Chapter 2

Probability Theory

Chapter 2 is a summary of the probability theory required by this text. It is assumed that the reader has taken an upper level, calculus-based probability course already, and this chapter is intended as a reference to be consulted if needed. Section 2.1 provides all the background necessary for Chapter 3 on population genetics. We recommend reading only this section before going on, and then returning to the later parts of Chapter 2 when and if necessary.

2.1 Elementary probability; random sampling

2.1.1 Probability spaces

A probability space consists of:

(i) a set Ω called the *outcome space*;

(ii) a class of subsets of Ω called events; and,

(iii) a rule $\mathbb{P}$ that assigns a probability $\mathbb{P}(U)$ to each event U.

This structure is a template for modeling any experiment with random outcome: the set Ω is a list of all the possible outcomes of the experiment; a subset U of Ω represents the 'event' that the outcome of the experiment belongs to U; and $\mathbb{P}(U)$ is the probability that event U occurs. The assignment, $\mathbb{P}$, is called a *probability measure*. It is required to satisfy the following properties, called the *axioms of probability:*

(P1) $0 \leq \mathbb{P}(U) \leq 1$ for all events U;

(P2) $\mathbb{P}(\Omega) = 1$;

(P3) If $U_1, U_2, \ldots, U_n$ are disjoint events, meaning they share no outcomes in common,

$$\mathbb{P}(U_1 \cup \cdots \cup U_n) = \mathbb{P}(U_1) + \cdots + \mathbb{P}(U_n). \tag{2.1}$$

(P4) More generally, if $U_1, U_2, U_3, \ldots$ is an infinite sequence of disjoint events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} U_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(U_i). \tag{2.2}$$

What exactly does it mean to say the probability of event U is some number $\mathbb{P}(U)$? Axioms (P1)—(P4) don't say! This should remind you of Euclidean geometry; its axioms, which are assumed properties of points and lines, leave points and lines undefined. But just as our intuition about physical space guides axioms of geometry, so does intuition about how we assign likelihoods to future events, when gambling or making predictions, guide axioms of probability. For example, in repeating an experiment over and over, we see that different events will occur with different frequencies, and we might view a probability as an ideal frequency of occurrence in repeated trials. Thus, assigning a probability of 1/2 to heads for a coin toss, would mean we expect, on average, half of future tosses to be heads. Axioms (P1)—(P4) can all be motivated in a simple manner from this viewpoint. As a frequency, $\mathbb{P}(U)$ will necessarily take a value between 0 and 1, as required by axiom (P1). Since Ω is the set of all possible outcomes, the fraction of trials in which the outcome falls in Ω is $\mathbb{P}(\Omega) = 1$, as required by axiom (P2). If U and V are two disjoint events, the fraction of times an outcome is in U or in V is the sum of the fraction of times it is in U and the fraction of times it is in V. Axiom (P3) generalizes this addition principle to arbitrary finite, disjoint unions and (P4) generalizes it to unions of infinite sequences of disjoint events. Axioms (P3) and (P4) are called *additivity* axioms. (Actually (P3) is a special case of (P4). We did not need to state it separately, but it is useful to distinguish the finite from the infinite case in developing probability theory.)

We should point out that many probabilists and statistician hotly contest interpreting probabilities as an objectively defined frequencies, as we did here to motivate axioms (P1)—(P4). They do generally accept these axioms, but view probabilities as subjective opinions subject to modification by evidence. We mean to take no sides on this issue. But the frequency approach provides simple intuition and is especially relevant to to repeated sampling, a procedure used over and over in modeling and statistical practice.

Examples 2.1.1.

(a) (*Roll of a fair die.*) Consider rolling a die and recording the number that is face up when the die comes to a rest. The outcome space of this experiment is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event of rolling precisely number i is the singleton subset $\{i\}$ of Ω. The die is fair if each of the 6 possible outcomes has the same probability, and since, by axioms (P3) and (P2) the sum of these 6 equal probabilities must be 1, we find $\mathbb{P}(\{i\}) = 1/6$ for each $1 \leq i \leq 6$. This completely determines $\mathbb{P}$. For example, the event of rolling an even number is represented by the subset $\{2, 4, 6\}$. Since it is the union of the singleton events $\{2\}$, $\{4\}$, and $\{6\}$, axiom (P3) implies

$\mathbb{P}(\{2,4,6\}) = 1/6 + 1/6 + 1/6 = 1/2$. By the same reasoning, for any subset U of $\{1,2,3,4,5,6\}$, $\mathbb{P}(U) = |U|/6$, where $|U|$ denotes the size of U. Granted, this example belabors the obvious, but it illustrates in very simple form the application of the probability axioms.

(b) (*General model for the roll of a die.*) The general model for the roll of a die, possibly biased, requires specifying $p_i = \mathbb{P}\big(\{i\}\big)$, the probability of rolling i, for each $i = 1, 2, \ldots, 6$. The only constraints on the vector $(p_1, \ldots, p_6)$ are that its entries be non-negative and that they sum to 1. By axiom (P3), for any event U, $\mathbb{P}(U)$ is the sum of the probabilities of the outcomes i belonging to U; for example, $\mathbb{P}(\{2,4,6\}) = p_2 + p_4 + p_6$ is the probability of an even roll.

(c) (*General discrete probability space.*) A probability space is called *discrete* if Ω is a finite set or a set, $\Omega = \{\omega_1, \omega_2, \ldots\}$, whose elements can be indexed by the natural numbers. (The latter type of set is called *countably infinite*.) When Ω is discrete, a probability space on Ω can be completely specified by assigning a probability, $p_\omega = \mathbb{P}(\{\omega\})$, to each singleton event $\{\omega\}$. For any event U in Ω which is not a singleton set, axioms (P3) and (P4) imply

$$\mathbb{P}(U) = \sum_{\omega \in U} \mathbb{P}\left(\{\omega\}\right) = \sum_{\omega \in U} p_\omega$$

because U is the disjoint union of the outcomes in U. (Here, the notation indicates that the sum is over all ω in U.) The only restrictions on the assignment, $\{p_\omega;\, \omega \in \Omega\}$, are that $0 \le p_\omega \le 1$ for all ω, which must hold because of axiom (P1), and that $1 = \mathbb{P}(\Omega) = \sum_{\omega \in \Omega} p_\omega$, which must hold because of axiom (P2). ◇

Three simple consequences of axioms (P1)—(P3) are repeatedly used.

(1) If U is an event, let U^c denote the complement of U, that is all elements in Ω that are not in U. Then, by the finite additivity axiom and axiom (P2), $1 = \mathbb{P}(\Omega) = \mathbb{P}\big(U \cup U^c\big) = \mathbb{P}(U) + \mathbb{P}(U^c)$. Hence,

$$\mathbb{P}(U^c) = 1 - \mathbb{P}(U). \tag{2.3}$$

In particular, since the empty set, $\emptyset$, equals Ω^c, $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 0$. (The empty set, $\emptyset$, should be thought of as the event that, in a trial of the experiment, *nothing* happens, which has zero probability.)

(2) If A and B are events and if A is a subset of B, then $\mathbb{P}(A) \le \mathbb{P}(B)$.

This is a consequence of axiom (P3). Since $B = A \cup [B - A]$ and A and $B - A$ are disjoint if $A \subset B$, $\mathbb{P}(B) = P(A) + \mathbb{P}(B - A)$. But $\mathbb{P}(B - A) \ge 0$, and it follows that $\mathbb{P}(B) \ge \mathbb{P}(A)$.

(3) (Inclusion-Exclusion Principle) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Indeed, since $A \cup B$ is the disjoint union of A and $B \cap A^c$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$, and since B is the disjoint union of $B \cap A$ and $B \cap A^c$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$, and hence $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Thus, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

The facts just stated depend only on axiom (P3). Axiom (P4) has the following important consequences; the proofs are omitted.

$$\text{if } A_1 \subset A_2 \subset \cdots, \quad \text{then} \quad \mathbb{P}(\bigcup_1^\infty A_n) = \lim_{n\to\infty} \mathbb{P}(A_n). \tag{2.4}$$

$$\text{if } A_1 \supset A_2 \supset \cdots, \quad \text{then} \quad \mathbb{P}(\bigcap_1^\infty A_n) = \lim_{n\to\infty} \mathbb{P}(A_n). \tag{2.5}$$

2.1.2 Random sampling

In this section and throughout the text, if U is a finite set, $|U|$ denotes its cardinality, that is, the number of elements in U.

A basic model in both statistical practice and probability theory is the *random sample*. Let $\mathcal{S}$ denote a finite set, called the *population*. *A (single) random sample from $\mathcal{S}$ is a random draw from $\mathcal{S}$ in which each individual has an equal chance to be chosen.* In the probability space for this experiment, the outcome space is $\mathcal{S}$, and the probability of drawing any particular element s of $\mathcal{S}$ is

$$\mathbb{P}(\{s\}) = \frac{1}{|\mathcal{S}|}.$$

$\mathbb{P}$ assigns probabilities uniformly to the singleton outcomes, so is sometimes called the *uniform* probability measure on $\mathcal{S}$. It follows that if U is any subset of $\mathcal{S}$,

$$\mathbb{P}(U) = \sum_{s\in U} \mathbb{P}(\{s\}) = \sum_{s\in U} \frac{1}{|\mathcal{S}|} = \frac{|U|}{|\mathcal{S}|}. \tag{2.6}$$

In many applications, each individual in the population bears an attribute or descriptive label. For example, in population genetics, each individual in a population of organisms is labeled by its genotype. In a bag of marbles, each of which is either red, yellow, or blue, color is a label. If x is a label attaching to individuals in $\mathcal{S}$, the *frequency* of x in $\mathcal{S}$ is the ratio,

$$f_x \stackrel{\triangle}{=} \frac{\text{number of individuals in } \mathcal{S} \text{ with label } x}{|\mathcal{S}|}.$$

Because random sampling is modeled by a uniform probability measure,

$$\mathbb{P}\left(\text{randomly selected individual bears label } x\right) = f_x. \tag{2.7}$$

This is simple, but is basic in population genetics.

Example 2.1.2. Random mating; Part I. One assumption behind the textbook analysis of Mendel's pea experiments—see Chapter 1—is that each individual chosen for a random cross is sampled randomly from the population. Consider a population of N pea plants, and suppose k of them have the genotype GG and ℓ of them genotype YY for pea color. Thus $N-k-\ell$ plants have the heterozygous genotype YG. Suppose the first parent for a cross is chosen by random sampling from the population. What is the probability it is heterozygotic? According to (2.1),

$$\mathbb{P}\Big(YG \text{ plant is selected}\Big) = f_{YG} = \frac{N-k-\ell}{N}.$$

$\diamond$.

In many applications, populations are sampled repeatedly. For example, statisticians will sample a large population multiple times to obtain data for estimating its structure. Models of population genetics often posit that populations evolve from one generation to the next by repeated sampling. Repeated random sampling may be performed *without replacement*, in which case the each sampled individual is removed from the population, or *with replacement*, in which case each sampled individual is returned to the population and may be sampled again.

Imagine sampling a population, $\mathcal{S}$, n times with replacement. The outcome space is then the set of all sequences of the form $(s_1, \ldots, s_n)$, where $s_i \in \mathcal{S}$ for each i. This set is denoted $\mathcal{S}^n$, it is called the n-fold product space of $\mathcal{S}$, and it contains $|\mathcal{S}|^n$ sequences. A random sample (with replacement) of size n is a sample in which every sequence in $\mathcal{S}^n$ is equally likely. This is a direct generalization of the definition of a single random sample given above. The probability measure modeling a random sample of size n is therefore just the uniform measure on $\mathcal{S}^n$:

$$\mathbb{P}\left(\{(s_1, \ldots, s_n)\}\right) = \frac{1}{|\mathcal{S}|^n}, \quad \text{for each } (s_1, \ldots, s_n) \in \mathcal{S}^n; \text{ and} \tag{2.8}$$

$$\mathbb{P}(V) = \frac{|V|}{|\mathcal{S}|^n}, \quad \text{for any subset } V \text{ of } \mathcal{S}^n. \tag{2.9}$$

This notion of random sample is consistent in the sense that a subsequence of m samples from a random sample of size $n > m$, is a random sample of size m. For example, consider just the outcome s_1 of the first draw. The event that $s_1 = s$ in a sample of size n is the subset of sequences in $\mathcal{S}^n$,

$$\Big\{(s_1, s_2, \ldots, s_n) : \, s_1 = s, \, s_j \in \mathcal{S}, \text{ for } 2 \le j \le n\Big\}.$$

The cardinality of this set is just the number of sequences $(s_2, \ldots, s_n)$ with length $n-1$ whose elements come from $\mathcal{S}$, and this number is $|\mathcal{S}|^{n-1}$. Hence,

$$\mathbb{P}\Big(\Big\{(s_1, s_2, \ldots, s_n) : \, s_1 = s, \, s_j \in \mathcal{S}, \text{ for } 2 \le j \le n\Big\}\Big) = \frac{|\mathcal{S}|^{n-1}}{|\mathcal{S}|^n} = \frac{1}{|\mathcal{S}|}.$$

This is exactly the probability of drawing s in a single sample.

A random sample from $\mathcal{S}$ without replacement is defined in a similar way. Of course, if $(s_1, \ldots, s_n)$ is a sample from $\mathcal{S}$ without replacement, n must be less than or equal to $\mathcal{S}$. Assume this is the case. Then the outcome space is the set Ω of all sequences $(s_1, \ldots, s_n)$ of elements of $\mathcal{S}$ with no repeats, and the cardinality of Ω is $|\Omega| = |\mathcal{S}|(|\mathcal{S}| - 1) \cdots (|\mathcal{S}| - n + 1)$. A random sample of size n without replacement is one in which every sequence in Ω is equally probable, so it is modeled by the uniform distribution on Ω.

In this text, the phrase 'random sampling', when used without further clarification, will always mean sampling *with replacement.*

Example 2.1.2, part II. The standard model for a random cross in Mendel's pea experiment assumes the plants to be crossed are chosen by a random sample of size 2. (Since this is random sampling with replacement, it is possible that a plant is crossed with itself.) In the set up of part I of this example above, what is the probability that a GG genotype is crossed with a YY genotype?

In this case, the outcome space is the set of all ordered pairs (s_1, s_2) where s_1 and s_2 belong to the population of plants. Let U be the event that one of these has genotype GG and the other genotype YY. Since there are k plants in the population of type GG and ℓ of type YY, there are $k \cdot \ell$ pairs (s_1, s_2) in which s_1 is type GG and s_2 is type YY. Likewise there are $\ell \cdot k$ pairs in which s_1 is type YY and s_2 is type GG. It follows that $\big|U\big| = 2k\ell$ and

$$\mathbb{P}(U) = \frac{2k\ell}{N^2} = 2f_{GG}f_{YY}. \qquad \diamond$$

Example 2.1.2, part III. Suppose we take a random sample of 10 plants from a population of N of Mendel's peas, of which, as in Part I of this example, k have genotype GG. What is the probability that exactly 6 of those sampled plants have genotype GG? The event V in this case is the set of sequences of 10 plants from the population for which a GG genotypes appears exactly 6 times. Now, there are $\binom{10}{6}$ ways the 6 positions at which GG types occur can be distributed among the 10 samples. For each such arrangement of these 6 positions, there are n choices of GG plants to place in each of these positions, and $N - k$ choices of non-GG plants to place in the remaining $10-6 = 4$ positions. Thus, $\big|V\big| = \binom{10}{6}k^6(N-k)^4$. By (2.9),

$$\mathbb{P}(V) = \frac{\binom{10}{6}k^6(N-k)^4}{N^{10}} = \binom{10}{6}\left(\frac{k}{N}\right)^6\left(1 - \frac{k}{N}\right)^4 = \binom{10}{6}f_{GG}^6(1 - f_{GG})^4.$$

(This problem is not so interesting biologically, but it serves as a review of binomial probabilities!) $\diamond$

By generalizing the argument of the last example, the probability that individuals labeled x appear m times in a random sample of size n from some population

$\mathcal{S}$ is the binomial probability

$$\binom{n}{m} f_x^m (1 - f_x)^{n-m}. \tag{2.10}$$

2.1.3 Randomly selecting a point from $[0, 1]$

The purpose of this section is to discuss an example of a probability space that is not discrete. Consider drawing a point at random from the interval $[0, 1]$. We would like to build a probability space modeling a selection procedure that does not favor any region of $[0, 1]$ over any other. For this, there is a natural candidate. For a subset U in $[0, 1]$, let $\mathbb{P}(U)$ be the total length of U. Thus $\mathbb{P}(U)$ is the fraction of $[0, 1]$ occupied by U, and since subsets of equal length have the same probability, it does not favor any particular region. For this reason, length measure on $[0, 1]$ is called the *uniform measure* on $[0, 1]$.

However, there are issues with this definition. First, how do we define $\mathbb{P}(U)$ for every subset U of $[0, 1]$? It is clear that if U is a union $\cup_1^\infty I_i$ of disjoint intervals, we should define $\mathbb{P}(U) = \sum_1^\infty \text{length}(I_i)$, if axiom (P4) is to hold. But the class of subsets of $[0, 1]$ is very rich, and includes many subsets that cannot be represented as unions of intervals. Second, even supposing we have a method to define $\mathbb{P}(U)$ for any U, can we then verify the additivity axiom, (P4)? The resolution of these problems is subtle. Assuming standard axioms of set theory used in analysis, it is **not** possible to extend length measure to all subsets of U in such way that (P4) holds and $\mathbb{P}(U)$ is unchanged by translation of U. A probability space satisfying all the axioms can be constructed, but only by restricting the class of subsets for which $\mathbb{P}(U)$ is defined. This theory is too advanced for this text, and we only want the reader to be aware of the issue. Fortunately, for applications it is rarely necessary to go beyond events which are unions of intervals, and then total length is well-defined and, by the advanced theory, perfectly valid to use.

As a probability measure, the length measure has a peculiar property. Let a be any point of $[0, 1]$. The event of selecting exactly a is the singleton set $\{a\}$, whose length is zero, and so the probability of observing any particular a is zero. Yet, if the experiment is run, some a is selected! We can weasel out of this conundrum by recognizing that physical quantities cannot be measured to arbitrary accuracy. If points are measured to a decimal accuracy of, say, five places, the result is really a draw from the finite set of decimal numbers of the form $0.r_1r_2r_3r_4r_5$ or the number 1 itself. There are $10^5 + 1$ such numbers in $[0, 1]$, and a uniform draw means they all have equal probabilities. So, why not just work with this discrete model in the first place? The answer is that the probability space given by length measure on $[0, 1]$ is independent of the accuracy we measure to, it is simpler to work with mathematically, and it is an excellent approximate model if measurements are very accurate.

2.1.4 Independence

Two events U and V in a probability space are said to be **independent** if

$$\mathbb{P}(U \cap V) = \mathbb{P}(U)\mathbb{P}(V). \tag{2.11}$$

This property is called 'independence' because of the theory of conditional probability. When $\mathbb{P}(V) > 0$, the conditional probability of U given V is defined as

$$\mathbb{P}\left(U\middle|V\right) \triangleq \frac{\mathbb{P}(U \cap V)}{\mathbb{P}(V)}, \tag{2.12}$$

and it represents the probability of U given that V has occurred. (Conditional probability will be reviewed in more depth presently.) If U and V are independent and both have positive probability, then $\mathbb{P}(U\big|V) = [\mathbb{P}(U)\mathbb{P}(V)]/\mathbb{P}(V) = \mathbb{P}(U)$, and likewise $\mathbb{P}(V\big|U) = \mathbb{P}(V)$; thus the probability of neither U nor V is affected by conditioning on the other event, and this is the sense in which they are independent.

Three events U_1, U_2, U_3 are said to be independent if they are pairwise independent—that is U_i and U_j are independent whenever $i \neq j$— *and, in addition*

$$\mathbb{P}(U_1 \cap U_2 \cap U_3) = \mathbb{P}(U_1)\mathbb{P}(U_2)\mathbb{P}(U_3). \tag{2.13}$$

The reason for the last condition is that we want independence of the three events to mean that any one event is independent of any combinations of the other events, for example, that U_1 is independent of U_2, of U_3, of $U_2 \cap U_3$, $U_2 \cup U_3$, etc. Pairwise independence is not enough to guarantee this, as the reader will discover by doing Exercise 2.4. However, adding condition (2.13) is enough—see Exercise 2.5.

The generalization to more than three events is straightforward. Events $U_1, \ldots, U_n$ are independent if

$$\mathbb{P}(U_{r_1} \cap \cdots \cap U_{r_k}) = \mathbb{P}(U_{r_1}) \cdots \mathbb{P}(U_{r_k}) \tag{2.14}$$

for every every k**,** $2 \leq k \leq n$**, and every possible subsequence** $1 \leq r_1 < r_2 < \cdots < r_k \leq n$**.** This condition will imply, for example, that event U_i is independent of any combination of events U_j for $j \neq i$.

Independence is fundamental in multiple random sampling with replacement. Imagine a random sample, $(s_1, \ldots, s_n)$, of size n from a population $\mathcal{S}$. Let B_1, B_2,$\ldots, B_n$ be subsets of $\mathcal{S}$ and let U_1 be the event that the first sample s_1 is in B_1, let U_2 be the event that the second sample is in B_2, and so on. Then, we claim that no matter what $B_1, \ldots, B_n$ are, the events $U_1, \ldots, U_n$ are independent. We summarize this state of affairs by saying that the successive samples in the sequence of n samples are independent. The converse is also true. These facts are so important, they deserve a formal statement.

Proposition 1 *In repeated random sampling (with replacement) the outcomes of the different samples are independent of one another. Conversely, if n single random samples from* $\mathcal{S}$ *are independent, they constitute a random sample of size n.*

We show this when $n = 3$; the notation is then simpler, and no new ideas are required for the general case. Let B_1, B_2, and B_3 be three subsets of $\mathcal{S}$. Consider U_1, the event that the first element, s_1, in a random sample (s_1, s_2, s_3), falls in B_1. It was shown above that s_1 by itself is a single random sample, namely that $\mathbb{P}(U_1) = |B_1|/|\mathcal{S}|$. Exactly the same reasoning applies to U_2 and U_3, so $\mathbb{P}(U_2) = |B_2|/|\mathcal{S}|$ and $\mathbb{P}(U_3) = |B_3|/|\mathcal{S}|$. Now consider $U_1 \cap U_2 \cap U_3$. This is the set of all sequences (s_1, s_2, s_3) such that $s_1 \in B_1$, $s_2 \in_2$, and $s_3 \in B_3$ and so this set contains $|B_1| \cdot |B_2| \cdot |B_3|$ sequences. By the definition of multiple random sample

$$\mathbb{P}\Big(U_1 \cap U_2 \cap U_3\Big) = \frac{\big|B_1\big| \cdot \big|B_2\big| \cdot \big|B_3\big|}{\big|\mathcal{S}\big|^3} = \frac{\big|B_1\big|}{\big|\mathcal{S}\big|} \cdot \frac{\big|B_2\big|}{\big|\mathcal{S}\big|} \cdot \frac{\big|B_3\big|}{\big|\mathcal{S}\big|} = \mathbb{P}(U_1)\mathbb{P}(U_2)\mathbb{P}(U_3).$$

This establishes (2.13) for any B_1, B_2, and B_3.

It is also necessary to show pairwise independence. But this is a consequence of what we have just shown, because, for instance, $U_1 \cap U_2 = U_1 \cap U_2 \cap \mathcal{S}$. It follows that $\mathbb{P}(U_1 \cap U_2) = \mathbb{P}(U_1 \cap U_2 \cap \mathcal{S}) = \mathbb{P}(\mathbb{P}(U_1)(U)2)\mathbb{P}(\mathcal{S}) = \mathbb{P}(U_1)\mathbb{P}(U_2)$, since $\mathbb{P}(\mathcal{S}) = 1$, proving independence of U_1 and U_2. The independence of U_1 and U_3 and of U_2 and U_3 is shown in the same way.

Assume conversely, that the three draws are independent. Let $U_1 = \{s_1\}$, $U_2 = \{s_2\}$, and $U_3 = \{s_3\}$. For each i, $\mathbb{P}(U_i) = 1/|\mathcal{S}|$. By independence,

$$\mathbb{P}\big(\{(s_1, s_2, s_3)\}\big) = \mathbb{P}(U_1 \cap U_2 \cap U_3) = \mathbb{P}(U_1)\mathbb{P}(U_2)\mathbb{P}(U_3) = \frac{1}{|\mathcal{S}|^3}.$$

Thus $\mathbb{P}$ is uniform measure on $\mathcal{S}^3$, which means it describes a random sample of size 3.

2.1.5 The law of large numbers for random sampling

Suppose we observe the fraction of times event U occurs as we repeat an experiment over and over. A law of large number is a theorem that states conditions under which this fraction tends to the probability $\mathbb{P}(U)$ of U, as the the number of trials tends to infinity. It establishes in what sense and under what hypotheses a frequency interpretation of probability, as described in Section 2.1.1, is valid,

For example, a law of large numbers holds for random sampling. Let $\mathcal{S}$ be a finite population, and imagine repeatedly sampling from $\mathcal{S}$ with replacement. If n is any positive integer and A is a subset of $\mathcal{S}$, let

$$f^{(n)}(A) = \frac{\text{number of sampled values in } A \text{ among the first } n \text{ samples}}{n}$$

denote the fraction of times A occurs in the first n samples. It is called the *empirical frequency of A in the first n trials.*

Theorem 1 *(*Strong Law of Large Numbers for random sampling.*) For a sequence of independent random samples from $\mathcal{S}$,*

$$\lim_{n\to\infty} f^{(n)}(A) = \frac{|A|}{|\mathcal{S}|} = \mathbb{P}(A), \quad \textit{with probability one.} \tag{2.15}$$

If x is a label, and if $f_x^{(n)}$ is the empirical frequency of individuals with label x in the first n samples, then Theorem 1 says,

$$\lim_{n\to\infty} f_x^{(n)} = f_x.$$

The proof of the strong law is advanced. There is also a *weak law of large numbers*, which has an elementary proof. Chebyshev's inequality (treated later in this chapter) implies that for any $a > 0$,

$$\mathbb{P}\left(\left|f_x^{(n)} - f_x\right| > a\right) \le \frac{f_x(1-f_x)}{na^2} \le \frac{1}{4na^2}. \tag{2.16}$$

(The last inequality is a consequence of the fact that if $0 \le p \le 1$, $p(1-p) \le 1/4$.) It follows that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|f_x^{(n)} - f_x\right| > a\right) = 0, \quad \text{for any } a > 0. \tag{2.17}$$

This is the weak law.

The difference between the strong and weak laws is subtle. The strong law implies the weak law, but the weak law in and of itself does not imply the strong law; we will not try to explain this point. Inequalities like that of Chebyshev, which imply the weak law, are very useful in practice because they give quantitative bounds on the probabilities of the difference between empirical frequencies and theoretical probabilities. Another inequality worth mentioning is Chernoff's inequality:

$$\mathbb{P}\big(\big|f^{(n)}(A) - f(A)\big| > a\big) < 2e^{-2na^2}. \tag{2.18}$$

This gives a much sharper bound than the Chebyshev inequality. Its proof, which uses the Markov inequality and moment generating functions, is omitted here.

The law of large numbers motivates an important approximation used in population genetics models. Consider repeated random sampling from a population of organisms in which the frequency of some genotype, call it L, is f_L. By the weak law of large numbers, the frequency of L in a large sample will be close to f_L with high probability. Therefore, when studying large samples from a population, we can replace the empirical frequency, which is random, by f_L, which is not, and obtain a good approximate model. This idea is at the heart of the *infinite population models* studied in Chapter 3.

2.1.6 Conditioning, Rule of total probabilities, Bayes' rule

Let U and V be two events, with $\mathbb{P}(V) > 0$. Recall that the conditional probability of U given V is defined as

$$\mathbb{P}\left(U\middle|V\right) \stackrel{\triangle}{=} \frac{\mathbb{P}(U \cap V)}{\mathbb{P}(V)},$$

and is interpreted as the probability U occurs knowing that V has occurred. Conditional probabilities are often used for modeling and for calculation when events are not independent. A very important tool for these applications is the *rule of total probabilities.*

Proposition 2 *Let the events $V_1, \cdots, V_n$ a form a disjoint partition of the outcome space Ω, that is, $\Omega = V_1 \cup \cdots V_n$ and $V_i \cap V_j = \emptyset$ if $i \neq j$. Assume that $\mathbb{P}(V_i) > 0$ for all i. Then for any event U,*

$$\mathbb{P}(U) = \mathbb{P}(U\big|V_1)\mathbb{P}(V_1) + \mathbb{P}(U\big|V_2)\mathbb{P}(V_2) + \cdots + \mathbb{P}(U\big|V_n)\mathbb{P}(V_n). \tag{2.19}$$

This is easy to derive. For any i, the definition of conditional probability implies

$$\mathbb{P}(U \cap V_i) = \mathbb{P}(U\big|V_i)\mathbb{P}(V_i). \tag{2.20}$$

Since $V_1, \ldots, V_n$ form a disjoint partition of Ω, $U = [U \cap V_1] \cup [U \cap V_2] \cup \cdots \cup [U \cap V_n]$, and it follows from axiom (P3) for probability spaces that

$$\begin{aligned} \mathbb{P}(U) &= \mathbb{P}(U \cap V_1) + \cdots + \mathbb{P}(U \cap V_n) \\ &= \mathbb{P}(U\big|V_1)\mathbb{P}(V_1) + \cdots + \mathbb{P}(U\big|V_n)\mathbb{P}(V_n). \end{aligned}$$

Example 2.1.3. In a cross of Mendel's pea plants, each parent plant contributes one allele at each locus. Thus a cross of a GG with another GG produces a GG offspring and the cross of a GG with a YY produces a GY offspring. What happens if a parent is YG? In this case, it is assumed that the parent passes on each allele with equal probability, so it contributes Y with probability 0.5 and G with probability 0.5. This can be expressed as a statement about conditional probabilities:

$$\mathbb{P}(G\big|GY) = \mathbb{P}(Y\big|GY) = 0.5.$$

where here $\mathbb{P}(G\big|GY)$ is shorthand for probability that a parent contributes G to an offspring *given* that it has genotype GY. Now suppose a plant is selected at random from a population and used to fertilize another plant. What is the probability this selected plant contributes allele G to the offspring? Let G represent the event that it contributes G. Let GG, GY, and YY represent the event that the selected plant has genotype GG, GY, or YY, respectively, and recall that random sampling means

$\mathbb{P}(GG) = f_{GG}$, $\mathbb{P}(GY) = f_{GY}$, $\mathbb{P}(YY) = f_{YY}$. Then clearly, $\mathbb{P}(G\big|GG) = 1$ and $\mathbb{P}(G\big|YY) = 0$. By the rule of total probabilities,

$$\mathbb{P}(G) = \mathbb{P}(G\big|GG)\mathbb{P}(GG) + \mathbb{P}(G\big|GY)\mathbb{P}(GY) + \mathbb{P}(G\big|YY)\mathbb{P}(YY) = f_{GG} + (0.5)f_{GY},$$

is the probability the randomly selected parent passes G to an offspring. ◇

The rule of total probabilities generalizes to conditional probabilities. Again, let $V_1, \ldots, V_n$ be a disjoint partition of Ω. Assume $\mathbb{P}(W) > 0$. Then

$$\mathbb{P}(U\big|W) = \sum_{i=1}^{n} \mathbb{P}\left(U\big|V_i \cap W\right) \mathbb{P}(V_i\big|W). \tag{2.21}$$

The proof is very similar to that for the ordinary rule of total probabilities. Write out all conditional expectations using the definition (2.12) and use the fact that $\mathbb{P}(U\cap W) = \sum_{i=1}^{n} \mathbb{P}(U\cap V_i \cap W)$. The details are left as an exercise. This conditioned version of the total probability rule will be used for analyzing Markov chains in Chapter 4.

Another important formula using conditional probabilities is **Bayes' rule**. By the total probability rule, $\mathbb{P}(U) = \mathbb{P}(U\big|V)\mathbb{P}(V) + \mathbb{P}(U\big|V^c)\mathbb{P}(V^c)$, where V^c is the complement $V^c = \Omega - V$ of V. Also, $\mathbb{P}(U \cap V) = \mathbb{P}(U\big|V)\mathbb{P}(V)$. Combining these formulae gives Bayes' rule:

$$\mathbb{P}(V\big|U) = \frac{\mathbb{P}(U \cap V)}{\mathbb{P}(U)} = \frac{\mathbb{P}(U\big|V)\mathbb{P}(V)}{\mathbb{P}(U\big|V)\mathbb{P}(V) + \mathbb{P}(U\big|V^c)\mathbb{P}(V^c)}. \tag{2.22}$$

Example 2.1.4. Suppose, in the previous example, that the randomly chosen parent plant contributes G? What is the probability that the parent was GY? Keeping the same notation, this problem asks for $\mathbb{P}(GY\big|G)$, and by Bayes' rule it is

$$\mathbb{P}(GY\big|G) = \frac{\mathbb{P}(G\big|GY)\mathbb{P}(GY)}{\mathbb{P}(G)} = \frac{(0.5)f_{GY}}{f_{GG} + (0.5)f_{GY}} = \frac{f_{GY}}{2f_{GG} + f_{GY}}. \qquad \diamond$$

Example 2.1.5. This example is a preview of population genetics models incorporating selection. Consider a diploid species that has three genotypes Aa, Aa, and aa, at a locus. Let U denote the event a (randomly chosen) individual survives from birth to reproductive maturity. We can quantify the effects genotype has on the probability of survival by the conditional probabilities, $w_{AA} = \mathbb{P}(U|AA)$, $w_{Aa} = \mathbb{P}(U|Aa)$, $w_{aa} = \mathbb{P}(U|aa)$, where, for example, $\mathbb{P}(U|AA)$ is the probability of survival from birth to maturity of an individual with genotype AA. Consider a generation in which the genotype frequencies at birth are f_{AA}, f_{Aa}, and f_{aa}.

a) What is the probability of U?

b) What is the probability that an individual has genotype AA, given that it has survived to reproductive maturity?

The answer to a) is obtained from the rule of total probabilities:

$$\mathbb{P}(U) = \mathbb{P}(U|AA)f_{AA} + \mathbb{P}(U|Aa)f_{Aa} + \mathbb{P}(U|aa)f_{aa} = w_{AA}f_{AA} + w_{Aa}f_{Aa} + w_{aa}f_{aa}.$$

Question b) asks for $\mathbb{P}(AA|U)$, and this is calculated by Bayes' rule:

$$\mathbb{P}(AA|U) = \frac{\mathbb{P}(U|AA)f_{AA}}{P(U)} = \frac{w_{AA}f_{AA}}{w_{AA}f_{AA} + w_{Aa}f_{Aa} + w_{aa}f_{aa}}.$$

2.1.7 Exercises.

2.1. Population I ($\{k_1, \ldots, k_N\}$) and population II ($\{\ell_1, \ldots, \ell_M\}$) are, respectively, populations of male and female gametes of Mendel's peas. Alice intends to perform a cross between the two populations. Assume she randomly chooses a pair of gametes, one from population I and one from II. The outcome space of this experiment is the set of pairs (k_i, ℓ_j), where $1 \leq i \leq N$ and $1 \leq j \leq M$. Assume that all pairs are equally likely to be chosen.

a) Let U_i be the event that the gamete from population I is k_i. Explicitly identify this event as a subset of the outcome space and determine its probability. Let V_j be the event the gamete from population II is ℓ_j. Show U_i and V_j are independent.

b) Assume that r of the gametes in population I and s of the gametes in population II carry the allele G for green peas and that the remainder carry allele Y. What is the probability that both gametes selected have genotype G? What is the probability that one of the selected gametes has genotype G and the other has genotype Y?

2.2. If junk DNA (see Chapter 1) has no function, no selective pressure has acted on it and all possible sequences should be equally likely in a population, because it mutates randomly as time passes. Thus, a hypothetical model for a randomly selected piece of junk DNA that is N base pairs long is the uniform distribution on all DNA sequences of length N. This is equivalent to a random sample of size N from the population of DNA letters $\{A, T, G, C\}$. (This model is a preview of the independent sites model with equal base probabilities—see Chapter 6.) Assume this model to do the following problem.

A four bp-long DNA segment is selected at random and sequenced. Find the probability that the selected sequence contains exactly two adenine (A) bases.

2.3. Randomly select two four-base-long DNA segments $x_1x_2x_3x_4$ and $y_1y_2y_3y_4$ and align them as follows: $\begin{smallmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{smallmatrix}$.

a) Assume that the selection of the x-sequence and the selection of the y-sequence are independent, and that both are random samples of size 4 from the DNA alphabet. Construct a probability space for the aligned pair of sequences.

b) What is the probability that both the x-sequence and the y-sequence begin with A? What is the probability that x_1 and y_1 are equal? What is the probability that an A is aligned with A exactly twice in the aligned pair of sequences? What is the probability that $x_1 = y_1$, $x_2 = y_2$, $x_3 = y_3$, $x_4 = y_4$?

2.4. Denoting heads by 1 and tails 0, the outcome space of four tosses of a coin is the set of all sequences $(\eta_1, \eta_2, \eta_3, \eta_4)$ of 0's and 1's. The uniform probability measure on this space models four independent tosses of a fair coin. Let U_1 be the event that the first two tosses are heads, U_2 the event that the last two tosses are heads, and U_3 the event the second and third tosses produce one head and one tail in either order. Show that these events are pairwise independent, but not independent.

2.5. Assume that U_1, U_2, U_3 are independent. Show that U_3 and $U_1 \cap U_2$ are independent and that U_3 and $U_1 \cup U_2$ are independent.

2.6. Three coins are in a box. Two are fair and one is loaded; when flipped, the loaded coin comes up heads with probability 2/3. A coin is chosen by random sampling from the box and flipped. What is the probability that it comes up heads? Given that it comes up heads, what is the conditional probability that it is the loaded coin?

2.7. *Probability space model for sampling without replacement.* Let $\mathcal{S}$ be a population of size N and let $n \leq N$. The outcome space for a random sample of size n drawn from $\mathcal{S}$ without replacement is the set Ω of sequences $(s_1, s_2, \ldots, s_n)$ in which $s_i \in \mathcal{S}$ for all i and in which $s_1, \ldots, s_n$ are all different. The probability model for random sampling without replacement is the uniform probability measure on Ω.

What is the probability of any single sequence in Ω? If k individuals in $\mathcal{S}$ bear the label x and $\ell \leq k$, what is the probability exactly ℓ individuals with label x appear in the sample of size n?

2.2 Random Variables

Suppose you are tossing a coin. Label the outcome of the next toss by the variable X, setting $X = 1$ if the toss is heads, $X = 0$ if tails. Then X is an example of a *random variable*, a variable whose value is not fixed, but random. In general, a variable Y that takes a random value in a set $\mathcal{E}$ is called an $\mathcal{E}$-valued random variable. *By convention, the term* random variable *by itself, with no explicit mention of $\mathcal{E}$, means a random variable taking values in a subset of the real numbers.* Random variables that are not real-valued do appear in some applications in this text; for example, random variables modeling the bases appearing along a DNA sequence take values

in the DNA alphabet, $\{A, C, G, T\}$. It will always be clear from context when a random variable is not real-valued.

In Section 2.1, we used probability spaces to model random phenomena. By interpreting outcomes as random variables, we could instead construct *random variable models*. The two approaches are really equivalent. In the random variables approach, we are just labeling the outcome by X and replacing the outcome space, Ω, by the set, $\mathcal{E}$, of possible values of X. But random variables are usually the preferred option. They provide a better framework for mathematical operations on outcomes, for defining expected values, and for stating limit laws. And it is easier to use them to model complex phenomena involving many interacting random components, each described by its own random variable.

In advanced probability theory, random variables are defined as *functions* on probability spaces. Thus a random variable assigns to each possible outcome, ω, some value $X(\omega)$, which might represent a special attribute of ω. This viewpoint may be useful to keep in mind, but is not explicitly used in this text. In stating random variable models we usually omit any explicit mention of a probability space. However, it is important to understand the probability space theory outlined in Section 2.1; the axioms of probability are used to handle all probability and conditional probability calculations with random variables.

2.2.1 Discrete Random Variable

A random ,X, taking values in a finite set, $\mathcal{E} = \{s_1, \ldots, s_N\}$, or a countably infinite set, $\mathcal{E} = \{s_1, s_2, \ldots\}$, is said to be **discrete**. In this case, the function

$$p_X(x) \stackrel{\triangle}{=} \mathbb{P}(X = x), \qquad x \text{ in } \mathcal{E}. \tag{2.23}$$

is called the **probability mass function** of X. *Modeling the outcome of a random experiment as a discrete random variable is equivalent to specifying its probability mass function.* Once this function is given, the additivity properties of probability determine the outcome of any other event concerning X, because if U is any subset of $\mathcal{E}$,

$$\mathbb{P}(\{X \in U\}) = \sum_{x \in U} p_X(x). \tag{2.24}$$

(The notation on the right-hand-side means that the sum is taken over all x in U.) Note in particular that $1 = \mathbb{P}(X \in \mathcal{E}) = \sum_{x \in \mathcal{E}} p_X(x)$.

In general, any function p on a discrete set $\mathcal{E}$ that satisfies $0 \leq p(x) \leq 1$, for each $x \in \mathcal{E}$, and $\sum_{x \in \mathcal{E}} p(x) = 1$, is called a probability mass function and is a potential model for an $\mathcal{E}$-valued random variable.

Example 2.2.1. Bernoulli random variables. X is a Bernoulli random variable with parameter p if X takes values in $\{0, 1\}$ and

$$p_X(0) = 1 - p, \qquad p_X(1) = p. \tag{2.25}$$

For later reference, it is useful to note that the definition (2.25) can be written in functional form as

$$p_X(s) = p^s(1-p)^{1-s} \qquad \text{for } s \text{ in the set } \{0,1\}. \qquad \diamond \tag{2.26}$$

The Bernoulli random variable is the coin toss random variable. By convention, the outcome 1 usually stands for head and 0 for tails. It is common also to think of Bernoulli random variables as modeling trials that can result either in success ($X = 1$) or failure ($X = 0$). In this case the parameter p is the probability of success. The language of success/failure trials is often used when discussing Bernoulli random variables.

The term *Bernoulli random variable* is also used for any random variable taking on only two possible values, even if these two values differ from 0 and 1. We shall indicate explicitly when this is the case. Otherwise 0 and 1 are the default values.

Example 2.2.2. Let $\mathcal{S}$ be a finite set, which may be thought of as a population. An $\mathcal{S}$-valued random variable X is said to be uniformly distributed on $\mathcal{S}$ if its probability mass function is $p_X(s) = 1/|\mathcal{S}|$ for all $s \in \mathcal{S}$. Obviously, X then *models a single random sample, as defined in the previous section*, in the language of random variables. $\diamond$

Usually, we want to analyze the outcomes of several random trials at once; often they simply represent repetitions of the same experiment. The overall outcome can then be described as a sequence $X_1, \ldots, X_n$ of random variables. To keep the discussion simple, assume they all take values in the same discrete set $\mathcal{E}$. Then the function,

$$p_{X_1 \cdots X_n}(x_1, \ldots, x_n) \stackrel{\triangle}{=} \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right), \qquad x_1, \ldots, x_n \in \mathcal{E} \tag{2.27}$$

is called the joint probability mass function of $(X_1, \ldots, X_n)$. It determines the probabilities of any event involving $X_1, \ldots, X_n$ jointly; if V is any subset of $\mathcal{E} \times \cdots \times \mathcal{E}$ (the n-fold product),

$$\mathbb{P}\Big((X_1, \ldots, X_n) \in V\Big) = \sum_{(x_1,\ldots,x_n) \in V} p_{X_1 \cdots X_n}(x_1, \ldots, x_n).$$

In particular, the probability mass function of each X_i is the so-called, i^{th} marginal of $p_{X_1 \cdots X_n}$:

$$p_{X_i}(x) = \sum_{\{(x_1,\ldots,x_n);\, x_i = x\}} p_{X_1 \cdots X_n}(x_1, \ldots, x_n),$$

where the sum is over all possible sequences in which the i^{th} term is held fixed at $x_i = x$. For example, when $n = 2$, $p_{X_1}(x) = \sum_{x_2 \in \mathcal{E}} p_{X_1, X_2}(x, x_2)$.

A set of random variables $X_1, \ldots, X_n$ with values in $\mathcal{E}$ is said to be independent if

- the events $\{X_1 \in U_1\}, \{X_2 \in U_2\}, \ldots, \{X_n \in U_n\}$ are independent for every choice of subsets $U_1, U_2, \ldots, U_n$ of $\mathcal{E}$.

This situation occurs very commonly in applied and statistical modeling. When it does, the joint distribution of $X_1, \ldots, X_n$ is particularly simple. If $x_1, \ldots, x_n$ are elements in $\mathcal{E}$, the events $\{X_1 = x_1\}, \ldots, \{X_n = x_n\}$ are independent, and hence

$$\begin{aligned} p_{X_1 \cdots X_n}(x_1, \cdots, x_n) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\ &= \mathbb{P}(X_1 = x_1)\, \mathbb{P}(X_2 = x_2) \cdots \mathbb{P}(X_n = x_n) \\ &= p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_n}(x_n) \end{aligned} \tag{2.28}$$

for any choice of $x_1, \ldots, x_n$ in $\mathcal{E}$. In words, the joint probability mass function is the product of the probability mass functions of the individual random variables. In fact, this condition characterizes independence.

Theorem 2 *The set $X_1, \ldots, X_n$ of discrete random variables with values in a discrete space $\mathcal{E}$ is independent if and only if (2.28) is true for all choices $x_1, \ldots, x_n$ of values from $\mathcal{E}$.*

We know already that independence implies (2.28), so to prove Theorem 2 requires showing the converse. The reader should prove the case $n = 2$ as an exercise; once this case $n = 2$ is understood, the general case can be proved by induction on n.

Example 2.2.3. Independent Bernoulli random variables. Flip a coin n times and record the result as the sequence $(X_1, \ldots, X_n)$, where, as usual, $X_i = 1$ in the event of heads and $X_i = 0$ in the event of tails. Assume the flips are independent and p is the probability of heads on each flip. Then $X_1, \ldots, X_n$ are independent Bernoulli random variables, each with parameter p. Using (2.26) their joint distribution is

$$\begin{aligned} \mathbb{P}(X_1 = s_1, \ldots, X_n = s_n) &= p^{s_1}(1-p)^{1-s_1} \cdots p^{s_n}(1-p)^{1-s_n} \\ &= p^{s_1 + \cdots + s_n}(1-p)^{n-(s_1+\cdots+s_n)}, \end{aligned} \tag{2.29}$$

for any sequence $(s_1, \ldots, s_n)$ of 0's and 1's. ◇

Examples, such as the last one, featuring independent random variables all sharing a common probability mass function, are so important that they get a special name.

Definition. The random variables in a finite or infinite sequence are said to be **independent and identically distributed**, abbreviated **i.i.d.**, if they are independent and if they all have the same probability mass function.

As an example, suppose $\mathcal{S}$ is a finite population, and let $X_1, \ldots, X_n$ be independent, each uniformly distributed on $\mathcal{S}$. It follows from using (2.28) that, every sequence of $(s_1, \ldots, s_n)$ of possible values of $X_1, \ldots, X_n$ is equally likely. *Therefore, $X_1, \ldots, X_n$ is a model, expressed in the language random variables, for a random sample of size n from $\mathcal{S}$, as defined in Section 2.1.* This is a very important point. I.i.d. sequences of random variables generalize random sampling to possibly non-uniform distributions. Indeed, many texts adopt the following terminology.

Definition. Let p_X be a probability mass function on a discrete set S. A **random sample of size n from distribution** p_X is a sequence $X_1, \ldots, X_n$ of independent random variables, each having probability mass function p_X.

To avoid confusion with random sampling as we have originally defined it, that is, with respect to a uniform measure, we shall generally stick to the i.i.d. terminology when p_X is not uniform.

Example 2.2.4. I.i.d. site model for DNA. Let $X_1, \ldots, X_n$ denote the successive bases appearing in a sequence of randomly selected DNA. The i.i.d. site model assumes that $X_1, \ldots, X_n$ are independent, each with the same distribution given by the parameters.

$$p_A = \mathbb{P}(X_i = A), \quad p_C = \mathbb{P}(X_i = C), \quad p_G = \mathbb{P}(X_i = G), \quad p_C = \mathbb{P}(X_i = T).$$

The i.i.d. site model with equal base probabilities imposes the additional assumption that each X_i is uniformly distributed on $\{A, G, C, T\}$, that is, $p_A = p_C = p_G = p_T = 1/4$. This latter model is the same as a random sample of size n from the DNA alphabet. It is the random variable version of probability space model for junk DNA introduced in Exercise 2.2. ◇

2.2.2 Basic discrete random variables

We have already defined discrete Bernoulli and discrete uniform random variables. There are several other important types repeatedly used in probabilistic modeling.

Binomial random variables. Let n be a positive integer and let p be a number in $[0, 1]$. A random variable Y has the **binomial distribution** with parameters n, p if Y takes values in the set of integers $\{0, 1, \ldots, n\}$ and has the probability mass function

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k \text{ in } \{0, 1, \ldots, n\}. \tag{2.30}$$

The binomial distribution is the probability model for the number of successes in n independent trials, where the probability of success in each trial is p. To see this, let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables with $p = \mathbb{P}(X_i = 1)$; then, the event $X_i = 1$ represents a success on trial i, and the sum $\sum_1^n X_i$ counts the number

of successes. If $(s_1, \ldots, s_n)$ is a sequence of 0's and 1's in which 1 appears exactly k times, then we know from (2.29) that

$$\mathbb{P}\Big((X_1, \ldots, X_n) = (s_1, \ldots, s_n)\Big) = p^k(1-p)^{n-k}$$

There are $\binom{n}{k}$ such sequences, because each such sequence corresponds to a choice of k positions among n at which 1 appears. Thus

$$\mathbb{P}\Big(\sum_1^n X_i = k\Big) = \binom{n}{k} p^k (1-p)^k.$$

Geometric random variables. A random variable Y has the geometric distribution with parameter p, where $0 < p < 1$, if the possible values of Y are the positive integers $\{1, 2, \ldots\}$, and

$$\mathbb{P}(Y = k) = (1-p)^{k-1} p \qquad k = 1, 2, \cdots \tag{2.31}$$

Like the binomial random variable, the geometric random variable has an interpretation in terms of success/failure trials or coin tossing. Let $X_1, X_2, \ldots$ be an infinite sequence of independent Bernoulli random variables, each with parameter p. *The geometric random variable with parameter p models the time of the first success in such a sequence.* Indeed, the first success occurs in trial k if and only if $X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1$. By independence, this occurs with probability

$$\mathbb{P}(X_1 = 0) \cdots \mathbb{P}(X_{k-1} = 0) \cdot \mathbb{P}(X_k = 1) = (1-p)^{k-1} p,$$

which is exactly the expression in (2.31).

The Bernoulli trial interpretation of geometric random variables can simplify calculations. Suppose we want to compute $\mathbb{P}(Y > j)$ for a geometric random variable. This probability is the infinite sum $\sum_{k=j+1}^{\infty} \mathbb{P}(Y = k) = \sum_{k=j+1}^{\infty} (1-p)^{k-1} p$. But $Y > j$ is equivalent to the event that there are no successes in the first j independent Bernoulli trials, and, as the trials are independent and the probability of failure is $1 - p$,

$$\mathbb{P}(Y > j) = (1-p)^j, \tag{2.32}$$

without having to do a sum. Of course, the sum is not hard to do using the formula for summing a geometric series; the reader should show directly that the series $\sum_{k=j+1}^{\infty} (1-p)^{k-1} p$ equals $(1-p)^j$.

Geometric random variables have an interesting property, called the *memoryless property*, which follows easily from (2.32). If X is geometric with parameter p,

$$\mathbb{P}\left(X > k + j \middle| X > k\right) = \frac{\mathbb{P}(X > k+j)}{\mathbb{P}(X > k)} = \frac{(1-p)^{k+j}}{(1-p)^j} = (1-p)^j = \mathbb{P}(X > j) \tag{2.33}$$

To understand what this says, imagine repeatedly playing independent games, each of which you win with probability p and lose with probability $1-p$. Let X be the first trial which you win; it is a geometric random variable. Now suppose you have played k times without success ($X > k$), and you want to know the conditional probability of waiting at least j additional trials before you win. Property (2.33) says that the X has no memory of the record of losses; the conditional probability of waiting an additional j trials for a success, given that you have lost the first k trials, is the same the probability of waiting for at least j trials in a game that starts from scratch. This may sound odd at first, but it is an immediate consequence of the independence of the plays.

Remark. Some authors define the geometric random variable to take values in the natural numbers $0, 1, 2, \ldots$ with probability mass function $\mathbb{P}(X = k) = (1-p)^k p$, $k \geq 0$.

Poisson random variables. A random variable Z has the Poisson random distribution with parameter $\lambda > 0$, if the possible values of Z are the natural numbers $0, 1, 2, \ldots$ and

$$\mathbb{P}\,(Z{=}k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots \tag{2.34}$$

Poisson random variables arise often in applications as limits of binomials, when the number of trials is large but the probability of success per trial is small. This will be explained in Chapter 5.

Multinomial distributions. We start with an example. A box contains marbles of three colors, red, green, and blue. Let p_1 be the probability of drawing a red, p_2 the probability of drawing a green, and p_3 the probability of drawing a blue. Of course, $p_1 + p_2 + p_3 = 1$. Sample the box n times with replacement, assuming all samples are independent of one another, and let Y_1 be the number of red marbles, Y_2 the number of greens, and Y_3 the number of blues in the sample. The random variables Y_1, Y_2, Y_3 each have a binomial distribution, but they are not independent—indeed, they must satisfy $Y_1 + Y_2 + Y_3 = n$—and so we cannot use (2.28) to compute their joint distribution. However the individual draws are independent. Let the result X_i of draw i be r, g, or b, according to the color of the marble drawn. If $(s_1, \ldots, s_n)$ is a specific sequence consisting of letters from the set $\{r, b, y\}$, and if this sequence contains k_1 letter r's, k_2 letter g's, and k_3 letter b's,

$$\mathbb{P}\Big((X_1, \ldots, X_n){=}((s_1, \ldots, s_n)\Big) = \mathbb{P}(X_1{=}s_1) \cdots \mathbb{P}(X_n{=}s_n) = p_1^{k_1} p_2^{k_2} p_3^{k_3}. \tag{2.35}$$

On the other hand, there are a total of $\dfrac{n!}{k_1!k_2!k_3!}$ different sequences of sequences of length n with k_1 red, k_2 green, and k_3 blue marbles. Thus,

$$\mathbb{P}\,(Y_1{=}k_1, Y_2{=}k_2, Y_3{=}k_3) = \frac{n!}{k_1!k_2!k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3}, \tag{2.36}$$

for any non-negative integers k_1, k_2, k_3 such that $k_1 + k_2 + k_3 = n$.

The general multinomial distribution is defined by a generalization of formula (2.36). To state it, recall the general notation,

$$\binom{n}{k_1 \cdots k_r} \triangleq \frac{n!}{k_1! \cdots k_r!}.$$

Fix two positive integers n and r with $0 < r < n$. Suppose that for each index i, $1 \leq i \leq r$, a probability p_i is given satisfying $0 < p_i < 1$, and assume also that $p_1 + \cdots + p_n = 1$. The random vector $Z = (Y_1, \cdots, Y_r)$ is said to have the multinomial distribution with parameters $(n, r, p_1, \ldots, p_r)$ if

$$\mathbb{P}\left(Y_1 = k_1, \ldots, Y_r = k_r\right) = \binom{n}{k_1 \cdots k_r} p_1^{k_1} \cdots p_r^{k_r}, \tag{2.37}$$

for any sequence of non-negative integers $k_1, \ldots, k_r$ such that $k_1 + \cdots + k_r = n$.

The interpretation of the multinomial distribution is just a generalization of the experiment with three marbles. Suppose a random experiment with r possible outcomes $1, \ldots, r$ is repeated independently n times. If, for each i, $1 \leq i \leq r$, p_i is the probability that i occurs and Y_i equals to the number of times outcome i occurs, then $(Y_1, \ldots, Y_r)$ has the multinomial distribution with parameters $(n, r, p_1, \ldots, p_r)$.

2.2.3 Continuous random variables.

A function f defined on the real numbers is called a *probability density function* if $f(x) \geq 0$ for all x, and

$$\int_{-\infty}^{\infty} f(x)\, dx = \mathbb{P}(X < \infty) = 1. \tag{2.38}$$

Given such an f, we say that X is a continuous random variable with probability density f if

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\, dx, \qquad \text{for any } a \leq b. \tag{2.39}$$

If X is a continuous random variable, we often write f_X to denote its density. *Modeling a continuous random variable means specifying its probability density function.*

Using principle (2.4) about limits of increasing events, we can extend (2.39) to the case in which either of a or b is infinite. Thus,

$$\mathbb{P}(X \leq b) = \lim_{a \downarrow -\infty} \mathbb{P}(a \leq X \leq b) = \lim_{a \downarrow -\infty} \int_a^b f_X(x)\, dx = \int_{-\infty}^b f_X(x)\, dx \tag{2.40}$$

Similarly,

$$\mathbb{P}(X \geq a) = \int_a^{\infty} f_X(x)\, dx \tag{2.41}$$

Taking $b \to \infty$ in (2.40) yields $\mathbb{P}(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)\,dx = 1$, as it should be, and this is the reason for imposing condition (2.38) in the definition of a probability density function.

The range of a continuous random variable is truly an uncountable continuum of values. Indeed, if b is any single point, (2.39) implies $\mathbb{P}(X=b) = \int_b^b f_X(x)\,dx = 0$. It follows that if $S = \{s_1, s_2, s_3, \dots\}$ is any discrete subset of real numbers,

$$\mathbb{P}(X \in S) = \sum_{i=1}^{\infty} \mathbb{P}(X=s_i) = 0$$

Thus the range of a continuous random variable cannot be reduced to any discrete set.

Because $\mathbb{P}(X=c) = 0$ for any c if X is a continuous random variable, $\mathbb{P}(a < X \leq b)$, $\mathbb{P}(a \leq X < b)$, and $\mathbb{P}(a < X < b)$ are all equal to $\mathbb{P}(a \leq X \leq b)$ and hence all these probabilities are described by the basic formula (2.38).

Definite integrals of non-negative functions compute areas under curves. Thus formula (2.39) says that $\mathbb{P}(a < X \leq b)$ is the area of the region bounded by the graph, $y = f_X(x)$, of the density function, by the x-axis, and by $x = a$ and $x = b$. This viewpoint is helpful in working with continuous random variables and probability densities. In particular, if dx is interpreted as a 'small' positive number, and f_X is continuous in an interval about x,

$$\mathbb{P}(x \leq X \leq x + dx) \approx f_X(x)\,dx,$$

because the region between $y=f(x)$ and the x-axis from x to $x+dx$ is approximately a rectangle with height $f_X(x)$ and width dx. This is a useful heuristic in applying intuition developed for discrete random variables to the continuous case. Many formulas for discrete random variables translate to formulas for continuous random variables by replacing $p_X(x)$ by $f_X(x)\,dx$ and summation by integration.

So far, we have introduced separate concepts, the probability mass function and the probability density function, to model discrete and continuous random variables. There is a way to include them both in a common framework. For any random variable, discrete or continuous, define its *cumulative (probability) distribution function* (c.d.f.) by

$$F_X(x) \stackrel{\triangle}{=} \mathbb{P}\left(X \leq x\right). \tag{2.42}$$

Knowing F_X, one can in principle compute $\mathbb{P}(X \in U)$ for any set U. For example, the probability that X falls in the interval $(a, b]$ is

$$\mathbb{P}\left(a < X \leq b\right) = \mathbb{P}\left(X \leq b\right) - \mathbb{P}\left(X \leq a\right) = F_X(b) - F_x(a). \tag{2.43}$$

The probability that X falls in the union of two disjoint intervals $(a, b] \cup (c, d]$ is then $\mathbb{P}(X \in (a, b]) + \mathbb{P}(X \in (c, d]) = F_X(b) - F_X(a) + F_x(d) - F_X(a)$, and so on.

Since (a, b) is the union over all integers n of the intervals $(a, b - 1/n]$,

$$\mathbb{P}(a < X < b) = \lim_{n\to\infty} \mathbb{P}(a < X \leq b-1/n) = \lim_{n\to\infty} F_X(b-1/n) - F_X(a) = F(b-) - F(a), \tag{2.44}$$

where $F(b-) = \lim_{x\to b-} F(x)$.

For a discrete random variable taking values in the finite or countable set S, $F_X(b) - F_X(a) = \sum_{y\in S, a<y\leq b} p_X(y)$. In particular, if $s \in S$, and $p_X(s) > 0$, then

$$0 < p_X(s) = \mathbb{P}(X = s) = \mathbb{P}(X \leq s) - \mathbb{P}(X < s) = F_X(s) - F_X(s-).$$

Thus the c.d.f. of X jumps precisely at the points of S and the size of the jump at any s is the probability that $X = s$. In between jumps, F_X is constant. Thus the probability mass function of X can be recovered from F_X.

If X is a continuous random variable, then $F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy$. The fundamental theorem of calculus then implies

$$F_X'(x) = f_X(x) \tag{2.45}$$

at any continuity point x of $f_X(x)$. Thus, we can also recover the density of a continuous random variable from its c.d.f.

It is also possible to define cumulative distribution functions which are combinations of both jumps and differentiable parts, or which are continuous, but admit no probability density. These are rarely encountered in applications.

2.2.4 Basic continuous random variables

The uniform distribution. This is the random variable version of the model described in Section 2.1.3 for randomly and uniformly selecting a point from an interval. Here it is defined for an arbitrary interval (α, β), where $\alpha < \beta$. A random variable, X, is said to be *uniformly distributed* on (α, β) if its density function has the form

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \text{if } \alpha < x < \beta; \\ 0, & \text{otherwise.} \end{cases}$$

If this is the case, $\mathbb{P}\big((X \in (\alpha, \beta)\big) = 1$ and X will have no preference as to its location in (α, β). Indeed, if $\alpha \leq a < b \leq \beta$,

$$\mathbb{P}(a < X \leq b) = \int_a^b \frac{1}{\beta - \alpha}\, dy = \frac{b-a}{\beta-\alpha},$$

and this answer depends only on the length of (a, b), not its position within (α, β).

The exponential distribution. The exponential distribution is a popular model for waiting times between randomly occurring events. It is the continuous analogue

of the geometric distribution. A random variable is said to have the *exponential distribution with parameter λ, where $\lambda > 0$*, if its density is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Since the density function is zero for $x \leq 0$, an exponentially distributed random variable can only take positive values. Thus, if X is exponentially distributed with parameter λ, and if $0 \leq a < b$,

$$\mathbb{P}(a < X \leq b) = \int_a^b \lambda e^{-\lambda x}\, dx = e^{-\lambda a} - e^{-\lambda b}.$$

In particular, if $a \geq 0$,

$$\mathbb{P}(X > a) = \int_a^\infty \lambda e^{-\lambda x}\, dx = e^{-\lambda a}. \tag{2.46}$$

Conversely, suppose that $\mathbb{P}(X > a) = e^{-\lambda a}$ for $a > 0$. Then $F_X(a) = 1 - \mathbb{P}(X > a) = 1 - e^{-\lambda a}$. By differentiating both sides and applying (2.45), $f_X(a) = F_X'(a) = \lambda e^{-\lambda a}$ for $a > 0$, and so X must be exponential with parameter λ. Thus, (2.46) is an equivalent characterization of exponential random variables.

Like geometric random variables, exponential random variables have a memoryless property. Indeed, by (2.46),

$$\mathbb{P}\left(X > t + s \middle| X > s\right) = \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda t}. \tag{2.47}$$

That is, the probability of waiting an additional t units of time, given that you have been waiting s units already, is the same as the probability of waiting t units of time starting from time zero.

The normal distribution. A random variable Z is said to be **normally distributed** or **Gaussian**, if it has a probability density of the form

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty, \tag{2.48}$$

Here $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The parameters μ and σ^2 are respectively the mean and variance of the random variable with density $\phi(x; \mu, \sigma^2)$; (mean and variance are reviewed in section 2.3). The factor of $\sigma\sqrt{2\pi}$ in the denominator of the density insures that $\int_{-\infty}^{\infty} \phi(y; \mu, \sigma^2)\, dy = 1$, as required for a probability density function.

Often, we use the shorthand notation, $X \sim N(\mu, \sigma^2)$, to identify X as a normal random variable with mean m and variance σ^2.

If $\mu = 0$ and $\sigma^2 = 1$, then Z is said to be a *standard* normal random variable. A conventional notation for the density of a standard normal r.v.. is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

and a conventional notation for its associated cumulative distribution function is

$$\Phi(z) \stackrel{\triangle}{=} \mathbb{P}(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

Unlike the uniform and exponential distributions, $\Phi(z)$ admits no closed form expression in terms of elementary transcendental and algebraic functions. Therefore, to compute probabilities of the form $\mathbb{P}(a < Z < b) = \Phi(b) - \Phi(a)$, where Z is standard normal, requires using either tables or a calculator or computer with a built in normal distribution function.

The importance of the normal distribution function stems from the Central Limit Theorem, which is stated below in Section 2.4.

Normal random variables have a very important scaling property. If Z is a standard normal r.v. then $\sigma Z + \mu \sim N(\mu, \sigma^2)$. Conversely, if $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma$ is standard normal. A proof of this fact follows shortly. First we show how it can be used to calculate probabilities for any normal random variable from tables for the standard normal. Suppose, for example, that X is normal with mean 1 and variance 4, and we want to know $\mathbb{P}(X \leq 2.5)$. Towards this end, define $Z = (X-\mu)/\sigma = (X-1)/2$. Since the event $X \leq 2.5$ is the same as $Z \leq (2.5-1)/2$, or, equivalently, $Z \leq .75$, and since Z is standard normal, $\mathbb{P}(X \leq 2.5) = \mathbb{P}(Z \leq .75) = \Phi(.75)$. Tables for Φ show that $\Phi(.75) = 0.7734$.

The general case is a simple extension of this argument. Let X be normal with mean μ and variance σ^2. For any $a < b$,

$$a < X < b \quad \text{if and only if} \quad \frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}.$$

But $(X - \mu)/\sigma$ is a standard normal r.v. Hence

$$\mathbb{P}\left(a < X < b\right) = \Phi\left((b-\mu)/\sigma\right) - \Phi\left((a-\mu)/\sigma\right).$$

To prove $Z = (X - \mu)/\sigma$ is standard normal if $X \sim N(\mu, \sigma^2)$, it is necessary to show that the density of Z is $(2\pi)^{-1} e^{-x^2/2}$. But

$$\begin{aligned} f_Z(x) &= F_Z'(x) = \frac{d}{dx}\,\mathbb{P}\big(Z \leq x\big) = \frac{d}{dx}\,\mathbb{P}\Big(\frac{X-\mu}{\sigma} \leq x\Big) \\ &= \frac{d}{dx} F_X(\mu + \sigma\, x) = \sigma\, F_X'(\mu + \sigma\, x). \end{aligned}$$

The last step of this calculation used the chain rule. Now, since $X \sim N(\mu, \sigma^2)$, $F_X'(y) = f_X(y) = (2\pi\sigma)^{-1} e^{-(x-\mu)^2/\sigma^2}$. It follows that

$$f_Z(x) = \sigma\,(2\pi\sigma)^{-1} e^{-(\mu+\sigma x-\mu)^2/2\sigma^2} = (2\pi)^{-1} e^{-x^2/2},$$

as required.

The Gamma distribution. A random variable is said to have the **gamma distribution** with parameters $\lambda > 0$ and $r > 0$ if its density is

$$f(x) = \begin{cases} \Gamma^{-1}(r)\lambda^r x^{r-1} e^{-\lambda x}, & \text{if } x > 0; \\ 0, & \text{otherwise} \end{cases}$$

where $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx$. It can be shown by repeated integration by parts that $\Gamma(n) = (n-1)!$ for positive integers n. Exponential random variables are gamma random variables with $r = 1$.

2.2.5 Joint density functions

The notion of continuous random variable has a generalization to the multivariable case.

Definition. Random variables X, Y are jointly continuous if there is a non-negative function $f(x, y)$, called the joint probability density of (X, Y), such that

$$\mathbb{P}\left(a_1 < X \leq b_1, a_2 < Y \leq b_2\right) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)\, dydx, \tag{2.49}$$

for any $a_1 < b_1$, and $a_2 < b_2$.

If (X, Y) have joint density f, then in fact for a region U in the (x, y)-plane,

$$\mathbb{P}\left((X, Y) \in U\right) = \int\!\!\int_U f(x, y)\, dxdy. \tag{2.50}$$

In a rigorous treatment of probability this rule is derived as a consequence of (2.49). Here we just state as a fact that it is valid for any region U such that the integral is well-defined; this requires some technical restrictions on U, which are not necessary to worry about for applications.

Example 2.2.5. Let (X, Y) have joint density

$$f(x, y) = \begin{cases} \frac{1}{2}, & \text{if } 0 < x < 2 \text{ and } 0 < y < 1; \\ 0, & \text{otherwise} \end{cases}$$

The density is zero except in the rectangle $(0, 2) \times (0, 1) = \{(x, y) : 0 < x < 2, 0 < y < 1\}$, so the probability that (X, Y) falls outside this rectangle is 0.

Let U be the subset of $(0, 2) \times (0, 1)$ for which $y > x$, as in Figure 2.1.

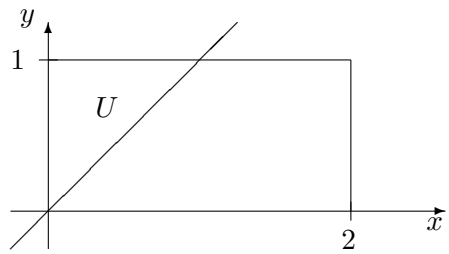

Figure 2.1

The area of U is $1/2$. Since the f has the constant value $1/2$ over U, the double integral of f over U is the $1/2 \times \text{area}(U) = 1/4$. Thus

$$\mathbb{P}(Y > X) = \int\int_U f(x,y)\,dxdy = \frac{1}{4}. \qquad \diamond$$

The definition of joint continuity extends beyond the case of just two random variables using multiple integrals of higher order; $X_1, \ldots, X_n$ are jointly continuous with joint density f if

$$\mathbb{P}\left((X_1, \ldots, X_n) \in U\right) = \int\cdots\int_U f(x_1, \ldots, x_n)\,dx_1 \cdots dx_n.$$

Theorem 2 characterizing independence of discrete random variables generalizes to the continuous case.

Theorem 3 *The jointly continuous random variables $X_1, \cdots, X_n$ are independent if and only if their joint density function f factors as*

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) \tag{2.51}$$

The density function of Example 2.2.5 is equal to $f_1(x)f_2(y)$, where $f_1(x) = 1/2$ on $(0,2)$ and 0 elsewhere, and $f_2(y) = 1$ on $(0,1)$ and 0 elsewhere. The function f_1 is the density of a random variables uniformly distributed on $(0,2)$ and f_2 the density of a random variable uniformly distributed on $(0,1)$. Hence X and Y in Example 2.2.5 are independent random variables, X being uniformly distributed on $(0,2)$ and Y on $(0,1)$.

2.2.6 Exercises.

Exercises 2.8—2.11 deal with successive generations of a population of 15 individuals. Each individual of generation $t+1$ is the child of a parent from generation t. Reproduction is asexual; essentially, a parent reproduces a copy of itself. These exercises anticipate the Wright-Fisher model introduced in Chapter 4. Each of the 15 children of generation $t+1$ is the child of a randomly selected parent of generation t, and each of the 15 parents is selected independently. The solutions to these Exercises make use of the basic discrete random variables introduced in Section 2.2.

2.8. Suppose in the first generation there are 10 individuals of type A and 5 of type a. Each child inherits the type of its parents. Thus the second generation is effectively a random sample of size 15 from the first generation.

a) Let X_i equal one if child i is type A and let it equal zero otherwise, for $1 \leq i \leq 15$. What is the probability that X_i equals one?

b) Let X be the total number of type A individuals in the second generation. What is the probability distribution of X?

2.9. Again the first generation has 10 type A and 5 type a individuals. This time however, mutation can occur. An A parent gives birth to an A child with probability .8 and to an a child with probability .2; an a type gives birth to an a child with probability .9 and to an A child with probability 0.1. Let X_i and X be defined as in Exercise 2.1.

a) Compute the probability X_i equals one. (Condition on the type of the parent.)

b) Determine the probability distribution of X.

2.10. a) Find the probability that a given individual leaves no progeny in the next generation.

b) Find the probability that an individual in the first generation leaves no descendants in the sixth generation.

2.11 Consider a population with 8 individuals of type A, 4 of type B and 3 of type C. Produce a new population of 15 individuals by random selection as in Exercise 2.8; again, each child inherits the type of its parent. Let X_A, X_B, and X_C denote the numbers of A, B, and C in the new population. What is the probability that $X_A = 5$, $X_B = 6$, and $X_C = 4$?

2.12. Let $X \sim N(3, 16)$. Find the probability that X is greater than or equal to 8.

2.13. Let X be an exponential random variable with parameter λ. For $k = 1, 2, 3, \ldots$, let $Y = k$ if $k - 1 < X < k$. Show that Y is a geometric random variable with parameter $p = 1 - e^{-\lambda}$.

2.3 Expectation

In this section all random variables are real-valued.

2.3.1 Expectations; basic definitions

Definition. Let X be a discrete random variable with values in S, and let p_X denote its probability mass function. The expected value of X, also called the mean of X, is

$$\mathbb{E}[X] \stackrel{\triangle}{=} \sum_{s \in S} s p_X(s), \tag{2.52}$$

if the sum exists. (When S is an infinite set, the sum defining $\mathbb{E}[X]$ is an infinite series and may not converge.) We shall often use μ_X to denote $E[X]$.

The expected value of X is an average of the possible values X can take on, where each possible values s is weighted by the probability that $X = s$. Thus, the expected value represents the value we expect X to have on average. This is made more precise in the discussion of the law of large numbers.

The expected value of a continuous random variable is again a weighted average of its possible values. Following the heuristic discussed in Section 2.2, to arrive at the appropriate definition we replace the probability mass function p_X in formula (2.52) by $f_X(x)\,dx$ and replace the sum by an integral.

Definition Let X be a continuous random variable. Then,

$$E[X] \stackrel{\triangle}{=} \int_{-\infty}^{\infty} x f_X(x)\,dx, \tag{2.53}$$

if the integral exists.

Example 2.3.1. Let X be a Bernoulli random variable X with parameter p. Since $p_X(0) = 1 - p$ and $p_X(1) = p$, its expectation is

$$\mathbb{E}[X] = 0 \cdot (1 - p) + 1 \cdot p = p. \tag{2.54}$$

This is intuitive; if you win a dollar for heads and win nothing for tails (a nice game to play!), you expect to earn an average of p dollars per play. ◇

Example 2.3.2. Probabilities as expectations. Let U be an event. The indicator of U is the random variable,

$$\mathbf{1}_U = \begin{cases} 1, & \text{if } U \text{ occurs;} \\ 0, & \text{otherwise} \end{cases}$$

This is a Bernoulli random variable with $p = \mathbb{P}(\mathbf{1}_U = 1) = \mathbb{P}(U)$. Hence

$$E[\mathbf{1}_U] = \mathbb{P}(U). \tag{2.55}$$

This trick of representing the probability of U by the expectation of its indicator is used frequently. ◇

Example 2.3.3. Let X be uniformly distributed on the interval $(0, 1)$. Since the density of X equals one on the interval $(0, 1)$ and 0 elsewhere,

$$E[X] = \int_{-\infty}^{\infty} f_X(x)\, dx = \int_0^1 x\, dx = \frac{1}{2}.$$

The answer makes perfect sense: X shows no preference as to where it lies in $(0, 1)$, and so its average value should be $1/2$. ◇

Example 2.3.4. Let Y is exponential with parameter λ. An application of integration by parts to compute the anti-derivative shows,

$$E[Y] = \int_0^{\infty} x\lambda e^{\lambda x}\, dx = -\left(\frac{1}{\lambda} + x\right) e^{-\lambda x}\Big|_0^{\infty} = \frac{1}{\lambda}.$$

This gives a nice physical interpretation of the meaning of the parameter λ; if Y is a waiting time, λ is the inverse of the expected time to wait. ◇

Example 2.3.5. If $X \sim N(\mu, \sigma^2)$, then $E[X] = \mu$. This is easy to see if $\mu = 0$ by the symmetry of the density function. If $\mu \neq 0$, then $X - \mu \sim N(0, \sigma^2)$, so $E[X - \mu] = 0$, again showing $E[X] = \mu$. ◇

The expectations of the other basic discrete and continuous random variables are listed in a table appearing in Section 2.3.4.

2.3.2 Elementary theory of expectations.

In probability, one repeatedly faces the following problem: given a random variable X with known distribution and a function g, compute $E[g(X)]$. To do this directly from the definition would require first calculating the probability mass function or probability density of $g(X)$, whichever is appropriate, and then applying (2.52) or (2.53). But there is an easier way, sometimes called the **law of the unconscious statistician**, presumably because it allows one to compute without thinking too hard.

Theorem 4 *a) If X is discrete and $\mathbb{E}[g(X)]$ exists,*

$$\mathbb{E}[g(X)] = \sum_{s \in S} g(s) p_X(s). \tag{2.56}$$

b) If X is continuous and $\mathbb{E}[g(X)]$ exists,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \tag{2.57}$$

c) More generally,

$$\mathbb{E}[h(X_1, \ldots, X_n)] = \sum_{s_1 \in S} \cdots \sum_{s_n \in S} h(s_1, \ldots, s_n) p_Z(s_1, \ldots, s_n), \tag{2.58}$$

if p_Z is the joint probability mass function of $(X_1, \ldots, X_n)$, and

$$\mathbb{E}[h(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) f(x_1, \cdots, x_n)\, dx_1 \ldots dx_n, \tag{2.59}$$

if f is the joint density function of $X_1 \ldots, X_n$.

The law of the unconscious statistician has several important consequences that are used repeatedly.

Theorem 5 *(Linearity of expectation.) Assuming all expectations are defined*

$$\mathbb{E}\left[c_1 X_1 + \cdots + c_n X_n\right] = c_1 \mathbb{E}[X_1] + \cdots + c_n \mathbb{E}[X_n]. \tag{2.60}$$

Linearity of expectations is extremely useful. Often, a complicated random variable can be written as a sum of simpler random variables with easy-to-compute expectations. Then formula (2.60) can be used to compute its expectation. This works even if we cannot compute an explicit formula for the density or probability mass function of the sum.

Example 2.3.6. The mean of a binomial r.v. Let X be binomial with parameters n and p. According to the definition in (2.52), $E[X] = \sum_0^n k \binom{n}{k} p^k (1-p)^{n-k}$, which looks a bit complicated. However, we know that if $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli with parameter p, then $Y_1 + \cdots + Y_n$ is binomial with parameters n and p. Since $E[X_i] = p$ for each i,

$$E[X] = E[Y_1] + \cdots + E[Y_n] = np. \qquad \diamond$$

The linearity property extends to integrals of of random variables. For each r in an interval $[a, b]$, let $Y(r)$ be a random variable, and consider $X = \int_a^b Y(r)\, dr$. Then, under some technical conditions,

$$E[X] = \int_a^b E[Y(r)]\, dr. \tag{2.61}$$

The endpoints a or b or both could be infinite in this identity.

A rigorous statement of the conditions under which this holds (it can fail!) is beyond the level of this text. But as a guideline, if you can make sense of the integral defining Z and if

$$\int_a^b \Big|\, E[Y(r)] \,\Big|\, dr < \infty,$$

then (2.61) works. It is not hard to understand why (2.61) should be true, at least from an intuitive viewpoint. An integral is a limit of sums, and because the expectation of a sum is a sum of expectations, the linearity property of expectation extends from sums to integrals.

Example 2.3.7. As an application of the interchange of integration and expectation, we derive a useful alternative formula for the expectation of a positive random variable. Let Z be a random variable satisfying $\mathbb{P}(Z \geq 0) = 1$. Recall that if U is an event, the $\mathbf{1}_U$ is the random variable which equals 1 if U occurs and 0, otherwise, and $E[\mathbf{1}_U] = \mathbb{P}(U)$—see Example 2.3.2. Thus

$$\mathbf{1}_{\{x \leq Z\}} = \begin{cases} 1, & \text{if } x \leq Z; \\ 0, & \text{if } x > Z, \end{cases}$$

and hence $\displaystyle\int_0^\infty \mathbf{1}_{\{x \leq Z\}}\, dx = \int_0^Z 1\, dx = Z$. Taking expectations on both sides,

$$\begin{aligned} E[Z] &= E\Big[\int_0^\infty \mathbf{1}_{\{x \leq Z\}}\, dx\Big] = \int_0^\infty E\big[\mathbf{1}_{\{x \leq Z\}}\big]\, dx \\ &= \int_0^\infty \mathbb{P}(Z \geq x)\, dx. \end{aligned} \tag{2.62}$$

Using the final expectation to represent or to compute $E[Z]$ is often very useful. $\diamond$

The next result, which also follows from the law of the unconscious statistician, is a generalization to expectations of the probability formula $\mathbb{P}(U_1 \cap \cdots \cap U_n) = \mathbb{P}(U_1)\mathbb{P}(U_1)\cdots\mathbb{P}(U_n)$ for independent events.

Theorem 6 *(Products of independent random variables.) Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$E\left[g_1(X_1)g(X_2)\cdots g_n(X_n)\right] = E[g_1(X_1)]\cdots E[g(X_n)] \tag{2.63}$$

whenever $E[g_i(X_i)]$ is defined and finite for each i.

We show why this theorem is true in the case $n = 2$ and X_1 and X_2 are independent and discrete. In this case Theorem 2 says that the joint probability mass

function of (X_1, X_2) is $p_Z(s_1, s_2) = p_{X_1}(s_1)p_{X_2}(s_2)$.. Thus, using formula (2.58)

$$\begin{aligned} E\left[g_1(X_1)g_2(X_2)\right] &= \sum_{s_1\in S}\sum_{s_2\in S} g_1(s_1)g_2(s_2)p_{X_1}(s_1)p_{X_2}(s_2) \\ &= \left[\sum_{s_1} g_1(s_1)p_{X_1}(s_1)\right]\left[\sum_{s_2} g_1(s_2)p_{X_2}(s_2)\right] \\ &= E[g(X_1)]E[g_2(X_2)]. \qquad \diamond \end{aligned}$$

2.3.3 Variance and Covariance

The variance of a random X variable measures the average square distance of X from its mean $\mu_X = E[X]$:

$$\mathrm{Var}(X) \stackrel{\triangle}{=} E\left[(X-\mu_X)^2\right]. \tag{2.64}$$

The size of the variance indicates how closely the outcomes of repeated, independent trials of X cluster around its expected value.

There are several basic identities to keep in mind when working with the variance. First

$$\mathrm{Var}(cX) = E\left[(cX - c\mu_X)^2\right] = c^2 E\left[(X-\mu_X)^2\right] = c^2\mathrm{Var}(X). \tag{2.65}$$

Second, using linearity of expectations,

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - 2E[\mu_X X] + E[\mu_x^2] = E[X^2] - 2\mu_X E[X] + \mu_X^2 \\ &= E[X^2] - \mu_X^2. \end{aligned} \tag{2.66}$$

The last two steps in the derivation use the fact that μ_X is a constant, so that $E[X\mu_X] = \mu_X E[X] = \mu_X^2$ and $E[\mu_X^2] = \mu_X^2$.

Example 2.3.8. If X is Bernoulli with parameter p, $\mathrm{Var}(X) = E[X^2] - \mu_X^2 = \left(0^2 p_X(0) + 1^2 p_X(1)\right) - p^2 = p - p^2 = p(1-p)$. $\diamond$

Example 2.3.9. If $Y \sim N(\mu, \sigma^2)$, then $\mathrm{Var}(Y) = \sigma^2$. This can be derived using moment generating functions—see Section 2.3.4. $\diamond$

The **covariance** of two random variables is:

$$\mathrm{Cov}(X,Y) \stackrel{\triangle}{=} E\left[(X-\mu_X)(Y-\mu_Y)\right]. \tag{2.67}$$

Similarly to (2.66), $\mathrm{Cov}(X,Y) = E[XY] - \mu_X\mu_Y$.

The covariance of X and Y is defined whenever $\mathrm{Var}(X) < \infty$ and $\mathrm{Var}(Y) < \infty$. In this case, the following important inequality is always true:

$$\mathrm{Cov}(X,Y) \le \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}. \tag{2.68}$$

The *correlation* between X and Y is defined to be

$$\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

By (2.68), $\left|\ \mathrm{Cor}(X,Y)\ \right| \leq 1$ for any X and Y.

Correlation is a measure of the statistical interaction between random variable For example if Y tends, on average, to be greater than μ_Y when $X > \mu_X$, then $\mathrm{Cov}(X,Y) > 0$ and hence $\mathrm{Cor}(X,Y) > 0$, whereas if Y tends to be less than μ_Y when $X > \mu_X$, $\mathrm{Cov}(X,Y) < 0$ and $\mathrm{Cor}(X,Y) < 0$.

Two random variables are **uncorrelated** if $\mathrm{Cov}(X,Y) = 0$. It is a very important fact that

$$\text{if } X \text{ and } Y \text{ are independent then they are uncorrelated.} \tag{2.69}$$

This is a consequence of the product formula of Theorem 6 for independent random variables: if X and Y are independent then

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_y)\right] = E\left[X-\mu_X\right]E\left[Y-\mu_Y\right] = 0,$$

because $E[X-\mu_X] = \mu_X - \mu_X = 0$.

There is also an important formula for the variance of a finite sum of random variables. Let $Y = \sum_1^n X_i$. From linearity $E[Y] = \sum_1^n \mu_i$, where μ_i is short hand for μ_{X_i}. Then

$$(Y-\mu_Y)^2 = \left(\sum_1^n X_i - \mu_i\right)^2 = \sum_1^n (X_i-\mu_i)^2 + \sum_{1\leq i,j\leq n, i\neq j} (X_i-\mu_i)(X_j-\mu_j).$$

So taking expectations on both sides, and using the linearity property of expectation and the definitions of variance and covariance,

$$\mathrm{Var}(\sum_1^n X_i) = \sum_1^n \mathrm{Var}(X_i) + \sum_{1\leq i,j\leq n, i\neq j} \mathrm{Cov}(X_i,X_j). \tag{2.70}$$

If $X_1,\ldots,X_n$ are all uncorrelated, which is true if they are independent, it follows that

$$\mathrm{Var}(\sum_1^n X_i) = \sum_1^n \mathrm{Var}(X_i). \tag{2.71}$$

This is a fundamental formula.

Example 2.3.10. Variance of the binomial. Let $X_1,\ldots,X_n$ be i.i.d. Bernoulli random variables, each with probability p of equaling 1. Then we know that $\sum_1^n X_i$ is binomial with parameters n and p. We have also shown that $\mathrm{Var}(X_i) = p(1-p)$ for a Bernoulli random variable. Therefore the variance of a binomial random variable with parameters n and p is

$$\mathrm{Var}(\sum_1^n X_i) = \sum_1^n \mathrm{Var}(X_i) = \sum_1^n p(1-p) = np(1-p). \qquad \diamond \tag{2.72}$$

2.3.4 Chebyshev's inequality and the Law of Large Numbers

Expectations and variances can be used to bound probabilities. The most basic bound is called **Markov's inequality,** which states that if X is a non-negative random variable and $a > 0$, then

$$\mathbb{P}(X \geq a) \leq \frac{E[X]}{a}. \tag{2.73}$$

This is a consequence of two simple facts: first, if Y and Z are two random variables such that $Y \geq Z$ with probability one, then $E[Z] \leq E[Y]$; second, if U is an event $\mathbb{P}(U) = E[\mathbf{1}_U]$, where $\mathbf{1}_U$ is the indicator of U (see Section 2.3.1). Now, if X is a positive random variable and $a > 0$,

$$\mathbf{1}_{\{X \geq a\}} \leq \frac{X}{a},$$

because $X \geq a$, when $\mathbf{1}_{\{X \geq a\}} = 1$. Taking expectations on both side, gives Markov's inequality.

Chebyshev's inequality is a consequence of Markov's inequality. Let Y be a random variable with finite mean μ and variance σ^2. By applying Markov's inequality $(Y - \mu_Y)^2$,

$$\mathbb{P}\left(\left|Y - \mu\right| \geq a\right) = \mathbb{P}\left((Y-\mu)^2 \geq a^2\right) \leq \frac{E\left[(Y-\mu)^2\right]}{a^2}.$$

Since $E[(Y-\mu)^2] = \mathrm{Var}(Y)$,

$$\mathbb{P}\left(\left|Y - \mu\right| \geq a\right) \leq \frac{\mathrm{Var}(Y)}{a^2}. \tag{2.74}$$

which is called Chebyshev's inequality. It gives a quantitative bound on the probability of deviation of Y from its mean, just in terms of its variance.

Let $X_1, X_2, \ldots$ be uncorrelated random variables all having mean μ and variance σ^2. Chebyshev's inequality leads directly to the weak law of large numbers for this case. Let

$$\hat{X}^{(n)} \triangleq \frac{1}{n}\sum_1^n X_i$$

denote the empirical mean of $X_1, \ldots, X_n$. By linearity of expectation, its mean is

$$E[\hat{X}^{(n)}] = \frac{1}{n}\sum_1^n E[X_i] = \frac{1}{n}\sum_1^n \mu = \mu.$$

By (2.65) and (2.71), its variance is

$$\mathrm{Var}(\hat{X}^{(n)}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_1^n X_i\right) = \frac{1}{n^2}\sum_1^n \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Now apply Chebyshev's inequality:

$$\mathbb{P}\left(\left|X^{(n)} - \mu\right| > a\right) \leq \frac{\mathrm{Var}(X^{(n)})}{a^2} = \frac{\sigma^2}{n}. \tag{2.75}$$

By letting $n \to \infty$ in this inequality, we get the following basic result

Theorem 7 *If $X_1, X_2, \ldots$ are uncorrelated random variables with a common mean μ and a common variance*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|X^{(n)} - \mu\right| > a\right) = 0, \tag{2.76}$$

.

This theorem is an example of a weak law of large numbers. It is worthwhile stating its application to binomial random variables separately.

Corollary 1 *For each positive integer n, let $Y^{(n)}$ be a binomial random variable with parameters n and p. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{Y^{(n)}}{n} - p\right| > a\right) = 0, \tag{2.77}$$

.

To derive this result, recall from Example 2.3.6 that the expectation of $Y^{(n)}$ is np and hence the expectation of $Y^{(n)}/n$ is p. Thus from Chebyshev's theorem, from the scaling properties of the variance (see equation (2.65)), and from the formula for the variance of a binomial random variable derived in Example 2.1.10,

$$\mathbb{P}\left(\left|\frac{Y^{(n)}}{n} - p\right| > a\right) \leq \frac{\mathrm{Var}(Y^{(n)}/n)}{a^2} = \frac{\mathrm{Var}(Y^{(n)})}{n^2a^2} = \frac{p(1-p)}{na^2}\ .$$

This tends to zero as $n \to \infty$, proving the claim. This calculation also proves the inequality stated in Section 2.1.5, formula (2.16), in the presentation of the weak law of large numbers for random sampling. In that framework, $f_x^{(n)}$ denoted the empirical frequency,

$$\frac{N_x^{(n)}}{n},$$

where $N^{(n)}$ represents the number of times an individual with label x is drawn in n independent random samples of a population $\mathcal{S}$. But $N^{(n)}$ is a binomial random variable with parameters n and $p = f_x$, where f_x, the frequency of x in the population, is the probability of drawing an individual with label x. By the calculation we have just done,

$$\mathbb{P}\left(\left|f_x^{(n)} - f_x\right| > a\right) \leq \frac{f_x(1-f_x)}{na^2},$$

as stated in (2.16).

2.3.5 The Moment Generating Function.

Definition. The moment generating function of a random variable X is by

$$M_X(t) \stackrel{\triangle}{=} E\left[e^{tX}\right]$$

defined for all real numbers t such that the expectation is finite.

Notice that $M_X(0)$ is always defined, and $M_X(0) = E[e^0] = E[1] = 1$. $M_X(t)$ is called the moment generating function because its derivatives at $t = 0$ can be used to compute the expectations $E[X^n]$ for any positive integer power of X, and these expectations are called the moments of X. In order to compute derivatives at $t = 0$, assumed that $M_X(t)$ is finite for all values t in some interval containing 0. This assumption is enough to guarantee that $M_X(t)$ has derivatives at $t = 0$ of all orders. Then, because $d^n/dt^n(e^{tx}) = x^n e^{tx}$,

$$\begin{aligned}\frac{d^n}{dt^n}M_X(t) &= \frac{d^n}{dt^n}E\left[e^{tX}\right] = E\left[\frac{d^n}{dt^n}e^{tX}\right] \\ &= E\left[X^n e^{tX}\right].\end{aligned}$$

Setting $t = 0$ gives

$$M^{(n)}(0) = E[X^n e^0] = E[X^n], \tag{2.78}$$

where $M_X^{(n)}(t)$ denotes the derivative of order n of $M_X(t)$. This calculation required interchanging expectation and derivation, which requires a justification we do not give, but is valid under the assumption that $M_X(t)$ is finite in an interval about $t = 0$.

Example 2.3.11. Exponential random variables The moment generating function of an exponential random variable with parameter λ is

$$E[e^{tX}] = \int_0^t \lambda e^{tx} e^{-\lambda x}\, dx = \frac{\lambda}{\lambda - t}, \qquad t < \lambda.$$

By repeated differentiation, $\dfrac{d^n}{dt^n}\dfrac{\lambda}{\lambda - t} = \dfrac{\lambda n!}{(\lambda - t)^{n+1}}$. Hence, the n^{th} moment of the exponential is $E[X^n] = n!/\lambda^n$. ◇

Moment generating functions are particularly suited for studying sums of independent random variables because of Theorem 6. Assume $X_1, \ldots X_n$ are independent, and let $Z = X_1 + \cdots + X_n$. The identity $e^{tZ} = e^{tX_1 + \cdots + tX_n} = e^{tX_1} e^{tX_2} \cdots e^{tX_n}$ is elementary. Now apply the product formula of Theorem 6.

$$M_Z(t) = E\left[e^{tX_1} e^{tX_2} \cdots e^{tX_n}\right] = E[e^{tX_1}]E[e^{tX_2}] \cdots E[e^{tX_n}] = M_{X_1}(t) \cdots M_{X_n}(t). \tag{2.79}$$

Thus, the moment generating function of a sum of independent random variables is the product of the moment generating functions of the summands. In particular,

suppose the random variables $X_1, \ldots, X_n$ are i.i.d. Then they all have the same moment generating function $M_X(t)$, and so

$$M_Z(t) = E[e^{tX_1}]E[e^{tX_2}] \cdots E[e^{tX_n}] = M_X^n(t). \tag{2.80}$$

Example 2.3.12. Bernoulli and binomial random variables. The moment generating function of a Bernoulli random variable X with probability p of success is

$$M(t) = e^{t \cdot 0}(1-p) + e^t p = 1 - p + pe^t.$$

Let $Y = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are i.i.d. Bernoulli random variables with probability p of success. Then Y is a binomial random variable with parameters p and n. Using (2.80) and the m.g.f. of the binomial r.v. , the m.g.f. of Y is

$$M_Y(t) = \left(1 - p + pe^t\right)^n. \tag{2.81}$$

Using $M_Y(t)$ and formula (2.78) for computing moments, it is not hard to recover the formulas we have already derived for the mean and variance of the binomial random variable:

$$\begin{aligned} E[X] &= M'(0) = n\left(1 - p + pe^t\right)^{n-1} pe^t \,\Big|_{t=0} = np, \quad \text{and} && (2.82) \\ \mathrm{Var}(X) &= E[X^2] - \mu_X^2 = M''(0) - (np)^2 \\ &= n(n-1)\left(1 - p + pe^t\right)^{n-2} (pe^t)^2 + n\left(1 - p + pe^t\right)^{n-1} pe^t - (np)^2 \,\Big|_{t=0} . \\ &= np(1-p) \qquad \diamond && (2.83) \end{aligned}$$

Moment generating functions have another very important property: they characterize the cumulative probability distribution functions of random variables.

Theorem 8 *Let X and Y be random variables and assume that there is an interval (a, b), where $a < b$, such that $M_X(t)$ and $M_Y(t)$ are finite and equal for all t in (a, b). Then $F_X(x) = F_Y(x)$ for all x, where F_X and F_Y are the respective cumulative distribution functions of X and Y. In particular, if X and Y are discrete, they have the same probability mass function, and if they are continuous, they have the same probability density function.*

Example 2.3.13. Sums of independent normal r.v.'s. The moment generating function of a normal random variable with mean μ and variance σ^2 is $M(t) = e^{\mu t + \sigma^2 t^2/2}$. We will not demonstrate this here, only apply it as follows. Let X_1 be normal with mean μ_1 and variance σ_1^2 and let X_2 be normal with mean μ_2 and variance σ_2^2. Suppose in addition that they are independent. Then, according to Theorem 6, the moment generating function of $X_1 + X_2$ is

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = e^{\mu_1 t + \sigma_1^2 t^2/2} e^{\mu_1 t + \sigma_1^2 t^2/2} = e^{(\mu_1+\mu_2)t + (\sigma_1^2+\sigma_2^2)t^2/2}.$$

However, the last expression is the moment generating function of a normal random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Thus, by Theorem 8, $X_1 + X_2$ must be normal with this mean and variance. $\diamond$

The last example illustrates a special case of a very important theorem.

Theorem 9 *If $X_1, \ldots, X_n$ are independent normal random variables, then $\sum_1^n X_i$ is normal with mean $\sum_1^n \mu_{X_i}$ and variance $\sum_1^n Var(X_i)$.*

The following table summarizes the means, variances, and moment generating functions of the basic random variables.

Table of means, variances, and moment generating functions.

Distribution	Mean	Variance	M.g.f.
Bernoulli 0-1(p)	p	$p(1-p)$	$1-p+pe^t$
Binomial(n,p)	np	$np(1-p)$	$\left(1-p+pe^t\right)^n$
Poisson(λ)	λ	λ	$e^{-\lambda+\lambda e^t}$
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Uniform(α,β)	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$	$\frac{e^{t\beta}-e^{t\alpha}}{t(\beta-\alpha)}$
Exponential(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
Normal(μ,σ^2)	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$
Gamma(λ,r)	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	$\frac{\lambda^r}{(\lambda-t)^r}$

2.3.6 Conditional Distributions and Conditional Expectations

Let X and Y be two discrete random variables. The **conditional probability mass function** of X given $Y = y$ is the function

$$p_{X|Y}(x|y) \stackrel{\triangle}{=} \mathbb{P}(X=x\big|Y=y), \qquad \text{where } x \text{ ranges over the possible values of } X.$$

The conditional expectation of X given $Y = y$ is

$$E[X\big|Y=y] \stackrel{\triangle}{=} \sum_x x p_{X|Y}(x|y).$$

The concepts are generalized to continuous random variables by replacing probability mass functions by probability densities. If X and Y are jointly continuous random variables with joint density $f_{(X,Y)}$, then the **conditional density of X given $Y = y$** is

$$f_{X|Y}(x|y) \stackrel{\triangle}{=} \begin{cases} \frac{f_{(X,Y)}(x,y)}{f_Y(y)}, & \text{if } f_Y(y) > 0; \\ 0, & \text{if } f_Y(y) = 0; \end{cases}$$

here $f_Y(y)$ is the density of Y. The conditional expectation of X given $Y = y$ is

$$E[X\big|Y=y] \stackrel{\triangle}{=} \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx.$$

The law of the unconscious statistician—see Theorem 4—holds for conditional expectations. In the discrete and continuous cases, respectively,

$$E[g(X)\big|Y=y] = \sum_x g(x) p_{X|Y}(x|y) \quad \text{and} \quad E[g(X)\big|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\, dx.$$

The rule of total probabilities generalizes to expectations and provides a very useful tool for computation.

Theorem 10 *For discrete and continuous random variables, respectively,*

$$E[g(X)] = \sum_y E[g(X)\big|Y=y] p_Y(y) \tag{2.84}$$

$$E[g(X)] = \int_{-\infty}^{\infty} E[g(X)\big|Y=y] f_Y(y)\, dy. \tag{2.85}$$

This result is particular useful if a problem is defined directly in terms of conditional distributions.

Example 2.3.14. Assume that X and Y are such that Y is exponential with parameter λ and for every $y > 0$, the conditional distribution of X given $Y = y$ is that of a random variable uniformly distributed on $(0, y)$. This is another way of saying that $f_{(X,Y)}(x|y) = 1/y$ if $0 < x < y$, and is 0 otherwise. Find $E[X]$.

The mean of a random variable uniformly distributed on $(0, y)$ is $y/2$. Hence, we find easily that $E[X|Y=y] = y/2$. Thus, using (2.85),

$$E[X] = \int_0^{\infty} E[X\big|Y=y] \lambda e^{-\lambda y}\, dy = \int_0^{\infty} \frac{y}{2} e^{-\lambda y}\, dy = \frac{1}{2\lambda}.$$

Notice that the last integral is just one-half the expected value of Y. $\diamond$

We will explain formula (2.85); formula (2.84) follows from a similar, but even easier argument. Using first the definition of the conditional expectation and then the definition of conditional density,

$$\begin{aligned}\int_{-\infty}^{\infty} E[g(X)\big|Y=y]f_Y(y)\,dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)\,dx\,f_Y(y)\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\frac{f(x,y)}{f_Y(y)}\,dx\,f_Y(y)\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)f(x,y)\,dx\,dy \\ &= E[g(X)].\end{aligned}$$

There is a useful an intuitive interpretation of this formula. Notice that the conditional expectation $E[g(X)|Y=y]$ defines a function of y. Let $E[g(X)|Y]$ denote this function evaluated at Y, so it is a random variable. This is subtle, and it helps to consider an example. For instance in Example 2.3.14, X and Y were random variables such that $E[X|Y=y] = y/2$. Thus, in this case, $E[X|Y] = Y/2$. This is a random variable, whose value is the conditional expectation of X given whatever value Y takes on. In general, $E[g(X)|Y] = h(Y)$ where $h(y)$ is defined as $h(y) = E[X|Y=y]$. We claim then that both (2.85) and (2.84) are equivalent to

$$E[g(X)] = E\big[E[X|Y]\big]. \tag{2.86}$$

We show this for continuous random variables. Letting $h(y) = E[X|Y=y]$,

$$E\big[E[X|Y]\big] = E[h(Y)] = \int_{-\infty}^{\infty} h(y)f_Y(y)\,dy = \int_{-\infty}^{\infty} E[X|Y=y]f_Y(y)\,dy = E[X],$$

where the last equality is a consequence of (2.85).

It is also possible to apply conditioning the the calculation of variance. The *conditional variance* of X given $Y = y$ is defined as

$$\mathrm{Var}(X|Y=y) = E[X^2|Y=y] - (E[X|Y=y])^2.$$

which just generalizes the identity, $\mathrm{Var}(X) = E[X^2] - (E[X])^2$ from expectation to conditional expectation. In line with the notation described in the previous paragraph,

$$\mathrm{Var}(X|Y) = E\big[X^2|Y\big] - \big(E[X|Y]\big)^2.$$

Theorem 11

$$Var(X) = Var\Big(E[X|Y]\Big) + E\Big[\,Var(X|Y)\Big] \tag{2.87}$$

This formula decomposes $\mathrm{Var}(X)$ into the sum of two terms, the first is the variance of $E[X|Y]$ about its mean, $E[E[X|Y]] = E[X]$, and the second is the mean value of $\mathrm{Var}(Y|X)$. It is not hard to prove this formula. By definition,

$$\mathrm{Var}\Big(E[X|Y]\Big) = E\Big[\big(E[X|Y]\big)^2\Big] - \Big(E\big[E[X|Y]\big]\Big)^2 = E\Big[\big(E[X|Y]\big)^2\Big] - \big(E[X]\big)^2,$$

and

$$E\Big[\mathrm{Var}(X|Y)\Big] = E\Big[E\big[X^2|Y\big]\Big] - E\Big[\big(E[X|Y]\big)^2\Big] = E[X^2] - E\Big[\big(E[X|Y]\big)^2\Big].$$

When added, the terms $E\Big[\big(E[X|Y]\big)^2\Big]$ cancel and what is left is

$$\mathrm{Var}\Big(E[X|Y]\Big) + E\Big[\mathrm{Var}(X|Y)\Big] = E[X^2] - \big(E[X]\big)^2 = \mathrm{Var}(X).$$

2.4 The Central Limit Theorem

The Central Limit Theorem explains the importance of the normal distribution. Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. Our goal is to understand the probability distribution of the sum $\sum_1^n X_i$ for large n. To do this we will scale the sum by additive and multiplicative factors to create a random variable with mean 0 and variance 1. We know that the sum $\sum_1^n X_i$ has mean $n\mu$, and so

$$\sum_1^n X_i - n\mu$$

has a mean of zero. We also know that $\mathrm{Var}\big(\sum_1^n X_i\big) = n\sigma^2$. Therefore, if we define

$$Z^{(n)} \triangleq \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}},$$

we see that $E[Z^{(n)}] = 0$ and

$$\begin{aligned}\mathrm{Var}(Z^{(n)}) &= E\left[\left(\frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}}\right)^2\right] = \frac{1}{n\sigma^2}E\left[\left(\sum_1^n X_i - n\mu\right)^2\right] \\ &= \frac{1}{n\sigma^2}\mathrm{Var}\left(\sum_1^n X_i\right) = \frac{1}{n\sigma^2}n\sigma^2 = 1.\end{aligned}$$

The Central Limit Theorem states that $Z^{(n)}$ looks more and more like a standard normal r.v. as $n \to \infty$. Recall the notation $\Phi(x)$ for the cumulative distribution function of the standard normal, as defined on page 25.

Theorem 12 The Central Limit Theorem. *Suppose $X_1, X_2, \ldots$ are independent, identically distributed random variables with common mean μ and variance σ^2. Then for all $-\infty \le a < b \le \infty$,*

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} \le b\right) = \int_a^b e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = \Phi(b) - \Phi(a). \tag{2.88}$$

This is an amazing theorem because the limit does not depend on the common distribution of the random variables in the sequence $X_1, X_2, \ldots$.

Remarks.

1. Let Z be a standard normal random variable (mean 0 and variance 1). Then the statement (2.88) of the Central Limit Theorem is equivalent to

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} \le b\right) = \mathbb{P}(a < Z \le b), \quad \text{for all } -\infty \le a < b \le \infty. \tag{2.89}$$

While this is an obvious way to rewrite the Central Limit Theorem, it expresses more directly the idea that $\frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}}$ is approximately a standard normal random variable when n is large.

2. The inequalities in the event $\{a < \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} \le b\}$ appearing in (2.88) are conventional in the statement of the Central Limit Theorem, but do not actually matter. It can be shown that if (2.88) is true, then

$$\mathbb{P}\left(a < \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} < b\right), \quad \mathbb{P}\left(a \le \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} < b\right), \quad \text{and} \quad \mathbb{P}\left(a <\le \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} \le b\right)$$

all have the same limit $\Phi(b) - \Phi(a)$. It can also be shown that (2.88) implies

$$\lim_{n\to\infty} E\Big[f\big(\frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}}\big)\Big] = E[f(Z)]$$

for any bounded and continuous function f, where Z is a standard normal random variable. We will not prove these claims, but they are important to know.

Historically, the Central Limit Theorem was first proved for binomial random variables. For each n, let Y_n be binomial with parameters n and p. Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli random variables with parameter p. Then we know that for each n, the sum $\sum_1^n X_i$ is binomial with parameters n and p. Thus,

$$\mathbb{P}\left(a < \frac{Y_n - np}{\sqrt{np(1-p)}} \le b\right) = \mathbb{P}\left(a < \frac{\sum_1^n X_i - np}{\sqrt{np(1-p)}} \le b\right).$$

By applying the Central Limit Theorem to the right-hand side, we obtain the following result.

Theorem 13 *(DeMoivre-Laplace CLT) For each integer n, let Y_n be a binomial random variable with parameters n and p, p being fixed. Then for any $-\infty \leq a < b \leq \infty$.*

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{Y_n - np}{\sqrt{np(1-p)}} \leq b\right) = \int_a^b e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \mathbb{P}(a < Z \leq b), \tag{2.90}$$

where Z is standard normal.

Paraphrasing Theorem 13, $(Y_n - np)/\sqrt{np(1-p)}$ is approximately standard normal for large n. The question is how large should n be for the approximation to be accurate? The general rule of thumb is that the approximation is accurate if $np(1-p) \geq 10$.

Example 2.3.15 Let X be binomial with $p = .5$ and $n = 50$. Find an approximate value of $\mathbb{P}(22 \leq X \leq 28)$.

Since X is discrete, and the central limit theorem approximates it by a continuous random variable, the approximation will be more accurate if we use the following *continuity correction* of the limits:

$$\mathbb{P}(22 \leq X \leq 28) = \mathbb{P}(21.5 < X < 28.5).$$

From Theorem 13 with $n = 50$ and $p = .5$, we have that $(X - 25)/\sqrt{12.5}$ is approximately standard normal. Thus

$$\mathbb{P}(21.5 < X < 28.5) = \mathbb{P}\left(\frac{21.5 - 25}{\sqrt{12.5}} < \frac{X - 25}{\sqrt{12.5}} < \frac{28.5 - 25}{\sqrt{12.5}}\right) \approx \Phi(.99) - \Phi(-.99),$$

Using tables, this turns out to be 0.6778. ◇

2.5 Continuous Limits of Discrete Random Variables and Approximation

Mathematical descriptions of real physical processes are almost never exact, and this is true for probabilistic as well as deterministic models. The probability distribution of the random outcome, X, of a physical variable will generally not take a simple form, and it could be influenced by small factors whose properties are not understood precisely. Rather, science aims at a good approximate model, $\widetilde{X}$, which can be characterized exactly, and which captures the essential features of X in the sense that $\mathbb{P}(\tilde{X} \in A) \approx \mathbb{P}(X \in A)$ for a wide class of events A. There are a variety of ways to make the notion of approximate model precise, but it suffices for now just to convey the idea: $\tilde{X}$ approximates X in distribution if the error made in computing probabilities using $\tilde{X}$ in place of X is small.

One approximation common in both applied probability and statistics, is to model a discrete random variable by one that is continuous. At first this may seem like a nonsensical thing to do, given the qualitative difference between the two types. But in fact, we have seen two instances already in this chapter. The first was discussed in Section 2.1.3 in the framework of probability space models, but it can be reformulated readily in terms of random variables. Consider a measurement, made to an accuracy of n decimal places, that results in a random number in the interval $[0, 1)$. Call this measurement X_n. If, as in the last paragraph of Section 2.1.3, each possible value of X_n has an equal probability, then X_n is uniformly distributed over the finite set of decimal numbers $0.\xi_1 \ldots \xi_n$ between 0 and 1. There are 10^n such numbers evenly distributed over $[0, 1)$, and they are closely spaced, even for moderate values of n. Thus, it is not surprising that X_n should be well approximated by a random variable U that is uniformly distributed over $(0, 1)$. In fact it can be shown that for any $0 \le a < b \le 1$, $\mathbb{P}(a < X_n < b)$ and $\mathbb{P}(a < U < b) = b - a$ differ by at most $2 \cdot 10^{-n}$. It follows that

$$\lim_{n\to\infty} \mathbb{P}(a < X_n \le b) = P(a < U \le b), \quad \text{for any } -\infty \le a < b \le \infty. \tag{2.91}$$

This statement adds even more insight: in a sense, the approximation becomes exact in the limit as $n \to \infty$. The connection between U and X_n goes beyond this limit statement. For any given n, imagine observing U and then rounding it down to the nearest decimal number of the form $0.\xi_1 \ldots \xi_n$; call the result U_n. Then U_n has exactly the same distribution as X_n.

The second example is the DeMoivre-Laplace Central Limit Theorem stated as Theorem 13. In this theorem, Y_n, being a binomial random variable, is discrete, and hence so is

$$Z_n := \frac{Y_n - np}{\sqrt{np(1-p)}}\,.$$

Since Y_n takes integer values, the distance between successive values of Z_n are a distance $1/\sqrt{np(1-p)}$ apart, running from a minimum of $\sqrt{n}\sqrt{p/(1-p)}$, when $Y_n = 0$, to a maximum of $\sqrt{n}\sqrt{(1-p)/p}$, when $Y_n = n$. So, as with in the previous example, the values of Z_n are becoming more and more closely spaced as n increases. The Central Limit Theorem says that $\lim_{n\to\infty} \mathbb{P}(a < Z_n \le b) = \mathbb{P}(a < Z \le b)$ for all $a < b$, where Z is standard normal, implying that Z_n is well-approximated in distribution by Z when n is large.

The use of a continuous distribution to approximate a discrete one is an important step in several of the major applications in this text. In this section, we explain the general nature of this approximation in greater detail, so that the reader has a better grasp on how it works, both quantitatively and qualitatively. The examples above indicate the usual framework for this discussion. There will be a family of discrete models, $\{X_n;\, n \ge 1\}$ indexed by a parameter n. As n becomes larger, the values of X_n become more and more closely spaced and the probabilities that X_n

takes on any specific value tend to 0 as $n \to \infty$; these are essentially necessary conditions for approximation by a continuous random variable. The approximation itself will be expressed as a limit statement, as in (2.91) or the Central Limit Theorem: for some continuous random variable V,

$$\lim_{n\to\infty} \mathbb{P}\big(a < X_n \le b\big) = \mathbb{P}\big(a < V \le b\big) \quad \text{for all } -\infty \le a < b \le \infty.$$

When this is true, we say that X_n *converges in distribution* to V; this is standard terminology in probability theory. In applications, convergence in distribution is used to justify using V as an approximate model for X_n, when n is reasonably large.

2.6 Notes

The material in this chapter is standard and may be found, mostly, in any undergraduate probability text. A standard and very good text is that of Sheldon M. Ross, *A First Course in Probability*, Prentice-Hall. There are many editions of this text; any of them is a good reference.