Chapter 2

Probability Theory

Chapter 2 reviews the probability theory necessary to understand this text. Topics covered include: probability spaces and axioms of probability, independence, the basic random variables of applied probability, expectation, conditional probability and conditional expectation, the law of large numbers and the central limit theorem. Where feasible, examples from biology anticipating later applications are used to illustrate the concepts. Also, special emphasis is given to random sampling and to the approximation of discrete by continuous random variables, which are both important in biological applications.

This chapter is not an introduction to probability for beginners. It is assumed that readers of this text have already studied probability at the undergraduate level, and this chapter is meant to serve as a reference and refresher. It can be read as needed. For understanding Chapter 3, it suffices to read Section 2.1 and review binomial random variables in Subsection 2.2.1. Chapters 4, 5, and 6 make serious use of the random variable theory covered in sections 2.2 and 2.3.

2.1 Probability basics and random sampling

2.1.1 Probability spaces

Probability theory provides a mathematical framework—a set of basic definitions and computational rules—for discussing random phenomena. This framework is encapsulated in the notion of a probability space, which may be thought of as a universal template for modeling experiments with random outcomes. A probability space consists of three components: (i) a set called the outcome space; (ii) subsets of the outcome space called events; and, (iii) a rule, called a probability measure, that assigns probabilities to events and obeys the axioms listed below as (P1)—(P3). These are the natural elements one would include in attempting to describe mathematically an experiment with uncertain outcome. We discuss each of them in turn. The outcome space, traditionally denoted Ω, is simply a list of all possible outcomes of the experiment. For example, a roll of a single die has six possible outcomes, which one may label by the number of pips showing on the top face: hence, one may represent the outcome space of a die roll as Ω = {1, 2, 3, 4, 5, 6}. We want, of course, to assign probabilities to outcomes, but we want to define probabilities for more general
results as well. For instance, we might be interested in the probability that a die roll is an even number. This is the same as asking that the outcome belongs to the subset \( \{2, 4, 6\} \) of the outcome space, and so it is convenient to identify the event of an even roll with this subset. In the setting of a general outcome space \( \Omega \), we speak of its as events, and we say that event \( U \) occurs in a trial if the outcome falls in \( U \). The final component of a probability space, the probability measure, is a function \( P \) that assigns to each event \( U \) a number \( P(U) \) representing the probability that \( U \) occurs. This assignment cannot be arbitrary. To qualify as a probability measure, \( P \) must satisfy the following properties, called the axioms of probability:

(P1) \( 0 \leq P(U) \leq 1 \) for all events \( U \);

(P2) \( P(\Omega) = 1 \), and;

(P3) If \( U_1, U_2, \ldots, U_n \) are disjoint events, meaning they share no elements in common,
\[
P(U_1 \cup \cdots \cup U_n) = P(U_1) + \cdots + P(U_n).
\]

(P3') More generally, if \( U_1, U_2, U_3, \ldots \) is an infinite sequence of disjoint events, then
\[
P\left( \bigcup_{i=1}^{\infty} U_i \right) = \sum_{i=1}^{\infty} P(U_i).
\]

These axioms are just about the simplest properties, consistent with intuition, to place on probability assignments. To say an event occurs with probability 1 means it is certain to happen, and since, by definition \( \Omega \) includes all possible outcomes, we require (P2). On the other hand, \( P(U_1 \cup U_2) \) is the probability that the outcome falls either in \( U_1 \) or in \( U_2 \), and so if \( U_1 \) and \( U_2 \) are disjoint it is natural to require \( P(U_1 \cup U_2) \) to be the sum of \( P(U_1) \) and \( P(U_2) \); this is what (P3) says, except more generally for any finite disjoint union. Axiom (P3') is a juiced up version of (P3) that is imposed on probability measures on outcome spaces that have an infinite number of elements. Axiom (P3') really includes (P3), as a special case, but we state them separately to emphasize the finite case, which is at the heart of elementary probability computations.

There are two simple consequences of the probability axioms that are repeatedly used.

1. If \( U \) is an event, let \( U^c \) denote the complement of \( U \), that is all elements in \( \Omega \) that are not in \( U \). Then
\[
P(U^c) = 1 - P(U)
\]

2. (Inclusion-Exclusion Principle) \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \)

It is left to the reader as an exercise to deduce these from the probability axioms.

2.1.2 Examples; random sampling.

The examples introduced in this section anticipate biological models to be discussed later in the text. We will return to them repeatedly in this chapter. The following notation will be used; if \( A \) is a finite set, \( |A| \) denotes the number of elements in \( A \).
Most of the examples in this section have the following common structure. First, the set of outcomes $\Omega$ is finite. Let us write $\Omega = \{\omega_1, \ldots, \omega_N\}$, so that $N = |\Omega|$. Second, the probability measure is defined so that each of the $N$ outcomes has an equal probability to occur. Expressed mathematically,

$$P(\{\omega_i\}) = \frac{1}{N} \left( = \frac{1}{|\Omega|} \right), \quad 1 \leq i \leq N.$$ 

(In this equation, the notation $\{\omega_i\}$ denotes the set consisting of the single outcome $\omega_i$ and it represents the event that exactly $\omega_i$ occurs; this is the correct way to write it, but often we will be less pedantic and write simply $P(\omega_i)$ instead.) Let $U$ be any event contained in $\Omega$. It can be viewed as the union of its elements: $U = \cup_{\omega \in U} \{\omega\}$. Then, by the additivity axiom (P3),

$$P(U) = \sum_{\omega \in U} P(\{\omega\}) = \sum_{\omega \in U} \frac{1}{N} = \frac{|U|}{N}. \quad (2.4)$$

In probability theory, $P$ defined this way is called the uniform probability measure on $\Omega$.

Finite probability spaces with equally likely outcomes are, of course, fundamental to probability theory, theoretical and applied. For our purposes, the first, very important application is to random sampling. Consider a finite set of items, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$, where $\theta_i$ denotes item $i$. A single random sample from $\Theta$ is a random choice of an element from $\Theta$ in such a way that each has equal probability $1/|\Theta|$ of being selected. The probability space model for a random selection is thus simply the uniform probability measure on $\Theta$. In the sampling context, the set $\Theta$ is called a population. It might be a population of living organisms, but the term is applied to any finite collection of interest, be it a collection of marbles, lottery numbers in an urn, or whatever is subject to being sampled.

**Example 2.1.1.** Suppose we have a population of $N$ gametes of Mendel’s pea plants (see Chapter 1). Suppose $n_1$ of them bear the allele $G$ for green colored peas and $n_2 = N - n_1$ bear the allele $Y$ for yellow colored peas. Consider the experiment of choosing a gamete by random sampling. This is the basis of the random mating model at the heart of population genetics. An explicit probability space for this experiment may be constructed as follows, following the prescription of the random sampling model just stated. Label the gametes of the population with the integers from 1 to $N$. Take as the outcome space $\Omega = \{1, 2, \ldots, N\}$, where number $i$ stands for the outcome that gamete $i$ is chosen. Then the probability measure describing gamete choice is the uniform measure, defined by requiring that $P_0(\{i\}) \triangleq 1/N$ for each $i$.

Assuming this model, what is the probability that the randomly sampled gamete carries allele $G$? This is easy to answer. The event of sampling a $G$ corresponds to the subset $A$ of $\Omega$ consisting of all those individuals bearing allele $G$, and, by assumption, $|A| = n_1$. Hence, using the rule (2.4) for uniform probability measures,

$$P(\text{choose a gamete with allele } G) = P(A) = \frac{|A|}{N} = \frac{n_1}{N}.$$
This answer, I hope, would have been obvious even had we just asked the question outright at the beginning of the section, before introducing any formal notions from probability. However, I hope that, because of the simplicity of the example, the reader sees clearly how the probability space formalism works.

Example 2.1.2: Sampling an urn. An urn containing $N$ marbles, $n_1$ of which are green and $n_2 = N - n_1$ of which are yellow. Consider a single random sample from this urn. Although we have replaced gametes by marbles, this is exactly the same situation as the first example and the probability space is the same. The probability of drawing a green marble is $n_1/N$.

Example 2.1.3. Gamete example revisited. Consider the setup of Example 2.1.1. We draw a gamete at random from a population of $N$ gametes, $n_1$ of which carry allele $G$, and we record whether it carries allele $Y$ or allele $G$. As $Y$ and $G$ are the only outcomes, we could model this experiment using the simple outcome space $\Omega = \{G, Y\}$. By the result of Example 2.1.1, the appropriate probability measure on this space is defined by $P(G) = n_1/N$ and $P(Y) = 1 - (n_1/N) = n_2/N$.

Example 2.1.4. Coin toss. Toss a coin once and record whether the result is heads ($H$) or tails ($T$). Assume the probability of heads is $p$. The natural probability space is: $\Omega = \{H, T\}$ and $P(H) = p$, $P(T) = 1 - p$. Abstractly, this is the same as the probability space of the previous example; we have only replaced $G$ by $H$, $Y$ by $T$, and $n_1/N$ by $p$.

These examples show that the random sampling models we will encounter in population genetics are equivalent to simple examples, such as coin tossing, drawing marbles from urns, etc., encountered in a standard elementary probability course.

Calculating the probability of what kind of an individual is chosen in a random sample, as in example 2.1.1, arises repeatedly, and so it is useful to have some general notation for it. Consider an abstract population $\Theta = \{\theta_1, \ldots, \theta_N\}$. Assume that each individual bears a label from a set $S$. In the gamete example, the label would be the allele for pea color, so $S = \{G, Y\}$; in the example 2.1.2 of selecting a marble, the label would be the marble’s color, and again $S = \{G, Y\}$. In this course, the label will typically be a number, a vector of numbers, or, especially in genetics, genotypes denoted by strings of letters.

For each $x$ in $S$, the frequency of $x$ in $\Theta$ is defined to be

$$f_x \triangleq \frac{\text{number of individuals in } \Theta \text{ with label } x}{|\Theta|}.$$ 

Equation (2.4) says that for a single random sample

$$P(\text{selected individual bears label } x) = f_x. \tag{2.5}$$

In many applications, especially for statistics, one is interested in sampling a population randomly many times. When sampling repeatedly, we can either return the sampled individual to the population each time, which allows it to be selected again, or remove each sampled individual from the population permanently. The first alternative is called sampling with replacement, the second, sampling without replacement. For large
populations and sample sizes that are small compared to the population size, the probability that an individual is sampled more than once is small relative to the probability it is sampled once, and the two schemes give similar results with high probability. We will focus on sampling with replacement as it is simpler to model. From now on, sampling will always mean sampling with replacement, unless explicitly stated to the contrary.

We will now formulate a precise model, in the language of probability spaces, for a random sample of size \( n \). Let \( \Theta \) denote the population, as before, and assume \( |\Theta| = N \). The outcome of a sequence of \( n \) samples is a sequence \( (\eta_1, \ldots, \eta_n) \), where each \( \eta_i \) is the individual selected from \( \Theta \) on draw \( i \). Since we are sampling with replacement, the same individual can appear repeatedly in the sequence. Therefore, the outcome space is the set \( \Omega \) of all sequences of length \( n \) of elements of \( \Theta \). Since there are \( N \) possible values for each \( \eta_i \) and \( n \) components of each sequence, the total number of sequences in \( \Omega \) is \( |\Omega| = N^n \).

A random sample of size \( n \) from \( \Theta \) (with replacement) is a sampling such that each sequence in \( \Omega \) has an equally likely chance of occurring. Thus it is described by the uniform probability measure on \( \Omega \),

\[
P\left(\{(\eta_1, \ldots, \eta_n)\}\right) = \frac{1}{N^n} \quad \text{for each } (\eta_1, \ldots, \eta_n) \text{ in } \Omega.
\]

This is a natural definition for a sampling scheme which is somehow the “most” random one could imagine; no sequence is privileged above others by having a higher probability of occurring. In the next section, we shall give an equivalent definition using the concept of independence.

**Example 2.1.5. Random sampling of a gamete population, continued.** This example is a continuation of Example 2.1.1, dealing with a population of \( N \) gametes of Mendel’s peas, in which \( n_1 \) bear allele \( G \) and \( n_2 = N - n_1 \) allele \( Y \). Consider now a random sample of size 10 from this population. What is the probability that exactly 4 of the 10 sampled alleles carry \( G \)?

If the individuals of the population are denoted by the numbers \( 1, \ldots, N \), the outcome space \( \Omega \) for the random sample of size 10 is the set of all subsequences \( (\eta_1, \ldots, \eta_{10}) \), where each \( \eta_i \in \{1, \ldots, N\} \), and \( |\Omega| = N^{10} \). For convenience, label the individuals so that individuals \( 1, 2, \ldots, n_1 \) are precisely those with allele \( G \). The event \( A \) that exactly 4 of the 10 sampled gametes carry \( G \) is the subset of all sequences of length 10 in \( \Omega \), exactly 4 elements of which are in \( \{1, \ldots, n_1\} \). Since the measure describing a random sample is the uniform measure, \( P(A) = |A|/N^{10} \). Hence to solve the problem, we need to calculate \( |A| \). There are \( \binom{10}{4} \) choice of 4 positions in the sequence \( (\eta_1, \ldots, \eta_{10}) \) at which individuals of type \( G \) occur. For each such choice of 4 draws, there are \( n_1^4 \) different ways to select elements of \( \{1, \ldots, n_1\} \) to appear in these draws, and there are \( (N - n_1)^6 \) ways to choose elements not in \( \{1, \ldots, n_1\} \) (individuals with \( Y \)) to appear in the remaining draws. Therefore \( |A| = \binom{10}{4} n_1^4 (N - n_1)^6 \).

It follows that

\[
P(A) = \binom{10}{4} n_1^4 (N - n_1)^6 \frac{1}{N^{10}} = \binom{10}{4} \left(\frac{n_1}{N}\right)^4 \left(\frac{N - n_1}{N}\right)^6 = \binom{10}{4} f_G^4 f_Y^6,
\]

where, in the last expression, we have used that the frequency \( f_G \) of \( G \) alleles is \( n_1/N \) and the frequency \( f_Y \) of \( Y \) alleles is \( (N - n_1)/N \). This answer for \( P(A) \) is an example of a binomial probability, which will be reviewed in more detail later in the chapter.
Example 2.1.6. Coin tossing, continued. Toss a fair coin \(n\) times, recording heads with a 1, and tails with a 0. The outcome space \(\Omega\) for this experiment is the set of all possible sequences of 0’s and 1’s of length \(n\), and \(|\Omega| = 2^n\). If the results of any toss do not affect the outcome of other tosses, then, since the coin is fair, there is no reason for any one sequence of tosses to be more or less likely than any other. Thus, a natural model for this experiment is the uniform probability measure, in which each sequence occurs with probability \(1/2^n\). Notice that this model is equivalent to a random sample of size \(n\) from a population consisting of two elements 0 and 1. In effect, we make our random selection by tossing a coin.

Example 2.1.7. A model for junk DNA. Biologist Bob is studying a stretch of DNA, hypothesized to be junk DNA. He has an experiment which samples and sequences 10 contiguous bases from single stranded DNA in his sample. (This is very artificial, but bear with me for the sake of an example.) He is interested in somehow testing the junk DNA hypothesis, and to do this he wants to translate the junk DNA assumption into a probability model for the 10 base sequences produced by his experiment. Here is a natural idea. If the DNA is truly junk DNA with no function, no selective pressure has acted on it, and as mutations have occurred to it over time, the bases in the sequence should appear as if they are chosen at random. In other words, assuming the junk DNA hypothesis, the results of Bob’s experiment should be equivalent to a random sample of size 10 from the DNA alphabet \(\{A, T, G, C\}\).

We will revisit this model in more generality in later chapters. There, it is called the independent site model with equal base probabilities. In this terminology, the equal base probability property is simply this. Let \(x\) stand for any one of the bases in the DNA alphabet, and let \(i \leq 10\) denote any position along the sequence. Then, under the model \(P(\text{base } x \text{ appears at site } i) = 1/4\). To see this note that because every sequence is equally likely in the random sample model, the different bases appear at site \(i\) with equal probabilities, and as there are 4 bases, this probability is \(1/4\). To repeat this argument more formally, consider the event \(U\) that base \(C\) appears as the first base in the sequence. The number of DNA base sequence of length 10 that start with \(C\) is the total number of DNA sequences of length 9 (because one element only is fixed at \(C\)), and this number is \(4^9\). The total number of DNA sequences of length 10 is \(4^{10}\). Hence \(P(U) = 4^9/4^{10} = 1/4\). Clearly this argument is independent of the base under consideration and its position in the sequence.

The independent site property for this model will be explained in the next section.

The equal base probability property of the last example reflects a general fact about random sampling. Consider a random sample \((\eta_1, \ldots, \eta_n)\) of size \(n\) from \(\Theta\). For any \(i, 1 \leq i \leq n\), draw \(i\) considered by itself is a single random sample. In other words,

\[
P(\theta \text{ is chosen on draw } i) = P(\{\eta_i = \theta\}) = \frac{1}{|\Theta|} \quad \text{for all } 1 \leq i \leq n \text{ and all } \theta \in \Theta. \tag{2.6}
\]

The argument is a generalization of that employed in the example. The number of sequences of elements of \(\Theta\) of length \(n\) in which precisely element \(\theta\) appears in position \(i\) is the number of sequences of length \(n-1\), and this is \(|\Theta|^{n-1}\). The total number of sequences of length \(n\) is \(|\Theta|^n\). Hence, \(P(\theta \text{ is chosen on draw } i) = |\Theta|^{n-1}/|\Theta|^n = 1/|\Theta|\).
2.1.3 Independence

Let $U$ and $V$ be two events and assume $\mathbb{P}(V) > 0$. The conditional probability of $U$ given $V$ is defined to be

$$
\mathbb{P}(U \mid V) \triangleq \frac{\mathbb{P}(U \cap V)}{\mathbb{P}(V)}.
$$

(2.7)

This quantity is interpreted as the probability that event $U$ occurs, knowing that $V$ has occurred.

Two events $U$ and $V$ are said to be independent if

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

(2.8)

If $U$ and $V$ satisfy this condition and if $\mathbb{P}(V) > 0$ and $\mathbb{P}(U) > 0$, then one can verify immediately from definition (2.7) that $\mathbb{P}(U \mid V) = \mathbb{P}(V)$ and likewise $\mathbb{P}(V \mid U) = \mathbb{P}(V)$; thus the probability of neither $U$ nor $V$ is affected by conditioning on the other event. This explains the terminology “independence.”

The events of a set $\{U_1, \ldots, U_n\}$ are said to be independent if

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

(2.9)

for every $k$, $2 \leq k \leq n$, and every possible subsequence $1 \leq r_1 < r_2 < \cdots < r_k \leq n$. For example, the events in $\{U_1, U_2, U_3\}$ are independent if and only if

$$
\mathbb{P}(U_1 \cap U_2) = \mathbb{P}(U_1)\mathbb{P}(U_2), \quad \mathbb{P}(U_1 \cap U_3) = \mathbb{P}(U_1)\mathbb{P}(U_3), \quad \mathbb{P}(U_2 \cap U_3) = \mathbb{P}(U_2)\mathbb{P}(U_3)
$$

and, in addition,

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

(2.10)

The condition (2.10) implies that the probability of any one of the events $U_1, U_2, U_3$ is not affected by combined knowledge of the other two. For example $U_1$ is independent of the intersection $U_2 \cap U_3$, and also of the union $U_2 \cup U_3$. The reader is asked to deduce this in Problem 2.1.6. Pairwise independence of the events is not sufficient to draw these conclusions; see Problem 2.1.5.

Independence is an essential feature of random sampling, as defined in the previous section. Consider a random sample of size 2 from a population $\Theta$ and let $\eta_1$ denote the individual selected on the first draw and $\eta_2$ the individual selected on the second draw. Let $E_1$ and $E_2$ be subsets of $\Theta$ and consider the event $\{\eta_1 \in E_1\}$ that the first draw is an individual in $E_1$ and the event $\{\eta_2 \in E_2\}$ that the second draw is from $E_2$. We know from (2.6) that each $\eta_i$ is a single random sample and hence

$$
\mathbb{P}(\eta_1 \in E_1) = \frac{|E_1|}{|\Theta|}, \quad \mathbb{P}(\eta_2 \in E_2) = \frac{|E_2|}{|\Theta|}.
$$

Also, the cardinality of the intersection of these two events is the number of sequences the first element of which is in $E_1$ and the second in $E_2$, and so

$$
| \{\eta_1 \in E_1\} \cap \{\eta_2 \in E_2\} | = |E_1| \cdot |E_2|.
$$

As the total number of samples of size 2 is $|\Theta|^2$

$$
\mathbb{P}(\eta_1 \in E_1 \cap \{\eta_2 \in E_2\}) = \frac{|E_1|\cdot|E_2|}{|\Theta|^2} = \mathbb{P}(\eta_1 \in E_1)\mathbb{P}(\eta_2 \in E_2).
$$
Hence the two events \( \{ \eta_1 \in E_1 \} \) and \( \{ \eta_2 \in E_2 \} \) are independent. The argument is valid for any subsets \( E_1 \) and \( E_2 \), and so, in a sense, the results of the first and second draws are independent of each other. This result generalizes to random samples of any length; the different draws of a random sample of size \( n \) are independent of each other.

**Theorem 1** Consider a random sample of size \( n \) from a population \( \Theta \). Then the outcomes of the different draws are all independent. This means that if \( \eta_1, \ldots, \eta_n \) are the results of the \( n \) draws and if \( E_1, \ldots, E_n \) is any sequence of subsets of \( \Theta \),

\[
P(\{ \eta_1 \in E_1 \} \cap \cdots \cap \{ \eta_n \in E_n \}) = P(\eta_1 \in E_1) \cdots P(\eta_n \in E_n).
\] (2.11)

In particular, the different draws of a random sample are all independent.

According to this theorem and to (2.6), a random sample of size \( n \) consists of \( n \) independent single random samples. This is an alternate way to define a random sample of size \( n \), and is the basis of an extensive generalization of the idea of a random sample that will be defined in the section of this chapter on random variables.

**Example 2.1.8; coin tossing, continued.** This is a continuation of example 2.1.6, which defines a probability space for \( n \) tosses of a fair coin, by assuming each sequence of tosses is equally likely. We saw that this model was equivalent to a random sample of size \( n \) from a population of size 2. By Theorem 1 the coin tosses are independent of one another. In fact, the usual way to define this model is as \( n \) independent tosses of a fair coin.

**Example 2.1.9; junk DNA model, continued.** Consider the model formulated in Example 2.1.7 for a random sequence of junk DNA 10 bases long. This model supposed that the sequence was obtained by a random sample of size 10 from the DNA alphabet. By Theorem 1 the random choice of bases at the different sites are independent from one another. This explains the term “independent sites” in the name we gave this model, “the independent sites model with equal base probabilities.”

### 2.1.4 Conditioning, Rule of total probabilities, Bayes’ rule

This section reviews basic applications of conditional probabilities.

Let \( \Omega \) be an outcome space with a probability \( \mathbb{P} \). Suppose that \( V_1, \ldots, V_n \) are all mutually disjoint and that they cover \( \Omega \), that is \( \Omega = V_1 \cup \cdots \cup V_n \); we say that \( V_1, \ldots, V_n \) is a disjoint partition of \( \Omega \). Assume \( \mathbb{P}(V_i) > 0 \) for each \( i \). The rule of total probabilities is a formula for computing the probability of an event \( U \) using the conditional probabilities \( \mathbb{P}(V_i) \), \( 1 \leq i \leq n \). It says:

\[
\mathbb{P}(U) = \mathbb{P}(U \mid V_1)\mathbb{P}(V_1) + \mathbb{P}(U \mid V_2)\mathbb{P}(V_2) + \cdots + \mathbb{P}(U \mid V_n)\mathbb{P}(V_n).
\] (2.12)

The natural data for building a probability model for a random experiment is often translates most directly into conditional probabilities. One then uses the rule of total probabilities to derive the probability measure for the model.

**Example 2.1.10.** Suppose that there are 100 marbles, each marked either \( G \) or \( Y \), distributed in in two bags, bag I containing 60 marbles, of which 35 are \( G \)’s, and bag II containing 40
marbles, of which 10 are $G$'s. The experiment consists of drawing a bag at random, each with equal probability one-half, drawing a marble at random from the chosen bag, and recording whether it is $G$ or $Y$. To describe this experiment probabilistically we need to compute the probability that a $G$ marble is drawn.

The answer is a direct application of the total probability rule. Let $G$ denote also the event that a $G$ marble is drawn. Let $V$ be the event that bag I is chosen; its complement $V^c$ is the event that Bag II is chosen. Since the marble is chosen from the selected bag by a random sample $P(G \mid V) = \frac{35}{60} = \frac{7}{12}$ and $P(G \mid V^c) = \frac{10}{40} = \frac{1}{4}$. Events $V$ and $V^c$ are disjoint and cover all possibilities. By definition of the experiment, $P(V) = P(V^c) = \frac{1}{2}$.

Thus, from the rule of total probabilities,

$$P(G) = \frac{35}{60} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{5}{12}.$$ 

\[ \diamond \]

The derivation of the total probability rule (2.12) is not hard. Observe that for any $i$

$$P(U \cap V_i) = P(U \mid V_i)P(V_i). \quad (2.13)$$

This follows directly from the definition of conditional probability. Further since $V_1, \ldots, V_n$ make a disjoint partition of $\Omega$, $U = U \cap V_1 \cup \cdots \cup U \cap V_n$, and the sets in this union are disjoint. By using additivity of probabilities, as in equation (2.1), and then the previous identity (2.13),

$$P(U) = P(U \cap V_1) + \cdots + P(U \cap V_n) = P(U \mid V_1)P(V_1) + \cdots + P(U \mid V_n)P(V_n).$$

We turn next to Bayes' rule. Observe first that if $V$ is an event, the pair of events $V$ and $V^c$, where $V^c = \Omega \setminus V$ is the set-theoretic complement of $V$, is a disjoint partition of $\Omega$. By the total probability rule, $P(U) = P(U \mid V)P(V) + P(U \mid V^c)P(V^c)$. Also, using (2.13) with $V_i$ replaced by $V$, $P(U \cap V) = P(U \mid V)P(V)$. Thus

$$P(V \mid U) = \frac{P(U \cap V)}{P(U)} = \frac{P(U \mid V)P(V)}{P(U \mid V)P(V) + P(U \mid V^c)P(V^c)}. \quad (2.14)$$

This is Bayes' rule. The next example illustrates a typical application.

**Example 2.1.10, continued.** Consider the experiment of the previous example. Suppose a $G$ marble is drawn. Conditioned on this event, what is the probability that bag I was selected in the first step, that is, what is $P(V \mid G)$? We saw above that $P(V) = P(V^c) = 1/2$, and $P(G \mid V) = \frac{35}{60}$, $P(G \mid V^c) = \frac{1}{4}$. Using Bayes' rule,

$$P(V \mid G) = \frac{(35/60)(1/2)}{(35/60)(1/2) + (1/4)(1/2)} = \frac{7}{10}.$$ 

(Notice that the calculation of the denominator recapitulates the derivation of $P(G)$ above.)
2.1.5 The Law of Large Numbers for Empirical Frequencies.

Consider a population $\Theta$ of individual each of whom carries a label $x$ from some set $S$. For each label $x$, we defined the frequency $f_x$ of $x$ in the population to be the number of individuals with label $x$ divided by the total number of individuals in $\Theta$. Given knowledge of $f_x$, probability theory tells us how to compute probabilities of drawing different combinations of labels if we do random sampling of $\Theta$. However, in real life scientists study target populations for which the frequencies are unknown and they want to estimate them as accurately as possible. For example, one might be interested in the distribution of randomly selected, DNA sequences. If one could sequence the DNA of every individual in the whole population, the frequency $f_x$ for each different sequence $x$ could be computed exactly. But this is impractical. Instead, one must make inferences about the true frequencies from a relatively small sample.

What is a good way to estimate $f_x$? One natural idea is to take a random sample of some size $n$ and estimate $f_x$ with the observed frequency of $x$ in this sample. This is called the empirical frequency of the sample. It is denoted and defined explicitly by

$$ \hat{f}_x(n) = \frac{\text{number of times } x \text{ occurs in the } n \text{ samples}}{n} \quad (2.15) $$

For large $n$, we can expect this to be a good estimate of the true population frequency because random sampling draws every individual with equal probability and so ought to produce a representative sample of the population. The following, very important result supports the idea that $\hat{f}_x(n)$ is “good” estimate of $f_x$.

**Theorem 2** Law of Large Numbers for Empirical Probabilities Consider an sequence of independent single random samples of a population $\Theta$ of labeled individuals. For any label $x$,

$$ \lim_{n \to \infty} \hat{f}_x(n) = f_x \quad \text{with probability one.} $$

This theorem is a special case of a much more general Law of Large Numbers that will be stated later on for random variables. The theorem is actually true even if some restricted dependence between trials is allowed, but such refinements are beyond the scope of this discussion. However, as stated, this Theorem does not tell one how good $\hat{f}_x(n)$ is as an estimate of $f_x$ for a given $n$. We will state some results on this later using Chebyshev’s inequality.

The Law of Large Numbers is conceptually very important. It supports a frequency interpretation of probability; namely, the probability of an event is its limiting empirical frequency in an infinite sequence of independent trials.

2.1.6 Problems.

Some of these problems require elementary combinatorial probability not covered in the text, but assumed known by the reader.

2.1.1. Population I is a population of $N$ gametes of Mendel’s peas, $r$ of which have genotype $G$ for pea color. Population II contains $M$ gametes, $s$ of which have genotype $G$. Gametes
not of genotype \( G \) are of genotype \( Y \). Alice performs a cross between the two populations by selecting a gamete from population I to mate with a gamete from population II

a) Assuming that Alice chooses the gametes so that every pair consisting of one gamete from I and one gamete from II is equally likely to be chosen. Construct a probability space to describe this experiment; define an explicit outcome space, state the probability of each outcome, and give a formula for the probability of a set \( U \) of pairs in terms of its cardinality \( |U| \).

b) Using the probability space defined in a), show that if we restrict attention to the gamete from population I, each gamete from population I is equally likely to be chosen. Thus the model of a) implies that the gamete choice from population I amounts to a single random sample of that population. The same argument shows that the gamete choice from population II amounts to a random sample from population II.

c) If \( U_i \) is the event that gamete \( i \) is selected from population I and \( V_j \) is the event that gamete \( j \) is selected from population II, show that \( U_i \) and \( V_j \) are independent. More generally, show that draws from the two populations are independent.

d) 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 \( Y \)?

2.1.2. A four bp-long DNA segment is selected at random and sequenced. Assume the independent site/equal base probability model of example 2.1.7, (with the difference that the segment has length 4 instead of 10.) Find the probability that the selected sequence contains exactly two adenine (A) bases.

2.1.3. Randomly select two four-base-long long DNA segments \( x_1x_2x_3x_4 \) and \( y_1y_2y_3y_4 \) and align them as follows: 

\[
x_1 \ y_1 \ x_2 \ y_2 \ x_3 \ y_3 \ x_4 \ y_4
\]

a) Assume that the selection of the \( x \)-sequence and the selection of the \( y \)-sequence are independent. Assume that the independent site/equal base probability model applies to each sequence. Construct a probability space for the aligned pair of sequences; specify the outcome space and the probability of each individual outcome.

b) What is the probability that both the \( x \)-sequence and the \( y \)-sequence begin with \( A \)? What is the probability that \( x_1 = y_1 \) and \( x_1 \) is 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 \( x_1 = y_1, x_2 = y_2, x_3 = y_3, x_4 = y_4 \)?

2.1.4. For the coin toss model of example 2.1.6, show in complete detail that the events, heads on the first toss \( (U_1) \), tails on the second toss \( (V_2) \), and tails on the third toss \( (V_3) \) are independent of one another.

2.1.5 Consider 4 tosses of a fair coin (example 2.1.6 with \( n = 4 \)). Let \( U \) be the event of heads on the first two tosses, let \( V \) be the event of heads on tosses 3 and 4. We know \( \mathbb{P}(U) = \mathbb{P}(V) = 1/4 \). Let \( W \) be the event that there is exactly one head and one tail on tosses 2 and 3; thus \( W \) is the set of all sequences of 0's and 1's of the form \( (x_1,1,0,x_4) \) or
(x_1, 0, 1, x_4); thus P(W) = 1/2. Show that the events U, V, W are pairwise independent but not independent. Hint: Show the $U \cap V \cap W$ is empty.

2.1.6 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.

2.1.7. 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?

2.2 Random variables

Consider one toss of a fair coin and assign a value of 1 to heads and 0 to tails. Let us label the potential outcome of a toss by the letter X. Before the toss takes place, X cannot be assigned a definite value; it will take on two possible values, 1, with probability one-half, if the toss is heads, or 0, again with probability one-half, if the toss is tails. Therefore it makes sense to call X a random variable. In general, a random variable Y is a variable whose value is not fixed, but may take on any one of a range of values with prescribed probabilities.

Random variables offer an alternative to probability modeling. Instead of building a probability space for a random trial, one can model the outcome as a random variable. For example, a random variable X which takes the value 1 with probability p and the value 0 with probability 1 − p models a coin toss in which the probability of heads is p. In a sense, this is hardly different from the probability space model for a coin toss; we have merely given the outcome of the toss a label, namely X. Actually, any random variable model, no matter how sophisticated, can always be rewritten purely in terms of probability spaces. But random variables are in general easier to use, especially in building sophisticated models, and easier to analyze. Also, random variables are the right framework for defining the important concepts of mean (expected value) and variance, and for stating the important limit laws of probability theory. This section will review the basic theory of random variables.

2.2.1 Discrete Random Variables. Probability Mass Functions, Independence, Random Samples

A random variable taking values in a finite set $S = \{s_1, \ldots, s_N\}$, or a countably infinite set $S = \{s_1, s_2, \ldots\}$, is called a discrete random variable. Officially, random variables are real-valued and so S should be a subset of the real numbers. However, it is convenient for biological sequence models to extend the concept to variables taking random values in a non-numerical alphabet. For example, if X denotes the first base in a randomly chosen DNA segment, we may think of X as a random variable with values in $\mathcal{A} = \{A, G, C, T\}$.

Convention: In this text, random variables are assumed to be real-valued, in accordance with general usage, unless it is explicitly specified otherwise or it is clear from the context.

Let X be a discrete random variable with values in S. The probability mass function of X is the function

$$p_X(x) \triangleq P(X = x), \quad x \text{ in } S,$$ (2.16)
2.2. RANDOM VARIABLES

specifying the probabilities with which \( X \) takes its different possible values. In effect, the probability mass function is the probability model for the outcome of \( X \). If \( U \) is a subset of \( S \), the notation \( \{ X \in U \} \) stands for the event that the value of \( X \) is in \( U \), and, by the additivity property of probabilities

\[
P(\{ X \in U \}) = \sum_{x \in U} p_X(x).
\]

(2.17)

(The notation on the right-hand-side means that the sum is taken over all \( x \) in \( U \).) Note in particular that 1 = \( P(X \in S) = \sum_{x \in S} p_X(x) \). In general, any function \( p \) on \( S \) satisfying \( 0 \leq p(x) \leq 1 \) and \( \sum_{x \in S} p(x) = 1 \) is called a probability mass function.

**Example 2.2.1. Bernoulli random variables.** Let \( X \) be a random variable with values in \( \{0, 1\} \) with the probability mass function

\[
p_X(0) = 1 - p, \quad p_X(1) = p.
\]

(2.18)

We shall call such an \( X \) a Bernoulli random variable with parameter \( p \). For later reference, it is useful to note that the definition (2.18) can be written in functional form as

\[
p_X(s) = p^s(1 - p)^{1-s} \quad \text{for } s \text{ in the set } \{0, 1\}.
\]

(2.19)

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. We shall often use the language of success/failure trials when discussing Bernoulli random variables.

The term *Bernoulli random variable* is also used for any random variable taking on only two possible values \( s_0 \) and \( s_1 \), even if these two values differ from 0 and 1. We shall indicate explicitly when values different than 0 and 1 are used, as in the next example. Otherwise \( s_0 = 0 \) and \( s_1 = 1 \) are the default values.

**Example 2.2.2. Bernoulli model for Example 2.1.1.** Recall the experiment of example 2.1.1; a gamete is randomly drawn from a population of size \( N \), \( n_1 \) of which bear genotype \( G \). The outcome of this experiment can be modeled as a random variable \( X \) with values in the alphabet \( \{G, Y\} \) and with probability mass function

\[
p_X(G) = 1 - \frac{n_1}{N}, \quad p_X(Y) = \frac{N - n_1}{N}.
\]

We will show next how to re-express Examples 2.1.6 and 2.1.7 using random variables. Each outcome in these examples consists of independent repetitions of single trial—\( n \) repetitions of a coin toss in Example 2.1.6, and 10 evaluations of bases at different sites along a random DNA segment in Example 2.1.7. If each trial is modeled by a single random variable, the complete model of the experiment will be a finite sequence of random variables. Therefore to present these models requires extending the notion of a probability mass function to sequences, and defining what it means for random variables to be independent.
Given discrete random variables $X_1, X_2, \ldots, X_n$, each with values in $S$, arrange them in a random vector $Z = (X_1, \ldots, X_n)$. The joint probability mass function of $X_1, \ldots, X_n$ (alternatively, the probability mass function of $Z$) is the function

$$p_Z(x_1, \ldots, x_n) \triangleq \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n), \quad x_1, \ldots, x_n \in S, \quad (2.20)$$

which expresses the joint probabilities of all different possibilities for the outcomes of $X_1, \ldots, X_n$.

The set of random variables $X_1, \ldots, X_n$ with values in $S$ 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 $S$.

For example, if $X_1, \ldots, X_n$ is an independent set and $x_1, \ldots, x_n$ are points in $S$, the events $\{X_1 = x_1\}, \ldots, \{X_n = x_n\}$ are independent, and hence

$$p_Z(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \mathbb{P}(X_2 = x_2) \cdots \mathbb{P}(X_n = x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \cdots p_{X_n}(x_n) \quad (2.21)$$

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

**Theorem 3** The set $X_1, \ldots, X_n$ of discrete random variables with values in $S$ is independent if and only if (2.21) is true for all choices $x_1, \ldots, x_n$ of values from $S$.

We shall not present the proof that (2.21) implies independence. 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; random variable version of the coin toss model of Example 2.1.6 and an extension to biased coins.** Example 2.1.6 gave a probability space model for $n$ independent tosses of a fair coin. To model this with random variables instead, simply let $X_i = 1$ if toss $i$ results in heads, and $X_i = 0$, if it results in tails. Then the model is: the random variables $X_1, \ldots, X_N$ are independent, Bernoulli random variables each with parameter $p = 1/2$. The parameter $p = 1/2$ is chosen because in the probability model, the probability of heads on any toss is $1/2$. Using (2.21), if $s_1, \ldots, s_N$ is a sequence of 0’s and 1’s,

$$\mathbb{P}(X_1 = s_1, \ldots, X_N = s_N) = \mathbb{P}(X_1 = s_1) \cdots \mathbb{P}(X_N = s_N) = \left(\frac{1}{2}\right)^N$$

It is easy using random variables to extend the model to a biased coin, one in which the probability of heads is some value $p$ not equal to $1/2$. Simply let $X_1, \ldots, X_N$ be independent random variables each with mass function $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$. In this case, to obtain the joint mass function use (2.21) and the formula (2.19) for the mass function of a Bernoulli. Then

$$\mathbb{P}(X_1 = s_1, \ldots, X_N = s_N) = p^{s_1}(1-p)^{1-s_1} \cdots p^{s_N}(1-p)^{1-s_N} = p^{s_1+\cdots+s_N}(1-p)^{N-(s_1+\cdots+s_N)}. \quad (2.22)$$
2.2. RANDOM VARIABLES

To appreciate the convenience of random variables, the reader should, as an exercise, develop directly a probability space model for \( N \) flips of a biased coin.

\[ 2.23 \]

Example 2.2.4; random variable version of the independent site, equal base probability model of Example 2.1.7. Recall that in Example 2.1.7, a ten bp segment of DNA is randomly sampled and sequenced. Let \( X_i \) denote the base at site \( i \) along the segment—as per the convention of Chapter 1, the sequence is read in the 5’ – 3’ direction. The independent site/equal base probability model is equivalent to the assertion that \( X_1, \ldots, X_{10} \) are independent random variables with values in the DNA alphabet and each has the identical probability mass function

\[
p(A) = p(G) = p(C) = p(T) = \frac{1}{4}. \quad \diamond
\]

I.i.d. Sequences and Random Samples

The examples of the previous section each featured a sequence of independent random variables all sharing a common probability mass function. Such sequences are so important that they get a special name.

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

**Example 2.2.5.** The i.i.d. site model for DNA. This model generalizes example 2.2.4. Let \( X_1, \ldots, X_N \) denote the bases along a randomly selected, single-stranded DNA segment, read in the 5’ – 3’ direction. The i.i.d. site model is simply the assertion that \( X_1, \ldots, X_N \) is an i.i.d. sequence. The assumption (2.23) of equal base probabilities has been dropped. In fact, the i.i.d. site model allows any possible probability mass function

\[
p(A) = q_1, \quad p(G) = q_2, \quad p(C) = q_3, \quad p(T) = q_4,
\]

so it is really a family of probability models, indexed by the probabilities \( q_1, q_2, q_3, q_4 \) specifying the base probabilities. Notice that, despite the notation, there are really only three free parameters in the model; once \( q_1, q_2, \) and \( q_3 \) are given, \( q_4 \) is determined by \( q_4 = 1 - q_1 - q_2 - q_3 \).

\[ 2.24 \]

**Example 2.2.6.** Random variable formulation of a random sample of size \( n \). Let \( \Theta \) be a population each of whose members is labeled with a letter from a set \( S \). Consider a random sample of size \( n \) from the population. To express the random sample in terms of random variables, simply let \( X_i \) be the label of the \( i \)th individual sampled. The definition of a random sample in section 2.1.5 implies that \( X_1, \ldots, X_n \) are i.i.d., each with probability mass function

\[
p(x) = \mathbb{P}(X_i = x) = f_x, \quad x \in S,
\]

where \( f_x \) for \( x \in S \) are the population frequencies defined in section 2.1.
This last example shows that a multiple random sample of a population is a special case of a model with i.i.d random variables, or conversely, i.i.d. sequences are a generalization of random sampling. For this reason, statisticians often use random sampling terminology to refer to i.i.d. sequences, as in the following definition.

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

With this definition we could describe the model of Example 2.2.5 as a random sample of size \( N \) from the probability mass function defined by (2.24).

### 2.2.2 The basic discrete random variable types

The discrete random variables from which models are built are generally either Bernoulli, binomial, geometric, Poisson, or uniform. To say that a discrete random variable has a uniform distribution simply means that it takes on each of its possible values with equal probability. The Bernoulli random variable, modeling a simple coin toss, has already been described. The binomial, geometric, and Poisson are defined next. We explain how each of them can be interpreted in terms of independent sequences of Bernoulli random variables.

**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

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

(2.25)

The binomial distribution is important because it is the probability model for the number of successes (heads) in \( n \) trials (tosses), where the probability of success in each trial is \( p \). With parameters \( n \) and \( p \). To understand this, let \( X_1, X_2, \ldots, X_n \) be independent Bernoulli random variables with \( p = \Pr(X_i = 1) \) for each \( i \). An outcome of 1 for \( X_i \) represents a success on trial \( i \). The total number of success in the \( n \) trials is therefore the sum \( \sum_{i=1}^{n} X_i \), and we claim that it is a binomial random variable with parameters \( n \) and \( p \). The event \( \{\sum_1^n X_i = k\} \) occurs if and only if

\[
(X_1, \ldots, X_n) \in U \triangleq \{(s_1, \ldots, s_n) \in \{0, 1\}^n \mid s_1 + \cdots + s_n = k\}
\]

But for any \( (s_1, \ldots, s_n) \in U \), (so that \( s_1 + \cdots + s_n = k \)), we obtain from formula (2.22) for the joint mass function of independent Bernoulli’s,

\[
\Pr((X_1, \ldots, X_n) = (s_1, \ldots, s_n)) = p^{s_1+\cdots+s_n} (1 - p)^{n-(s_1+\cdots+s_n)} = p^k (1 - p)^{n-k},
\]

independent of \( (s_1, \ldots, s_n) \in U \). The cardinality of \( U \) in the number of ways to place exactly \( k \) 1’s in a sequence of length \( n \) and so is \( \binom{n}{k} \), the number of ways to choose \( k \) from \( n \). Hence,

\[
\Pr((X_1, \ldots, X_n) \in U) = |U| p^k (1 - p)^{n-k} = \binom{n}{k} p^k (1 - p)^{n-k},
\]
2.2. RANDOM VARIABLES

which is the binomial distribution.

**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

$$P(Y = k) = (1 - p)^{k-1} p \quad k = 1, 2, \ldots$$

(2.26)

Like the binomial random variable, the geometric random variable is also related to sequences of independent Bernoulli random variables. Indeed, let $X_1, X_2, \ldots$ be an infinite sequence of Bernoulli random variables, each with parameter $p$. Think of this sequence as indicating a sequence of success/failure trials. What is the probability that the first success occurs in trial $k$? This is just the probability of observing the sequence of outcomes $X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1$, and by independence, this probability is

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

which is the same as the expression in (2.26). Thus, the geometric random variable with parameter $p$ represents the time to the first success in a sequence of independent Bernoulli trials.

The Bernoulli trial interpretation can simplify calculations. Suppose we want to compute $P(Y > j)$ for a geometric random variable. This probability is the infinite sum

$$\sum_{j+1}^{\infty} P(Y = k) = \sum_{j+1}^{\infty} (1 - p)^{k-1} p.$$ 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 $1 - p$,

$$P(Y > j) = (1 - p)^j,$$

(2.27)

without having to do a sum. Of course, the sum is not hard to do; the reader should show directly that $\sum_{j+1}^{\infty} P(Y = k) = \sum_{j+1}^{\infty} (1 - p)^{k-1} p$ sums to $(1 - p)^j$.

Geometric random variables have an interesting property, called the memoryless property, which follows easily from (2.27). If $X$ is geometric with parameter $p$,

$$P(X > k + j \mid X > k) = \frac{P(X > k+j)}{P(X > k)} = \frac{(1 - p)^{k+j}}{(1 - p)^j} = (1 - p)^j = P(X > j)$$

(2.28)

Suppose you are playing a game of independent trials and $p$ is the probability that you win on trial $i$. Let $X$ be the first trial which you win; it is a geometric random variable. You have played $k$ times without success ($X > k$). But the game does not remember this. The probability of waiting an additional $j$ trials conditioned on $X > k$ is just $P(X > j)$, the probability of waiting at least $j$ without any conditioning.

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

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

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

(2.29)
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 4 when it is used.

**Multinomial distributions** This section on multinomial random variables may be skipped on a first reading. It is needed only infrequently in the text. We start with an example. Suppose that we have a box with three colors of marbles, 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, and let \( Y_1 \) be the number of reds drawn, \( Y_2 \) the number of greens drawn, and \( Y_3 \) the number of blues. The random variables \( Y_1, Y_2, Y_3 \) are not independent; indeed, they must satisfy \( Y_1 + Y_2 + Y_3 = n \). What is their joint distribution? It is:

\[
P(Y_1 = k_1, Y_2 = k_2, Y_3 = k_3) = \frac{n!}{k_1!k_2!k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3},
\]  

(2.30)

for any non-negative integers \( k_1, k_2, k_3 \) such that \( k_1 + k_2 + k_3 = n \). To see this, consider a specific sequence of \( n \) draws that results in \( k_1 \) red, \( k_2 \) green, and \( k_3 \) blue marbles. Since the draws are independent their probabilities multiply, and so the probability of any such sequence is \( p_1^{k_1} p_2^{k_2} p_3^{k_3} \). On the other hand there are a total of \( \frac{n!}{k_1!k_2!k_3!} \) different sequences of draws giving \( k_1 \) red, \( k_2 \) green, and \( k_3 \) blue marbles. Thus the total probability is given as in (2.30).

The general multinomial distribution is a generalization of formula (2.30). To state it, recall the general notation,

\[
\begin{pmatrix} n \\ k_1 \cdots k_r \end{pmatrix} \triangleq \frac{n!}{k_1! \cdots k_r!}.
\]

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 \( p_1 + \cdots + p_r = 1 \). 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

\[
P(Y_1 = k_1, \ldots, Y_r = k_r) = \begin{pmatrix} n \\ k_1 \cdots k_r \end{pmatrix} p_1^{k_1} \cdots p_r^{k_r},
\]  

(2.31)

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. Consider a random experiment with \( r \) possible outcomes \( s_1, s_2, \ldots, s_r \) and for each integer \( i, 0 \leq i \leq r \), let \( p_i \) be the probability of outcome \( s_i \). Consider \( n \) independent repetitions of the experiment, and let \( Y_1 \) be the number of trials that result in outcome \( s_1 \), \( Y_2 \) be the number of trials that result in outcome \( s_2 \), etc. 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.

When a random variable is not discrete its probabilistic behavior can no longer be described by a probability mass function, because its possible values are no longer restricted to a
2.2. RANDOM VARIABLES

discrete, and hence finite or countable, set. Rather, to characterize a general random variable \( X \), we define its (cumulative probability) distribution function,

\[
F_X(x) \triangleq \mathbb{P}(X \leq x).
\] (2.32)

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

\[
\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a).
\] (2.33)

The probability that \( X \) falls in the union of two disjoint intervals \((a, b]\) \( \cup \) \((c, d]\) is then \( \mathbb{P}(X \in (a, b]\) \( \cup \) \((c, d]\) = \( F_X(b) - F_X(a) + F_X(d) - F_X(a) \), and so on. For a discrete random variable taking values in the finite or countable set \( S \),

\[
F_X(b) - F_X(a) = \sum_{y \in S, a < y \leq b} p_X(y).
\] (2.34)

In particular, if \( s \in S \) and \((a, b]\) contains \( s \) but no other point of \( S \), then \( p_X(s) = F_X(b) - F_X(a) \), so \( p_X \) can be reconstructed from \( F_X \).

If \( F_X \) cannot be written in the form in (2.34), \( X \) will not be discrete. In particular, \( X \) is called a continuous random variable if there is a non-negative function \( f_X(x) \), called the probability density of \( X \), such that

\[
F_X(x) = \int_{-\infty}^{x} f_X(y) \, dy, \quad \text{for all } x.
\] (2.35)

By the Fundamental Theorem of Calculus,

\[
F_X'(x) = f_X(x).
\] (2.36)

For any \( a < b \), we can combine (2.34) and (2.35) to obtain

\[
\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_{a}^{b} f_X(y) \, dy.
\] (2.37)

Also by letting \( x \to \infty \) in (2.35),

\[
\int_{-\infty}^{\infty} f_X(x) \, dx = \mathbb{P}(X < \infty) = 1.
\] (2.38)

All probability density functions must satisfy this condition. Conversely, any non-negative function \( f \) satisfying (2.35) is called a probability density function, since we may define a random variable having \( f \) as its density.

If \( X \) is a continuous random variable, then for any point \( s \) and any \( h > 0 \),

\[
\mathbb{P}(X = s) \leq \mathbb{P}(s - h < X \leq s) = \int_{s-h}^{s} f_X(y) \, dy.
\]

Letting \( h \downarrow 0 \), the right-hand side tends to 0, and so \( \mathbb{P}(X = s) = 0 \). Thus the possible values of \( X \) cannot be restricted to a countable set \( S = \{s_1, s_2, \ldots\} \), for \( \mathbb{P}(X \in S) = \sum \mathbb{P}(X = s_i) = 0 \).
(It sounds like a contradiction to have a random variable with the property that $P(X = s) = 0$ for all $s$. Surely, when the trial that $X$ models occurs, there is some outcome! The resolution to this paradox is to think of continuous random variables as approximations to discrete random variables that can take on many, closely spaced values. Such approximations are discussed in the next section and arise in Chapter 4.)

Definite integrals of non-negative functions compute areas under curves. Thus formula (2.34) says that $P(a < X \leq b)$ is the area of the region defined by the graph $y = f_X(x)$ of the density function, the $x$-axis, and $x = a$ and $x = b$. It is vital to understand this point to understand continuous random variables. In particular, if $\triangle x$ is small, and $f_X$ is continuous in an interval about $x$,

$$P(x < X \leq x + \triangle x) \approx f_X(x)\triangle x,$$

as the right hand side is the area of the rectangle approximating the region under the density curve from $x$ to $x + \triangle x$. Comparing (2.37) to (2.34) shows that $f_X(y)dy$ plays a role for continuous random variables analogous to that of $p_X(y)$ for discrete random variables. This is a useful heuristic in applying intuition about discrete random variables to the continuous case many formulas for discrete random variables translate to corresponding formulas for continuous random variables by replacing $p_X(y)$ by $f_X(y)dy$, and summation by integration, just as in the transition from (2.34) to (2.37).

### 2.2.4 Basic continuous random variables.

**The uniform distribution.** This is the simplest continuous distribution. Consider a random variable $X$ which can take on any value in the finite interval $(\alpha, \beta)$. It is said to be uniformly distributed on $(\alpha, \beta)$ 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, $X$ will have no preference as to its location in $(\alpha, \beta)$. Indeed, if $\alpha \leq a < b \leq \beta$,

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

and this answer depends only on the length of $(a, b)$ not its position within $(\alpha, \beta)$.

**The exponential distribution.** The exponential distribution is a popular model for waiting times, that is, times between randomly occurring events. A random variable is said to have the exponential distribution with parameter $\lambda$, 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 0 for $x \leq 0$, an exponentially distributed random variable can only take positive values. Thus, if $X$ is exponentially distributed with parameter $\lambda$, and if $0 \leq a < b$,

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

In particular, if \( a \geq 0 \),
\[
\mathbb{P}(X > a) = \int_a^\infty \lambda e^{-\lambda x} \, dx = e^{-\lambda a}.
\] (2.39)

On the other hand suppose that \( X \) is a random variable for which this identity is true. Then \( F_X(a) = 1 - \mathbb{P}(X > a) = 1 - e^{-\lambda a} \). By differentiating both sides and using (2.36), we get \( f_X(a) = F'_X(a) = \lambda e^{-\lambda a} \) for \( a > 0 \), and so \( X \) must be exponential with parameter \( \lambda \). In short, (2.39) is an equivalent way to characterize exponential random variables.

Formula (2.39) also implies an exponential random variable has a memoryless property similar to that of a geometric random variable. Indeed,
\[
\mathbb{P}(X > t + s \mid X > s) = \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t}.
\] (2.40)

In fact, the relationship between exponential and geometric random variables is close, as we will see in Example 2.2.8 below.

The normal distribution A random variable \( Z \) is said to be normally distributed or Gaussian, if it has a probability density of the form
\[
\phi(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty,
\] (2.41)

Here \(-\infty < \mu < \infty \) and \( \sigma^2 > 0 \). The parameters \( \mu \) and \( \sigma^2 \) are respectively the mean and variance of the random variable with density \( \phi(x; \mu, \sigma^2) \); (the 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; see (2.38).

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
\[
\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.
\]

The notation \( \Phi(z) \) is used for the cumulative distribution function of the standard normal density:
\[
\Phi(z) \triangleq \mathbb{P}(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx.
\]

Unlike the uniform and exponential distributions, \( \Phi(z) \) admits no closed form expression for 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.

Normal random variables have a very important scaling property. If \( Z \) is a standard normal r.v. then \( W = \sigma Z + \mu \) will be a normal random variable with mean \( \mu \) and variance \( \sigma^2 \). And conversely, if \( X \) is normal with mean \( \mu \) and variance \( \sigma^2 \), then \((X - \mu)/\sigma\) is a standard normal random variable. We will show that this is true shortly. First we show how it can be used to calculate probabilities for any normal random variable using the distribution function \( \Phi \) for the standard normal.
We start with an example. Let $X$ be normal with mean 1 and variance 4 and find $\Pr(X \leq 2.5)$. We know that $Z = (X - \mu)/\sigma = (X - 1)/2$ is standard normal. Now, the event that $X \leq 2.5$ is the same as the event that $Z \leq (2.5 - 1)/2$, or, equivalently, $Z \leq .75$. Hence $\Pr(X \leq 2.5) = \Pr(Z \leq .75) = \Phi(.75) = 0.7734$.

The general case is a simple extension of this argument. Let $X$ be normal with mean $\mu$ and variance $\sigma^2$. 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

$$\Pr(a < X < b) = \Phi((b - \mu)/\sigma) - \Phi((a - \mu)/\sigma).$$

We now show that if $X$ is normal with mean $\mu$ and variance $\sigma^2$, then $Z = (X - \mu)/\sigma$ will be normal with mean 0 and variance 1. This is an exercise in manipulating integrals and in the definition of a probability density. We need to show that $f_Z(x)$ has the density given in (2.42) or, equivalently,

$$F_Z(x) = \Pr(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$ 

But, the density of $X$ is $f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$. Thus

$$\Pr(Z \leq x) = \Pr((X - \mu)/\sigma \leq x) = \Pr(X \leq \sigma x + \mu) = \int_{-\infty}^{\sigma x + \mu} \frac{1}{\sigma \sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2} dy.$$ 

In the last integral, make the substitution, $u = (y - \mu)/\sigma$. The result is

$$\Pr(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du,$$

as required.

**Gamma distribution** This distribution will also be used. A random variable is said to have the **gamma distribution** with parameters $\lambda > 0$ and $r > 0$ if its density has the form

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

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 $r = 1$. 

[Gamma distribution](#)
2.2. RANDOM VARIABLES

2.2.5 Approximation

Continuous random variables arise often in applications as approximations to random variables which are really discrete. This may sound odd considering the very different nature of the distributions of continuous and discrete random variables, so we need to specify in what sense one random variable approximates another. Without trying to be too precise, we will compare two random variables $X$ and $Y$ through the magnitude of the difference between their distribution functions,

$$| \mathbb{P}(X \leq x) - \mathbb{P}(Y \leq x) | = | F_X(x) - F_Y(x) |.$$

We will show examples in which one random variable is continuous, while the other is discrete, yet we can make this difference small for all $x$. When this is possible, typical probability computations with $X$ (expected value, probability of being in a set) will give answers very close to the same computations done with $Y$, and so one can use an approximating continuous random variable in place of a discrete one. But why do so in the first place? It is basically a matter of conceptual simplicity and calculational convenience. A continuous density model is easier to state and, because one has calculus at one’s disposal, easier to work with.

The examples that follow will be important in Chapter 4. The reader eager to get to the review of expectation and its applications in section 2.2.5 may skip them until that time.

Example 2.2.7. Approximation of a discrete uniform r.v. on an evenly spaced partition.

This example is relevant to models of DNA sequencing statistics explored in Chapter 4. Let the 'true' random variable $Y_N$ we are modeling take values in the set

$$S_N \triangleq \{ \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1 \}$$

obtained by partitioning $[0, 1]$ into $N$ equal subintervals. Suppose that $Y_N$ is uniformly distributed, that is, $Y_N$ takes on any one of its possible values $\frac{i}{N}$ with probability $\frac{1}{N}$. It is intuitively reasonable that if $N$ is large, $Y_N$ should be well approximated by a continuous random variable $U$ uniformly distributed on $[0, 1]$. Indeed, we can show that for all $b$,

$$| \mathbb{P}(Y_N \leq b) - \mathbb{P}(U \leq b) | \leq \frac{1}{N}.$$  (2.43)

Thus, replacing $Y_N$ by $U$ does not lead to large errors in probability computations. To see why (2.43) is true, note first that we only need to consider values of $b$ in $(0, 1)$: if $b \leq 0$, both $\mathbb{P}(Y_N \leq b)$ and $\mathbb{P}(U \leq b)$ are zero, while if $b > 1$, both probabilities are 1. Therefore let $b$ be a point in $(0, 1)$. Suppose that exactly the first $k$ points of the partition $S_N \triangleq \{ \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1 \}$ are less than or equal to $b$. This means that

$$\frac{k}{N} \leq b < \frac{k+1}{N},$$

and in particular that $0 \leq b - (k/N) < 1/N$. But,

$$\mathbb{P}(Y_N \leq b) = \frac{\text{no. points in } S_N \text{ less than or equal to } b}{N} = \frac{k}{N}.$$
and $\mathbb{P}(U \leq b) = b$. Thus,

$$\left| \mathbb{P}(Y_N \leq b) - \mathbb{P}(U \leq b) \right| = \left| \frac{k}{N} - b \right| < \frac{1}{N}.$$  

For the examples to follow, it is convenient to note

$$\left| \mathbb{P}(X \leq x) - \mathbb{P}(Y \leq x) \right| = \left| (1 - \mathbb{P}(X > x)) - (1 - \mathbb{P}(Y > x)) \right| = \left| \mathbb{P}(Y > x) - \mathbb{P}(X > x) \right|.$$  

The final bound does not depend on $j$, so it is true for all $x$. Notice that $\lim_{p \to 0^+} \ln(1/(1 - p)) = \ln 1 = 0$. Thus, when $p$ is small, (2.44) shows that the distributions functions of $Y$ and $X$ closely approximate one another.

**Example 2.2.9. Approximation of exponential by geometric.** In the previous example we started with a geometric random variable and found an exponential random variable to try to match it. The accuracy of the match was determined by the parameter $p$ of the geometric random variable. In this example, we start with a geometric random variable and approximate it by *scaled* geometric random variables in such a way that we can make the approximation as accurate as we like.

Let $X$ denote an exponential random variable with parameter $\lambda$. It is helpful to imagine that $X$ models a random waiting time. For example think of $X$ as the time of arrival of your next e-mail. Suppose now that you only check your email at intervals of length $1/N$; $N$ is an index of how frantic you are about checking email—high $N$ means frequent checking. Let $Y_N$ denote the first time you actually see the next email message. It is a discrete random variable taking values in the set $S_N = \{1/N, 2/N, \ldots\}$ and $Y_N = k/N$ if the message arrives in the time interval $((k-1)/N, k/N]$. Thus

$$\mathbb{P}(Y_N = k/N) = \mathbb{P}((k-1)/N < X \leq k/N) = e^{-\lambda (k-1)/N} - e^{-\lambda k/N} = e^{-(k-1)\lambda/N} \left(1 - e^{-\lambda/N}\right).$$
2.2. RANDOM VARIABLES

If we let $p_N \triangleq 1 - e^{-\lambda/N}$, we see that

$$\mathbb{P}(NY_N = k) = \mathbb{P}(Y_N = k/N) = (1 - p_N)^{k-1}p_N.$$  

This is says that $NY_N$ is a geometric random variable; we say that $Y_N$ itself is a scaled geometric random variable, since it is taking values at the points $1/N, 2/N, \ldots$ instead of at integer-valued points. For comparison to the previous example, observe that $\lambda = N \ln(1/(1 - p_N))$.  

When $N$ is large, it is intuitively clear that $Y_N$ and $X$ should be good approximations of each other. In fact, using an argument very similar to that of the previous example, one can derive that

$$|\mathbb{P}(Y_N > x) - \mathbb{P}(X > x)| < \frac{\lambda}{N}$$  

for all $x$. (2.45)

The difference in the distribution functions thus tends to 0 as $N$ tends to infinity. The details of showing (2.45) are left as an exercise.

2.2.6 Joint density functions

**Definition.** The 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

$$\mathbb{P}(a_1 < X \leq b_1, a_2 < Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y) \, dy \, dx,$$  

(2.46)

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

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

$$\mathbb{P}((X, Y) \in U) = \int_U f(x, y) \, dx \, dy$$  

(2.47)

if the double integral over the region $U$ is well-defined. Of course, the double integral may be interpreted as the volume of the region between the graph of $z = f(x, y)$ and the $(x, y)$-plane over the set $U$.

**Example 2.2.10** 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.
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 $1/2 \times \text{area}(U) = 1/4$. Thus

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

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

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

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

**Theorem 4** The jointly continuous random variables $X_1, \ldots, X_n$ are independent if and only if their joint density function $f$ factors as

$$
f(x_1, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n) \quad (2.48)
$$

The density function of Example 2.2.10 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 2.2.10 are independent random variables, $X$ being uniformly distributed on $(0,2)$ and $Y$ on $(0,1)$.

### 2.3 Expectation and Its Applications

It is important to remember our convention that $S$ is a subset of the real numbers, unless otherwise specified. In this section, we do not otherwise specify! The definition of expectation requires working with real-valued random variables.
2.3. EXPECTATION AND ITS APPLICATIONS

2.3.1 Definition

**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

$$E[X] \triangleq \sum_{s \in S} sp_X(s),$$

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

The expected value of $X$ is an average of the possible values $X$ can take on, where each possible value $s$ 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.

As a simple review example, consider a Bernoulli random variable $X$ with parameter $p$. Since $p_X(0) = 1 - p$ and $p_X(1) = p$, the expectation is

$$E[X] = 0 \cdot (1 - p) + 1 \cdot p = p.$$  \hspace{1cm} (2.50)

To understand this result intuitively, suppose that you win one dollar if the toss results in head (probability $p$) and win nothing if the toss results in tails. (A nice game to play!). Then $E[X] = p$ is the expected winnings per play.

Using Bernoulli random variables, one can express the probability of any event as an expectation, and this is an important and useful technique. Indeed, let $U$ be an event. Define the indicator random variable of $U$ by

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

This is a Bernoulli random variable with $p = P(I_U = 1) = P(U)$. Hence

$$E[I_U] = P(U).$$  \hspace{1cm} (2.51)

To define the expected value of a continuous random variable, we again want to average the possible values of $X$ by the probabilities they have of occurring. Following our heuristic, we replace the probability mass function $p_X$ in the (2.49) by $f_X(x) dx$ and replace the sum by an expectation.

**Definition** Let $X$ be a continuous random variable. Then,

$$E[X] \triangleq \int_{-\infty}^{\infty} xf_X(x) dx,$$

if the integral exists.

For example if $X$ is uniformly distributed on the interval $(0, 1)$, the density of $X$ equals one on the interval $(0, 1)$ and 0 elsewhere, so

$$E[X] = \int_{-\infty}^{\infty} f_X(x) dx = \int_{0}^{1} x dx = \frac{1}{2}.$$
CHAPTER 2. PROBABILITY THEORY

The answer makes perfect sense, because if $X$ shows no preference as to where it lies in $(0, 1)$, its average value should be $1/2$. As another example, if $Y$ is exponential with parameter $\lambda$, then applying integration by parts shows

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

Final, if $Z$ is a normal random variable with probability density $e^{-(x-\mu)^2/2\sigma^2}/\sqrt{2\pi}$, then

$$E[Z] = \int_{-\infty}^{\infty} x e^{-(x-\mu)^2/2\sigma^2} \frac{dx}{\sigma\sqrt{2\pi}} = \mu,$$

so the parameter $\mu$ is indeed the mean of $Z$. The reader should verify this equation as an exercise.

2.3.2 Law of Large Numbers

Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables. All random variables have the same distribution and hence the same expected value $\mu$; $(E[X_i] = \mu$ for each $i$). For each $n$, define the empirical mean of $X_1, \ldots, X_n$, as $\frac{1}{n} \sum_{i=1}^{n} X_i$.

This is the average of the first $n$ observed values of $X$ and generalizes the concept of empirical frequency introduced in section 2.1.5. The very important Law of Large Numbers says that in the long run, the empirical mean tends to the theoretical mean $\mu$ and hence justifies the definition of expected value in equations (2.49) and (2.52).

Theorem 5 Let $X_1, X_2, \ldots$ be an i.i.d. sequence. with a common finite mean $\mu$, assumed to be finite. Then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu \quad \text{with probability one.} \quad (2.53)$$

2.3.3 Properties of expectation.

Let $X$ be a random variable, and consider the random variable $g(X)$ obtained by applying function $g$ to $X$. To compute $E[g(X)]$ directly from the definition requires first finding the probability mass function or probability density of $g(X)$, according to whether it is discrete or continuous and then applying (2.49) or (2.52). But there is an easier way, sometimes called the law of the unconscious statistician, presumably because it allows one to compute expectations without thinking as hard.

Theorem 6 a) If $X$ is discrete and $E[g(X)]$ exists,

$$E[g(X)] = \sum_{s \in S} g(s)p_X(s). \quad (2.54)$$
2.3. EXPECTATION AND ITS APPLICATIONS

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

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)\,dx.$$  \hfill (2.55)

c) More generally,

$$\mathbb{E}[h(X_1,\ldots,X_n)] = \sum_{s_1 \in S} \cdots \sum_{s_n \in S} h(s_1,\ldots,s_n)p_Z(s_1,\ldots,s_n),$$  \hfill (2.56)

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

$$\mathbb{E}[h(X_1,\ldots,X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1,\ldots,x_n)f(x_1,\ldots,x_n)\,dx_1\cdots dx_n,$$  \hfill (2.57)

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

$$\mathbb{E}[c_1X_1 + \cdots + c_nX_n] = c_1\mathbb{E}[X_1] + \cdots + c_n\mathbb{E}[X_n].$$  \hfill (2.58)

**Theorem 8** (Products of independent random variables.) Let $X_1,\ldots,X_n$ be independent random variables. Then

$$\mathbb{E}[g_1(X_1)g_2(X_2)\cdots g_n(X_n)] = \mathbb{E}[g_1(X_1)]\cdots\mathbb{E}[g_n(X_n)]$$  \hfill (2.59)

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

Linearity of expectations is extremely useful. When the probability mass function of density of a random variable is complicated or difficult to compute exactly, it is hard to find its expectation directly from formula (2.49) or (2.52). But if a random variable $X$ can be represented as the sum of much simpler random variables, then linearity can be used to calculate its expectation easily, without finding the probability mass function or density of $X$.

**Example 2.3.1. The mean of a binomial r.v.** Let $X$ be binomial with parameters $n$ and $p$. According the definition in (2.49), $\mathbb{E}[X] = \sum_0^n k\binom{n}{k}p^k(1-p)^{n-k}$, which looks a bit complicated. However, we know that $X_1 + \cdots + X_n$ is binomial with parameters $n$ and $p$ if $X_1,\ldots,X_n$ are independent Bernoulli with parameter $p$. Since $\mathbb{E}[X_i] = p$ for each $i$,

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np.$$  \hfill $\diamond$

The linearity property has an important extension to integrals of random variables, which will be applied in Chapter 4. Suppose that for each point $r$ in an interval $[a,b]$, $Z(r)$ is a random variable. We can integrate the random random variables with respect to $r$ to obtain
a new random variable, \( Z = \int_a^b Z(r) \, dr \). The linearity property of expectations says that one can exchange expectation and integration:

\[
E \left[ \int_a^b Z(r) \, dr \right] = \int_a^b E[Z(r)] \, dr
\] (2.60)

To understand why this is true, recall that the definite integral is a limit of Riemann sums,

\[
\int_a^b Z(r) \, dr = \lim_{n \to \infty} \sum_{j=1}^n Z(r^n_j) \Delta_n r,
\]

where, for each \( n \), \( a = r^n_0 < r^n_1 < \ldots < r^n_n = b \) is a partition of \([a, b]\) into \( n \) equal subintervals of length \( \Delta_n r \). But by the linearity property, the expectation of the Riemann sum is

\[
E \left[ \sum_{j=1}^n Z(r^n_j) \Delta_n r \right] = \sum_{j=1}^n E[Z(r^n_j)] \Delta_n r.
\]

The right hand side is a Riemann sum and as \( n \to \infty \) it converges to the integral \( \int_a^b E[Z(r)] \, dr \).

Theorem 8 is a generalization to expectations of the probability formula

\[
P(U_1 \cap \cdots \cap U_n) = P(U_1)P(U_2) \cdots P(U_n)
\]

for independent events. We will show why it is true for the case of two independent, discrete random variables as an exercise using the law of the unconscious statistician. If \( X_1 \) and \( X_2 \) are independent and discrete, Theorem 3 says that their joint probability mass function is

\[
p_Z(s_1, s_2) = p_{X_1}(s_1)p_{X_2}(s_2).
\]

Thus, using formula (2.56)

\[
E[g_1(X_1)g_2(X_2)] = \sum_{s_1 \in S} \sum_{s_2 \in S} g_1(s_1)g_2(s_2)p_{X_1}(s_1)p_{X_2}(s_2)
\]

\[
= \left[ \sum_{s_1} g_1(s_1)p_{X_1}(s_1) \right] \left[ \sum_{s_2} g_2(s_2)p_{X_2}(s_2) \right]
\]

\[
= E[g_1(X_1)]E[g_2(X_2)].
\]

### 2.3.4 Variance and Covariance

The variance measures an average of the square distance of \( X \) from \( \mu_X = E[X] \):

\[
\text{Var}(X) \triangleq E \left[ (X - \mu_X)^2 \right].
\] (2.61)

The size of the variance indicates how close 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

\[
\text{Var}(cX) = E \left[ (cX - c\mu_X)^2 \right] = c^2 E \left[ (X - \mu_X)^2 \right] = c^2 \text{Var}(X).
\] (2.62)
2.3. EXPECTATION AND ITS APPLICATIONS

Second, using linearity of expectations,

$$\text{Var}(X) = E[X^2] - 2E[\mu_X X] + E[\mu_X^2] = E[X^2] - 2\mu_X E[X] + \mu_X^2$$

$$= E[X^2] - \mu_X^2. \quad (2.63)$$

The last two steps in the derivation use the fact that $\mu_X$ is a constant, so that $E[\mu_X X] = \mu_X E[X] = \mu_X^2$ and $E[\mu_X^2] = \mu_X^2$. As an example, the variance of a Bernoulli random variable with parameter $p$ is

$$E[X^2] - \mu_X^2 = (0^2 p_X(0) + 1^2 p_X(1)) - p^2 = p - p^2 = p(1 - p).$$

Another fundamental fact is that the variance of a normal random variable with density $e^{-(x - \mu)^2/2\sigma^2}/\sigma\sqrt{2\pi}$, is the parameter $\sigma^2$. The review of this fact is again left as an exercise.

The covariance of two random variables is:

$$\text{Cov}(X,Y) \triangleq E[(X - \mu_X)(Y - \mu_Y)]. \quad (2.64)$$

Similarly to (2.63), one can show that $\text{Cov}(X,Y) = E[XY] - \mu_X \mu_Y$.

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

if $X$ and $Y$ are independent then they are uncorrelated. \quad (2.65)

This is a consequence of the product formula of Theorem 8 for independent random variables: if $X$ and $Y$ are independent then

$$\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[X - \mu_X] E[Y - \mu_Y] = 0,$$

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 $\mu_i$ is short hand for $\mu_{X_i}$. 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,

$$\text{Var}(\sum_1^n X_i) = \sum_1^n \text{Var}(X_i) + \sum_{1 \leq i,j \leq n, i \neq j} \text{Cov}(X_i, X_j). \quad (2.66)$$

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

$$\text{Var}(\sum_1^n X_i) = \sum_1^n \text{Var}(X_i). \quad (2.67)$$
Example 2.3.2. 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 $\text{Var}(X_i) = p(1-p)$ for a Bernoulli random variable. Therefore the variance of a binomial random variable with parameters $n$ and $p$ is

$$\text{Var}(\sum_1^n X_i) = \sum_1^n \text{Var}(X_i) = n p(1-p). \qquad (2.68)$$

### 2.3.5 The Moment Generating Function.

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

$$M_X(t) \triangleq E[e^{tX}]$$

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 derivative 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 be able to do this it is only necessary to assume that $M_X(t)$ is finite for all values $t$ in some interval containing 0. Then, because $d^n/dt^n(e^{tx}) = x^n e^{tx},$

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

Setting $t = 0$ gives

$$M_X^{(n)}(0) = E[X^n e^0] = E[X^n], \quad (2.69)$$

where $M_X^{(n)}(t)$ denotes the derivative of order $n$ of $M_X(t)$.

**Example 2.3.3. Exponential random variables** The moment generating function of an exponential random variable with parameter $\lambda$ is

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

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

Moment generating functions are particularly suited for studying sums of independent random variables because of Theorem 8. Assume $X_1, \ldots, X_n$ are independent, and let $Z = X_1 + \cdots + X_n$. The identity $e^{tX} = e^{tX_1 e^{tX_2} \cdots e^{tX_n}}$ is elementary. Now apply the product formula of Theorem 8.

$$M_Z(t) = E[e^{tX_1} e^{tX_2} \cdots e^{tX_n}] = E[e^{tX_1}] E[e^{tX_2}] \cdots E[e^{tX_n}] = M_{X_1}(t) \cdots M_{X_n}(t). \quad (2.70)$$
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

$$M_Z(t) = E[e^{tX_1}]E[e^{tX_2}] \cdots E[e^{tX_n}] = M_X^n(t). \quad (2.71)$$

**Example 2.3.4. Bernoulli and binomial random variables.** The moment generating function of a Bernoulli random variable $X$ with probability $p$ of success is

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

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 (2.71) and the m.g.f. of the binomial r.v. , the m.g.f. of $Y$ is

$$M_Y(t) = (1 - p + pe^t)^n. \quad (2.72)$$

Using $M_Y(t)$ and formula (2.69) 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:

$$E[X] = M'(0) = n \left(1 - p + pe^t\right)^{n-1} pe^t \big|_{t=0} = np, \quad \text{and}$$  
$$\text{Var}(X) = E[X^2] - \mu^2_X = M''(0) - (np)^2$$

$$= n(n-1) \left(1 - p + pe^t\right)^{n-2} (pe^t)^2 + n \left(1 - p + pe^t\right)^{n-1} pe^t - (np)^2 \big|_{t=0}.$$  

$$= np(1-p) \quad \diamond \quad (2.74)$$

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

**Theorem 9** 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 2.3.5. Sums of independent normal r.v.’s.** The moment generating function of a normal random variable with mean $\mu$ and variance $\sigma^2$ 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 $\mu_1$ and variance $\sigma_1^2$ and let $X_2$ be normal with mean $\mu_2$ and variance $\sigma_2^2$. Suppose in addition that they are independent. Then, according to Theorem 8, the moment generating function of $X_1 + X_2$ is

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

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 9, $X_1 + X_2$ must be normal with this mean and variance.

The last example illustrates a special case of a very important theorem.
**Theorem 10** If \( X_1, \ldots, X_n \) are independent normal random variables, then \( \sigma_i X_i \) is normal with mean \( \mu_i \) and variance \( \text{Var}(X_i) \).

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

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean</th>
<th>Variance</th>
<th>M.g.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli (0,1)</td>
<td>( p )</td>
<td>( p(1-p) )</td>
<td>( 1 - p + pe^t )</td>
</tr>
<tr>
<td>Binomial ((n, p))</td>
<td>( np )</td>
<td>( np(1-p) )</td>
<td>((1-p+p^t)^n)</td>
</tr>
<tr>
<td>Poisson ((\lambda))</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( e^{-\lambda + \lambda e^t} )</td>
</tr>
<tr>
<td>Geometric ((p))</td>
<td>( \frac{1}{p} )</td>
<td>( \frac{1-p}{p^2} )</td>
<td>( \frac{pe^t}{1-(1-p)e^t} )</td>
</tr>
<tr>
<td>Uniform ((\alpha, \beta))</td>
<td>( \frac{\alpha + \beta}{2} )</td>
<td>( \frac{(\beta - \alpha)^2}{12} )</td>
<td>( \frac{t(\beta - \alpha)}{\lambda^2} )</td>
</tr>
<tr>
<td>Exponential ((\lambda))</td>
<td>( \frac{1}{\lambda} )</td>
<td>( \frac{1}{\lambda^2} )</td>
<td>( \frac{\lambda}{\lambda - t} )</td>
</tr>
<tr>
<td>Normal ((\mu, \sigma^2))</td>
<td>( \mu )</td>
<td>( \sigma^2 )</td>
<td>( e^{\mu t + \sigma^2 t^2/2} )</td>
</tr>
<tr>
<td>Gamma ((\lambda, r))</td>
<td>( \frac{r}{\lambda} )</td>
<td>( \frac{r}{\lambda^2} )</td>
<td>( \frac{\lambda^r}{(\lambda - t)^r} )</td>
</tr>
</tbody>
</table>

### 2.3.6 Tchebysheff’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

\[
P(X \geq a) \leq \frac{E[X]}{a}.
\]  

(2.75)

To show why this is true, we use two simple facts. First, if \( Y \) and \( Z \) are two random variables such that \( Y \geq Z \) with probability one, then \( E[Z] \leq E[Y] \). The second fact is stated in section 2.3.1, equation (2.51): \( P(U) = E[I_U] \), where \( I_U \) is the indicator random variable for event \( U \). Let \( U \) be the event \( \{X \geq a\} \). Then \( I_U = 1 \) if \( U \) occurs and 0 otherwise. We claim that \( I_U \leq a^{-1}X \) for all outcomes. Indeed, if \( U = \{X \geq a\} \) is the outcome, then \( X/a \geq 1 \), while \( I_U = 0 \); if \( U \) is not the outcome, \( I_U = 0 \), while \( X/a \geq 0 \) because \( X \) is assumed to be always non-negative. It follows then that

\[E[I_U] \leq E[X/a].\]
The left-hand side is just \( P(U) = P(X \geq a) \) and the right-hand side is \( E[X]/a \), and so we obtain Markov’s inequality.

**Tchebysheff’s inequality** is a consequence of Markov’s inequality. Let \( Y \) be a random variable with finite mean \( \mu_Y \) and variance \( \sigma^2 = \text{Var}(X) \). Let \( X = (Y - \mu_Y)^2 \). By applying Markov’s inequality to \( X \),

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

However, \( E[X] = E[(Y - \mu_Y)^2] = \text{Var}(Y) \). Thus we get Tchebysheff’s inequality,

\[
P\left( |Y - \mu_Y| \geq a \right) \leq \frac{\text{Var}(Y)}{a^2}.
\] (2.76)

This very useful bound shows how the variance of \( Y \) controls how close \( Y \) tends to be to its mean; in particular, if \( a \) is fixed, the smaller \( Y \) is, the smaller is the probability that \( Y \) differ by more than \( a \) from \( \mu_Y \), while if \( Y \) is fixed, the probability that \( |Y - \mu_Y| \) is larger than \( a \) decays at least as fast as \( 1/a^2 \) as \( a \to \infty \).

We have stated the all-important Law of Large Numbers several in several contexts. Using Tchebysheff’s inequality, we are now able to quantify how well the empirical mean approximates the expected value. Let \( X_1, X_2, \ldots \) be uncorrelated random variables all having mean \( \mu \) and variance \( \sigma^2 \). Let

\[
\hat{X}^{(n)} \triangleq \frac{1}{n} \sum_{1}^{n} X_i
\]
denote the empirical mean of \( X_1, \ldots, X_n \). The first step is to compute the mean and variance of \( \hat{X}^{(n)} \). For the mean, we use linearity of the expectation.

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

For the variance, we use (2.62 and (2.67):

\[
\text{Var}(\hat{X}^{(n)}) = \frac{1}{n^2} \text{Var} \left( \sum_{1}^{n} X_i \right) = \frac{1}{n^2} \sum_{1}^{n} \text{Var}(X_i) = \frac{n \sigma^2}{n^2} = \frac{\sigma^2}{n}.
\]

Now apply Tchebysheff’s inequality:

\[
P \left( |\hat{X}^{(n)} - \mu| > a \right) \leq \frac{\text{Var}(X^{(n)})}{a^2} = \frac{\sigma^2}{a^2}.
\] (2.77)

This is a fundamental identity. It implies that

\[
\lim_{n \to \infty} P \left( |\hat{X}^{(n)} - \mu| > a \right) = 0,
\] (2.78)

a result called the weak law of large numbers (it is subtly different from the Large Number Law stated previously which says that the empirical means tends to \( \mu \) with probability one.)
2.3.7 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

$$p_{X|Y}(x|y) \triangleq \mathbb{P}(X = x \mid Y = y),$$

where $x$ ranges over the possible values of $X$.

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

$$E[X \mid Y = y] \triangleq \sum_x x p_{X|Y}(x|y).$$

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

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

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

$$E[X \mid Y = y] \triangleq \int x f_{X|Y}(x|y) \, dx.$$

The law of the unconscious statistician—see Theorem 6—holds for conditional expectations. In the discrete and continuous cases, respectively,

$$E[g(X) \mid Y = y] = \sum_x g(x) p_{X|Y}(x|y), \quad E[g(X) \mid Y = y] = \int g(x) p_{X|Y}(x|y) \, dx.$$

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

**Theorem 11** For discrete and continuous random variables, respectively,

$$E[g(X)] = \sum_y E[g(X) \mid Y = y] p_Y(y) \quad \text{and} \quad E[g(X)] = \int E[g(X) \mid Y = y] f_Y(y) \, dy. \quad (2.79)$$

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

**Example 2.3.6.** Assume that $X$ and $Y$ are such that $Y$ is exponential with parameter $\lambda$ 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 $E[X \mid Y = y] = y/2$. Thus, using (2.79),

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

Notice that the last integral is just one-half the expected value of $Y$. \hfill \diamond
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 \( \mu \) and variance \( \sigma^2 \). Our goal is to understand the probability distribution of the sum \( \sum_{i=1}^{n} X_i \) 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 \( \sum_{i=1}^{n} X_i \) has mean \( n \mu \), and so

\[
\sum_{i=1}^{n} X_i - n \mu
\]

has a mean of zero. We also know that the variance of \( \sum_{i=1}^{n} X_i = n \sigma^2 \). Therefore, if we define

\[
Z^{(n)} \triangleq \frac{\sum_{i=1}^{n} X_i - n \mu}{\sigma \sqrt{n}},
\]

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

\[
\text{Var}(Z^{(n)}) = E \left[ \left( \frac{\sum_{i=1}^{n} X_i - n \mu}{\sigma \sqrt{n}} \right)^2 \right] = \frac{1}{n \sigma^2} E \left[ \left( \sum_{i=1}^{n} X_i - n \mu \right)^2 \right]
\]

\[
= \frac{1}{n \sigma^2} \text{Var} \left( \sum_{i=1}^{n} X_i \right) = \frac{1}{n \sigma^2} n \sigma^2 = 1.
\]

The Central Limit Theorem states that \( Z^{(n)} \) looks more and more like a standard normal r.v. as \( n \to \infty \).

**Theorem 12 The Central Limit Theorem.** Suppose \( X_1, X_2, \ldots \) are independent, identically distributed random variables with common mean \( \mu \) and variance \( \sigma^2 \). Then

\[
\lim_{n \to \infty} P \left( a < \frac{\sum_{i=1}^{n} X_i - n \mu}{\sigma \sqrt{n}} < b \right) = \int_{a}^{b} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}.
\] (2.80)

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

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 \( \sum_{i=1}^{n} X_i \) is binomial with parameters \( n \) and \( p \). Thus,

\[
P \left( a < \frac{Y_n - np}{\sqrt{np(1-p)}} < b \right) = P \left( a < \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} < b \right).
\]

By applying the Central Limit Theorem to the right-hand side, we obtain the following result.
Theorem 13 (DeMoivre-Laplace CLT) Suppose for each integer \( n \) that \( Y_n \) is a binomial random variable with parameters \( n \) and \( p \), \( p \) being fixed. Then

\[
\lim_{n \to \infty} P \left( a < \frac{Y_n - np}{\sqrt{np(1-p)}} < b \right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx. \tag{2.81}
\]

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 2.4.1 Let \( X \) be binomial with \( p = .5 \) and \( n = 50 \). What is \( P(22 \leq X \leq 28) \)?

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:

\[
P(22 \leq X \leq 28) = P(21.5 < X < 28.5).
\]

From Theorem 2 with \( n = 50 \) and \( p = .5 \), we have that \((X - 25)/\sqrt{12.5}\) is approximately standard normal. Thus

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

Using tables, this turns out to be 0.6778.

\[\diamond\]

2.5 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, 2006.