PRIMER 3

Probability Theory

Probability theory, or the mathematical description of chance events, is arguably one of the areas of mathematics that has provided the most insight into biology. This primer serves as a basic introduction to probability theory, providing the background material necessary for Chapters 13–15.

P3.1 An Introduction to Probability

Before introducing the concept of probability, we must introduce the concept of a *trial*. A trial can be any sort of occurrence, like the birth of a child or the flowering of a plant. We are interested in trials that can have more than one possible outcome. For example, a baby might be a boy or a girl. A plant might produce any number of seeds from zero to thousands. Because more than one outcome is possible, we can consider the outcome to be a variable, specifically a *random variable*, which we denote by a capital letter (e.g., X). Once a trial has happened, the random variable takes on a specific value. For example, if a boy is born we could write X = "boy" if we wanted to describe the outcome in words, or we could write $X = 0$ if we were counting the number of girls.

Before the trial actually occurs, we can consider quantifying the chance or *probability* that a particular outcome will be realized. We denote a particular outcome of a random variable by a lower-case letter (e.g., x). Some outcomes will have a high chance of occurring and others will be extremely unlikely. We can write the probability that the random variable X takes on the value x as $P(X = x)$ or just $P(x)$. There are two ways to think about probabilities:

- Frequency interpretation: A probability is understood as the frequency of a particular outcome across the course of many trials.
- Subjective interpretation: A probability is understood as a subjective belief or opinion of the chance that a particular outcome will be realized.

For example, when a baby rhinoceros is born, you might think that there is roughly a 50% probability that the baby is male, $P(X$ = "male") = 0.5. This opinion might be

based on previous observations that about half of all mammals are male; this would be a frequentist's perspective. Alternatively, your opinion might be based on the idea that sex chromosomes should segregate 50:50 (i.e., half of all sperm should bear an *X* chromosome and half a *Y* chromosome); this would be a subjectivist's perspective. Most of us think about probabilities in both ways, depending on the situation.

Venn Diagram	Set Language	Set Notation
(a) 	The set	Ω
(b) 	Subset A (or event A)	A
(c) 	Complement of A	A^C
(d) 	Intersection of A and B	$A \cap B$ or AB
(e) 	Union of A and B	$A \cup B$
(f) 	A and B are disjoint (mutually exclusive)	$AB = \emptyset$
(g) 	A is a subset of B	$A \subset B$

Figure P3.1: Venn diagrams. The probability of an outcome can be represented as an area (shaded grey) within a Venn diagram, which is a square of area one. (a) If an outcome is certain to occur (with probability one), the entire square is shaded. (b) If outcome A has a probability less than one of occurring (say, p), a fraction p of the square is shaded. (c) The complement of A represents any outcome other than A, so that a fraction $1 - p$ of the square is shaded. (d) The probability that two outcomes, A and B, both occur is given by the area of their intersection. (e) The probability that A or B or both occurs is represented by the total area inside the shaded regions (their intersection must be counted only once). (f) If two outcomes A and B have no intersection, then they are mutually exclusive (they cannot both occur). (g) An outcome A is a subset of B if its area is entirely encompassed within the area of B, in this case B will always be observed when A occurs.

Because a probability represents the chance that a trial has a particular outcome, any probability must lie between 0 and 1 (or, equivalently, between 0% and 100%). An aid to visualizing probabilities is a "Venn diagram" (Figure P3.1), which is a square whose area is one. The area of the whole square (one) represents the probability that the trial has any outcome (including, potentially, that nothing happens). We can subdivide the square into subsets, where each subset represents a potential outcome and

the area of the subset represents the probability of observing that outcome. For example, if your sister is pregnant, you might think that there is a 1/7 chance that the baby will be born on any particular day of the week, e.g., $P(X = \text{“Monday”}) = 0.14$. In this example, the potential outcomes (days of the week) are “disjoint” or “mutually exclusive,” meaning that only one of the alternative outcomes is possible—the baby cannot be born both on a Monday and on a Tuesday. For trials with mutually exclusive outcomes, the Venn diagram can be partitioned into nonoverlapping subsets, and the following rule applies:

Rule P3.1: Probabilities of Mutually Exclusive Outcomes

If a trial can result in only one of a set of possible outcomes, the outcomes are said to be “mutually exclusive.” The probabilities of mutually exclusive outcomes sum to one:

$$\sum_{i=1}^{\#\text{ of outcomes}} P(X = x_i) = 1.$$

As a special case of Rule P3.1, one can always partition the outcomes of a trial into one outcome of interest, A, and its complement, A^C. The complement represents “not A,” and the probability of observing the complement is the probability of not observing A. For example, the baby might be born on a Monday (A) with probability 1/7 or on any other day of the week (A^C) with probability 6/7.

Rule P3.2: Complement Rule

The probability of an outcome plus the probability of its complement sum to one:

$$P(X = A) + P(X = A^C) = 1.$$

This rule is easy to visualize using a Venn diagram (Figure P3.1c). Rule P3.2 is extremely handy, because it is sometimes easier to calculate the probability of the complement of an outcome of interest. For example, if you are monitoring the populations of lizards on five islands and you want to know the probability that one, two, three, four, or five of the populations goes extinct over the course of a year, then the easiest way to calculate $P(X = \text{“one or more extinctions”})$ is to calculate the

complement $P(X =$ "no extinctions"). Rule P3.2 then tells us that $P(X =$ "one or more extinctions") equals one minus $P(X =$ "no extinctions").

The outcomes of a trial need not be mutually exclusive. For example, if you are observing interactions between two fish in a five-minute interval, you might observe no contact, aggressive contact, mating, or avoidance behavior, but you might very well see more than one of these outcomes in the same period (e.g., aggression and mating). If two outcomes are not mutually exclusive, then there is some probability that both will be observed. On a Venn diagram, the intersection of the two outcomes represents this probability. We can write the probability that both A and B occur using $P(X = A \cap B)$ where $\cap$ is called the intersection and represents "and" (see Figure P3.1d). Following convention, we can drop the "$X =$ " and the "$\cap$" in such probability statements and write $P(X = A \cap B)$ as $P(A\ B)$. $P(A\ B)$ is read as "the probability that the random variable X has both the outcome A *and* the outcome B."

The outcomes A and B are said to be *independent* if the probability of observing both, $P(A\ B)$, equals the product of each outcome's probability, $P(A)\ P(B)$. When outcomes A and B are independent, observing A provides no information about whether or not B will be observed. For example, imagine throwing two dice—as long as you don't have any tricks up your sleeve, the number showing on the first die will have no influence on the number showing on the second die; they will be independent events. In biology, independence is often assumed for trials involving different individuals who are separated in time and space and who have had no contact. For example, the day of the month in which a woman in Vancouver and a woman in New York start menstruating might reasonably be independent of one another, but this is not true for women living in close proximity (Preti et al. 1986). Mutually exclusive outcomes are never independent because their intersection $P(A\ B)$ is zero and not $P(A)\ P(B)$; for example, a single die thrown cannot show both a "two" and a "five" as these are mutually exclusive.

Often we are interested in knowing the probability of outcome A or B or both (that is, A and/or B). In a Venn diagram, this probability is represented by the total area of the subsets A and B. We write this probability as $P(A \cup B)$, where $\cup$ is called the "union" and represents "and/or" (see Figure P3.1e). For example, we might be interested in the probability that a forest patch is decimated by fire or disease or both. To calculate the union of two subsets, we could add together the two subsets, but then we would be counting their intersection twice. Thus, to find the union, we must subtract the intersection from the sum of the subsets.

Rule P3.3: Inclusion-Exclusion Rule

The probability of outcome A or B or both is the sum of each outcome's probability minus the probability that both occur:

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

CAUTION: It can be tempting to interpret $\cup$ as "and" whereas $\cup$ really represents "and/or." It can help to remember that $\cup$ represents the total area ("union") in a Venn diagram. Instead, it is $\cap$ that represents "and," where $\cap$ specifies the area in a Venn diagram within which both A and B occur (their intersection).

Because the intersection is known for two independent outcomes ($P(A)$ $P(B)$) and for two mutually exclusive outcomes (zero), we can calculate the probability of A and B as well as the probability of A or B or both:

Rule P3.4: Independent Outcomes

(a) If outcomes A and B are independent, the probability of observing both outcomes is the product of observing each separately:

$$P(A \cap B) = P(A\,B) = P(A)\,P(B).$$

(b) Using Rule P3.3, the probability of A or B (or both) is then

$$P(A \cup B) = P(A) + P(B) - P(A)\,P(B).$$

Rule P3.5: Mutually Exclusive Outcomes

(a) If outcomes A and B are mutually exclusive, the probability of observing both outcomes is zero:

$$P(A \cap B) = P(A\,B) = 0.$$

(b) Using Rule P3.3, the probability of A or B (or both) is then

$$P(A \cup B) = P(A) + P(B).$$

These rules are fairly intuitive, in part because we have experience with games involving these probability calculations. For example, if you take a randomly shuffled deck of 52 cards, the probability that the first card you turn over is the queen of spades is 1/52. Because 1/52 is the product of the probability of observing a queen, P("queen") = (1/13), and the probability of observing a spade, P("spade") = (1/4), these two outcomes are independent. Thus, we can calculate the probability that a

queen or a spade (or both) shows up from Rule P3.4 as $P(\text{“queen”} \cup \text{“spades”}) = P(\text{“queen”}) + P(\text{“spades”}) - P(\text{“queen”} \cap \text{“spades”}) = (1/13) + (1/4) - (1/52)$, which equals 16/52. We can get the same answer by counting the number of queens (4) and the number of spades that are not queens (12) out of 52 cards (=16/52), where this calculation avoids counting the intersection (the queen of spades) twice. As another example, the probability of getting a red card is 1/2, which equals the probability of getting a heart (1/4) plus the probability of getting a diamond (1/4). In this case, we don't have to subtract off the intersection, because "heart" and "diamond" are mutually exclusive outcomes (Rule P3.5).

Exercise P3.1: For each question, write the answer as *P*(insert appropriate description) = solution, and state any assumptions that you make.

(a) In a forest, imagine that 1% of trees are infected by fungal rot and 0.1% have owl nests. What is the probability that a tree has both fungal rot and an owl nest if the two are independent? If the two are mutually exclusive?

(b) Individuals of blood type O that are Rhesus negative are universal donors. If 46% of individuals have blood type O, if 16% of individuals are Rhesus negative, and if the two blood types are independent of one another, what is the probability that a randomly chosen individual is a universal donor (O−)?

(c) In a population, 46% of individuals have blood type O, 40% have blood type A, 10% have blood type B, and 4% have blood type AB. An individual with blood type A can receive transfusions from people with blood type O or A. What is the probability that a donor has the appropriate blood type for a patient of blood type A?

(d) Two independent studies are performed to test the same null hypothesis. What is the probability that one or both of the studies obtains a significant result and rejects the null hypothesis even if the null hypothesis is true? Assume that, in each study, there is a 0.05 probability of rejecting the null hypothesis.

Answers to the exercises are provided at the end of the primer.

P3.2 Conditional Probabilities and Bayes' Theorem

Unless two outcomes are independent, the probability of observing one outcome depends on whether the other outcome is observed. *Conditional probabilities* describe the relationship between outcomes.

Rule P3.6: Conditional Probability

Given that outcome B has occurred, we write the probability of observing outcome A as $P(A \mid B)$. The "|" can be read as "given that" or "conditional upon." By definition, $P(A \mid B)$ equals

$$P(A \mid B) = \frac{P(A\ B)}{P(B)}.$$

That is, the probability of observing A given that B has occurred, $P(A \mid B)$, is the fraction of cases in which B occurs, $P(B)$, that A also occurs, $P(A\ B)$.

For independent outcomes, $P(A \mid B) = P(A)$, because observing B provides no information about whether or not A has occurred.

For mutually exclusive outcomes, $P(A \mid B) = 0$, because observing B implies that A has not occurred.

Conditional probabilities can make it easier to determine the probability that two outcomes are both observed. The probability of both A and B occurring, $P(A\ B)$, is the probability of observing B times the probability that, among those cases in which B occurs, A occurs:

$$P(A\ B) = P(B)\ P(A \mid B). \qquad \text{(P3.1a)}$$

Rearranging (P3.1a) we get the definition for $P(A \mid B)$ given in Rule P3.6. Of course, the same reasoning allows us to write this joint probability as

$$P(A\ B) = P(A)\ P(B \mid A), \qquad \text{(P3.1b)}$$

which is the probability of observing A times the probability of observing B given that A has occurred.

These formulae look simple enough but they are extremely powerful. They immediately lead to one of the most important theorems in probability:

Rule P3.7: Bayes' Theorem

Because the joint probability of observing two outcomes, $P(A\ B)$, equals both $P(B)\ P(A \mid B)$ and $P(A)\ P(B \mid A)$, we can determine one conditional probability from the other using Bayes' theorem:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}.$$

As an example, suppose we want to calculate the probability that a person will die of lung cancer given that they smoke. We could study a cohort of individuals, determining which ones smoke and which ones don't and tracking them until they died. At that point we could calculate the fraction of smokers who died of lung cancer. In this example, we are trying to calculate the conditional probability *P*(death due to lung cancer | smoker). Using Bayes' rule, however, there is an alternative way to calculate this probability:

$$P(\text{death due to lung cancer} \mid \text{smoker}) = \frac{P(\text{smoker} \mid \text{death due to lung cancer})\, P\,(\text{death due to lung cancer})}{P(\text{smoker})}.$$

The probabilities on the right-hand side have already been estimated (Shopland 1995), allowing us to estimate the risk that a smoker dies of lung cancer without the above-mentioned study. *P*(smoker | death due to lung cancer) is estimated as the fraction of people that have died of lung cancer who are smokers. *P*(death due to lung cancer) is estimated from death records, and *P*(smoker) is estimated by polling an appropriate control population (a population similar in age drawn from similar environments). Using the data in Shopland (1995), *P*(smoker | death due to lung cancer) = 0.9, *P*(death due to lung cancer) = 0.3, and *P*(smoker) = 0.5, the probability that a smoker will die of lung cancer is estimated as (0.9)(0.3)/(0.5) = 0.54. Similar calculations for nonsmokers give a probability of death of only 0.06 (Exercise P3.2c). Thus, smokers have a nearly tenfold higher risk of dying of lung cancer compared to nonsmokers.

Bayes' theorem is widely used in scientific inference, using a methodology known as Bayesian analysis (see Hilborn and Mangel 1997). As described in Supplementary Material P3.1, Bayesian analysis allows scientists to infer aspects of the biological world that are hard to measure directly.

Exercise P3.2:

(a) If the probability of having green eyes is 10%, the probability of having brown hair is 75%, and the probability of having both green eyes and brown hair is 9%, what is the probability of having brown hair given that you have green eyes?

(b) Ability to taste phenylthiocarbamide (PTC) is thought to be determined by a single dominant gene with incomplete penetrance. Among North American Caucasians, there is a 70% chance of being able to taste PTC [P(taster) = 0.7]. If everybody who tastes PTC is a carrier [P(carrier | taster) = 1] and if 80% of the population carries the gene [P(carrier) = 0.8], what is the penetrance of the gene? That is, what is the probability of tasting PTC if you are a carrier, P(taster | carrier)?

(c) Write a formula for the risk of dying of lung cancer given that a person does not smoke in terms of P(smoker | death due to lung cancer), P(death due to lung cancer), and P(smoker). Estimate the risk of death due to lung cancer among nonsmokers using P(smoker | death due to lung cancer) = 0.9, P(death due to lung cancer) = 0.3, and P(smoker) = 0.5.

We can also use conditional probabilities to calculate the overall probability of outcome *A, P*(*A*), when *A* occurs in the context of a set of mutually exclusive outcomes, B_i, of a second random variable:

Rule P3.8: Law of Total Probability

Suppose that the outcomes, B_i, consist of n mutually exclusive events whose probabilities sum to one (i.e., $\sum_{i=1}^{n} P(B_i) = 1$). Then the probability of A is equal to the sum of the probabilities of A given each outcome B_i, weighted by the probability of each outcome B_i occurring:

$$P(A) = P(A \mid B_1)\, P(B_1) + P(A \mid B_2)\, P(B_2) + \cdots + P(A \mid B_n)\, P(B_n).$$

For example, Rule P3.8 can be used to calculate the overall probability of a randomly chosen individual contracting the flu, $P(A)$, when some individuals have had a flu shot (B_1) and others have not (B_2). If vaccinated individuals have a probability of infection of $P(A \mid B_1) = 0.01$ and nonvaccinated individuals have a probability of infection of $P(A \mid B_2) = 0.2$, then the overall probability of contracting the flu is $P(A) = 0.01\ P(B_1) + 0.2\ P(B_2)$ according to Rule P3.8. Thus, if 90% of the population were vaccinated ($P(B_1) = 0.9$, $P(B_2) = 0.1$), the probability of a randomly chosen individual contracting the flu would be 0.029.

An important concept that we will see repeatedly in this primer is the "expected value" or mean value of a random variable X. The expected value, denoted $E[X]$, can

be thought of as the average outcome that would be observed if the trial were repeated infinitely many times (see more precise Definitions P3.2 and P3.9 below). There is a useful formula for calculating the expected value of a random variable that we present here because of its analogy to the law of total probability.

Rule P3.9: Law of Total Expectation

Suppose that the outcomes, B_i, of a second random variable consist of n mutually exclusive events whose probabilities sum to one (i.e., $\sum_{i=1}^{n} P(B_i) = 1$). Then the expectation of a random variable X is equal to the sum of the expectation of X given each outcome B_i, weighted by the probability of each outcome B_i occurring:

$$E[X] = E[X \mid B_1]\, P(B_1) + E[X \mid B_2]\, P(B_2) + \cdots + E[X \mid B_n]\, P(B_n).$$

The law of total expectation is analogous to the law of total probability. Either the expected value of the random variable or the probability of a particular outcome can be calculated by summing the conditional values over all possible outcomes, B_i, of another random variable. For example, Rule P3.9 can be used to calculate the expected fitness (the "mean fitness") of a population consisting of three genotypes: *AA, Aa*, and *aa*. Here, the genotype is a second random variable with three mutually exclusive outcomes. The expected fitness is then $E[W] = E[W \mid AA]\, P(AA) + E[W \mid Aa]\, P(Aa) + E[W \mid aa]\, P(aa)$. In Chapter 3, we used subscripts to write the expected fitness conditional on being *AA* as $E[W \mid AA] = W_{AA}$. If we also assume that the genotype frequencies at time t are at Hardy-Weinberg proportions: $P(AA) = p(t)^2$, $P(Aa) = 2p(t)\, q(t)$, $P(aa) = q(t)^2$, the expected fitness becomes $E[W] = W_{AA}\, p(t)^2 + W_{Aa}\, 2\, p(t)\, q(t) + W_{aa}\, q(t)^2$, which equals the mean fitness in equation (3.12). The law of total expectation is particularly helpful when it is easier to describe the distribution of X conditional on the state of another factor.

P3.3 Discrete Probability Distributions

The first step in incorporating stochasticity into a model is to determine what process (or processes) has chance outcomes and then to describe the outcome of this process by a random variable. The next step is to describe the "probability distribution" for that random variable, which specifies how likely it is for the random variable to take on various values. In this section, we consider *discrete probability distributions*, where

the random variable has a discrete set of mutually exclusive outcomes (e.g., 0, 1, 2). In section P3.4, we describe random variables whose outcomes can be any point along a continuum (e.g., any real number between 0 and 1). In both cases, we show how important quantities like the mean and the variance can be derived. Key attributes of all of the distributions are summarized in tables at the end of the Primer.

We start with the simplest discrete probability distribution describing the outcome of a single Bernoulli trial:

Definition P3.1:

A Bernoulli trial has two possible outcomes, say "zero" and "one," where the probability of the outcome "one" equals $P(X = 1) = p$.

We will often refer to an outcome of one as a "success," despite the fact that outcome one is not always desirable (e.g., if "one" represents "death"). Because there are only two outcomes, observing zero is the complement of observing one. Thus, by Rule P3.2, $P(X = 0) + P(X = 1) = 1$, and the probability of observing zero is $P(X = 0) = 1 - p$ (Figure P3.1c). For example, p might be the probability of having a successful crop (outcome 1), and $1 - p$ would be the probability of having a crop failure (outcome 0).

In describing a probability distribution, we assume that the outcomes form a mutually exclusive set (e.g., success versus failure) and that we have described all possible outcomes. As a consequence, Rule P3.1 tells us:

Rule P3.10: The Sum of a Discrete Probability Distribution

The sum of $P(X = x_i)$ over all outcomes, x_i, equals one:

$$\sum_{x_i} P(X = x_i) = 1.$$

The notation $\sum_{x_i}$ in Rule P3.10 means the sum over all outcomes, x_i. For a Bernoulli trial, $P(X = 0) + P(X = 1) = (1 - p) + p$, which does equal one. For the distributions described in this primer, Rule P3.10 has been checked. If you want to develop a new probability distribution, however, you must confirm that your distribution obeys Rule P3.10.

Once a probability distribution has been specified, the distribution can be plotted. Typically, histograms are used, with the area of each bar representing the probability of

observing the outcome labeled on the horizontal axis (Figure P3.2).

Besides plotting a probability distribution, the two most important quantities that we might wish to know about a distribution are its mean and its variance. We write the *mean* (or *average*, or *expectation*) of a random variable X as μ or $E[X]$, calculated as follows:

Definition P3.2: The Mean of a Discrete Random Variable

The mean (or average) of a discrete random variable is the sum of the value of each outcome weighted by the probability of that outcome:

$$\mu = E[X] = \sum_{x_i} x_i P(X = x_i).$$

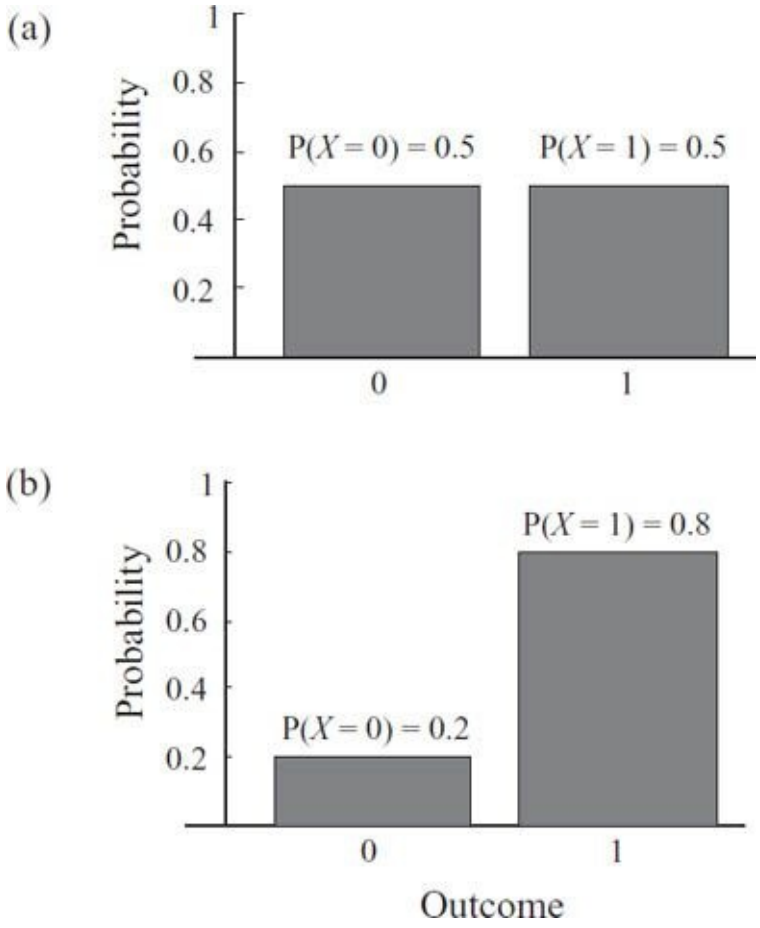

Figure P3.2: Histograms. The area of each bar represents the probability of observing outcome 0 or outcome 1 in a Bernoulli trial. Throughout, we will draw bars whose widths are equal and arbitrarily set to one, so that the height of the bar gives the probability of observing the outcome. The heights will then always sum to one. (a) Outcomes 0 and 1 are equally probable (as in a coin toss). (b) Outcome 1 is four times more likely than outcome 0.

One way to visualize the mean is to imagine balancing the histogram depicting the probability distribution on your finger, assuming that the weight of each bar is proportional to its height. The histogram will balance perfectly when you place your finger exactly at the mean. Thus, the mean is the "center of mass" of the probability distribution. For a Bernoulli trial, the mean equals p. This follows from Definition P3.2, which tells us that $E[X]$ is $0 \times P(X = 0) + 1 \times P(X = 1)$, which equals $0 \times (1 - p) + 1 \times (p) = p$. For example, if we let 0 represent crop failure and 1 represent crop success and if there is a 90% chance of a successful crop ($P(X = 1) = p = 0.9$), then the mean outcome will be 0.9. In any one year, the crop will either fail (outcome 0) or be successful (outcome 1), but we can think of the expected value as the average outcome that we would see after an indefinitely large number of years.

Often, we want to know how dispersed the random variable is around its mean. One measure of dispersion is the *variance*, which is often written as σ^2 or as Var[X]:

Definition P3.3: The Variance of a Discrete Random Variable

The variance of a discrete random variable is the expected value of $(X - \mu)^2$ over the probability distribution. It is calculated as the sum of the squared distance of each outcome from the mean, weighted by the probability of that outcome:

$$\text{Var}[X] = E[(X - \mu)^2] = \sum_{x_I} (x_i - \mu)^2 P(X = x_i).$$

For example, the variance of the distribution describing a Bernoulli trial is equal to $p(1 - p)$ because $E[(X - \mu)^2] = (0 - p)^2 P(X = 0) + (1 - p)^2 P(X = 1) = p^2(1 - p) + (1 - p)^2 (p)$, which factors into $p(1 - p)\{(1 - p) + p\} = p(1 - p)$.

There are several alternative ways to measure dispersion around the mean, the two most important being the standard deviation and the coefficient of variation. The *standard deviation* is the square root of the variance, represented by σ. The standard deviation has the same units as the random variable and the mean. In contrast, the variance is in terms of these units squared. The *coefficient of variation, CV*, equals the standard deviation divided by the mean, σ/μ, and is sometimes expressed as a percentage, $\sigma/\mu \times 100\%$. The *CV* is a dimensionless measure of the variability around the mean. It has the advantage of being the same regardless of the measurement scale used (e.g., centimeters or kilometers).

Table P3.1 lists several useful facts that can simplify matters when calculating expectations and variances. For example, we can use the rules of Table P3.1 to derive a

second formula for the variance. We can always expand the square in $E[(X-\mu)^2]$ as $E[X^2 - 2X\mu + \mu^2]$. According to Table P3.1, the expectation of a sum equals the sum of the expectations, yielding $E[X^2] + E[-2X\mu] + E[\mu^2]$. The mean μ is a constant parameter; all such constants can be factored out of expectations, leaving $E[X^2] - 2\,\mu\,E[X] + \mu^2\,E[1]$. Finally, because $E[X] = \mu$ and $E[1] = 1$, we can rewrite the variance as

$$\mathrm{Var}[X] = E[X^2] - \mu^2. \tag{P3.2}$$

The expectation and the variance are two descriptors of a probability distribution and are sometimes referred to as the first and second *central moments* of the distribution. This terminology reflects the fact that they are expectations of the first and second powers of the random variable, after subtracting the mean so that the distributions are "centered" around the mean. In some cases, you might be interested in knowing the skew or kurtosis (peakedness) of a distribution, which are quantities related to higher moments (the third and fourth moments, respectively). After becoming familiar with the material in this Primer, consult Appendix 5 for a general method for finding moments of a distribution using "moment generating functions."

In the following sections, we describe a number of probability distributions that commonly arise in biology. In each case, we provide an overview of the distribution, specify its mean and variance, and describe the contexts in which the distribution is likely to arise. Having a good intuitive sense for the context of each probability distribution is extremely useful. It makes it easier to solve many probabilistic problems that arise in biology by allowing you to make connections between the problem and known facts about probability distributions. Furthermore, in order to incorporate stochasticity into any biological model, you must first choose the most appropriate probability distribution, which is easier to do if you have a good sense of the different possibilities.

TABLE P3.1
Some useful rules of expectations and variances. The rules involving summations assume a discrete probability distribution, but analogous formulas involving integrals exist for continuous probability distributions.

Rule	Notes
$E[c] = c$	If c is a constant
$E[c\,X] = c\,E[X]$	If c is a constant
$E[g(X)] = \sum_{x_i} g(x_i)P(X = x_i)$	The expectation of the function $g(X)$ of a random variable
Geometric mean $X = \prod_{x_i} x_i^{P(X=x_i)}$	The geometric mean of a random variable

Harmonic mean $X = \frac{1}{\sum_{x_i} \frac{1}{x_i} P(X = x_i)}$	The harmonic mean of a random variable
$E[f(X,Y)] = \sum_{x_i} \sum_{y_j} f(x_i,y_j) P(X = x_i \cap Y = y_j)$	The expectation of a function $f(X, Y)$ involving two random variables
$E[X + Y] = E[X] + E[Y]$	The expectation of a sum is the sum of the expectations
$E[X\,Y] = E[X]\,E[Y]$	If X and Y are independent random variables, the expectation of a product is the product of the expectations
$\mathrm{Var}[c] = 0$	If c is a constant
$\mathrm{Var}[c\,X] = c^2\,\mathrm{Var}[X]$	If c is a constant
$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$	If X and Y are independent random variables
$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X\,Y]$	If X and Y are not independent
$\mathrm{Cov}[X\,Y] = E[X\,Y] - E[X]\,E[Y]$	$\mathrm{Cov}[X\,Y]$ describes the "covariance" between X and Y. It equals zero if X and Y are independent.
$\rho = \frac{\mathrm{Cov}[X,Y]}{\sigma_x \sigma_y}$	The "correlation" coefficient standardizes the covariance by the standard deviation of X and Y
$\mathrm{Cov}[X\ \ Y] = E[\mathrm{Cov}_i[X_i\ \ Y_i]] - \mathrm{Cov}[E_i[X_i]\,E_i[Y_i]]$	The *covariance decomposition theorem* calculates the covariance over a set of mutually exclusive classes, i. On the right, Cov[ ] and E[ ] are calculated across classes, weighted by the proportion of the population in each class, p_i, while Cov_i[ ] and E_i[ ] are the covariances and expectations within a class.

Exercise P3.3:

(a) Calculate the variance for the distribution describing a Bernoulli trial using the definition $\mathrm{Var}[X] = E[X^2] - \mu^2$, and show that the variance equals $p\,(1 - p)$.

(b) Imagine doing an experiment involving two independent Bernoulli trials. The

total number of successes could be 0, 1, or 2. Determine the probability of each outcome. Confirm that these probabilities sum to one (Rule P3.10). Determine the mean outcome using Definition P3.2. Determine the variance in the outcome using Definition P3.3.

(c) Show that you can obtain your answers more easily for the mean and variance of two Bernoulli trials using the following facts from Table P3.1: $E[X + Y] = E[X] + E[Y]$ and $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$, where X represents the outcome from one Bernoulli trial and Y represent the outcome from the second Bernoulli trial.

P3.3.1 Binomial Distribution

The binomial distribution generalizes a single Bernoulli trial to n independent trials. In each trial, there are two possible outcomes (say "zero" and "one"), where the probability of outcome "one" is p in every trial. The random variable in a binomial distribution is then the total number of ones observed in n trials, which takes on integer values from 0 to n.

Definition P3.4:

The binomial distribution describes the probability of observing a total of k "ones" in n independent Bernoulli trials:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

$\binom{n}{k}$ is read as "n choose k"; it equals $n!/(k!(n - k)!)$ and represents the number of different ways in which k "ones" can occur over the course of n trials (see Box P3.1 on page 559 at the end of this primer for more details). For example, $\binom{2}{1} = 2!/(1!\ 1!) = 2$, which reflects the fact that there are two ways to get a single "one" in two trials—the "one" can occur on the first trial or on the second trial. By definition, 0! equals one, so that $\binom{2}{0} = 2!/(0!\ 2!) = 1$, which reflects the fact that there is only one way to get zero "ones" in two trials—a "one" must not occur in the first trial or in the second trial.

For $p = 1/2$, the binomial distribution is symmetric and bell-shaped (Definition P1.6), while for p values near 0 or 1, the distribution becomes quite skewed (Figure P3.3).

The mean of a binomial random variable is

$$E[X] = n\,p. \tag{P3.3}$$

This follows from the fact that the binomial represents the sum of n random variables, each of which corresponds to a single Bernoulli trial (see Exercise P3.3). Because the expected value of a sum of random variables equals the sum of the expected values of each random variable (Table P3.1), and because $E[X] = p$ for each Bernoulli trial, the sum of n such trials has an expected value of $n\,p$.

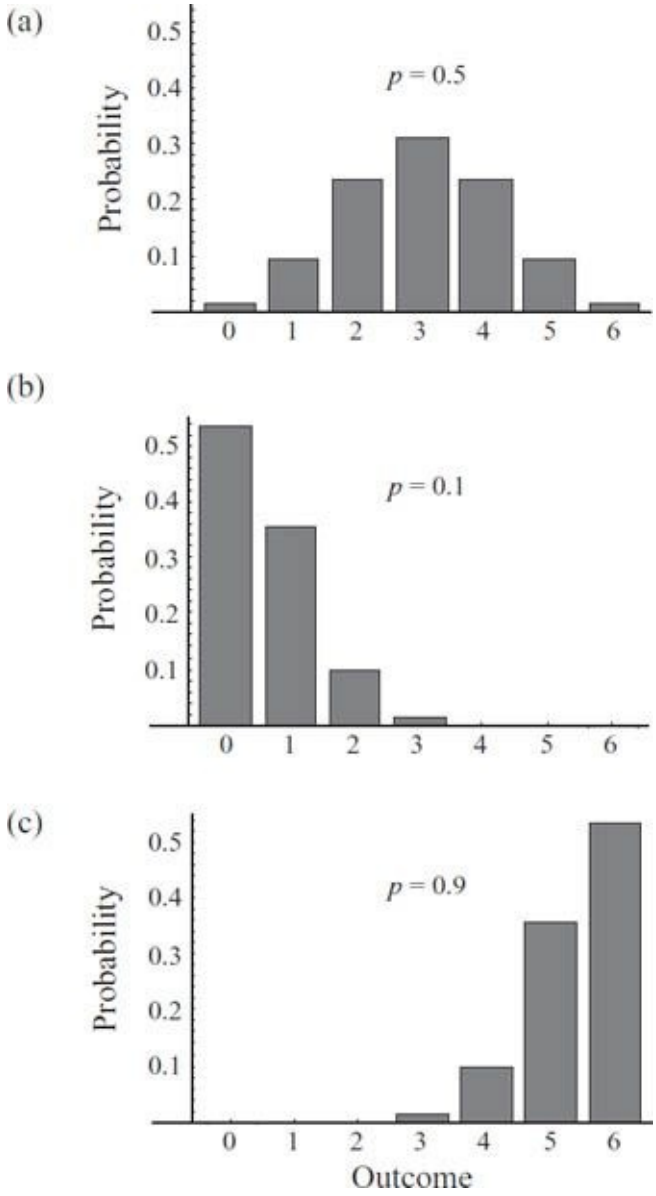

Figure P3.3: Binomial distribution (Definition P3.4). Each bar represents the probability of observing a particular number of successes (from zero to six) among six trials ($n = 6$). The probability of success is (a) $p = 0.5$, (b) $p = 0.1$, (c) $p = 0.9$.

Similarly, the variance of a binomial random variable equals

$$\mathrm{Var}[X] = np\,(1 - p). \tag{P3.4}$$

This follows from the fact that the variance of a sum of independent random variables is the sum of the variance of each random variable (Table P3.1) and from the fact that $\mathrm{Var}[X] = p(1 - p)$ for a single Bernoulli trial.

Examples

The binomial distribution arises when there are a number of independent trials and each trial results in one of two possible outcomes. For example, if there is a 50% probability of having a daughter ($p = 0.5$), the binomial distribution would describe the probability of observing a certain number of daughters in a family, e.g., three daughters among five children: $P(X = 3) = \binom{5}{3} (0.5)^3(0.5)^{5-3} = 5/16$. The binomial distribution also arises when a binary attribute (e.g., diseased or healthy, flowering or not flowering) is measured in a *sample* of size n taken from a population. This assumes that the sample is only a small fraction of the total population or that sampling occurs with replacement so that p does not change as we take our sample.

The binomial distribution is very helpful when interpreting data and simulations. For example, you might collect data on the number of frogs with and without limb defects near a nuclear reactor. Often, you will be collecting such data to estimate an unknown parameter p (e.g., the probability of limb defects). We can estimate p using equation (P3.3) by dividing both sides by n and replacing the expected number of ones, $E[X]$, with the observed number, x, giving an estimate of $\tilde{p} = x/n$, where we have written a tilde over the p to indicate that it is an estimate. We can also estimate $\mathrm{Var}[X/n]$, which describes the variance in this estimate of p. Given that $\mathrm{Var}[c\,X] = c^2\,\mathrm{Var}[X]$ (Table P3.1), the variance of X/n is $\mathrm{Var}[X]/n^2 = n\,p(1 - p)/n^2 = p(1 - p)/n$. Replacing p with its estimated value $\tilde{p}$, the variance of the estimate for p becomes $\tilde{p}(1 - \tilde{p})/n$. Taking the square root of the variance, we get the standard deviation of the estimate for p (referred to as the "standard error of the proportion," SE): $SE = \sqrt{\tilde{p}(1 - \tilde{p})/n}$. As a rule of thumb, the true value p has roughly a 95% chance of lying within two standard errors of the estimate $\tilde{p}$. (Note: Replacing p with $\tilde{p}$ in the variance introduces a bias, especially when $\tilde{p}$ is near 0 and 1. More exact treatments correct for this bias; see Zar (1998).)

P3.3.2 Multinomial Distribution

For some problems, each trial might have more than two possible outcomes. For example, you might want to classify the offspring of a cross as homozygous *AA*,

heterozygous *Aa*, or homozygous *aa*. The multinomial provides such an extension to the binomial distribution.

Definition P3.5:

The multinomial distribution describes the probability of observing $\{k_1, k_2, \ldots, k_c\}$ individuals in each of c discrete categories, where the probability of observing an outcome in category i is p_i.

$$P(X = \{k_1, k_2, \ldots, k_c\}) = \frac{n!}{(k_1!)(k_2!)\cdots(k_c!)} p_1^{k_1} p_2^{k_2} \cdots p_c^{k_c}.$$

The expected number in category i and the variance in this number are

$$E[X_i] = n\, p_i, \tag{P3.5a}$$

$$\mathrm{Var}[X_i] = n\, p_i\, (1 - p_i). \tag{P3.5b}$$

Example

If you were to survey plants within a tropical forest, the probability of observing a certain number of each species would be described by a multinomial distribution, with n equal to the total number of individuals that you sample and p_i equal to the proportion of plants of species i.

P3.3.3 Hypergeometric Distribution

In section P3.3.1, we mentioned that the binomial distribution arises when sampling n individuals from a population that has a proportion p of individuals of a certain type. Technically, this claim is true only if each individual sampled is replaced before the next individual is sampled, otherwise p will change as the sample is gathered, causing the outcome of each trial to depend on the outcomes of previous trials. If sampling occurs without replacement, the hypergeometric distribution describes the distribution of possible samples.

Definition P3.6:

The hypergeometric distribution describes the probability of observing k "ones" in a sample of size n, which is randomly drawn without replacement from a population of size N:

$$P(X = k) = \frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N}{n}},$$

where $N_1 = N\,p$ is the number of "ones" and $N_2 = N(1-p)$ is the number of "zeros" in the total population before sampling.

The probability distribution for a hypergeometric distribution looks complicated but it can be derived by counting up all of the types of samples that could occur (Box P3.1). The denominator represents the number of different ways (i.e., the number of combinations; Box P3.1) in which n individuals can be chosen without replacement from a population of size N, regardless of whether they are successes or failures. For example, there are three ways to chose two individuals ($n = 2$) from a population of size three ($N = 3$): either the first, the second, or the third individual can be left out. Out of all of these possibilities, we then need to count up all of those instances in which there were exactly k successes and $n - k$ failures. Moving to the numerator, the quantity $\binom{N_1}{k}$ is the number of ways (i.e., the number of combinations) in which k successes can be drawn (without replacement) from the subpopulation of N_1 successes without caring about the order in which they occur. For each of these, there are then $\binom{N_2}{n-k}$ different ways (i.e., combinations) in which the desired $n - k$ failures can be drawn (without replacement) from the subpopulation of N_2 failures. Thus the total number of ways in which we can obtain exactly k successes and $n - k$ failures is $\binom{N_1}{k}\binom{N_2}{n-k}$. Consequently, of the $\binom{N}{n}$ ways that we could sample n individuals from a population, only $\binom{N_1}{k}\binom{N_2}{n-k}$ of these will contain k successes and $n - k$ failures. The fraction of samples with k successes is thus given by Definition P3.6.

Using Definitions P3.2 and P3.3, the mean and variance of a hypergeometric random variable are

$$E[X] = n\,p, \tag{P3.6}$$

$$\mathrm{Var}[X] = n\,p\,(1-p)\frac{N-n}{N-1}. \tag{P3.7}$$

The mean is the same as a binomial random variable (P3.3). But, the variance is a factor $(N - n)/(N - 1)$ smaller than the variance of a binomial random variable. The variance decreases toward zero as the sample size approaches the population size (n

→ N), because the composition of the sample becomes nearly the same as the composition of the whole population. Conversely, if the sample size is very small relative to the population size ($n << N$), then $(N - n)/(N - 1)$ approaches one, and the hypergeometric distribution converges upon the binomial distribution.

Example

Imagine that you are studying the nesting behavior of puffins on an island, which contains $N = 100$ suitable nesting cavities. Of these nesting cavities, 30 are on a cliff face that is inaccessible to mammalian predators, while the remainder are on a grassy slope. You watch as the first $n = 20$ puffins choose cavities and begin nesting, and you observe that $k = 11$ choose cliff sites. Thus, among the first nesters, you observe a higher proportion (11/20 = 55%) using cliff sites than expected on the basis of the proportion of cliff sites (30/100 = 30%). The hypergeometric distribution can be used to determine the probability of observing exactly $k = 11$ nesting on the cliff:

$$P(X = 11) = \frac{\binom{30}{11}\binom{70}{9}}{\binom{100}{20}} = 0.0066.$$

Of greater interest is the probability that 11 or more early nesters choose cliff sites, which again can be calculated from the hypergeometric distribution: $P(X \geq 11) = \sum_{k=11}^{20} P(X = k) = 0.0085$. This probability is so low that you can conclude it is unlikely that early nesters are randomly choosing their nest sites and that they appear to prefer cliff sites.

P3.3.4 Geometric Distribution

All of the above distributions (binomial, multinomial, and hypergeometric) describe the number of outcomes that fall into different categories (e.g., cliff nesters vs noncliff nesters). Each outcome falls into some category, and these distributions predict "counts" in each category. In contrast, the distributions discussed next (geometric, negative binomial) were derived to describe how much time passes before a particular outcome (or set of outcomes) is observed. Despite this fundamental difference, the geometric distribution is again based on a series of Bernoulli trials.

Definition P3.7:

The geometric distribution describes the probability that, in a series of

Bernoulli trials, the first success is observed on the kth trial:

$$P(X = k) = p(1 - p)^{k-1}.$$

The geometric distribution is derived as follows. For the first success to occur on the kth trial requires that the previous $k - 1$ trials were unsuccessful. Assuming that each Bernoulli trial is independent of previous ones, the probability of $k - 1$ unsuccessful trials is the product of the probability that each trial is unsuccessful, which is $(1 - p)^{k-1}$ (Rule P3.4a). Following this series of failures, the kth trial will be successful with probability p. Because each trial is independent, we can multiply these terms together to get the geometric distribution.

Because at least one trial must occur to observe a success, k can be any integer greater than or equal to one. $P(X = k)$ always declines with increasing k because every trial that passes unsuccessfully decreases the probability by a factor $(1 - p)$. Thus, the event that there are no failures prior to the first success (i.e., $k = 1$) always has the highest probability (Figure P3.4).

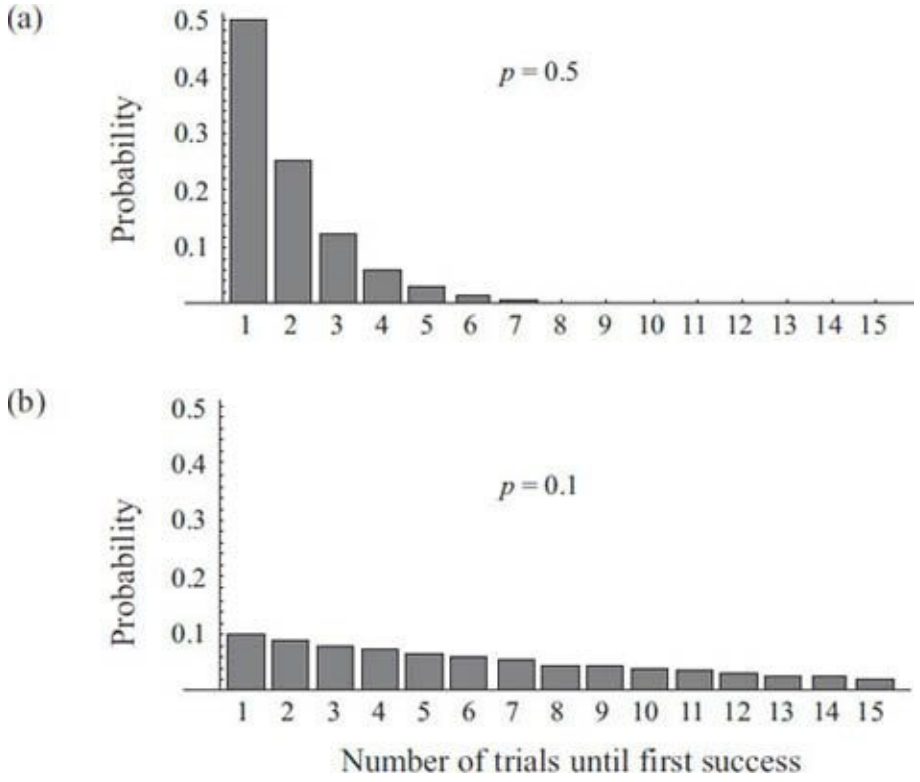

Figure P3.4: Geometric distribution (Definition P3.7). Each bar represents the probability that the first successful event occurs after a particular number of trials (from one to infinity). The probability of success in any one trial is (a) $p = 0.5$, (b) $p = 0.1$.

The mean of a geometric random variable is given by $E[X] = 1/p$. To get more

comfortable working with sums, it is worth deriving this fact. We start with Definition P3.2 giving the mean:

$$E[X] = \sum_{k=1}^{\infty} k\, p\, (1 - p)^{k-1}. \tag{P3.8}$$

This sum is not one of those listed in Appendix A1, but consider taking the derivative of both sides of A1.20 with respect to a, giving us $\sum_{i=1}^{\infty} i\, a^{i-1} = 1/(1 - a)^2$ (To see this, it might help to think about writing out the summation as $a + a^2 + a^3 + \cdots$). If we factor out p from (P3.8) and let $a = 1 - p$, the sum in (P3.8) can be written as $p\sum_{k=1}^{\infty} k\, a^{k-1}$, which equals $p/(1 - a)^2$. Plugging in $a = 1 - p$, the mean equals

$$E[X] = \frac{1}{p}. \tag{P3.9}$$

In a similar fashion, the variance of the geometric random variable is

$$\mathrm{Var}[X] = \frac{1 - p}{p^2}. \tag{P3.10}$$

Examples

The number of courtship displays made by a male before he successfully mates might be described by a geometric random variable. Here, each time a male displays is a Bernoulli trial resulting in a mating ("success") or not ("failure"). The key assumption for this process to be described by a geometric distribution is that the probability that a mating attempt succeeds remains constant over time and is not influenced by ("is independent of") the outcome of previous mating attempts. The geometric distribution might also describe the time until extinction of an endangered population that is censused yearly, if the probability of extinction is constant. Thus, with an annual extinction risk of 10% ($p = 0.1$), the expected time until extinction is ten years (i.e., $1/p = 1/0.1 = 10$). The variance in this case is pretty large (90 years squared). This means that the actual year in which the population goes extinct is very hard to predict, as suggested by Figure P3.4.

P3.3.5 Negative Binomial Distribution

The negative binomial distribution generalizes the geometric distribution and describes the waiting time until r "successes" have occurred:

Definition P3.8:

The negative binomial distribution describes the probability that, in a series of Bernoulli trials, the rth success is observed on the kth trial:

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}.$$

For the rth success to occur on the kth trial, there must have been $r-1$ successes in the previous $k-1$ trials. We have already described the probability of observing a certain number of successes out of a total number of trials: it is given by the binomial distribution. Thus, we can write the probability distribution for the negative binomial as the product of the binomial probability of observing $r-1$ successes out of $k-1$ trials, $\binom{k-1}{r-1} p^{r-1} (1-p)^{k-r}$, multiplied by p, the probability that the kth trial is a success.

Because at least r trials must occur to observe r successes, k can be any integer greater than or equal to r. Now, $P(X = k)$ does not always decline with an increasing numbers of trials (k). In fact, if we were waiting for a large number of successful outcomes, the negative binomial distribution has a bell shape (Figure P3.5).

We can think of the negative binomial distribution as describing the sum of r independent random variables: the sum of the waiting times before each of the r successful trials. Each of these waiting times follows a geometric distribution with mean $1/p$ and variance $(1-p)/p^2$. Using the fact that the expectation of a sum is the sum of the expectations, the number of trials until the rth success is expected to equal

$$E[X] = \frac{r}{p}. \tag{P3.11}$$

Similarly, because the variance of a sum of independent random variables is the sum of the variance of each random variable (Table P3.1), the variance of a negative binomial random variable is:

$$\mathrm{Var}[X] = \frac{r(1-p)}{p^2} \tag{P3.12}$$

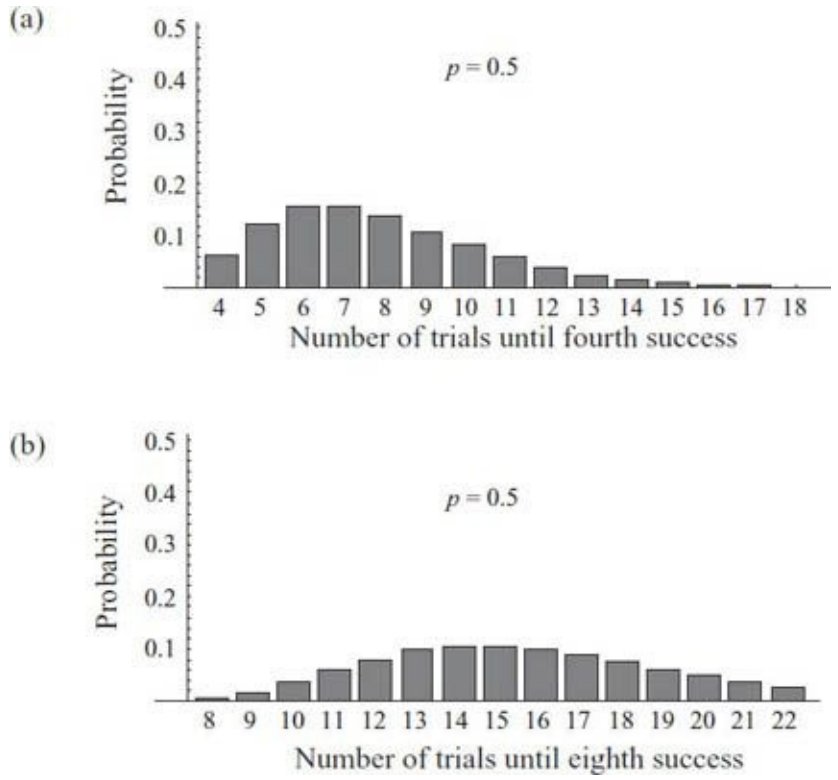

Figure P3.5: Negative binomial distribution (Definition P3.8). Each bar represents the probability that (a) the fourth successful event occurs after a particular number of trials (from four to infinity) and (b) the eighth successful event occurs after a particular number of trials (from eight to infinity). The probability of success in any one trial is $p = 0.5$ (Figure P3.4a describes the comparable probability distribution for the first successful event).

Example

If a predator must capture $r = 10$ prey before it can grow sufficiently large to reproduce, and if it has a 10% success rate per hunt ($p = 0.1$), the age of onset of reproduction would be described by the negative binomial distribution. On average, the predator must go on 100 hunts before it can reproduce, where the variance in this number is 900 hunts^2 ($SD = 30$ hunts).

P3.3.6 Poisson Distribution

The last of the discrete distributions that we will visit is the Poisson distribution. It differs fundamentally from the above distributions because it describes neither the numbers that fall into various categories (binomial, multinomial, and hypergeometric) nor the waiting time until a certain number of events have occurred (geometric, negative binomial). Rather, the Poisson distribution describes the number of events that occur in a given time period (or within a given area) when events occur randomly and independently over time (or space). The Poisson distribution naturally arises when counting the number of events witnessed during an observation period, such as the number of birds that stray onto an island within a year or the number of seedlings that germinate on a plot within a week.

Definition P3.9:

The Poisson distribution describes the probability of observing k events in a given space or time period when the expected number of events is μ and when each event occurs independently:

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$$

If you know the rate at which events occur per unit time (λ), then the expected number of events is $\mu = \lambda t$, where t is the time period of observation. Similarly, if you know the density of events per unit area (δ), then the expected number of events is $\mu = \delta A$, where A is the area under observation. The actual number of events that occurs, k, can be any integer from 0 to infinity. When μ is small, the Poisson distribution is skewed, and the probability of observing no events or only one is high. Alternatively, when μ is large, the Poisson distribution becomes bell shaped, with the most likely number of observations centered on μ (Figure P3.6).

The Poisson distribution has an unusual attribute in that its mean equals its variance:

$$\begin{aligned} E[X] &= \mu \\ &= \text{Var}[X]. \end{aligned} \tag{P3.13}$$

Another important attribute of the Poisson distribution is that the sum of a number of Poisson random variables is itself Poisson distributed (Supplementary Material P3.2). For example, if the number of hemlock seedlings, the number of cedar seedlings, and the number of fir seedlings that emerge within a square meter is Poisson distributed, then their sum, describing the total number of tree seedlings that emerge, will also be Poisson distributed.

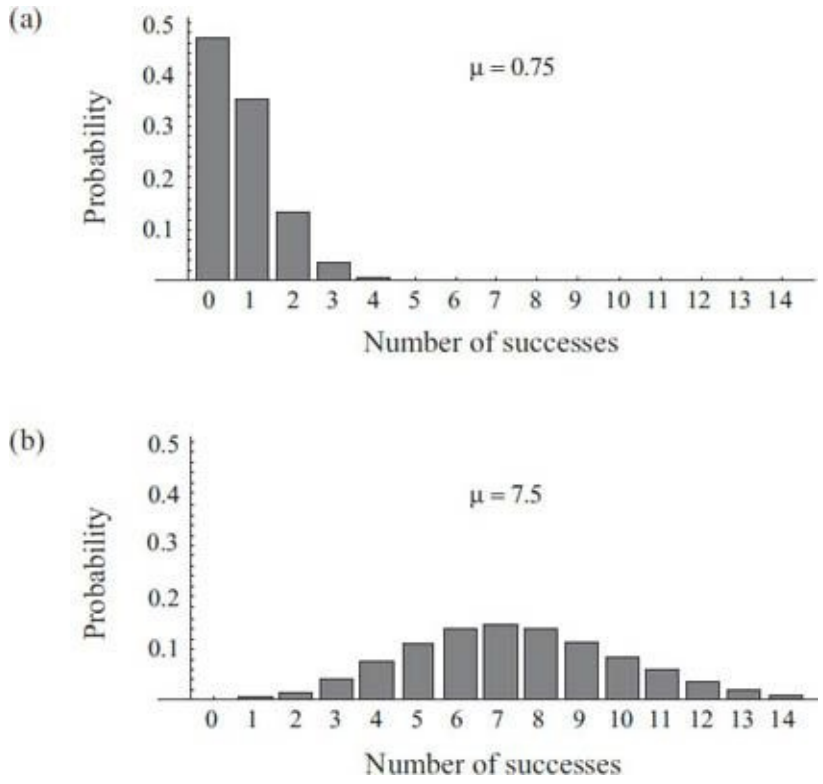

Figure P3.6: Poisson distribution (Definition P3.9). Each bar represents the probability of observing a certain number of events when the expected number is (a) $\mu = 0.75$ and (b) $\mu = 7.5$.

Examples

If hummingbirds arrive at a flower at a rate $\lambda = 0.2$ per minute, the expected number of visits in $t = 20$ minutes of observation would be $\mu = 4$. Assuming that hummingbirds arrive independently and randomly over time, we would expect the actual number of visits to be Poisson distributed with a mean and variance of 4. If the observed variance is significantly lower, this would call into question the assumption that hummingbird visits occur independently over time and indicates that birds tend to space out their visits, causing the visits to be more evenly distributed than expected under the Poisson distribution.

As another example, the Poisson distribution describes the number of new mutations that an individual is expected to carry. In the diploid human genome, there are about $A = 6.4 \times 10^9$ basepairs and the mutation rate per generation per basepair is approximately $\delta = 1.8 \times 10^{-8}$ (Kondrashov 2003). In this case, we are monitoring a particular area (A, here the stretch of DNA) for events that occur at a particular density (δ). The expected number of events is then $A \times \delta = 115.2$. According to a Poisson distribution, the variance should also equal 115.2, and the standard deviation should be $\sqrt{115.2} = 10.7$. Furthermore, if we plot the Poisson distribution with mean 115.2, we can predict that about 95% of us carry between 96 and 136 new mutations that were not present within our parents (Figure P3.7).

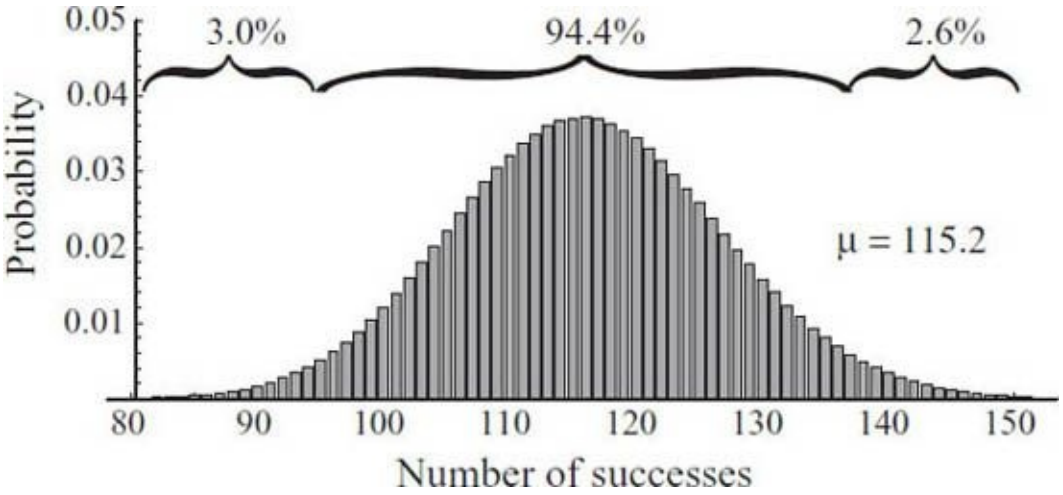

Figure P3.7: The distribution of the number of mutations according to a Poisson distribution with mean $\mu = 115.2$ (Definition P3.9), the number of mutations is expected to fall between 96 and 136 in 94.4% of cases. (From 0 to 95 accounts for 3.0% of the distribution; from 137 to infinity accounts for 2.6% of the distribution. These ranges were chosen to be the smallest possible ranges that accounted for >2.5% of the distribution each.)

Although we have used a Poisson distribution to describe the number of mutations, technically, this assumes that mutations can occur at any real-valued position and ignores the discrete nature of the nucleotides that make up chromosomes. In reality, a mutation alters the first nucleotide, the second nucleotide, . . . or the nth nucleotide in the sequence. We can capture the discrete nature of mutations using a binomial distribution to describe the probability of observing k mutations out of $n = 6.4 \times 10^9$ nucleotides, where the probability of a mutation (a "success") is $p = 1.8 \times 10^{-8}$ per generation. According to the binomial distribution, the mean number of mutations is $n\,p = 115.2$ and the variance is $n\,p\,(1 - p) \approx 115.2$. Interestingly, these are the same mean and variance predicted by the Poisson distribution. In fact, the binomial distribution converges upon a Poisson distribution whenever the probability of success, p, is small and the number of trials, n, is large. In this case, the variance of the binomial, $np\,(1 - p)$, is nearly equal to the mean, np, which is a property of the Poisson distribution. The higher moments also converge, as can be shown using moment generating functions (Appendix 5).

Exercise P3.4: Unlike the Poisson distribution, the sum of two independent random variables, each following a binomial distribution, is not generally binomial. Let X represent the outcome from n_1 Bernoulli trials each of which has a probability of success of p_1, and let Y represent the outcome from n_2 trials with probability of success p_2. What is the variance of the sum of these two random variables, $X + Y$? Show that $\mathrm{Var}[X + Y]$ cannot be factored into the form $n\,p\,(1 - p)$ and so does not equal the variance expected if the sum were binomial, unless $p_1 = p_2$. As an example, consider $n_1 = 100$ trials with $p_1 = 0$ and $n_2 = 100$ trials with $p_2 = 1$.

What is Var[$X + Y$]? For comparison, what variance would you expect from the binomial distribution with $n = 200$ trials and an average proportion of successes, $p = 1/2$?

P3.4 Continuous Probability Distributions

In the previous distributions, the possible outcomes were discrete (e.g., integers from 0 to n). What if you were interested in a random variable that could take on any real value (e.g., any point in time)? Random variables that can take on a continuum of possible values are known as *continuous random variables*. The procedures described above to calculate the total probability, mean, and variance for discrete random variables are similar to the procedures for continuous random variables, but there is one crucial difference. Imagine calculating the sum in Rule P3.10, $\sum_{x_i} P(X = x_i) = 1$, for a continuous random variable. Because there is a continuum of possible outcomes (e.g., all points in time), this sum would be infinitely large if $P(X = x_i)$ were finite for every possible value of x_i. Even for a continuous random variable, however, the total probability that the random variable takes on some value must be one, not infinity.

How is this discrepancy resolved? It is resolved by recognizing that, with a continuum of possible outcomes, the probability of any one particular outcome is not a finite number but is, instead, infinitesimally small. For example, if a continuous random variable lies between 0 and 1, the probability of it taking on the exact value of, say, 1/8 (i.e., 0.12500000 . . .) is essentially zero. The same is true for any other particular value. Because we cannot talk about the probability of any one outcome, we instead describe the probability that the random variable X falls within a small interval dx of x:

$$P(x < X < x + \mathrm{d}x) = f(x)\,\mathrm{d}x, \qquad \text{(P3.14)}$$

where $f(x)$ is known as the "*probability density function*" describing the probability distribution for a continuous random variable. Typically, $f(x)$ is drawn as a curve over the region of possible outcomes x (Figure P3.8). Equation (P3.14) can be interpreted as the area of a histogram with a height of $f(x)$ and a very small width of dx, which gives the probability that the random variable falls within a region from x to $x + \mathrm{d}x$. More generally, the area under the curve between any two points, a and b, equals the probability that the random variable falls between a and b (Figure P3.8):

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x. \qquad \text{(P3.15)}$$

Thus, the probability density function tells us the regions in which the random variable is likely to fall (high $f(x)$) or unlikely to fall (low $f(x)$).

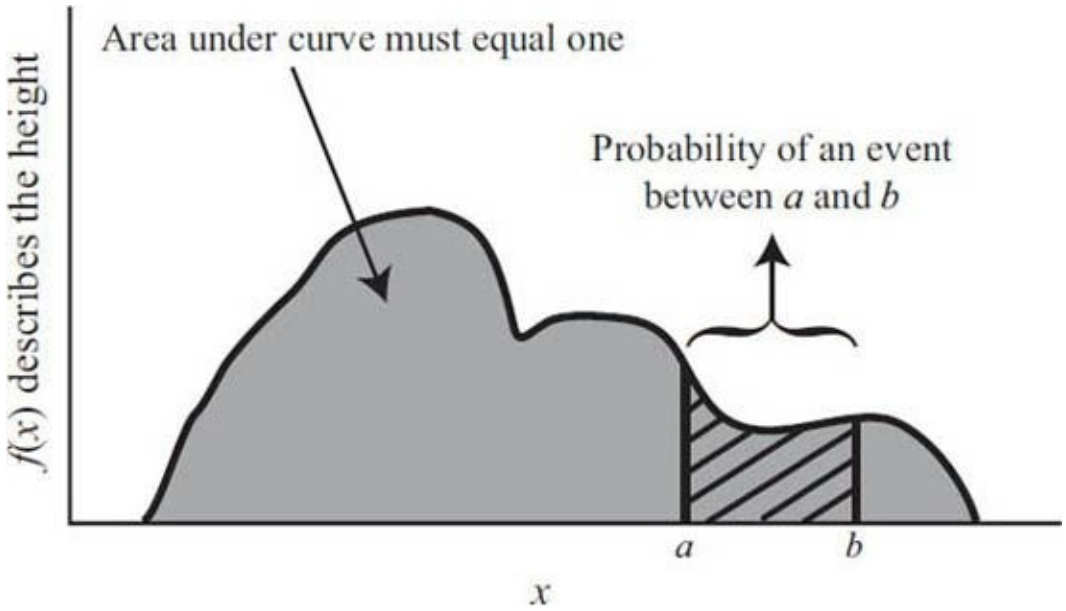

Figure P3.8: A probability density function with height $f(x)$. To represent a probability distribution, the area under the curve must equal one, and its height must never be negative. The probability of falling within any particular interval (e.g., between points a and b) is the area under the curve within this interval (hatched). Regions in which the curve is high ($f(x)$ is large) are more probable than regions in which the curve is low.

Replacing $P(X = x_i)$ with $f(x)\,\mathrm{d}x$ and summations with integrals, we can proceed to analyze continuous random variables as before. In particular, because the random variable must take on some outcome, we have the following rule:

Rule P3.11: The Integral of a Continuous Probability Distribution

The integral of $f(x)$ over the range of possible outcomes must equal one:

$$\int_{\min}^{\max} f(x)\,\mathrm{d}x = 1.$$

As with a discrete probability distribution, the mean of a continuous probability

distribution equals the average value that would be obtained from an infinite number of draws from the distribution. To calculate this average, we use a definition analogous to Definition P3.2:

Definition P3.10a: The Mean of a Continuous Random Variable

The mean of a continuous random variable is given by integrating the value of each outcome x weighted by the probability density function $f(x)$ over the range of possible outcomes:

$$\mu = E[X] = \int_{\min}^{\max} x\, f(x)\, dx.$$

Again, the mean can be visualized as the "center of mass" of the probability distribution represented by the curve $f(x)$.

The variance of a continuous random variable is calculated in a similar fashion:

Definition P3.10b: The Variance of a Continuous Random Variable

The variance of a continuous random variable is the integral of the squared distance of each outcome x from the mean, weighted by the probability density function $f(x)$:

$$\mathrm{Var}[X] = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \ .$$

Equivalently, the variance equals

$$\mathrm{Var}[X] = E[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 f(x)\, dx \right) - \mu^2.$$

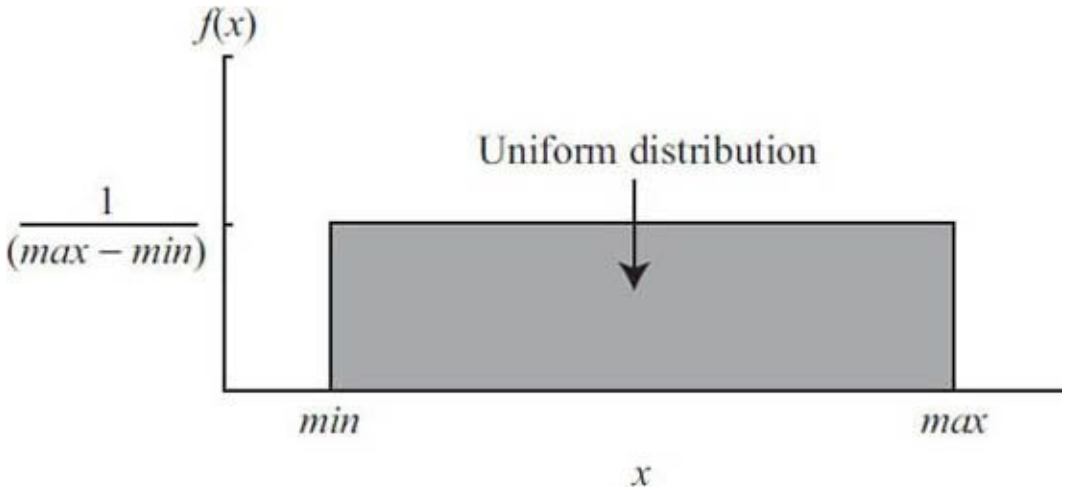

Figure P3.9: Uniform distribution (Definition P3.11). All outcomes between $x = min$ and $x = max$ are equally likely, and no outcome outside this region can occur.

We turn next to a description of some of the most important continuous probability distributions.

P3.4.1 Uniform Distribution

The uniform distribution is the simplest continuous probability distribution. It describes a random variable that is equally likely to fall at any point within a range from *min* to *max* (Figure P3.9). Within this range, the probability density function has a constant height $f(x) = h$, which is calculated from the fact that the integral over the range of possible values must equal one (Rule P3.12):

$$\int_{\min}^{\max} h \, dx = h \max - h \min = 1, \qquad \text{(P3.16)}$$

Solving for h, we get $h = 1/(\max - \min)$.

Definition P3.11:

The uniform distribution describes the probability density at x for a random variable when all outcomes between min and max are equally likely:

$$f(x) = \frac{1}{\max - \min} \quad \text{for } \min \leqslant x \leqslant \max$$

and $f(x) = 0$ for x outside of min and max.

Intuitively, the mean of a uniform probability density function occurs halfway between min and max. Indeed, applying Definition P3.9, we get

$$E[X] = \int_{\min}^{\max} x \frac{1}{\max - \min}\, dx$$

$$= \left(\frac{x^2}{2} \frac{1}{\max - \min}\right)\Bigg|_{\min}^{\max}$$

$$= \frac{(\max^2 - \min^2)}{2(\max - \min)} \qquad \text{(P3.17)}$$

$$= \frac{\max + \min}{2}.$$

The variance is not so easy to intuit, but it can be determined using Definition P3.10:

$$\begin{aligned} \mathrm{Var}[X] &= \int_{\min}^{\max} (x - \mu)^2 f(x)\, dx \\ &= \int_{\min}^{\max} \left(x - \frac{\max + \min}{2}\right)^2 \frac{1}{\max - \min}\, dx \\ &= \left(\frac{1}{3}\left(x - \frac{\max + \min}{2}\right)^3 \frac{1}{\max - \min}\right)\Bigg|_{\min}^{\max} \\ &= \frac{(\max - \min)^2}{12}. \end{aligned} \qquad \text{(P3.18)}$$

Examples

Imagine that you are studying mating behavior in *Drosophila*. The flies are in a cage and reproducing continuously. For your study, you watch the flies in ten-minute intervals (600 seconds) and record each time a mating takes place. You notice that out of 100 matings, none occur within the first 20 seconds. This makes you concerned that the initial handling might affect the behavior of the flies. To test this, you determine the probability that a mating occurs any time after 20 seconds:

$$\int_{20}^{600} \frac{1}{600} dx = \frac{29}{30}. \qquad \text{(P3.19)}$$

This is the probability that one mating is observed after 20 seconds, assuming that the probability of mating is uniformly distributed over the 600 seconds of observation. The probability that all 100 matings occur after 20 seconds would then be $(29/30)^{100} = 0.034$ (Rule P3.4a). As this is a small probability, you conclude that handling might well have an influence on the flies, making them initially less likely to mate.

As another example, imagine a chromosome that is 2 Morgans in length. (A Morgan gives the distance along a chromosome within which one recombination event, or *crossover*, occurs, on average.) Among those chromosomes containing a single crossover, the mean observed position of the crossover is at 1 Morgan with a variance of 1/2. If crossovers occurred uniformly across the chromosome (min = 0 and max = 2 Morgans), we would expect the mean position to be at 1 Morgan with a variance of $(2 - 0)^2/12 = 1/3$. Thus, there is more variance than expected based on a uniform distribution of crossover positions. If this increase in variance were significant, it would suggest that crossovers are less likely to occur near the middle of the chromosome.

P3.4.2 Exponential Distribution

The exponential distribution arises when measuring the time until an event first occurs in continuous time.

Definition P3.12:

The exponential distribution describes the probability density of the waiting time x until an event first occurs under the assumption that events occur at a constant rate α per unit time:

$$f(x) = \alpha e^{-\alpha x} \qquad \text{for } 0 \le x \le \infty$$

Here we write the waiting time as x rather than t to be consistent with the other probability distributions. We also use the rate parameter α to be consistent with the gamma distribution described next. In the biological literature, λ is often used as the rate parameter in place of α.

To derive the exponential distribution, consider calculating the probability $P(x)$ that

no events occur before time x. From the results of previous chapters, we can write a recursion equation for P over a small time step dx as $P(x + dx) = P(x)\ (1 - \alpha dx)$. In other words, the probability that the event has still not occurred at time $x + dx$ is just the probability that it had not occurred at time x, multiplied by the probability it does not occur in the time interval dx. As we saw in Box 2.6 of Chapter 2, we can rearrange this as $(P(x + dx) - P(x))/dx = \alpha\ P(x)$ and then take that limit as dx gets small to obtain the differential equation $dP/dx = -\alpha\ P(x)$. This differential equation has the form of the exponential growth model and can be solved to get $P(x) = e^{-\alpha x}$ (see Chapter 6). For the event to occur for the first time near time x (i.e., between time x and $x + dx$), we multiply the probability that the event does not happen before time x, $P(x) = e^{-\alpha x}$, by the probability that the event does occur in the short interval of time dx, which is αdx. This gives us $\alpha e^{-\alpha x}\ dx$, which equals the exponential probability distribution $f(x)$ times the time interval dx.

The exponential distribution starts at height, α, when $x = 0$ and declines exponentially with x at rate α (Figure P3.10). The total area under the curve correctly integrates to one (Rule P3.11):

$$\int_0^\infty \alpha\ e^{-\alpha x}\ dx = -\left.\frac{\alpha\ e^{-\alpha x}}{\alpha}\right|_0^\infty = 1. \tag{P3.20}$$

Using Definitions P3.10a and P3.10b, the mean and variance of an exponential random variable are:

$$E[X] = \frac{1}{\alpha}, \tag{P3.21}$$

$$\mathrm{Var}[X] = \frac{1}{\alpha^2} \tag{P3.22}$$

(see Exercise P3.6).

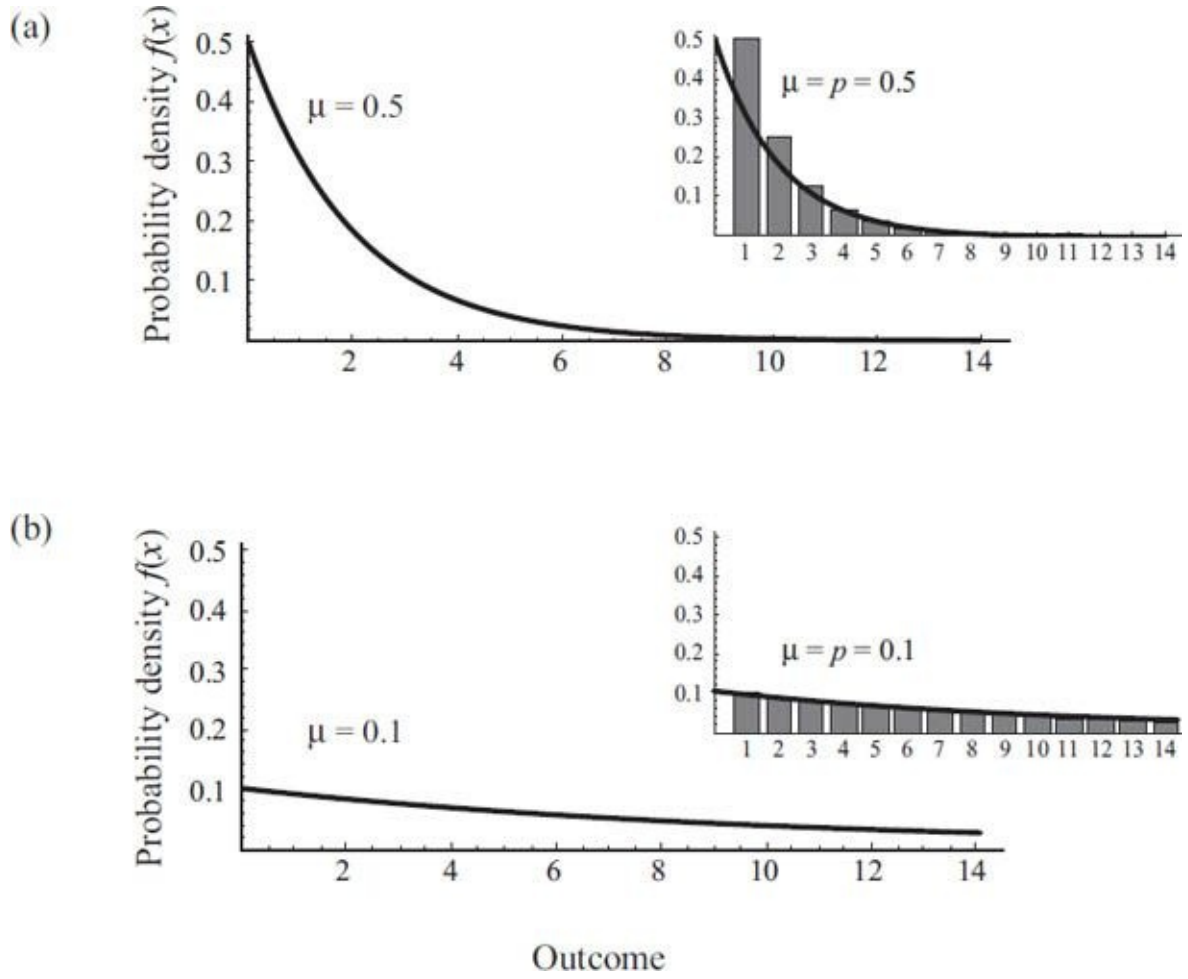

Figure P3.10: Exponential distribution (Definition P3.12). The probability density function, $f(x)$, declines exponentially with x. Although the expected outcomes are (a) $\mu = 0.5$ and (b) $\mu = 0.1$, the most likely outcomes are always near zero ($f(x)$ is highest at zero). The inset figures compare the exponential distribution (continuous curves) to the geometric distribution (discrete bars). The two distributions are similar when μ and p are similar and small.

The exponential distribution applies only if the rate of events per unit time is constant. For the rate to be constant, the probability that the event occurs in a short interval of time dx must always be the same, αdx. This is not the same as events occurring at regular intervals, which would exhibit a rate of zero except at these regular points in time. It is not necessary, however, for all events to be identical in kind; the exponential distribution continues to apply when events can be broken down into subcategories, just as we saw with the Poisson distribution. For example, the death rate from cancer might be α_1, the death rate from heart attacks might be α_2, and the death rate from other causes might be α_3. Yet if we are only interested in the time until death, its distribution would be exponential with a rate parameter equal to the total death rate $\alpha = \alpha_1 + \alpha_2 + \alpha_3$, as long as α is constant over time.

Examples

If individuals die at a constant rate α per unit time, the lifespan of the individual is

described by an exponential distribution. Here, the lifespan is measured in continuous time so that an individual could have, for example, a lifespan of 70.23853 years. This differs from the geometric distribution, which can also be used to describe the age at which an individual dies but which measures lifespan in discrete age classes (e.g., 70 or 71 years). Which distribution is most appropriate depends on the precision desired as well as the information provided about the chance of death. If the chance of death is given as an instantaneous death rate, α, then the exponential distribution should be used. If the chance of death is given as the probability of death within a year, p, then the geometric distribution should be used. That said, we can always convert a death rate α into the mortality risk in one year using the integral

$$p = P(\text{death per year}) = \int_{x=0}^{1} \alpha\, e^{-\alpha x}\, dx = 1 - e^{-\alpha}. \qquad \text{(P3.23)}$$

Equation (P3.23) provides a way of translating between an exponential distribution in continuous time and a geometric distribution in discrete time (see Exercise P3.5). As a numerical example, if the death rate is $\alpha = 0.1$ per year, then the probability of dying within a year is $p = 0.095$. These numbers predict similar mean life spans ($E[X] = 1/\alpha = 10$ years according to the exponential distribution versus $E[X] = 1/p = 10.5$ years according to the geometric distribution). In fact, whenever the rate of events is small, the exponential and geometric distributions are very similar in shape with $p \approx \alpha$ (Figure P3.10).

The exponential distribution also arises when measuring the distance traveled until a certain event occurs, assuming that the event occurs at a constant rate per distance. For example, if a bee is foraging and stops at flowers at a constant rate, α, per meter, then the distance until it stops at a flower would be exponentially distributed. If α were 0.05 per meter, then the mean distance traveled between flowers would be $E[X] = 1/\alpha = 20$ meters with a standard deviation of $SD[X] = 1/\alpha = 20$ meters. Essentially, we are measuring a waiting time in this example, but in terms of meters traveled rather than chronological time.

Exercise P3.5: Based on the exponential distribution, calculate the probability that an event occurs for the first time between the interval of time $k - 1$ and k. Rewrite your answer in terms of the annual mortality risk by replacing $e^{-\alpha}$ with $1 - p$ (see equation (P3.23)). Show that the result, which describes the probability that the event first occurs within the time interval between $k - 1$ and k, is the same as the geometric probability distribution.

Exercise P3.6: [Advanced]

(a) Calculate the mean for an exponential distribution using Definition P3.10a. Remember to restrict the range of x from 0 to positive infinity.

(b) Calculate the variance for an exponential distribution using Definition P3.10b. Remember to restrict the range of x from 0 to positive infinity.

(c) Calculate the mean and the variance for an exponential distribution using the fact that its moment generating function is $MGF(z) = \alpha/(\alpha - z)$ (see Appendix 5).

P3.4.3 Gamma Distribution

The gamma distribution generalizes the exponential distribution by describing the waiting time until β events occur:

Definition P3.13:

The gamma distribution describes the probability density that x amount of time passes before β events occur when each event occurs at a constant rate, α, per unit time:

$$f(x) = \frac{\alpha^{\beta}}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha x} \qquad \text{for } 0 \le x \le \infty,$$

Where $\Gamma(\beta) = \int_{y=0}^{\infty} e^{-y} y^{\beta-1} \, dy$ is known as the gamma function.

When $\beta = 1$, $\Gamma(1) = 1$ and the gamma distribution reduces to the exponential distribution.

The gamma distribution can be derived by drawing a connection to the Poisson distribution. For x to be the first time that β events have occurred, there must have been $\beta - 1$ events within the time interval from 0 to x. The probability of observing $\beta - 1$ events in a fixed time interval is described by the Poisson distribution, with an expected number of events equal to the rate of events times the time interval: $\mu = \alpha x$. This Poisson probability must then be multiplied by the probability that the βth and final event occurs between time x and $x + dx$ (i.e., α dx). The probability density function for observing β events for the first time near time x is therefore

$$f(x) = \underbrace{\left(\frac{e^{-\alpha x}(\alpha x)^{\beta-1}}{(\beta-1)!}\right)}_{\substack{\text{Poisson distribution of observing} \\ \beta - 1 \text{ events given a mean of } \alpha x}} \alpha. \tag{P3.24}$$

This derivation assumes that β is an integer, but the gamma distribution is typically written in a more general fashion that allows for any positive value of β (see Definition P3.13). To generalize (P3.24), we replace the factorial $(\beta - 1)!$, which is defined for integers only, by a new constant that is chosen to ensure that $f(x)$ integrates to one. Using equation (P3.24) in Rule P3.11, this constant must equal $\int_{x=0}^{\infty} e^{-\alpha x}\alpha^{\beta}x^{\beta-1}\mathrm{d}x$. This integral can be simplified by rewriting it in terms of $y = \alpha x$ (and hence $\mathrm{d}y = \alpha\, \mathrm{d}x$), giving $\int_{y=0}^{\infty} e^{-y}\alpha^{\beta}(y/\alpha)^{\beta-1}(\mathrm{d}y/\alpha)$. The α terms cancel out of this integral, leaving us with $\int_{y=0}^{\infty} e^{-y}y^{\beta-1}\,\mathrm{d}y$, which is known as the gamma function, $\Gamma(\beta)$. The gamma function generalizes factorials to any real number; when β is an integer, $\Gamma(\beta) = (\beta - 1)!$. For more facts involving gamma functions, see Abramowitz and Stegun (1972).

The gamma distribution in Definition P3.13 can be thought of as the continuous-time version of the negative binomial distribution. The main difference is that the probability density function is positive for the gamma distribution regardless of how little time has passed, because there is always some small chance that all β events occur in rapid succession. In contrast, we must wait until at least r trials have passed in discrete time before the probability of observing r events is positive with the negative binomial distribution.

The mean and variance of a gamma distribution can be calculated using Definition P3.10. It is easier, however, to use the fact that the gamma distribution represents the sum of β waiting times, each of which is exponentially distributed. Because the expectation of a sum of independent random variables is the sum of the expectations and the same is true for the variance (Table P3.1), we can multiply the mean and variance of the exponential distribution by β to get the mean and variance of the gamma distribution:

$$E[X] = \frac{\beta}{\alpha}, \tag{P3.25}$$

$$\mathrm{Var}[X] = \frac{\beta}{\alpha^2} \tag{P3.26}$$

(see Exercise P3.7).

Examples

Consider an experiment in which you wish to study $\beta = 100$ grooming events in a baboon colony. If the rate of grooming events is two per hour ($\alpha = 2$/hour), then you would expect the study to take 50 hours ($\beta/\alpha = 100/2$ hours) with a standard deviation of 5 hours ($\sqrt{\beta/\alpha^2}$). Furthermore, you can use the gamma distribution to tell you what the chances are that you complete the study by any given time. For example, if you were only able to have 60 hours of observation time, the probability that you will successfully observe 100 grooming events would be 97.2%, as calculated from the integral

$$\int_{x=0}^{60} \frac{\alpha^{\beta}}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha x} \, dx = 0.972.$$

As another example, imagine that you are collecting truffles in a forest and you find one truffle every 200 meters. The rate at which you encounter truffles is thus $\alpha = 1/(200$ meters). If you wish to collect 30 truffles ($\beta = 30$), you can expect to walk $E[X] = \beta/\alpha =$ 6000 meters. Again, this is fundamentally similar to a waiting time problem, where we are measuring the waiting time in terms of meters traveled.

The shape of the gamma distribution varies from L shaped when β is small to bell shaped when β is large (Figure P3.11, Exercise P3.7). Thus, β is often called the "shape" parameter for the gamma distribution. In contrast, α is called the "scale" parameter. Increasing or decreasing α while holding β constant does not change the shape of the distribution.

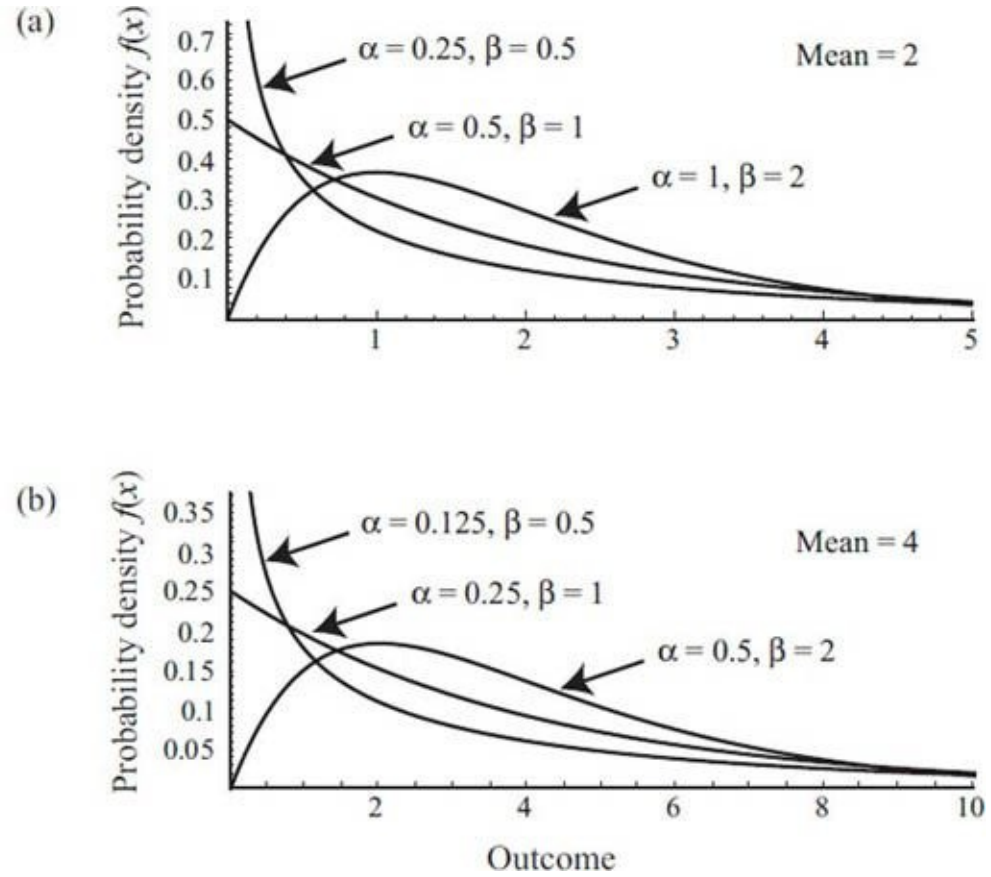

Figure P3.11: Gamma distribution (Definition P3.13). The probability density function, $f(x)$, is plotted for various values of α and β, holding the mean value constant at (a) $E[X] = 2$ and (b) $E[X] = 4$. When $\beta = 1$, the shape of the gamma distribution is the same as the exponential distribution. When $\beta < 1$, the distribution is more L-shaped, with substantial probability density near zero. When $\beta > 1$, the distribution is more bell-shaped. Because the same set of β values is used in (a) and (b), the shapes of the curves are the same, but the horizontal axis has expanded and the vertical axis has shrunk in (b) because the mean has doubled.

Because the gamma distribution is so flexible in shape, it is often used to describe the distribution of an unknown parameter. For example, the gamma distribution is used in analyzing DNA sequences to describe the variation in mutation rates among sites. The use of the gamma distribution is not rigorously justified in this case. This application does not involve the sum of waiting times, for example. Instead, the gamma distribution is used as a heuristic description of what the distribution of substitution rates might look like. Sequence data can then be used to estimate α and β, providing us with information about whether there is a large (β small) or small (β large) degree of variation among sites in mutation rate (Felsenstein 2004; Keightley 1994).

Exercise P3.7:

(a) Calculate the coefficient of variation for the gamma distribution.

(b) Rewrite the probability density function for the gamma distribution replacing

α and β in terms of the mean and coefficient of variation.

(c) What must the coefficient of variation be for the gamma distribution to reduce to the exponential distribution? Would smaller values of the coefficient of variation correspond to more L-shaped or more bell-shaped distributions?

P3.4.4 Normal (Gaussian) Distribution

Arguably the most important distribution in biology is the normal or Gaussian distribution:

Definition P3.14a:

The normal distribution is bell-shaped, with a probability density function that falls off exponentially with the squared distance to the mean:

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \qquad \text{for } -\infty \le x \le \infty.$$

The denominator ensures that the distribution integrates to one. The mean and variance of the normal distribution are

$$E[X] = \mu, \tag{P3.27}$$

$$\text{Var}[X] = \sigma^2. \tag{P3.28}$$

When σ^2 is small (low variance), the probability density function is very narrow and drops off rapidly in height away from the mean, so that most observations are expected to lie near the mean. Conversely, when σ^2 is large (high variance), the probability density function is very broad (Figure P3.12a).

Historically, the normal distribution has appeared in many different contexts. The normal distribution was first described by Abraham de Moivre (1667–1754), who used it to approximate the binomial distribution and to provide gambling advice to rich patrons. Others, including Pierre Simon de Laplace (1749–1827) and Carl Friedrich Gauss (1777–1855), noticed that measurement errors tend to be normally distributed. In the nineteenth century, Adolphe Quetelet (1796–1874) and Francis Galton (1822–1911) observed that the heights and weights of human and animal populations, along with many other characteristics, roughly follow a normal distribution.

Why does the normal distribution play such a ubiquitous role? The reason lies in one of the most important theorems in statistics first developed by Laplace and known as the central limit theorem.

Rule P3.12: Central Limit Theorem

The sum (or the average) of n independent and identically distributed random variables tends toward a normal distribution as n goes to infinity.

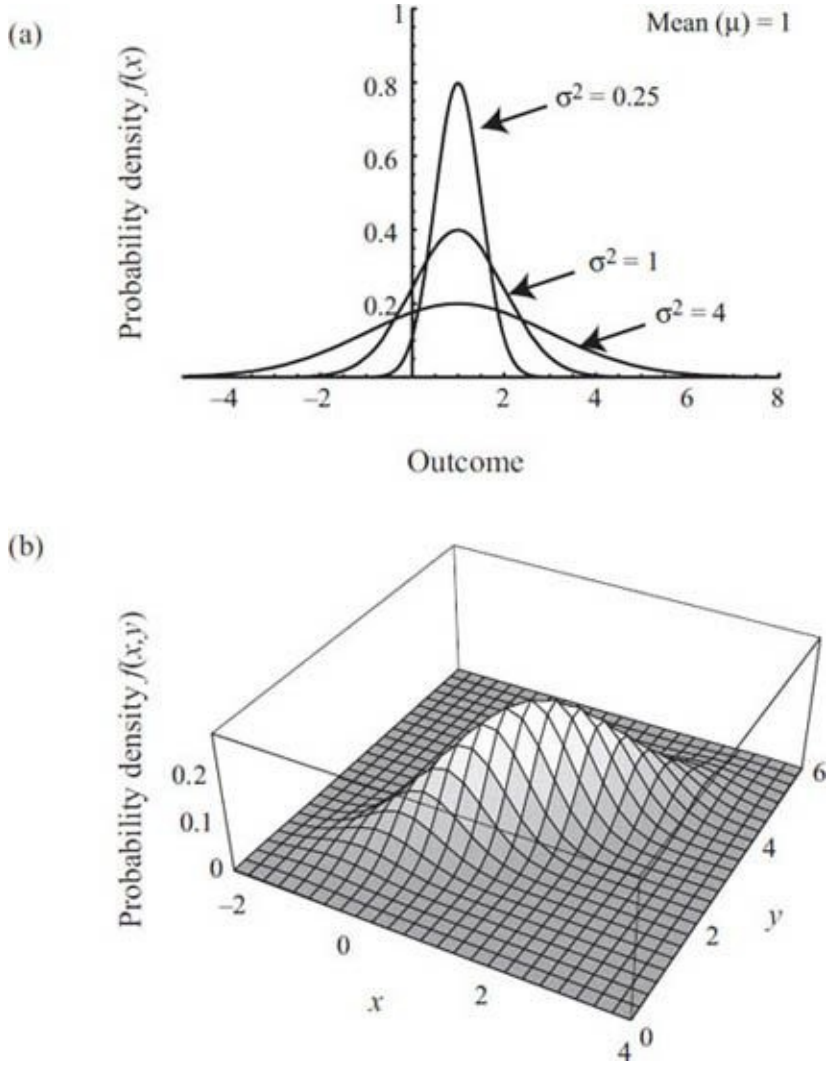

Figure P3.12: Normal distribution. (a) The probability density function, $f(x)$, of a normal distribution (Definition P3.14a) is plotted for various values of the variance, holding the mean value constant at $\mu = 1$. (b) The probability density function, $f(x,y)$, of a bivariate normal distribution (Definition P3.14b) is plotted assuming that the means are $\mu_x = 1$, $\mu_y = 3$, that the correlation between X and Y is $\rho = 0.9$, and that the variances are $\sigma_x^2 = \sigma_y^2 = 1$.

Technically, the central limit theorem requires that each random variable follow a

distribution with finite mean and variance, as is the case for the distributions that we have considered. Variants of the central limit theorem have also been proven relaxing the requirement that each random variable is drawn from the same distribution and that the random variables are entirely independent of one another. Basically, as long as enough random variables are combined and these variables are nearly independent, then the combined effect of the random variables looks nearly normal in shape.

The central limit theorem explains why many of the distributions described in this Primer are, under certain circumstances, bell shaped. First, the binomial distribution involves summing the outcome of n independent Bernoulli trials. Thus, the normal distribution provides an excellent approximation for a binomial distribution, as long as n is sufficiently large that multiple successes ($n\,p$) and multiple failures ($n\,(1-p)$) are expected (Figure P3.13). Similarly, the negative binomial distribution involves summing the waiting times needed for r events to occur and is nearly normal in shape when r is large (see Figure P3.5). The same holds for the gamma distribution if we wait for a sum total of β events to occur in continuous time (see Figure P3.11). Even the Poisson distribution is approximately normal in shape when the total number of events is expected to be large (see Figure P3.6). As we increase the number of events being summed, every one of these distributions becomes more bell shaped, that is, more closely approximated by a normal distribution. These distributions are never exactly normal. For example, a negative outcome is not possible with these distributions, whereas negative outcomes are always possible with the normal distribution. Nevertheless, the discrepancy between the true distribution and the normal distribution becomes smaller as more random variables are summed.

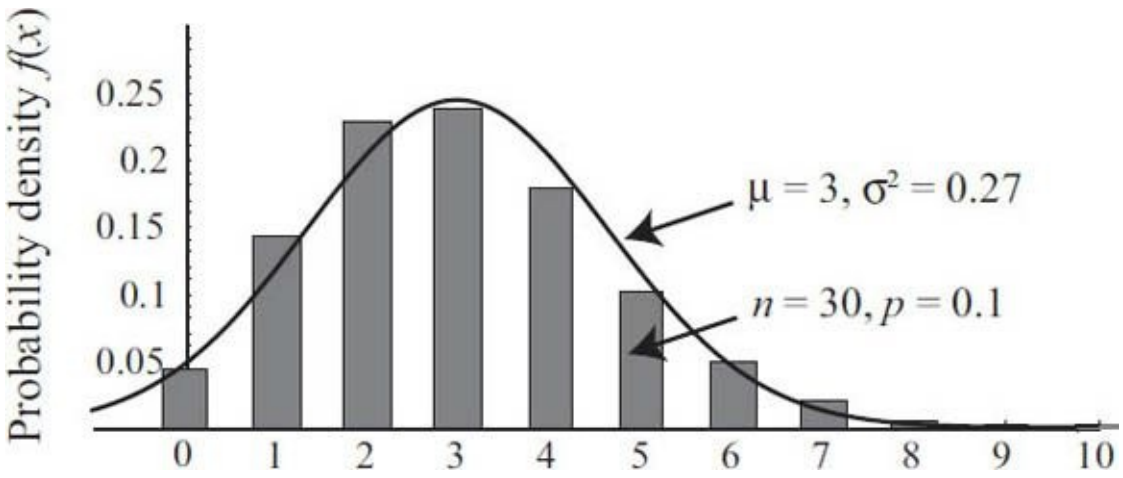

Figure P3.13: Approximating a binomial distribution with a normal distribution. When the number of trials, n, is large in a binomial distribution, its shape is approximately normal. Here we compare the normal distribution with mean $\mu = n\,p$ and variance $\sigma^2 = np\,(1-p)$ to a binomial distribution with parameters n and p. The fit is good in this example even though n is not very large.

The central limit theorem also helps explain why many traits follow a normal

distribution, because such traits are typically influenced by a large number of factors (genetic and/or environmental). In this case, the random variable that is being summed (or averaged) is the contribution of each factor to the trait.

An important generalization of the normal distribution is the multivariate normal distribution. Here we present the two-variable (bivariate) case:

Definition P3.14b:

The bivariate normal distribution is a probability distribution for two random variables. It is bell shaped for both random variables and again the probability density function falls off exponentially with the squared distance to the mean of either variable:

$$f(x,y) = \frac{\exp\left[-\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) \right) \Big/ 2(1-\rho^2) \right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

for $-\infty \le x \le \infty$ and $-\infty \le y \le \infty$.

Again the denominator ensures that the integral of the distribution over both variables equals one. The mean and variance of the X variable is calculated just as it was for a one-variable probability density (i.e., from Definitions P3.9 and P3.10), only now we must integrate over both x and y. The same is true for the Y variable. For the bivariate normal, the mean and variance of the two random variables X and Y are

$$E[X] = \mu_x,\ E[Y] = \mu_y, \tag{P3.29}$$

$$\mathrm{Var}[X] = \sigma_x^2,\ \mathrm{Var}[Y] = \sigma_y^2. \tag{P3.30}$$

Now, however, there is an additional parameter, $-1 \le \rho \le 1$, which is known as the *correlation* coefficient. Positive values of ρ mean that larger than average values of X tend to be associated with larger than average values of Y and vice versa. Negative values of ρ mean that larger than average values of X tend to be associated with smaller than average values of Y and vice versa (Figure P3.12b). The correlation coefficient is related to the covariance of two random variables by the equation $\rho = \mathrm{Cov}[X,Y]/(\sigma_x\ \sigma_y)$ (Table P3.1).

P3.4.5 Log-Normal Distribution

The log-normal distribution arises when describing the *product* of a large number

of independent and identically distributed random variables. If $Y = Y_1\ Y_2 \cdot \cdot \cdot\ Y_n$, then we expect $X = \ln(Y)$ to become normally distributed as the number of variables increases. This follows from the fact that $\ln(Y) = \ln(Y_1) + \ln(Y_2) + \cdot \cdot \cdot + \ln(Y_n)$ is the sum of a number of random variables. Thus, the central limit theorem applies to $X = \ln(Y)$ and says that X tends toward a normal distribution. We then say that $Y = e^x$ has a log-normal distribution:

Definition P3.15:

The log-normal distribution describes the distribution of a random variable Y whose natural logarithm is normally distributed:

$$f(y) = \frac{e^{-(\ln(y)-m)^2/(2s^2)}}{y\sqrt{2\pi s^2}} \qquad \text{for } 0 \le y \le \infty.$$

The log-normal distribution can be derived from a normal distribution with mean m and variance s^2 (Definition P3.14a) by replacing x with $\ln(y)$ and then using the fact that $dx = d(\ln(y)) = (1/y)\ dy$.

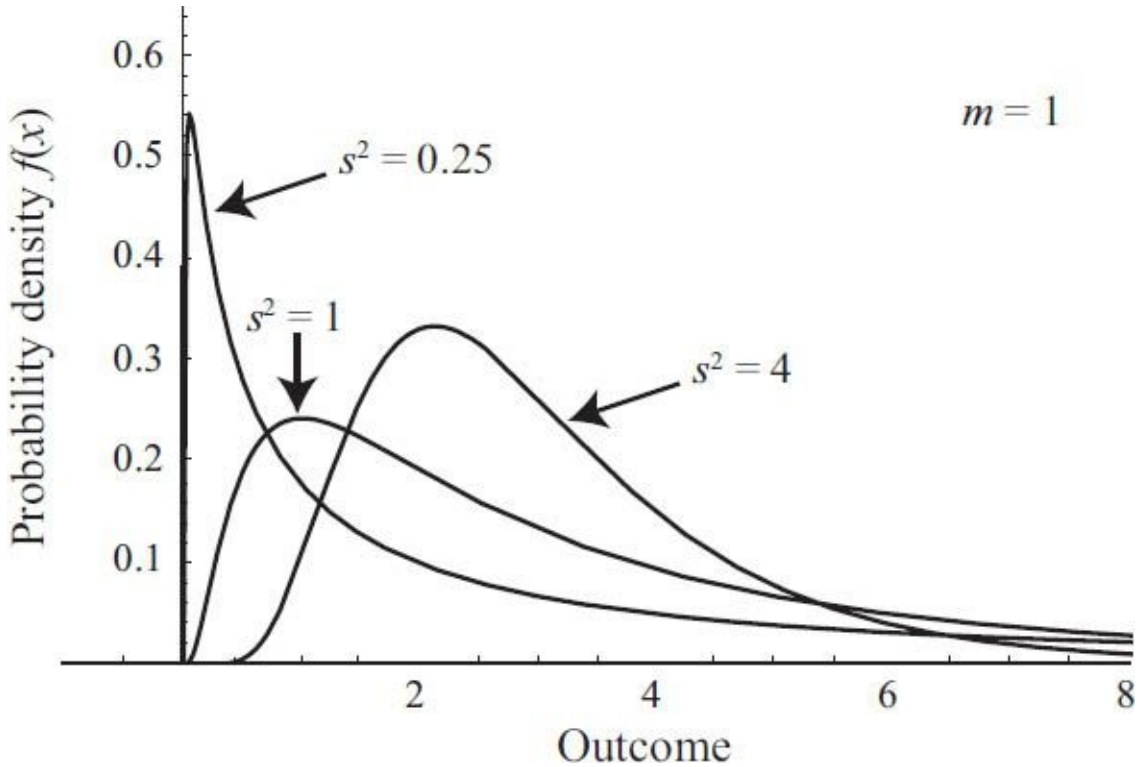

Figure P3.14: Log-normal distribution (Definition P3.15). The probability density function, $f(y)$, is plotted for various

values of s^2, holding constant $m = 1$. Note that the mean of the log-normal depends on both m and s; for the distributions drawn, the mean is 3.08 ($s^2 = 0.25$), 4.48 ($s^2 = 1$), and 20.09 ($s^2 = 4$). The mean is much larger than you might predict based on these graphs because the tail to the right is long and fat.

The mean and variance of the log-normal distribution equal

$$E[Y] = e^{m+(s^2/2)}, \tag{P3.31}$$

$$\mathrm{Var}[Y] = e^{2m+2s^2} - e^{2m+s^2} \tag{P3.32}$$

The log-normal distribution is asymmetrical with a long tail extending to the right, but it becomes more bell shaped as s^2 decreases (Figure P3.14).

Example

We would expect the abundance of a population to follow a log-normal distribution if growth rates (i.e., r in the exponential growth model) vary over time in an additive fashion, because the population size n is proportional to e^r. Indeed, population size surveys often reveal log-normal distributions in a variety of species, from diatoms to birds (Limpert et al. 2001). We would also expect survival times to be log-normally distributed if several factors have a multiplicative impact on survival (e.g., increasing or decreasing survival by a percentage). In fact, survival times after diagnosis with cancer have been shown to follow a log-normal distribution (Limpert et al. 2001).

P3.4.6 Beta Distribution

For the binomial distribution, we focused on a random variable describing the number of successes, k, out of n trials, where p and n were the parameters of the distribution. What if, instead, we wanted a distribution for the probability of success, p, given that we have observed k successes out of n trials? The appropriate distribution for the random variable, p, is the beta distribution with $a = k + 1$ and $b = n - k + 1$:

Definition P3.16:

The beta distribution describes a probability density for a proportion p:

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, p^{a-1}(1-p)^{b-1} \qquad \text{for } 0 \le p \le 1$$

where a and b are real and positive parameters and $\Gamma(a) = \int_{y=0}^{\infty} e^{-y} y^{a-1} dy$ is the gamma function.

To derive the beta distribution, we start by assuming that the probability density function for p is proportional to the binomial distribution, $n!/(k!(n-k)!)\ p^k\ (1-p)^{n-k}$. In other words, values of the random variable p that are more likely to yield k successes out of n trials are given greater probability density. Again, we can generalize this distribution to noninteger parameter values by replacing the binomial coefficient, $n!/(k!(n-k)!)$ (considered here to be a constant given the data), by a new constant chosen to ensure that the probability density function integrates to one. Using Rule P3.11, the constant by which we must divide $p^k(1-p)^{n-k}$ is $\int_{p=0}^{1} p^k\ (1-p)^{n-k}\ \mathrm{d}p$. *Mathematica* comes in handy for this integration and gives

$$\int_{p=0}^{1} p^k(1-p)^{n-k}\,\mathrm{d}p = \frac{\Gamma(k+1)\,\Gamma(n-k+1)}{\Gamma(n+2)}.$$

Finally, to obtain Definition P3.16, we rewrite the distribution in terms of the parameters, a and b, where $a = k+1$ and $b = n-k+1$, as it is more typically written.

The beta distribution has mean and variance

$$E[X] = \frac{a}{a+b}, \tag{P3.33}$$

$$\mathrm{Var}[X] = \frac{a\,b}{(a+b)^2(1+a+b)} \tag{P3.34}$$

(see Exercise P3.8).

The shape of the beta distribution is extremely flexible (Figure P3.15). It is bell shaped when a and b are similar in magnitude and large. It is a flat line when a and b are one. It is U shaped when a and b are similar in magnitude and smaller than one. The beta distribution can even be L shaped when b is much larger than a or J shaped when a is much larger than b.

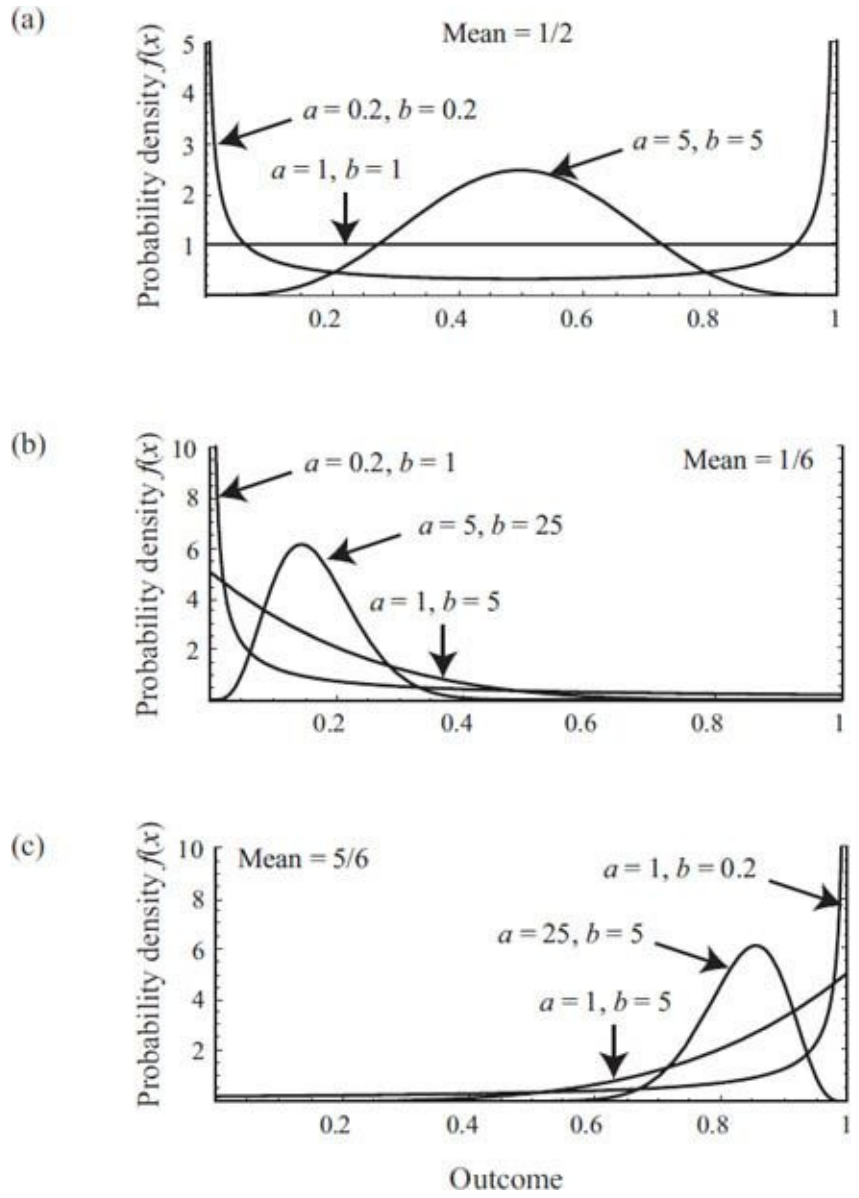

Figure P3.15: Beta distribution (Definition P3.16). The probability density function, $f(x)$, is plotted for various values of a and b, holding the mean value constant at (a) $E[X] = 1/2$, (b) $E[X] = 1/6$, and (c) $E[X] = 5/6$.

Example

As we shall see in Chapter 15, the beta distribution arises in genetic models describing the probability distribution of allele frequencies within a population (see equation (15.25); Crow and Kimura 1970; Turelli 1981). The most frequent context in which the beta distribution arises, however, is in Bayesian analysis where the beta distribution is often used as a prior probability distribution for parameters that lie between 0 and 1 (see Supplementary Material P3.1).

Exercise P3.8: Use the fact that $\int_{x=0}^{1} p^{a-1}(1-p)^{b-1}dp = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(a+1) = a\Gamma(a)$ to find the mean of the beta distribution.

P3.4.7 Dirichlet Distribution

Just as the multinomial distribution generalizes the binomial distribution for outcomes involving more than two possible states, the Dirichlet distribution generalizes the beta distribution:

Definition P3.17:

The Dirichlet distribution describes the probability density function for the proportions p_i in each of c discrete categories:

$$f(p_1, p_2, \ldots, p_c) = \frac{\Gamma(a_1 + a_2 + \cdots + a_c)}{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_c)} p_1^{a_1-1} p_2^{a_2-1} \cdots p_c^{a_c-1}$$

for $0 \le p_i \le 1$

Here the a_i are real and positive parameters (akin to $k + 1$ in the multinomial distribution, see Definition P3.5), and the proportions in each state must sum to one: $\sum_{i=1}^{c} p_i = 1$. For example, the Dirichlet distribution arises when describing the frequency distribution of multiple alleles at a locus.

P3.5 The (Insert Your Name Here) Distribution

While we have discussed a number of classic distributions that arise often, it is important to recognize that there are an unlimited number of probability distributions. For many problems, the appropriate distribution might be one of those discussed in this Primer. It is definitely possible, however, that the appropriate distribution is a new one. At that point you should take the plunge and describe your very own distribution. The rules and definitions described in this Primer allow you to check that your distribution correctly sums to one and to determine such things as the mean and variance of the distribution.

For example, you might want to model a population in which there are two types of males, and where females have a preference for one type over the other. To begin, you choose a simple procedure by which females decide on their mates. First, a female randomly encounters a male from the population. If she encounters a type-1 male, she mates with him. If, however, she encounters a type-2 male, she mates with him with

probability ϕ. If she remains unmated, she tries again, and the same rules apply. However, after two attempts, the female mates with any male she encounters. Say that you are particularly interested in knowing how often females mate in one, two, or three attempts. You could simulate the above process a number of times to answer this question, but it is much easier to develop the probability distribution.

If the proportion of males that are type 1 is p, then the probability that a female mates in the first attempt is $P(X = 1) = p + (1 - p)\,\phi$, which we will define as F. Of the remaining $1 - F$ females, a similar proportion mates in the second attempt. Thus, $P(X = 2) = (1 - F)\,F$. Any females that remain unmated then mate at the third attempt, so that $P(X = 3) = (1 - F)\,(1 - F)$. This completes our derivation of a new probability distribution:

$$\begin{aligned} P(X = 1) &= F, \\ P(X = 2) &= (1 - F)\,F, \\ P(X = 3) &= (1 - F)^2. \end{aligned} \tag{P3.35}$$

These probabilities sum to one and so obey Rule P3.10. The mean can be calculated using Definition P3.2 as

$$\begin{aligned} E[X] &= P(X = 1) + 2\,P(X = 2) + 3\,P(X = 3) \\ &= 3 - 3F + F^2. \end{aligned} \tag{P3.36}$$

If males of the preferred type are common (p high) or if females are inclined to mate even with males of the second type (ϕ high), then $F = p + (1 - p)\phi$ will be near one and the mean will be near one (Figure P3.16). After a little bit of algebra it is also possible to show that the variance of this distribution equals $\mathrm{Var}[X] = F(1 - F)\,(5 - 5F + F^2)$.

Equation (P3.35) describes the probability that a particular female mates at the first, second, or third attempt. We can use these probabilities within the multinomial distribution (Definition P3.2), however, to describe the number of females mating at the first, second, or third attempt within a population of n females. This would be a lot faster than simulating each female as she chooses her mate, especially if there were thousands of females!

The main distributions described in this Primer are summarized in Tables P3.2 (discrete probability distributions) and P3.3 (continuous probability distributions). These tables provide a quick reference describing the probability distribution, as well as its mean and variance. In addition, moment generating functions are given where they exist and are simple enough to be useful. As described in Appendix 5, moment generating functions provide a quick and relatively painless method for finding higher

moments of a distribution. Finally, Tables P3.2 and in P3.3 provide one-sentence descriptions of when we might expect each distribution to apply. Remember, however, that probability distributions are not written in stone. If you are interested in a process that is not well described by any probability distribution that you know, forge ahead and develop the appropriate distribution on your own. Who knows, you might go down in history as the person for whom a probability distribution is named!

Exercise P3.9: In equation (P3.35), we described the probability distribution for the number of mating attempts made by a female before she mates. Calculate the overall probability that she mates with a male of type 1. Relate this result to the mean number of trials (P3.36). Use the complement rule (Rule P3.2) to determine the probability that she mates with a male of type 2.

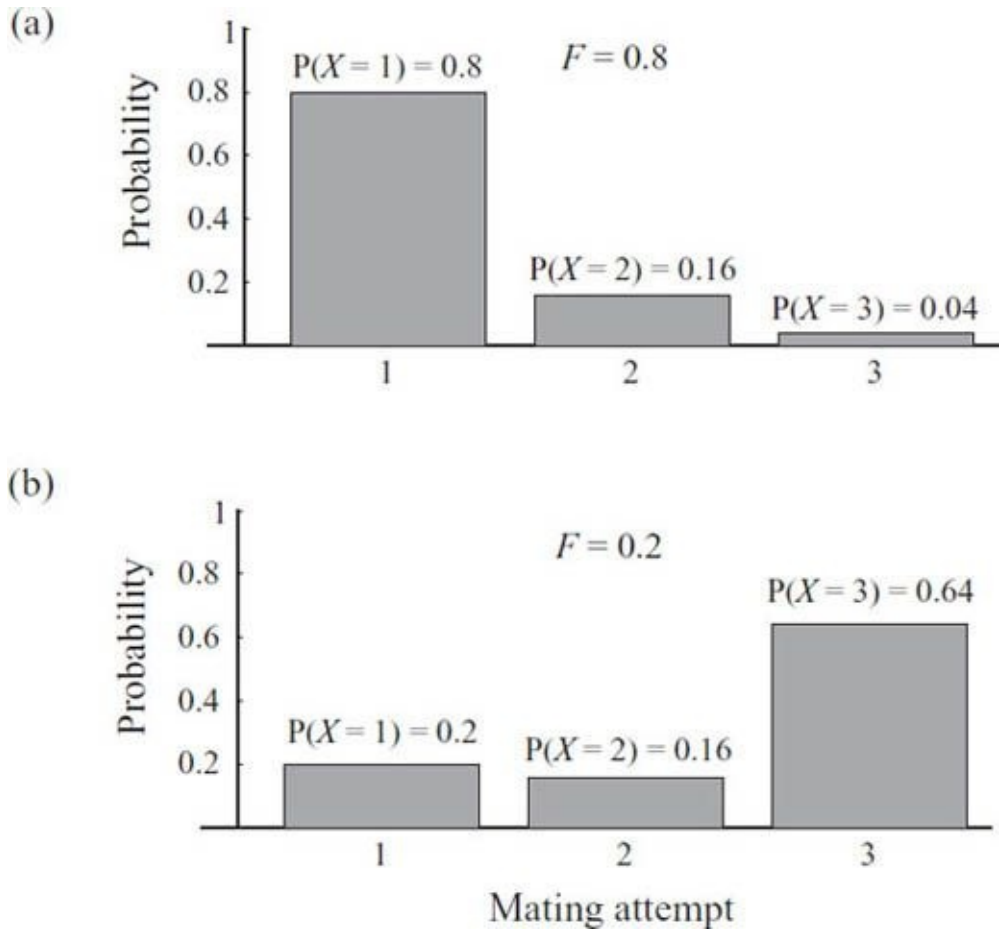

Figure P3.16: Mating probabilities. The probability that a female mates on her nth encounter with a male, according to the probability distribution defined by equation (P3.35). In (a), females are likely to mate with a randomly encountered male ($F = 0.8$), while in (b), females are more choosy ($F = 0.2$).

TABLE P3.2
Discrete probability distributions. Moment generating functions $MGF(z)$ are given where they exist in a useful form (see Appendix 5).

Binomial Distribution Parameters: n, p

Definition P3.4

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \qquad \text{for } k = 0,1,2, \ldots, n$$

$$E[X] = np$$

$$\text{Var}[X] = np(1-p)$$

$$MGF[z] = (1 - p + e^z p)^n$$

Circumstances: The binomial arises when there are n independent events, each of which can have two outcomes ("success" or "failure"). The probability of observing a total of k successes is $P(X = k)$.

Hypergeometric Distribution. Parameters: N, p, n, where $N_1 = p\,N$ and $N_2 = (1 - p)\,N$

Definition P3.6

$$P(X = k) = \frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N}{n}} \quad \text{for } k = 0,1,2, \ldots, \min(n, N_1)$$

$$E[X] = n\,p$$

$$\text{Var}[X] = n\,p(1 - p)\frac{N - n}{N - 1}$$

Circumstances: The hypergeometric describes the probability of observing k successes when n objects are sampled without replacement from a total pool of N objects, of which a fraction, p, represent a successful outcome.

Geometric Distribution. Parameter: p

Definition P3.7

$$P(X = k) = p\,(1-p)^{k-1} \qquad \text{for } k = 1, 2, 3, \ldots$$

$$E[X] = \frac{1}{p}$$

$$\text{Var}[X] = \frac{(1-p)}{p^2}$$

$$MGF[z] = \frac{e^z p}{1-(1-p)e^z}$$

Circumstances: The geometric arises when measuring the number of independent trials, k, until the first success.

Negative Binomial Distribution. Parameters: r, p.

Definition P3.8

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad \text{for } k = r, r+1, r+2, \ldots$$

$$E[X] = \frac{r}{p}$$

$$\text{Var}[X] = \frac{r(1-p)}{p^2}$$

$$MGF[z] = \left(\frac{e^2 p}{1-(1-p)e^z}\right)^r$$

Circumstances: The negative binomial arises when measuring the number of independent trials, k, until the rth success.

Poisson Distribution. Parameter: μ

Definition P3.9

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!} \qquad \text{for } k = 0,1,2,\ldots$$

$$E[X] = \mu$$

$$\text{Var}[X] = \mu$$

$$MGF[z] = e^{\mu(e^z-1)}$$

Circumstances: The Poisson arises when measuring the number of independent events, k, that occur in a certain period (or area) of observation. The Poisson also arises as an approximation to the binomial when n is large and p is small, in which case the mean and variance are well approximated by $\mu = np$.

TABLE P3.3
Continuous probability distributions. Moment generating functions $MGF(z)$ are given where they exist in a useful form (see Appendix 5).

Uniform Distribution. Parameters: min, max

Definition P3.11

$$f(x) = \frac{1}{\max - \min} \qquad \text{for } \min \le x \le \max$$

$$E[X] = \frac{\max + \min}{2}$$

$$\text{Var}[X] = \frac{(\max - \min)^2}{12}$$

$$MGF[z] = \frac{e^{\max z} - e^{\min z}}{(\max - \min)\, z}$$

Circumstances: The uniform distribution arises whenever you are interested in describing where an event occurs for events that have the same chance of occurring anywhere between two points (min and max).

Exponential Distribution. Parameters: α

Definition P3.12

$$f(x) = \alpha e^{-\alpha x} \qquad \text{for } 0 \le x \le \infty$$

$$E[X] = \frac{1}{\alpha}$$

$$\text{Var}[X] = \frac{1}{\alpha^2}$$

$$MGF[z] = \frac{\alpha}{\alpha - z} \qquad \text{for } z < \alpha$$

Circumstances: The exponential distribution arises when measuring the amount of time that passes, x, until an event occurs, measured in continuous time.

Gamma Distribution. Parameters: α, β

Definition P3.13

$$f(x) = \frac{\alpha^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha x} \qquad \text{for } 0 \le x \le \infty$$

$$E[X] = \frac{\beta}{\alpha}$$

$$\text{Var}[X] = \frac{\beta}{\alpha^2}$$

$$MGF[z] = \left(\frac{\alpha}{\alpha - z}\right)^\beta \qquad \text{for } z < \alpha$$

Circumstances: The gamma distribution arises when measuring the amount of time that passes, x, until β independent events occur, measured in continuous time.

Normal Distribution. Parameters: μ, σ^2

Definition P3.14a

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \qquad \text{for } -\infty \le x \le \infty$$

$$E[X] = \mu$$

$$\text{Var}[X] = \sigma^2$$

$$MGF[z] = e^{z\,\mu + z^2\sigma^2/2}$$

Circumstances: The normal distribution arises when several factors sum (or average) together to influence an outcome.

Bivariate Normal Distribution. Parameters: $\mu_x, \mu_y, \sigma_x, \sigma_y, \rho$

Definition P3.14b:

$$f(x,y) = \frac{\exp\left[-\frac{z}{2(1-\rho^2)}\right]}{2\,\pi\,\sigma_x\,\sigma_y\sqrt{1-\rho^2}}$$

$$\text{where } z = \left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\,\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) \text{ for } -\infty \le x \le \infty \text{ and } -\infty \le y \le \infty$$

$$E[X] = \mu_x, \quad E[Y] = \mu_y$$

$$\text{Var}[X] = \sigma_x^2, \quad \text{Var}[Y] = \sigma_y^2, \quad \text{Correlation}[X,Y] = \rho$$

Log-normal Distribution. Parameters: m, s^2

Definition P3.15

$$f(y) = \frac{e^{-(\ln(y)-m)^2/(2s^2)}}{y\sqrt{2\pi s^2}} \qquad \text{for } 0 \le y \le \infty$$

$$E[Y] = \ln\left(\frac{m^2}{\sqrt{m^2 + s^2}}\right)$$

$$\text{Var}[Y] = \sqrt{\ln\left(1 + \frac{s^2}{m^2}\right)}$$

Circumstances: The log-normal distribution arises when several factors have multiplicative effects on an outcome.

Beta Distribution. Parameters: *a, b*

Definition P3.16

$$f(p) = \frac{\Gamma(a + b)}{\Gamma(a)\,\Gamma(b)}\, p^{a-1}(1 - p)^{b-1}, \qquad \text{for } 0 \le p \le 1$$

$$E[X] = \frac{a}{a + b}$$

$$\text{Var}[X] = \frac{ab}{(a - b)^2\,(1 + a + b)}$$

Circumstances: The beta distribution arises when estimating an unknown probability or proportion, *p*.

Box P3.1: Counting and combinatorics

In the description of the binomial distribution, we encountered a quantity $\binom{n}{k}$, read as "*n* choose *k*." This quantity, often referred to as the *binomial coefficient*, is an integer that counts up the number of different ways in which *k* "ones" can occur over the course of *n* trials. There are other probability distributions that require us to count up the various ways in which things can happen, including the hypergeometric distribution (Definition P3.6) and the negative binomial distribution (Definition P3.8). Being able to enumerate the possible outcomes of a

process is invaluable in many other areas of mathematical modeling, as well. The area of mathematics devoted to understanding arrangements of sets of items is *combinatorics*. Here we describe the basics of combinatorics and derive the binomial coefficient.

As an example, consider counting up the number of different genetic strains of a DNA virus that are possible if the virus's genome is 1000 base pairs long. The easiest way to evaluate the total count is to begin with the first position and consider all of the possible nucleotides that might be present (4 for A, C, T, or G). For each one of these, we then move on to the second position and count up the number of possibilities (4 again). Thus, considering only the first two positions, there are $4 \times 4 = 16$ possibilities (AA, AC, AT, AG, CA, CC, CT, CG, TA, TC, TT, TG, GA, GC, GT, and GG). Proceeding onward using similar reasoning, there are $4 \times 4 \times 4 = 4^3$ possibilities for the first three positions (AAA, AAC, AAT, etc.), and each additional position gives 4 times as many possibilities. Over the entire genome, there will thus be 4^{1000} possible sequences. The different possibilities are referred to as *permutations*, and the above process enumerates all of the possible permutations of the genome. In particular, these are the permutations possible with *replacement*; after we assign a particular basepair (say, G) to the first site, we *replace* G back in the list of possibilities (A, C, T, or G) for the next site. More generally:

Rule P3.1.1: Enumerating Permutations with Replacement

Suppose that there are k positions, each of which can be occupied by any one of n different items. The total number of permutations possible is n^k. Because there are always n choices, this is the total number of permutations with replacement.

A related question focuses on enumerating all possible permutations without replacement. As an example, imagine you are studying a migratory bird species and that you have marked the entire population with leg bands so that you can identify each individual. Each season you record the order in which the birds arrive back on the breeding ground. If there are 25 birds in the population, let's count how many different ways the population of birds could return.

Again we can count the number of the possible choices for the first bird to return, the second bird to return, and so on. There are 25 possible choices for the first bird (bird 1, 2, 3, etc.). For each of these, however, there are only 24 possible choices for the second bird because the second bird to return cannot also be the first bird. That is, we do not *replace* the first bird in the list of possibilities

for subsequent trials. Thus, considering only the first two birds, there are 25×24 = 600 possible orderings. Considering the 23 possible birds that can return next, there are 25×24×23 = 13,800 possible orderings of the first three birds. Repeating this process and assuming that all 25 birds return, the total number of possible permutations without replacement is 25 × 24 × 23 × . . . × 2 × 1, which we can write as 25! ("25 factorial").

What if you were only interested in the order of the first k birds to arrive at the breeding ground? By the time you get to the kth bird, $25 - (k - 1)$ birds remain, so that there will be $25 \times 24 \times \ldots \times (25 - k + 1)$ possible orderings. This result can be written more compactly as

$$\frac{25!}{(25-k)!} = \frac{25 \times 24 \times \cdots \times (25-k+1) \times (25-k) \times \cdots \times 1}{(25-k) \times \cdots \times 1}$$

$$= 25 \times 24 \times \cdots \times (25-k+1).$$

If we watch all of the birds return ($k = 25$), this equals 25!/0!. By definition, 0! equals one, so we regain the result from the previous paragraph that there are 25! possible permutations without replacement of the entire set of birds. More generally:

Rule P3.1.2: Enumerating Permutations without Replacement

Suppose that there are n different items, each of which is placed in one of k positions, and once placed it is removed from the list of items. The total number of permutations without replacement is then $n \times (n - 1) \times (n - 2) \times \ldots \times (n - k + 1)$ or, equivalently, $n!/(n-k)!$. This result assumes that $n \geq k$ and that $0! = 1$, so that the number of permutations is $n!$ when $k = n$.

The number of permutations without replacement counts the various possibilities when the order in which the various items occur is of interest. For instance, in the bird example, the number of different *orderings* in which the first k birds can arrive is $25!/(25-k)!$. If, instead, we are interested in the identity of the first k birds, but we do not care about the order in which they appear, then we must group together all equivalent possibilities and count only the number of *unordered* possibilities. For example, if we label each bird from A to Y and observe two birds, one possible combination is bird M and bird G, where we do not care whether bird M or G arrived first. The total number of unordered

possibilities is referred to as the number of *combinations*.

Returning to the example of migratory birds, let us count the number of possible combinations of birds that might be found in the first three arrivals of the season. Our calculations above reveal that there are 25 × 24 × 23 = 13,800 ordered permutations, but some of these contain the same three birds in different orders. For example, suppose that the first three to arrive are birds D, S, and G (in that order). This is one of the 13,800 permutations just mentioned, but S, D, and G (in that order) is another of these. From the standpoint of enumerating the possible combinations, however, these two outcomes are the same (because the order does not matter). In fact, there are 3 × 2 × 1 = 3! different ordering of birds D, G, S that might occur and that would all be considered as the same combination ({D, G, S}, {D, S, G}, {G, D, S}, {G, S, D}, {S, D, G}, and {S, G, D}). Furthermore, this is true for any three specific birds that might be the first to arrive. Therefore, the number of combinations possible for the first three birds is given by the number of permutations of the arrivals (i.e., 25 × 24 × 23 = 13,800) divided by the number of equivalent orderings, 3!. The result, (25 × 24 × 23)/3!, equals 2300 and can be rewritten using Rule P3.1.2 as 25!/((25 − 3)! 3!), which, by definition, equals the binomial coefficient $\binom{25}{3}$. And, in general,

Rule P3.1.3: Enumerating Combinations without Replacement

Suppose that there are n different items, each of which is placed in one of k positions ($n \geq k$) and once placed it is removed from the list of items. The total number of combinations of the k items that are possible equals the binomial coefficient

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}.$$

At this point, we are ready to connect the number of combinations to the binomial distribution. Say that we want to know how many ways in which k "ones" can occur in n trials. Let us label the trials as A, B, C, etc., just as we labeled the birds above. Given that k ones occur, we want to keep track of the trial in which they occur. This amounts to assigning one of the n trial names to each of the k ones that have occurred. There is nothing to distinguish the k ones from one another, however, so we only wish to count the number of distinct combinations of k trial numbers. For example, if $k = 3$ ones occur in $n = 25$ trials, one possible outcome is that there was a "one" at trial D, trial G, and trial S.

Again, we must group together the 3! different permutations of trial numbers ({D, G, S}, {D, S, G}, etc.) just as we did in the bird example. As a result, there are $n!/((n-k)!\,k!) = \binom{n}{k}$ unordered ways in which the n trial names could be matched to the k ones that have occurred (Rule P3.1.3).

Further Reading

For an introductory text on probability theory, consult

- Pitman, J. 1997. *Probability*. Springer-Verlag, Berlin.
- Larsen, R. J. and M. L. Marx. 2001. *An Introduction to Probability and Its Applications*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.
- Taylor, H. M. and S. Karlin. 1998. *An Introduction to Stochastic Modeling*, 3rd ed. Academic Press, New York.

For a more advanced text on probability theory, consult

- Rice, J. A. 1995. *Mathematical Statistics and Data Analysis*, 2nd ed. Duxbury Press, Belmont, Calif.

References

Abramowitz, M., and I. A. Stegun. 1972. *Handbook of Mathematical Functions*. Dover Publications, New York.

Crow, J. F., and M. Kimura. 1970. *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

Hilborn, R., and M. Mangel. 1997. *The Ecological Detective*. Princeton University Press, Princeton, N.J.

Keightley, P. D. 1994. The distribution of mutation effects in *Drosophila melanogaster. Genetics* 138:1315–1322.

Kondrashov, A. S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat*. 21:12–27.

Limpert, E., W. A. Stahel, and M. Abbt. 2001. Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51:341–352.

Preti, G., W. B. Cutler, C. R. Garcia, G. R. Huggins, and H. J. Lawley. 1986. Human axillary secretions influence women's menstrual cycles: The role of donor extract of females. *Horm. Behav*. 20:474–482.

Shopland, D. R. 1995. Tobacco use and its contribution to early cancer mortality with a special emphasis on cigarette smoking. *Environ. Health Perspect. Suppl*. 103(S8):131–142.

Turelli, M. 1981. Temporally varying selection on multiple alleles: a diffusion analysis. J. Math. Biol. 13:115–129.

Zar, J. H. 1998. *Biostatistical Analysis*. Prentice-Hall, Upper Saddle River, N.J.

Answers to Exercises

Exercise P3.1

(a) According to Rule P3.4, if the two events are independent, P("tree has fungal rot" ∩ "tree has an owl nest") = (1/100)(1/1000) = 10^{-5}. According to Rule P3.5, if the two events are mutually exclusive, P("tree has fungal rot" ∩ "tree has an owl nest") =

0.

(b) According to Rule P3.4 for independent events, P(“blood type O” ∩ “Rhesus negative”) = (0.46)(0.16) = 0.0736. That is, roughly 7% of the population is expected to be O−.

(c) According to Rule P3.5 for mutually exclusive events, P(“blood type O” ∩ “blood type A”) = 0.46 + 0.40 = 0.86. That is, the probability that the donor is acceptable is 86%.

(d) According to Rule P3. 4 for independent events, P(“first study is significant” ∪ “second study is significant”) = (0.05) + (0.05) − $(0.05)^2$ = 0.0975. Thus, when two studies are performed, there is nearly a 10% chance that at least one of the studies will conclude, incorrectly, that a true hypothesis is false. An alternative way to reach the same answer is by using the complement Rule P3.2. The probability that at least one of the two studies obtains a significant result is equal to 1 − P (“neither study obtains a significant result”). Because the two studies are independent, we can use Rule P3.4 to calculate P(“neither study obtains a significant result”) = $(1 - 0.05)^2$. This second method gives the same probability, $1 - (1 - 0.05)^2$ = 0.0975, that one or both of the results is significant.

Exercise P3.2

(a) We start by rewriting the question in terms of probability statements. We want to know P(brown hair | green eyes) given that P(green eyes) = 0.10, P(brown hair) = 0.75, P(brown hair ∩ green eyes) = 0.09. Equation (P3.1a) then can be used to write P(brown hair ∩ green eyes) = P(brown hair | green eyes) P(green eyes). Thus, P(brown hair | green eyes) = 0.09/0.10 and there is a 90% chance of having brown hair given that you have green eyes.

(b) Using Bayes’ Rule P3.7, P(taster | carrier) = P(carrier | taster) P(taster)/P(carrier), which equals (1) (0.7)/(0.8) = 0.875.

(c) Using Bayes’ Rule P3.7, P(death due to lung cancer | not a smoker) = P(not a smoker | death due to lung cancer) P(death due to lung cancer) / P(not a smoker). Because P(not a smoker) is the complement of P(smoker), we can use the complement Rule P3.2 to write P(not a smoker) = 1 − P(smoker). The complement rule also applies to conditional statements, so that P(not a smoker | death due to lung cancer) = 1 − P(smoker | death due to lung cancer). Altogether, we get the formula: P(death due to lung cancer | not a smoker) = (1 − P(smoker | death due to lung cancer)) P(death due to lung cancer)/(1 − P (smoker)). Using the data, the risk of death due to lung cancer among non-smokers is P(death due to lung cancer | not a smoker) = (1 − 0.9) (0.3)/(1 − 0.5) = 0.06.

Exercise P3.3

(a) The expected value of X^2 equals $E[X^2] = 0^2 \times P(X = 0) + 1^2 \times P(X = 1) = 0^2 \times (1 - p) + 1^2 \times (p) = p$. Subtracting off the square of the mean of a Bernoulli trial, $\mu^2 = p^2$, gives the variance, $\text{Var}[X] = E[X^2] - \mu^2 = p - p^2$, which again equals $p(1 - p)$.

(b) With two Bernoulli trials, the probability of no successes is $P(X = 0) = (1 - p)^2$ (getting a failure on the first trial and then independently getting a failure on the second trial), the probability of getting a single success is $P(X = 1) = p(1 - p) + (1 - p)p = 2 p(1 - p)$ (having a success followed by a failure or vice versa), and the probability of getting two successes, is $P(X = 2) = p^2$. The sum of these probabilities is $(1 - p)^2 + 2 p(1 - p) + p^2$, which factors to one, as it should. The expected outcome is given by the formula, $E[X] = 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2)$, which evaluates to $0 \times (1 - p)^2 + 1 \times 2p\ (1 - p) + 2 \times p^2$, which equals $2p$. The variance of the outcome is slightly easier to calculate using $\text{Var}[X] = E[X^2] - \mu^2$. First, we calculate the expected value of X^2: $E[X^2] = 0^2 \times P(X = 0) + 1^2 \times P(X = 1) + 2^2 \times P(X = 2)$, which equals $2p(1 - p) + 4p^2$. We then subtract off μ^2, where μ is the mean of two Bernoulli trials, which we have already calculated as $\mu = E[X] = 2p$, leaving us with $\text{Var}[X] = 2\ p(1 - p)$.

(c) Because $E[X] = E[Y] = p$ for a single Bernoulli trial, the expected outcome from two independent Bernoulli trials is $E[X + Y] = E[X] + E[Y] = 2p$. Because $\text{Var}[X] = \text{Var}[Y] = p(1 - p)$ for a single Bernoulli trial, $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] = 2\ p(1 - p)$. These results are identical to those obtained in (b).

Exercise P3.4

Using the fact that the variance of a sum is the sum of the variance for independent random variables, $\text{Var}[X + Y] = n_1 p_1(1 - p_1) + n_2 p_2(1 - p_2)$. For general values of p_1 and p_2, $\text{Var}[X + Y]$ cannot be factored and so cannot be written in the form $np(1 - p)$. Only if the probability of success is the same for each trial, does the variance factor into the form of the variance of a binomial distribution, $n\ p\ (1 - p)$, where $n = (n_1 + n_2)$ and $p = p_1 = p_2$. When $n_1 = 100$ and $p_1 = 1$, $\text{Var}[X] = 0$. Similarly, when $n_2 = 100$ and $p_2 = 1$, $\text{Var}[Y] = 0$. In this example, we would always observe 100 failures and 100 successes, and $\text{Var}[X + Y] = 0$, which is much lower than the expected variance of a binomial, $n\ p\ (1 - p) = 50$, for the same total number of events ($n = 200$) and average probability of success ($p = 1/2$).

Exercise P3.5

According to the exponential distribution, the probability, $P(k-1 < X < k)$, that an event occurs between $k - 1$ and k is given by $\int_{x=k-1}^{k} \alpha\ e^{-\alpha x} = -e^{-\alpha k} + e^{-\alpha(k\ -\ 1)}$.

Replacing $e^{-\alpha}$ with $1 - p$ gives $-(1-p)^k + (1-p)^{k-1}$, which factors to $p(1-p)^{k-1}$. This is the same formula as the probability that an event is first observed at time step k in a geometric distribution (Definition P3.7).

Exercise P3.6

(a) The mean of the exponential distribution is given by (Definition P3.10a)

$$\mu = E[X] = \int_0^\infty x\,\alpha\,e^{-\alpha x}\,dx.$$

Integrating by parts (Rule A2.29 from Appendix 2) with $u = x$ and $v = -e^{-\alpha x}$, $\int x\,\alpha e^{-\alpha x}\,dx = -x\,e^{-\alpha x} - e^{-\alpha x}/\alpha$. In the limit as x goes to positive infinity the indefinite integral goes to zero (both $x\,e^{-\alpha x}$ and $e^{-\alpha x}$ approach 0 as x increases), while at $x = 0$ the indefinite integral becomes $-1/\alpha$. Thus the definite integral from $x = 0$ to infinity is $\mu = 1/\alpha$.

(b) The variance of the exponential distribution is given by (Definition P3.10b)

$$\mathrm{Var}[X] = E[X^2] - \mu^2 = \left(\int_0^\infty x^2\,\alpha\,e^{-\alpha x}\,dx\right) - \frac{1}{\alpha^2}$$

Integrating by parts (Rule A2.29), starting with $u = x^2$ and $v = -e^{-\alpha x}$, $\int x^2\,\alpha\,e^{-\alpha x}\,dx = -x^2\,e^{-\alpha x} - \int 2\,x\,e^{-\alpha x}\,dx$. From part (a), we know that $(2/\alpha)\int x\alpha e^{-\alpha x}\,dx = (2/\alpha)\,(-x\,e^{-\alpha x} - e^{-\alpha x}/\alpha)$. Evaluating the definite integral then gives $E[X^2] = 2/\alpha^2$, so that $\mathrm{Var}[X] = 2/\alpha^2 - 1/\alpha^2 = 1/\alpha^2$.

(c) The mean is given by $(d(MGF(z))/dz)|_{z=0}$, which equals $\alpha/(\alpha - z)^2|_{z=0} = 1/\alpha$. The variance can be calculated using $\mathrm{Var}[X] = E[X^2] - \mu^2$ (Definition P3.10b). $E[X^2]$ is given by $(d^2(MGF(z))/dz^2)|_{z=0}$, which equals $2\alpha/(\alpha - z)^3|_{z=0} = 2/\alpha^2$. Thus the variance equals $2/\alpha^2 - (1/\alpha)^2 = 1/\alpha^2$. Alternatively, the variance can be calculated directly from the central moment generating function $CMGF(z) = e^{-z/\alpha}\,\alpha/(\alpha - z)$ as $(d^2(CMGF(z))/dz^2)|_{z=0}$, which also equals $1/\alpha^2$.

Exercise P3.7

(a) $$CV = \sqrt{\frac{\mathrm{Var}[X]}{E[X]^2}} = \frac{1}{\sqrt{\beta}}.$$

(b) Rearranging (a), $\beta = 1/CV^2$ and thus $\alpha = 1/(\mu CV^2)$. This allows us to rewrite the probability density function for the gamma distribution as

$$f(x) = \frac{(\mu e CV^2)^{-(1/CV^2)}}{\Gamma\left(\frac{1}{CV^2}\right)} x^{(1/CV^2)-1} e^{-x/(\mu\, CV^2)}.$$

(c) The coefficient of variation for an exponential distribution equals 1. Smaller values of CV (i.e., larger values of β) correspond to more bell-shaped distributions. As the coefficient of variation goes to zero, the probability density function narrows, and most outcomes are observed near the mean, μ.

Exercise P3.8

Because $\int_{p=0}^{1} p^{a-1}(1-p)^{b-1}\, dp = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, it must be the case that $\int_{p=0}^{1} p\, p^{a-1}\, (1-p)^{b-1}\, dp = \Gamma(a+1)\Gamma(b)/\Gamma(a+b+1)$. Thus, the expectation of the beta distribution is $E[X] = \int_{p=0}^{1} p\, f(p)\, dp = \Gamma(a+1)\, \Gamma(b)\, \Gamma(a+b)/\Gamma(a+b+1)\, \Gamma(a)\, \Gamma(b)$. Because $\Gamma(a+1) = a\, \Gamma(a)$ and $\Gamma(a+b+1) = (a+b)\, \Gamma(a+b)$, this reduces to $E[X] = a/(a+b)$.

Exercise P3.9

In each trial a female has a chance p of encountering a male of type 1. The chance that she has not mated before the kth trial is $(1-F)^{k-1}$. Multiplying these two together and summing over all possible numbers of trials, we find that the probability that a female mates with a male of type 1 is $p + p\,(1-F) + p\,(1-F)^2$, which simplifies to $p\,(3 - 3\,F + F^2)$. This equals p times the mean number of trials, which in hindsight, makes sense. Because there are only two mutually exclusive outcomes (she mates with a male of type 1 or with a male of type 2), the probability that she mates with a male of type 2 is $1 - p\,(3 - 3\,F + F^2)$. Note that if we ignore the mechanics behind how a choice was made, whether a female mates with a male of type 1 or type 2 is described by a Bernoulli trial.