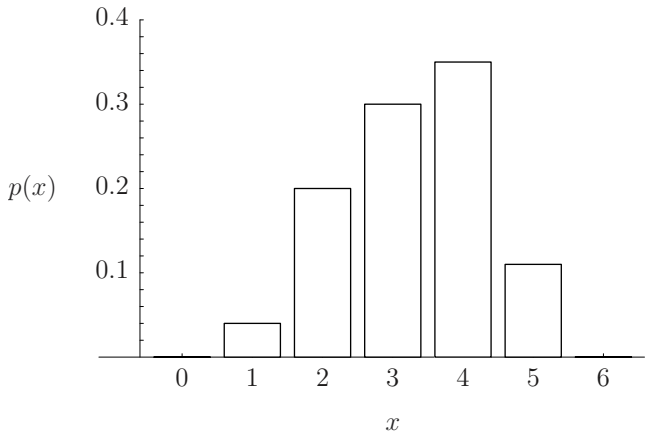

Figure 2.2: A histogram of the probability function of a discrete random variable. The heights of the bars — $0.04, 0.2, 0.3, 0.35$, and 0.11 — represent the probabilities that X is equal to 1, 2, 3, 4, and 5, respectively.

Random variables are often characterized by their moments, which summarize properties of the probability function. The expected value of a random variable X is its average over the

entire distribution, *i.e.* with its values weighted by the probability function:

$$E[X] = \sum_{i=1}^{\infty} x_i p(x_i). \tag{2.10}$$

This is sometimes referred to simply as the expectation of X or the mean of X. Roughly speaking, it is what we expect to see on average. More specifically, if we make a very large number of observations of the random variable, the mean (arithmetic average) of those observations will be close to the expected value. The variance of X is defined to be

$$\mathrm{Var}[X] = \sum_{i=1}^{\infty} (x_i - E[X])^2 p(x_i) \tag{2.11}$$

and is a measure of dispersion about the mean. It is the average of the squared deviations from the mean. Thus, the variance is always greater than or equal to zero, and is smaller when the distribution is more concentrated around the mean and larger when the probabilities associated with values far from the are higher. The mean and the variance are not sufficient to fully describe a probability distribution, but they do allow for the rough comparison of distributions. Very often, the symbol μ is used to denote the mean and the symbol σ^2 is used to denote the variance.

Returning to the example of the coin toss, let X be the number of heads when a coin is tossed. Thus X assumes the values 0 and 1 with $p(0) = p(1) = 1/2$. The mean of X is, using equation 2.10,

$$E[X] = (0)\frac{1}{2} + (1)\frac{1}{2} = \frac{1}{2}$$

and the variance of X is

$$\mathrm{Var}[X] = \left(0 - \frac{1}{2}\right)^2 \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \frac{1}{2} = \frac{1}{4}.$$

The toss of a fair coin is a special case of a Bernoulli random variable. In general, a Bernoulli random variable assumes the value 1 with probability p, and the value 0 with probability $1 - p$. The mean of a Bernoulli random variable is p, and its variance is $p(1 - p)$.

Now suppose that X is a continuous random variable such that $-\infty < x < \infty$. Then the probability function is given by $P\{x < X \leq x + dx\} = f_X(x)dx$. We are now in the realm of infinitesimal calculus. Although $f_X(x)$ is a continuous function, we can still think of $f_X(x)dx$ as a bar in a histogram: $f_X(x)$ is its height, dx is its width, and $f_X(x)dx$ is the total area of the bar, which is equal to the probability that X assumes a value between x and $x + dx$. For continuous random variables we call $f_X(x)$ the probability density function. The probability that X is between a and b is given by the integral

$$P\{a < X \leq b\} = \int_a^b f_X(x)dx.$$

The total probability, or the total area under the curve $f_X(x)$, must be equal to one:

$$\int_{-\infty}^{\infty} f_X(x)dx = 1,$$

which is the continuous version of equation 2.9.

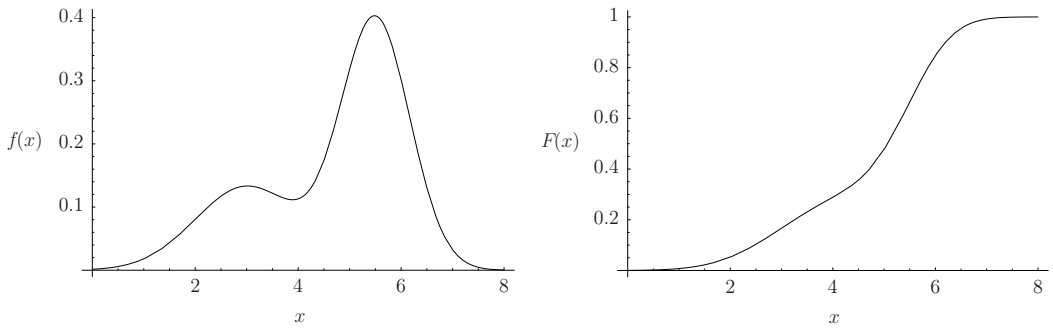

Figure 2.3: An example probability density function of a continuous random variable.

Random variables can be defined in terms of their cumulative distribution functions, or simply distribution functions, rather than by their probability densities. The distribution function of a continuous random variable X is equal to $P\{X \leq x\}$, or

$$F_X(x) \quad = \quad \int_{-\infty}^{x} f_X(x)dx.$$

From the fundamental theorem of calculus, $dF_X(x)/dx = f_X(x)$, and it is clear that we can study X using $F_X(x)$ rather than $f_X(x)$ is we choose. Importantly, we have

$$P\{a < X \leq b\} \quad = \quad \int_{a}^{b} f_X(x)dx \quad = \quad F_X(b) - F_X(a).$$

Figure 2.3 depicts the probability density function and the distribution function of a continuous random variable. The total area under the curve in the left panel, which depicts the density, is equal to one and probability that the random variable is in the vicinity of x is proportional to the height of the curve above each value on the x-axis. The right panel depicts the distribution function for this same density. It is common, as in figure 2.3, to omit the subscript X when it is understood.

The moments of a continuous random variable are computed as would be expected. The mean and the variance of X are given by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx, \tag{2.12}$$

and

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)dx. \tag{2.13}$$

We can compute the expected value of any function $g(X)$ of a random variable. For the continuous case, we have

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx. \tag{2.14}$$

Some very useful rules can be derived from equation 2.14, or its analogue for a discrete random variable, when $g(x)$ involves simple multiplication or addition of a constant number c:

$$E[c] = c, \tag{2.15}$$

$$E[cX] = cE[X], \tag{2.16}$$

$$E[X + c] = E[X] + c, \tag{2.17}$$

$$\mathrm{Var}[c] = 0, \tag{2.18}$$

$$\mathrm{Var}[cX] = c^2\mathrm{Var}[X], \tag{2.19}$$

$$\mathrm{Var}[X + c] = \mathrm{Var}[X]. \tag{2.20}$$

We can also use equation 2.14 to calculate the moments of the distribution. The r-th moment about zero is defined to be

$$E[X^r] = \int_{-\infty}^{\infty} x^r f_X(x)dx.$$

The first moment ($r = 1$) is just the expected value (equation 2.12). The variance is the second central moment, or the second moment about the mean. It can be expressed in terms of the first and second moments about zero

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 \tag{2.21}$$

which is valid for both discrete and continuous random variables. It is sometimes easier to calculate the variance using equation 2.21 than 2.11 or 2.13.

In many problems we are confronted with not just one, but two or more random variables. In coalescent theory, for example, we often need to model both genealogies and mutation. In fact, the genealogy itself is typically broken up into a series of coalescence times, which are first understood individually, then treated jointly. Equations 2.2 and 2.5 established the rules for computing the probabilities of joint events involving two discrete random variables, depending on whether the events are independent or non-independent. In addition, equations 2.6 and 2.7 show how the probability of an event can be computed using joint probabilites, directly in equation 2.6 or via conditional probabilities in equation 2.7. These equations can be applied directly discrete random variables. For a pair of continuous random variables X_1 and X_2, we have the joint density $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$ if they are independent, and $f_{X_1,X_2}(x_1, x_2) = f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2)$ in general. Finally, we have

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)dx_2 \tag{2.22}$$

$$= \int_{-\infty}^{\infty} f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2)dx_2 \tag{2.23}$$

for the unconditional or marginal density of a random variable, which correspond to equations 2.6 and 2.7.

The covariance of two random variables X_1 and X_2 is a joint moment which quantifies their non-independence. For the present case of two continuous random variables, we have

$$\begin{aligned} \text{Cov}[X_1, X_2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 - E[X_1])(x_2 - E[X_2])f_{X_1,X_2}(x_1, x_2)dx_1dx_2 \\ &= E[X_1X_2] - E[X_1]E[X_2], \end{aligned} \tag{2.24}$$

while for discrete random variables, we would simply replace the integral above with a sum. The covariance measures the association of the two random variables or the ability to predict values of one of them from values of the other. A positive covariance means that as X_1 increases, on average so does X_2. A negative covariance means that as X_1 increases, X_2 tends to decrease. The covariance of two independent random variables is equal to zero. To the useful identities, 2.15 through 2.20, we can add

$$E[X_1 + X_2] = E[X_1] + E[X_2], \tag{2.25}$$

$$\text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2] + 2\text{Cov}[X_1, X_2], \tag{2.26}$$

and note that the variance of the sum of two independent random variables is simply equal to the sum of the variances.

Sums of Random Variables

Many problems in coalescent theory can be represented using sums of random variables. For example, in Chapter 3 we will see that both the time to the most recent common ancestor of a sample and the total lengths of the genealogy of a sample are sums of random variables. If $Y = X_1 + X_2 + \ldots + X_k$, then using the theory above, and analogous to 2.25 and 2.26, it is not difficult to show that

$$E[Y] = \sum_{i=1}^{k} E[X_i], \tag{2.27}$$

$$\text{Var}[Y] = \sum_{i=1}^{k} \text{Var}[X_i] + \sum_{i=1}^{k}\sum_{j\neq i} \text{Cov}[X_i, X_j]. \tag{2.28}$$

Thus, regardless of whether the X_i are independent of one another, the expectation of the sum is equal to the sum of the individual expected values. However, the variance of Y does depend on whether the X_i are correlated. If they are independent of one another, then $\text{Cov}[X_i, X_j] = 0$ and equation 2.28 becomes simply

$$\text{Var}[Y] = \sum_{i=1}^{k} \text{Var}[X_i]. \tag{2.29}$$

That is, the variance of the sum of independent random variables is equal to the sum of the variances of its individual components.

In the case where the X_i are independent, the distribution of their sum can be found using a convolution. Consider the case of the sum of two random variables: $Y = X_1 + X_2$. In words, a convolution computes the probability that Y takes a particular value y by adding up the

probabilities of all possible pairs of values (x_1, x_2) such that $x_1 + x_2$ is equal to y. In the discrete case we have

$$P\{Y = y\} = \sum_i P\{X_1 = i\}P\{X_2 = y - i\},$$

and if X_1 and X_2 are continuous random variables, we have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(y - x_2) f_{X_2}(x_2) dx_2. \tag{2.30}$$

The distribution of $Y = X_1 + X_2 + \ldots + X_k$ can be found by applying these equations repeatedly. As we will see, equation 2.30 provides is a straightforward way to derive some of the fundamental results of coalescent theory.

A slightly more complicated case of a sum of random variables is when the number of items summed is itself a random variable. For instance, as we will see in Chapter 4, the number of mutations in the history of a sample is equal to the (random) number of mutations per generation summed over the total length of the history of the sample, which is itself a random variable. Thus, now we have $Y = X_1 + X_2 + \ldots + X_K$ where K is a random variable. If $X_1, X_2, \ldots, X_k$ are independent and identically distributed (i.i.d.), which means that the have the same probability function and are independent, then

$$E[Y] = E[K]E[X_i], \tag{2.31}$$

$$\mathrm{Var}[Y] = E[K]\mathrm{Var}[X_i] + \mathrm{Var}[K]E[X_i]^2. \tag{2.32}$$

These equations can be derived by conditioning on K, that is using equation 2.7 together with the equations 2.27 and 2.28 for sums of fixed numbers of random variables. Since this illustrates the very useful technique of conditioning on a random variable, it is shown in some detail:

$$\begin{aligned} E[Y] &= \sum_Y Y P\{Y\}, \\ &= \sum_Y Y \sum_K P\{Y|K\}P\{K\} \\ &= \sum_K E[Y|K]P\{K\} \\ &= \sum_K K E[X_i]P\{K\}, \end{aligned} \tag{2.33}$$

which gives equation 2.31. In addition,

$$\begin{aligned} E[Y^2] &= \sum_Y Y^2 P\{Y\}, \\ &= \sum_K E[Y^2|K]P\{K\} \\ &= \sum_K (\mathrm{Var}[Y|K] + E[Y|K]^2)P\{K\} \\ &= \sum_K (K\mathrm{Var}[X_i] + K^2 E[X_i]^2)P\{K\} \\ &= E[K]\mathrm{Var}[X_i] \ + \ E[K^2]E[X_i]^2. \end{aligned} \tag{2.34}$$

and with the formula for the variance, $\mathrm{Var}[Y] = E[Y^2] - E[Y]^2$, equation 2.32 is obtained.

2.1.2 Four Famous Probability Distributions

The Binomial Distribution and the Geometric Distribution

We have already seen one "famous" distribution in statistics: the Bernoulli distribution. This models the result of a single random trial in which the probability of success ($X = 1$) is p. The Bernoulli distribution forms the building block for two other important distributions: the binomial distribution and the geometric distribution. Both arise from the considering what happens if we perform a series of Bernoulli trials. For example, the result of n such trials could be zero successes or n successes or any number in between. This is how we would model the number of heads observed when a coin is tossed n times.

Pattern	Probability	# Successes, y	$P\{Y = y\}$
111	ppp	3	p^3
110	$pp(1-p)$	2	$3p^2(1-p)$
101	$p(1-p)p$		
011	$(1-p)pp$		
100	$p(1-p)(1-p)$	1	$3p(1-p)^2$
010	$(1-p)p(1-p)$		
001	$(1-p)(1-p)p$		
000	$(1-p)(1-p)(1-p)$	0	$(1-p)^3$

Table 2.1: Breakdown of the binomial$(3, p)$ distribution.

Table 2.1 lists every possible outcome of such an experiment when just three trials are performed ($n = 3$). The trials are done one at a time, so there is an order to the observed outcomes. Since the trials are independent, the probability of each pattern is equal to the product of the probabilities of each individual outcome. Let the random variable Y be the number of successes among the n trials. As table 2.1 shows, each pattern with y successes, and

$n-y$ failures, has probability $p^y(1-p)^{n-y}$. To compute the probability function of Y, we need to sum over all the possible orderings for each number of successes. This leads to the formula

$$P\{Y=y\} = \binom{n}{y} p^y (1-p)^{n-y} \qquad y = 0, 1, 2, \ldots, n \tag{2.35}$$

in which $\binom{n}{y}$ is the number of ways that y successes and $n-y$ failures can be ordered. Equation 2.35 is known as the binomial distribution, and we say that Y is a binomial(n,p) random variable.

Binomial coefficients count the number of ways of choosing y items from a total of n items, and are well-known numbers due of their role in the binomial theorem,

$$\begin{aligned}(p+q)^n &= p^n + np^{n-1}q + \ldots + \binom{n}{y} p^y q^{n-y} + \ldots + npq^{n-1} + q^n \\ &= \sum_{y=0}^{n} \binom{n}{y} p^y q^{n-y}. \end{aligned} \tag{2.36}$$

With $q = 1-p$, as in equation 2.35, we have $(p+q)^n = 1^n = 1$ and this verifies equation 2.6 for the binomial distribution, *i.e.* that the total probability is equal to one. Binomial coefficients are computed using

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}, \tag{2.37}$$

in which $k! = k(k-1)(k-2)\ldots 3 \times 2 \times 1$, and where by definition $0! = 1$. Table 2.2 lists some binomial coefficients, and shows that they also satisfy the recursion

$$\binom{n}{y} = \binom{n-1}{y} + \binom{n-1}{y-1} \tag{2.38}$$

for $n \geq 0$ and $0 \leq y \leq n$, and starting with $\binom{0}{0} = 1$. Table 2.2 is often called Pascal's triangle.

		y						
		0	1	2	3	4	5	6
	0	1						
	1	1	1					
	2	1	2	1				
n	3	1	3	3	1			
	4	1	4	6	4	1		
	5	1	5	10	10	5	1	
	6	1	6	15	20	15	6	1

Table 2.2: Some binomial coefficients, $\binom{n}{y}$.

A binomially-distributed random variable is just the sum of n independent Bernoulli random variables. We could calculate the variance of a binomial random variable with equations 2.10

and 2.11, but it is a bit easier to use the formulas for sums of random variables, 2.27 and 2.28. This gives

$$E[Y] \;=\; \sum_{i=1}^{n} p \;=\; np \tag{2.39}$$

and

$$\mathrm{Var}[Y] \;=\; \sum_{i=1}^{n} p(1-p) \;=\; np(1-p). \tag{2.40}$$

Figure 2.4 plots the histograms of two binomially-distributed random variables with different probabilities of success, p. When $p = 0.5$, as in figure 2.4(b), the distribution is symmetric about its expected value, and when $p \neq 0.5$, *e.g.* in figure 2.4(a), it is skewed to one side or the other. The number of threes observed when a die is rolled some number of times is an example of the latter, with $p = 1/6$. Thus, when more than one event is possible, but we are only interested in a particular outcome, $\{A\}$, all other events can be grouped as not-$\{A\}$ and the binomial distribution applies.

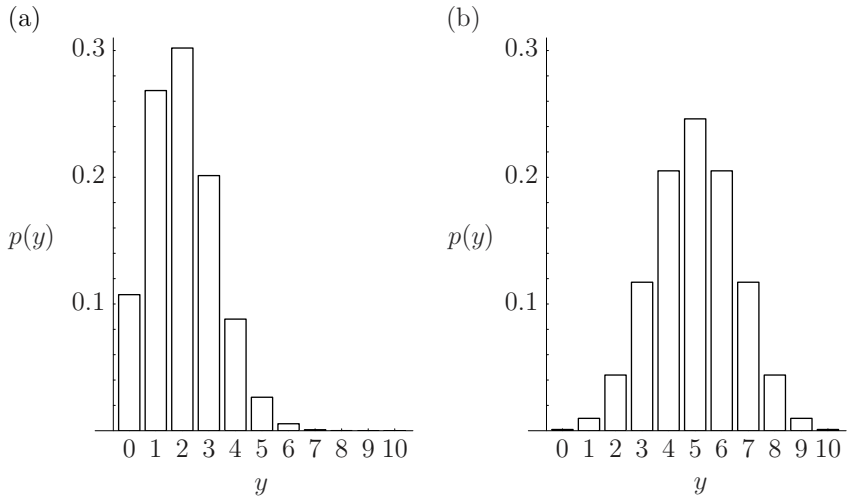

Figure 2.4: Histograms of two binomial random variables. In (a) $p = 0.2$ and in (b) $p = 0.5$. For both, $n = 10$.

Another way we can characterize a string of Bernoulli trials is to ask: how many trials must be performed before the first success is observed? For this, we must imagine a potentially infinite number of trials. If the trials are done on a regular basis, we can call this the waiting time, T, to the first success. Intuitively, it makes sense that this number should be inversely proportional to the probability of success. The smaller p is, the longer we should have to wait. For example, we expect about one out of every six rolls of a die to show a three, one out of every two coin tosses to come up heads. Specifically, the waiting time follows the geometric distribution

$$P\{T = t\} \;=\; (1-p)^{(t-1)}p \qquad\qquad t = 1, 2, \ldots \tag{2.41}$$

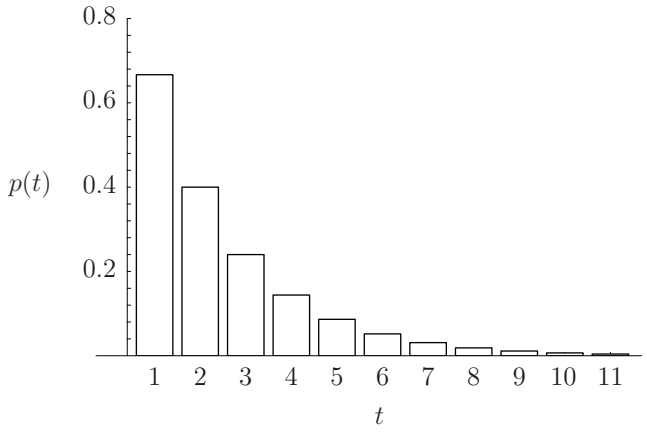

Figure 2.5: A geometric distribution with parameter $p = 0.4$.

The derivation of 2.41 is straightforward: for the first success to be in trial number t, there must be $t-1$ failures in a row followed by a single success.

Notice that while a binomial random variable is defined over a finite range $\{0, 1, \ldots, n\}$, a geometric random variable can assume an infinite number of values. This must be so, in order for the total probability to sum to one:

$$\sum_{t=1}^{\infty} P\{T = t\} \quad = \quad \sum_{t=1}^{\infty} (1-p)^{(t-1)} p \quad = \quad 1. \tag{2.42}$$

Equation 2.42 follows from the result for the sum of a geometric series,

$$\sum_{k=0}^{\infty} ar^k \quad = \quad \frac{a}{(1-r)} \tag{2.43}$$

when $|r| < 1$, after putting in $a = p$ and $r = 1 - p$. Equation 2.42 proves a simple yet profound fact, that if we wait long enough we are guaranteed to observe even the most unlikely outcomes.

By differentiating with respect to r twice, equation 2.43 can also be used to show that the mean and variance of a geometric random variable are given by

$$E[T] = \frac{1}{p} \qquad \text{and} \qquad \text{Var}[T] = \frac{1-p}{p^2}. \tag{2.44}$$

Figure 2.5 shows an example of a geometric distribution with an expected value of $E[T] = 2.5$. The most likely outcome is success in the first trial, which is can be seen from equation 2.41 since both p and $1 - p$ are between zero and one. The expected value of the waiting time confirms our intuition about the process. Taking the example of a die, in which the probability of success (*e.g.* roll a two) is equal to 1/6, then we have to toss the die six times on average to observe a success.

The Poisson Distribution and the Exponential Distribution

The binomial and geometric distributions describe two different aspects of a series of Bernoulli trials. If we think of the trials as taking place at regular intervals, then the binomial distribution

counts the number of successes in a given time period and the geometric distribution describes the waiting time between successes. However, when we observe a process over some very large number of trials and the probability of success in any one trial is very low, we sometimes prefer to measure time continuously rather than in discrete intervals. Many examples of this arise in coalescent theory; for example counting the number of mutations along some branch in a genealogy or the number of generations back to the common ancestor of a pair of sequences. In continuous-time approximations to such processes, events can happen at any moment and occur at some rate λ per unit of time, depending on how time is measured. Mathematically, these exist in the limit as the probability of success tends to zero and the number of trials per unit of time tends to infinity, such that λ remains constant.

We can get an intuitive sense of this by imagining dividing continuous time into very small increments, δt, so that a length of time, t, comprises $t/\delta t$ such steps (trials). Then $\lambda \delta t$ is the probability that an event occurs in a single step, and the waiting time to the first event is described by a geometric distribution

$$P\{T = t + \delta t\} \quad = \quad \lambda \delta t (1 - \lambda \delta t)^{(t/\delta t)}. \tag{2.45}$$

Although the rate λ is not a probability, but can assume any positive value, $\lambda \delta t$ can always be made small by letting δt approach zero. For example, T might be the waiting time until an automobile accident at an intersection, which for the sake of argument we can assume is uniformly busy day and night. If three accidents occur there per week, on average, then the chance of an accident occurring in the span of one minute is equal to $3/(7 \times 24 \times 60) \approx 0.0003$. Note also that if t =3:27PM, then in continuous time the event might have happened at any time between 3:27PM, and 3:28PM.

In a similar way, by discretizing time, we can use a binomial distribution to represent the number of events that occur over some period of time t. Any fixed t could be treated, but there is no loss in considering the number of events over a unit length of time $t = 1$. The number of steps in the interval is equal to $n_{\delta t} = 1/\delta t$, in which the subscript signifies the dependence of $n_{\delta t}$ on the value of δt, *i.e.* on how finely we choose to divide time. The fact that $n_{\delta t}$ might not be an integer will not matter in the limit as δt tends to zero ($n_{\delta t}$ tends to infinity) where single steps become infinitesimal. The number of events in an interval of unit length is given by the binomial distribution

$$P\{X = k\} \quad = \quad \binom{n_{\delta t}}{k} \left(\frac{\lambda}{n_{\delta t}}\right)^k \left(1 - \frac{\lambda}{n_{\delta t}}\right)^{n_{\delta t} - k}. \tag{2.46}$$

In the limit as δt tends to zero, 2.45 and 2.46 converge to an exponential and a Poisson distribution, respectively, two distributions which are vital to the coalescent. Convergence follows from the Taylor series expansion of the exponential function

$$e^{-x} \quad = \quad 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \ldots . \tag{2.47}$$

We have already seen an example of a series representation of a continuous functionin in equation 2.43, which gave the power series of $a/(1 - r)$. Equation 2.47 shows that when x is small,

$$e^{-x} \quad \approx \quad 1 - x. \tag{2.48}$$

Alternatively, we can write

$$e^{-x} \quad = \quad 1 - x + o(x)$$

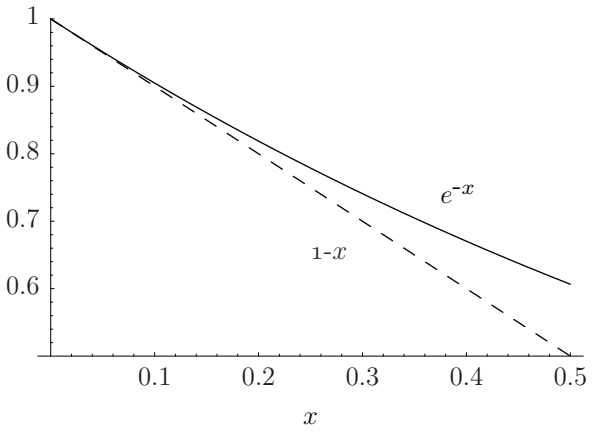

Figure 2.6: Graph comparing e^{-x} and $1 - x$.

in which the notation $o(x)$ captures all the terms that approach zero faster than x (*e.g.* x^p, where $p > 1$). That is, a function $h(x)$ is $o(x)$ if

$$\lim_{x \to 0} \frac{h(x)}{x} \quad = \quad 0, \tag{2.49}$$

and we will see this notation again, *e.g.* in Chapter 3. Figure 2.6 displays the accuracy of the approximation in equation 2.48 and shows that it works reasonably well even for values of x that are not tiny. For example, even when x is as large as 0.1, e^{-x} is only 0.54% larger than $1 - x$. Note that in this case, the next term in the series, $x^2/2$, which is equal to 0.005 when $x = 0.1$, accounts for most of the error.

The above shows that in the limit as δt tends to zero, with the values of λ and t assumed to be fixed (*i.e.* constant),

$$\lim_{\delta t \to 0} (1 - \lambda \delta t)^{(t/\delta t)} \quad = \quad e^{-\lambda t},$$

and the geometric distribution in equation 2.45 becomes $P\{t \leq T < t + \delta t\} = f_T(t)\delta t$, in which

$$f_T(t) \quad = \quad \lambda e^{-\lambda t} \qquad\qquad t \geq 0. \tag{2.50}$$

Equation 2.50 above specifies an exponential distribution with parameter λ. Again, the parameter λ is the rate at which events occur per unit time, and T is the waiting time to the first event. The mean and the variance of an exponentially-distributed random variable are

$$E[T] = \frac{1}{\lambda} \qquad \text{and} \qquad \text{Var}[T] = \frac{1}{\lambda^2} \tag{2.51}$$

which can be derived using 2.12 and 2.13. Figure 2.7 shows an example of an exponential distribution. The exponential distribution always has this characteristic shape. It has the property, in common with its discrete analogue the geometric distribution, that the smallest values are the most likely, while large values are possible but rare.

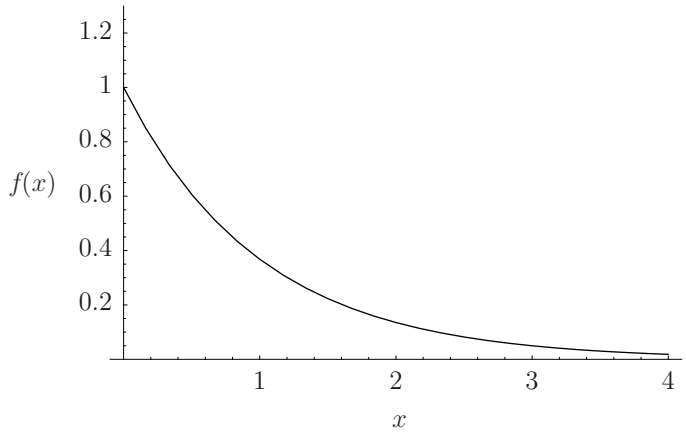

Figure 2.7: An exponential distribution with parameter $\lambda = 1$.

Note that we are free to choose the units in which time is measured. That is, an exponential random variable can be rescaled by any constant factor to yield a new exponential random variable with an appropriately rescaled parameter. For example, if we wish to measure time in new units that are C times longer than the old units, we perform a change of variable, $s = t/C$, so that $t = Cs$ and $dt = Cds$. Then, we have

$$\begin{aligned} f_S(s)ds &= \lambda e^{-\lambda t}dt \\ &= \lambda C e^{-\lambda C s}ds \end{aligned} \tag{2.52}$$

which is again an exponential distribution but with parameter λC.

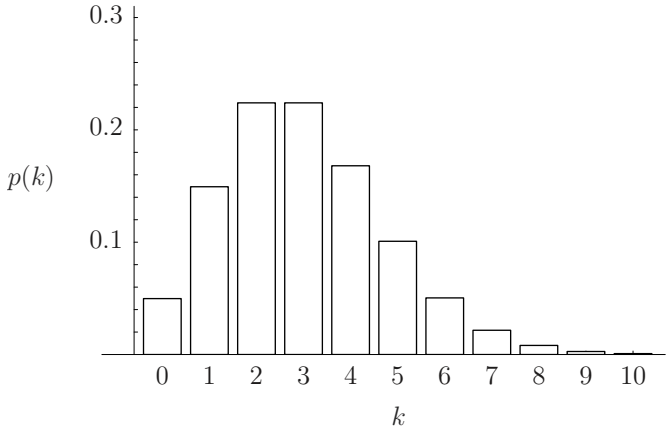

Figure 2.8: Poisson distribution with $\lambda = 3$.

Returning to the binomial distribution in equation 2.46 and dividing time more and more

finely by letting δt tend to zero gives another limiting distribution. Since $n_{\delta t} = 1/\delta t$, letting δt tend to zero is the same as letting $n_{\delta t}$ tend to infinity, and we have

$$\begin{aligned} P\{X = k\} &= \lim_{\delta t \to 0} \binom{n_{\delta t}}{k} \left(\frac{\lambda}{n_{\delta t}}\right)^k \left(1 - \frac{\lambda}{n_{\delta t}}\right)^{n_{\delta t} - k} \\ &= \frac{\lambda^k}{k!} \lim_{n_{\delta t} \to \infty} \frac{n_{\delta t}!}{(n_{\delta t} - k)!} \frac{1}{(n_{\delta t} - \lambda)^k} \left(1 - \frac{\lambda}{n_{\delta t}}\right)^{n_{\delta t}} \\ &= \frac{\lambda^k}{k!} e^{-\lambda} \qquad\qquad k = 0, 1, 2, \ldots \end{aligned} \tag{2.53}$$

This is the Poisson distribution with parameter λ. It applies to the number of events that occur in one unit of time, where λ is the rate at which events occur on this time scale. We can use this to approximte the distribution of the number of heads when a very unfair coin (biased towards tails) is tossed a great many times. The number of occurrences in an arbitrary length of time, t, follows a Poisson distribution with parameter λt. The mean and the variance of a Poisson(λ) distributed random variable are the same:

$$E[K] = \mathrm{Var}[K] = \lambda.$$

From the total probability, we have the useful fact

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}, \tag{2.54}$$

which we have already used in equation 2.48. Figure 2.8 shows the histogram of a Poisson random variable. When λ is small, most of the probability mass is at zero, and when λ is large, the Poisson distribution tends to be symmetric with mode near λ.

2.2 Poisson Processes

Like their discrete-time analogues, the binomial and geometric distributions, in continuous time the Poisson distribution describes the number of events in some interval and the exponential distribution describes the times between events. As will be seen in Chapters 3 and 4, the Poisson distribution and the exponential distribution form the backbone of the neutral coalescent. In this section, we will take some time to place these distributions, and to explore them in more detail, within the more general setting of Poisson processes. This provides a good framework for understanding and deriving many of the fundamental results of the coalescent, and even for beginning future work. We will review the theory of Poisson processes and derive some of the results most relevant to coalescent theory. Several of the results in Chapters 3 and 4 are direct applications of results given below.

A Poisson process is a counting process, $K(t)$, which meets the following criteria:

1. $K(0) = 0$,
2. The process has stationary and independent increments,
3. $P\{K(t) = 1\} = \lambda t + o(t)$,
4. $P\{K(t) \geq 2\} = o(t)$.

In words, the process starts at time zero with count equal to zero, then jumps to one, two, three, etc., at random future times. The rate λ is constant over time, and events in non-overlapping time intervals are independent of one another. The chance that an event occurs over a very small period of time t is equal to λt, and the chance that two or more events occur at the same time is equal to zero. Really, the only thing we have added to the model just above, in Section 2.1, is the notion that we might continue to monitor the process, count events, etc., forever. The example of car accidents at a uniformly busy intersection might satisfy the above criteria.

We already know a good deal about Poisson processes from Section 2.1. First, the number of events over time zero to t is Poisson distributed,

$$P\{K(t) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad k = 0, 1, 2, \ldots$$

In fact, it does not matter when we start counting,

$$P\{K(t+s) - K(s) = k\} = P\{K(t) = k\} \qquad s, t \geq 0,$$

so the number of events in any time period of length t follows the Poisson distribution with paramter λt. Second, the waiting time to the first event is exponentially distributed,

$$f_T(t) = \lambda e^{-\lambda t} \qquad t > 0. \tag{2.55}$$

However, now we can say that equation 2.55 holds regardless of when we begin waiting. This so-called memoryless property of the exponential distribution also holds for the series of Bernoulli trials described by the geometric distribution. For example, since the probability of heads is constant at 1/2 and successive tosses are independent, the number of tosses it will take to see heads again is independent of what the last toss yielded (and the one before that, and the one before that, ...). Importantly for Poisson processes, the waiting times between successive events are i.i.d. exponential random variables.

Another distribution, the gamma, arises in the context of Poisson processes. The waiting time W until the nth event, which is the sum of n i.i.d. exponential waiting times, follows the gamma distribution

$$f_W(w) = \lambda e^{-\lambda w} \frac{(\lambda w)^{n-1}}{(n-1)!} \qquad w > 0. \tag{2.56}$$

This can be derived by taking $n - 1$ successive convolutions of exponential distributions 2.55. The mean and the variance of W are given by

$$E[W] = \frac{n}{\lambda} \qquad \text{and} \qquad \text{Var}[W] = \frac{n}{\lambda^2}. \tag{2.57}$$

Although equation 2.56, as written with the factorial $(n-1)!$, requires n to be an integer, and there is no interpretation of a non-integer n in terms of Poisson processes, the gamma distribution is defined for any positive n:

$$f_W(w) = \lambda e^{-\lambda w} \frac{(\lambda w)^{n-1}}{\Gamma(n)} \qquad w > 0. \tag{2.58}$$

The function $\Gamma(n)$ is called the gamma function and is defined by

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx. \tag{2.59}$$

If n is an integer, then $\Gamma(n) = (n-1)!$. We can use equation 2.59, with the change of variable $x = \lambda w$, to show that the gamma distribution 2.58 integrates to one. Surprisingly, the moments given in equation 2.57 hold even when n is not an integer.

The coalescent is well framed by the theory of Poisson processes because it considers the times to, and types of, events that each have a very small probability of occurring in any single generation, *i.e.* common ancestor events between sequences, mutation events, migration events if they are rare, and so on. So far, we have considered only a single Poisson process, or a single type of event, but to get to the coalescent we need to allow for the possibility of several different kinds of events. If each of these has a very low probability of occurring, then they will each form a Poisson process and the co-occurrence of two or more events in a single generation will be negligible. The following subsections present some results relevant to the coalescent for multiple Poisson processes with parameters λ_1, λ_2, ..., λ_n. Results are derived for the case of two Poisson processes, but can be extended to any number of processes.

2.2.1 Poisson Process Results for the Coalescent

The Sum of Independent Poissons

If X_1 and X_2 are independent and Poisson distributed with parameters λ_1 and λ_2, what is the distribution of $Y = X_1 + X_2$? One way to answer this question is to take the convolution of X_1 and X_2:

$$\begin{aligned} P\{Y=k\} &= \sum_{i=0}^{k} P\{X_1 = i\} P\{X_2 = k-i\} \\ &= \sum_{i=0}^{k} \frac{\lambda_1^i}{i!} e^{-\lambda_1} \frac{\lambda_2^{k-i}}{(k-i)!} e^{-\lambda_2} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda_1^i \lambda_2^{k-i} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} (\lambda_1 + \lambda_2)^k. \end{aligned} \tag{2.60}$$

Thus, Y is Poisson-distributed with parameter $\lambda_1 + \lambda_2$. The last step in getting to equation 2.60 uses the binomial theorem, given by equation 2.36. More processes can be added simply by performing additional convolutions. Thus, the sum of independent Poisson processes is another Poisson process whose rate is equal to the sum of the individual rates.

The Probability that the First Event is of a Particular Type

What is the probability that the first event is of type one (*i.e.* from the rate λ_1 process)? This is the same as the probability that the time to an exponential (λ_1) event is smaller than the time to an exponential (λ_2) event. Let T_1 and T_2 be the times to each event. We can solve this by conditioning on the value of T_1,

$$P\{T_1 < T_2\} = \int_0^\infty P\{T_2 > t\} f_{T_1}(t) dt.$$

We know that $f_{T_1}(t) = \lambda_1 e^{-\lambda_1 t}$, and we can compute

$$\begin{aligned} P\{T_2 > t\} &= \int_t^\infty \lambda_2 e^{-\lambda_2 t} dt \\ &= e^{-\lambda_2 t}. \end{aligned} \tag{2.61}$$

Therefore, we have

$$\begin{aligned} P\{T_1 < T_2\} &= \int_0^\infty e^{-\lambda_2 t} \lambda_1 e^{-\lambda_1 t} dt \\ &= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)t} dt \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2}. \end{aligned} \tag{2.62}$$

The probability that the first event is of a particular type is given simply by the relative rate of that event (*i.e.* as a fraction of the total rate).

The Time to the First Event among Independent Poissons

What is the distribution of the time to the first event? In other words, what is the distribution of $T = \min(T_1, T_2)$? That is, what is the distribution of the smaller of T_1 and T_2 when two Poisson processes are running simultaneously? This is easiest to compute by considering the probability that $T = \min(T_1, T_2)$ is greater than some value t, which can only be true if both T_1 and T_2 are greater than t. Because the processes are independent, and using equation 2.61, we have

$$\begin{aligned} P\{T > t\} &= P\{\min(T_1, T_2) > t\} \\ &= P\{T_1 > t\} P\{T_2 > t\} \\ &= e^{-\lambda_1 t} e^{-\lambda_2 t} \\ &= e^{-(\lambda_1+\lambda_2)t}, \end{aligned} \tag{2.63}$$

and this is precisely what would be obtained from integrating the exponential density implied by equation 2.60 from t to infinity. In fact, there is a one-to-one correspondence between distribution functions, *i.e.* the cumulative distribution $P\{T \leq t\} = 1 - P\{T > t\}$, and probability densities, so equation 2.63 proves that $T = \min(T_1, T_2)$ is exponential with rate $\lambda_1 + \lambda_2$.

The Number of Events Required to See a Particular Outcome

Suppose that for the moment we are not interested in the times between events, but only in counting the events themselves. In Chapter 4 for example, we will be interested in the number of mutation events K that occur before the common ancestor event between a pair of sequences. In this case, we are waiting until events happen and simply noting their types while disregarding their arrival times. We can make use of the results just derived to answer the following question. In our system with two types of events, what is the distribution of the number of events that

occur up to the time that an event of type 1 occurs? From equation 2.62 we know that the probability that the next event is of type 1 is equal to $\lambda_1/(\lambda_1 + \lambda_2)$ and the probability that it is of type 2 is just one minus this, or $\lambda_2/(\lambda_1 + \lambda_2)$. Since the sum of the processes is a Poisson process with the sum of the rates (see equation 2.60), the process is memoryless and the probabilities above apply to every event. Thus, the events form a series of Bernoulli trials with probability of success $p = \lambda_1/(\lambda_1 + \lambda_2)$, since we are interested in the type-1 event. Therefore, the number of events that have occurred when the first event of type 1 occurs must follow the geometric distribution

$$P\{K = k\} = \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{k-1} \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{2.64}$$

Again, this can be extended to the case of n possible outcomes or processes.

Tying All This Together: A Filtered Poisson Process

We can bring these results full circle and demonstrate something we knew at the start, namely that the time to an event of type 1 is exponentially distributed. This is not just for the sake of completeness. We will encounter this problem again in Chapter 7 and will use the general method of conditioning one random variable to obtain the distribution or other properties of another random variable many times throughout this book.

First of all, we know that the number of events K up to and including the first type-1 event is geometrically distributed according to equation 2.64. Second, although we did not derive this, we also know that the time to the kth event is gamma distributed as in equation 2.56, with $\lambda = \lambda_1 + \lambda_2$. Thus, we know the distribution of T given K and we know the distribution of K, and this is all we need to make use of equation 2.7 or 2.23. The only difference is minor, which is that here one of the random variables is continuous and one is discrete. We simply sum the conditional density of T given K over all possible values of K, weighted by the probability function of K. We have

$$\sum_{k=1}^{\infty} (\lambda_1 + \lambda_2) e^{-(\lambda_1+\lambda_2)t} \frac{[(\lambda_1 + \lambda_2)t]^{k-1}}{(k-1)!} \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{k-1} \frac{\lambda_1}{\lambda_1 + \lambda_2} = \lambda_1 e^{-\lambda_1 t},$$

and this equality can be demonstrated with the aid of equation 2.54. More generally, we can imagine applying a "filter" to a Poisson process with rate λ by observing it but rejecting (not counting) events with some constant probability $1 - p$. Above, we had $\lambda = \lambda_1 + \lambda_2$ and $p = \lambda_1/(\lambda_1 + \lambda_2)$, while in general the Poisson process might include more types of events or even just a single type of event, and the filter need not be a function of the type of event. The method above, of course, shows that the time to the first accepted event in this filtered Poisson process is exponentially distributed with parameter λp. Of course, the filtered process is a Poisson process with rate λp.

2.2.2 Convolutions of Exponential Distributions

Now consider a related problem in which the rate of a Poisson process changes with each event. This arises in the study of genealogies because the rate of coalescence changes every time a coalescent event occurs. Let λ_i be the rate of the process between the $(i-1)$st event and the ith event, and consider the distribution of the time to the nth event. This is the sum of n exponentially distributed random variables. If the rates did not change with each event, that is if $\lambda_i = \lambda$ for all i, then the sum of the waiting times would be gamma distributed as in equation 2.56. However, if $\lambda_i \neq \lambda_j$ for $i \neq j$, then it is necessary to take a series of convolutions

(see equation 2.30) of these individual exponential random variables. Beginning with just two, we have

$$\begin{aligned} f_{T_1+T_2}(t) &= \int_0^t f_{T_1}(s) f_{T_2}(t-s) ds \\ &= \int_0^t \lambda_1 e^{-\lambda_1 s} \lambda_2 e^{-\lambda_2 (t-s)} ds \\ &= \lambda_1 \lambda_2 e^{-\lambda_2 t} \int_0^t e^{-(\lambda_1 - \lambda_2) s} ds \\ &= \frac{\lambda_1}{\lambda_1 - \lambda_2} \lambda_2 e^{-\lambda_2 t} (1 - e^{-(\lambda_1 - \lambda_2) t}) \\ &= \frac{\lambda_1}{\lambda_1 - \lambda_2} \lambda_2 e^{-\lambda_2 t} + \frac{\lambda_2}{\lambda_2 - \lambda_1} \lambda_1 e^{-\lambda_1 t}. \end{aligned} \tag{2.65}$$

Equation 2.65 is simply a weighted sum of the distributions of T_1 and T_2. Therefore, the same steps above are applied repeatedly in the convolution of larger numbers of exponential random variables, as long as their parameters are all distinct. A series of $n-1$ such convolutions gives

$$f_{\sum_{i=1}^n T_i}(t) = \sum_{i=1}^n \lambda_i e^{-\lambda_i t} \prod_{j=1, j \neq i}^n \frac{\lambda_j}{\lambda_j - \lambda_i} \tag{2.66}$$

for the distribution of the total waiting time $\sum_{i=1}^n T_i$. We will use this result in the next chapter.

2.3 Exercises

1. If a fair coin is tossed four times, which is the more likely result: four heads or two heads and two tails?

2. Compute the probability obtaining a two in a single roll of a fair die, given that neither a six nor a four is obtained.

3. Compute the expected value and the variance of the following random variable:

$$X = \begin{cases} 3 & \text{with probability } 0.4 \\ \\ -3 & \text{with probability } 0.6 \end{cases}$$

4. Let $Y_n = X_1 + \cdots + X_n$, where the X_i are i.i.d. random variables with the distribution given in excercise 1. Compute $E[Y_n]$ and $\mathrm{Var}[Y_n]$?

5. What is the distribution of Y_2 in excercise 2?

6. A coin is tossed and a die is rolled at the same time, and this can be repeated indefinitely. What is the expected number of tails at the time the die first shows a six?

7. If the conditional moments $E[Y^r|X]$ are known and $P\{X\}$, how can $\mathrm{Var}[Y]$ be computed? what is the correct formula for $\mathrm{Var}[Y]$ computed by conditioning on X?

8. What are the expected value and the variance of the random variable whose distribution is given by equation 2.66?

9. What is the probability density function for the maximum of two exponential random variables with the same parameter λ?

10. What is the expected value of the maximum of k exponential random variables, each with the same parameter λ?