Chapter 1

Probability Theory

"You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."

Sherlock Holmes
The Sign of Four

The subject of probability theory is the foundation upon which all of statistics is built, providing a means for modeling populations, experiments, or almost anything else that could be considered a random phenomenon. Through these models, statisticians are able to draw inferences about populations, inferences based on examination of only a part of the whole.

The theory of probability has a long and rich history, dating back at least to the seventeenth century when, at the request of their friend, the Chevalier de Meré, Pascal and Fermat developed a mathematical formulation of gambling odds.

The aim of this chapter is not to give a thorough introduction to probability theory; such an attempt would be foolhardy in so short a space. Rather, we attempt to outline some of the basic ideas of probability theory that are fundamental to the study of statistics.

Just as statistics builds upon the foundation of probability theory, probability theory in turn builds upon set theory, which is where we begin.

1.1 Set Theory

One of the main objectives of a statistician is to draw conclusions about a population of objects by conducting an experiment. The first step in this endeavor is to identify the possible outcomes or, in statistical terminology, the sample space.

Definition 1.1.1 The set, S, of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

If the experiment consists of tossing a coin, the sample space contains two outcomes, heads and tails; thus,

$$S = \{\mathrm{H}, \mathrm{T}\}.$$

If, on the other hand, the experiment consists of observing the reported SAT scores of randomly selected students at a certain university, the sample space would be

the set of positive integers between 200 and 800 that are multiples of ten—that is, $S = \{200, 210, 220, \ldots, 780, 790, 800\}$. Finally, consider an experiment where the observation is reaction time to a certain stimulus. Here, the sample space would consist of all positive numbers, that is, $S = (0, \infty)$.

We can classify sample spaces into two types according to the number of elements they contain. Sample spaces can be either countable or uncountable; if the elements of a sample space can be put into 1–1 correspondence with a subset of the integers, the sample space is countable. Of course, if the sample space contains only a finite number of elements, it is countable. Thus, the coin-toss and SAT score sample spaces are both countable (in fact, finite), whereas the reaction time sample space is uncountable, since the positive real numbers cannot be put into 1–1 correspondence with the integers. If, however, we measured reaction time to the nearest second, then the sample space would be (in seconds) $S = \{0, 1, 2, 3, \ldots\}$, which is then countable.

This distinction between countable and uncountable sample spaces is important only in that it dictates the way in which probabilities can be assigned. For the most part, this causes no problems, although the mathematical treatment of the situations is different. On a philosophical level, it might be argued that there can only be countable sample spaces, since measurements cannot be made with infinite accuracy. (A sample space consisting of, say, all ten-digit numbers is a countable sample space.) While in practice this is true, probabilistic and statistical methods associated with uncountable sample spaces are, in general, less cumbersome than those for countable sample spaces, and provide a close approximation to the true (countable) situation.

Once the sample space has been defined, we are in a position to consider collections of possible outcomes of an experiment.

Definition 1.1.2 An *event* is any collection of possible outcomes of an experiment, that is, any subset of S (including S itself).

Let A be an event, a subset of S. We say the event A occurs if the outcome of the experiment is in the set A. When speaking of probabilities, we generally speak of the probability of an event, rather than a set. But we may use the terms interchangeably.

We first need to define formally the following two relationships, which allow us to order and equate sets:

$$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B; \qquad \text{(containment)}$$

$$A = B \Leftrightarrow A \subset B \text{ and } B \subset A. \qquad \text{(equality)}$$

Given any two events (or sets) A and B, we have the following elementary set operations:

Union: The union of A and B, written $A \cup B$, is the set of elements that belong to either A or B or both:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

Intersection: The intersection of A and B, written $A \cap B$, is the set of elements that belong to both A and B:

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

Complementation: The complement of A, written A^c, is the set of all elements that are not in A:

$$A^c = \{x: x \notin A\}.$$

Example 1.1.3 (Event operations) Consider the experiment of selecting a card at random from a standard deck and noting its suit: clubs (C), diamonds (D), hearts (H), or spades (S). The sample space is

$$S = \{\mathrm{C}, \mathrm{D}, \mathrm{H}, \mathrm{S}\},$$

and some possible events are

$$A = \{\mathrm{C}, \mathrm{D}\} \quad \text{and} \quad B = \{\mathrm{D}, \mathrm{H}, \mathrm{S}\}.$$

From these events we can form

$$A \cup B = \{\mathrm{C}, \mathrm{D}, \mathrm{H}, \mathrm{S}\}, \quad A \cap B = \{\mathrm{D}\}, \quad \text{and} \quad A^c = \{\mathrm{H}, \mathrm{S}\}.$$

Furthermore, notice that $A \cup B = S$ (the event S) and $(A \cup B)^c = \emptyset$, where $\emptyset$ denotes the *empty set* (the set consisting of no elements). ||

The elementary set operations can be combined, somewhat akin to the way addition and multiplication can be combined. As long as we are careful, we can treat sets as if they were numbers. We can now state the following useful properties of set operations.

Theorem 1.1.4 *For any three events, A, B, and C, defined on a sample space S,*

a. Commutativity	$A \cup B = B \cup A$, $A \cap B = B \cap A$;
b. Associativity	$A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$;
c. Distributive Laws	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
d. DeMorgan's Laws	$(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

Proof: The proof of much of this theorem is left as Exercise 1.3. Also, Exercises 1.9 and 1.10 generalize the theorem. To illustrate the technique, however, we will prove the Distributive Law:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

(You might be familiar with the use of Venn diagrams to "prove" theorems in set theory. We caution that although Venn diagrams are sometimes helpful in visualizing a situation, they do not constitute a formal proof.) To prove that two sets are equal, it must be demonstrated that each set contains the other. Formally, then

$$A \cap (B \cup C) = \{x \in S: x \in A \text{ and } x \in (B \cup C)\};$$

$$(A \cap B) \cup (A \cap C) = \{x \in S: x \in (A \cap B) \text{ or } x \in (A \cap C)\}.$$

We first show that $A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$. Let $x \in (A \cap (B \cup C))$. By the definition of intersection, it must be that $x \in (B \cup C)$, that is, either $x \in B$ or $x \in C$. Since x also must be in A, we have that either $x \in (A \cap B)$ or $x \in (A \cap C)$; therefore,

$$x \in ((A \cap B) \cup (A \cap C)),$$

and the containment is established.

Now assume $x \in ((A \cap B) \cup (A \cap C))$. This implies that $x \in (A \cap B)$ or $x \in (A \cap C)$. If $x \in (A \cap B)$, then x is in both A and B. Since $x \in B, x \in (B \cup C)$ and thus $x \in (A \cap (B \cup C))$. If, on the other hand, $x \in (A \cap C)$, the argument is similar, and we again conclude that $x \in (A \cap (B \cup C))$. Thus, we have established $(A \cap B) \cup (A \cap C) \subset A \cap (B \cup C)$, showing containment in the other direction and, hence, proving the Distributive Law. □

The operations of union and intersection can be extended to infinite collections of sets as well. If $A_1, A_2, A_3, \ldots$ is a collection of sets, all defined on a sample space S, then

$$\bigcup_{i=1}^{\infty} A_i = \{x \in S : x \in A_i \text{ for some } i\},$$

$$\bigcap_{i=1}^{\infty} A_i = \{x \in S : x \in A_i \text{ for all } i\}.$$

For example, let $S = (0, 1]$ and define $A_i = [(1/i), 1]$. Then

$$\begin{aligned} \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} [(1/i), 1] &= \{x \in (0,1] : x \in [(1/i), 1] \text{ for some } i\} \\ &= \{x \in (0,1]\} = (0,1]; \\ \bigcap_{i=1}^{\infty} A_i = \bigcap_{i=1}^{\infty} [(1/i), 1] &= \{x \in (0,1] : x \in [(1/i), 1] \text{ for all } i\} \\ &= \{x \in (0,1] : x \in [1,1]\} = \{1\}. \end{aligned}$$

(the point 1)

It is also possible to define unions and intersections over uncountable collections of sets. If Γ is an index set (a set of elements to be used as indices), then

$$\bigcup_{a \in \Gamma} A_a = \{x \in S : x \in A_a \text{ for some } a\},$$

$$\bigcap_{a \in \Gamma} A_a = \{x \in S : x \in A_a \text{ for all } a\}.$$

If, for example, we take $\Gamma = \{\text{all positive real numbers}\}$ and $A_a = (0, a]$, then $\cup_{a \in \Gamma} A_a = (0, \infty)$ is an uncountable union. While uncountable unions and intersections do not play a major role in statistics, they sometimes provide a useful mechanism for obtaining an answer (see Section 8.2.3).

Finally, we discuss the idea of a partition of the sample space.

Definition 1.1.5 Two events A and B are *disjoint* (or *mutually exclusive*) if $A \cap B = \emptyset$. The events $A_1, A_2, \ldots$ are *pairwise disjoint* (or *mutually exclusive*) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Disjoint sets are sets with no points in common. If we draw a Venn diagram for two disjoint sets, the sets do not overlap. The collection

$$A_i = [i, i+1), \quad i = 0, 1, 2, \ldots,$$

consists of pairwise disjoint sets. Note further that $\cup_{i=0}^{\infty} A_i = [0, \infty)$.

Definition 1.1.6 If $A_1, A_2, \ldots$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, \ldots$ forms a *partition* of S.

The sets $A_i = [i, i+1)$ form a partition of $[0, \infty)$. In general, partitions are very useful, allowing us to divide the sample space into small, nonoverlapping pieces.

1.2 Basics of Probability Theory

When an experiment is performed, the realization of the experiment is an outcome in the sample space. If the experiment is performed a number of times, different outcomes may occur each time or some outcomes may repeat. This "frequency of occurrence" of an outcome can be thought of as a probability. More probable outcomes occur more frequently. If the outcomes of an experiment can be described probabilistically, we are on our way to analyzing the experiment statistically.

In this section we describe some of the basics of probability theory. We do not define probabilities in terms of frequencies but instead take the mathematically simpler axiomatic approach. As will be seen, the axiomatic approach is not concerned with the interpretations of probabilities, but is concerned only that the probabilities are defined by a function satisfying the axioms. Interpretations of the probabilities are quite another matter. The "frequency of occurrence" of an event is one example of a particular interpretation of probability. Another possible interpretation is a subjective one, where rather than thinking of probability as frequency, we can think of it as a belief in the chance of an event occurring.

1.2.1 Axiomatic Foundations

For each event A in the sample space S we want to associate with A a number between zero and one that will be called the probability of A, denoted by $P(A)$. It would seem natural to define the domain of P (the set where the arguments of the function $P(\cdot)$ are defined) as all subsets of S; that is, for each $A \subset S$ we define $P(A)$ as the probability that A occurs. Unfortunately, matters are not that simple. There are some technical difficulties to overcome. We will not dwell on these technicalities; although they are of importance, they are usually of more interest to probabilists than to statisticians. However, a firm understanding of statistics requires at least a passing familiarity with the following.

Definition 1.2.1 A collection of subsets of S is called a *sigma algebra* (or *Borel field*), denoted by $\mathcal{B}$, if it satisfies the following three properties:

a. $\emptyset \in \mathcal{B}$ (the empty set is an element of $\mathcal{B}$).

b. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation).

c. If $A_1, A_2, \ldots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

The empty set $\emptyset$ is a subset of any set. Thus, $\emptyset \subset S$. Property (a) states that this subset is always in a sigma algebra. Since $S = \emptyset^c$, properties (a) and (b) imply that S is always in $\mathcal{B}$ also. In addition, from DeMorgan's Laws it follows that $\mathcal{B}$ is closed under countable intersections. If $A_1, A_2, \ldots \in \mathcal{B}$, then $A_1^c, A_2^c, \ldots \in \mathcal{B}$ by property (b), and therefore $\cup_{i=1}^{\infty} A_i^c \in \mathcal{B}$. However, using DeMorgan's Law (as in Exercise 1.9), we have

$$\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c = \bigcap_{i=1}^{\infty} A_i. \tag{1.2.1}$$

Thus, again by property (b), $\cap_{i=1}^{\infty} A_i \in \mathcal{B}$.

Associated with sample space S we can have many different sigma algebras. For example, the collection of the two sets $\{\emptyset, S\}$ is a sigma algebra, usually called the trivial sigma algebra. The only sigma algebra we will be concerned with is the smallest one that contains all of the open sets in a given sample space S.

Example 1.2.2 (Sigma algebra–I) If S is finite or countable, then these technicalities really do not arise, for we define for a given sample space S,

$$\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}.$$

If S has n elements, there are 2^n sets in $\mathcal{B}$ (see Exercise 1.14). For example, if $S = \{1, 2, 3\}$, then $\mathcal{B}$ is the following collection of $2^3 = 8$ sets:

$$\begin{array}{lll} \{1\} & \{1,2\} & \{1,2,3\} \\ \{2\} & \{1,3\} & \emptyset \\ \{3\} & \{2,3\} & \end{array}$$

‖

In general, if S is uncountable, it is not an easy task to describe $\mathcal{B}$. However, $\mathcal{B}$ is chosen to contain any set of interest.

Example 1.2.3 (Sigma algebra–II) Let $S = (-\infty, \infty)$, the real line. Then $\mathcal{B}$ is chosen to contain all sets of the form

$$[a, b], \quad (a, b], \quad (a, b), \quad \text{and} \quad [a, b)$$

for all real numbers a and b. Also, from the properties of $\mathcal{B}$, it follows that $\mathcal{B}$ contains all sets that can be formed by taking (possibly countably infinite) unions and intersections of sets of the above varieties. ‖

We are now in a position to define a probability function.

Definition 1.2.4 Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function P with domain $\mathcal{B}$ that satisfies

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$.
2. $P(S) = 1$.
3. If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

The three properties given in Definition 1.2.4 are usually referred to as the Axioms of Probability (or the Kolmogorov Axioms, after A. Kolmogorov, one of the fathers of probability theory). Any function P that satisfies the Axioms of Probability is called a probability function. The axiomatic definition makes no attempt to tell what particular function P to choose; it merely requires P to satisfy the axioms. For any sample space many different probability functions can be defined. Which one(s) reflects what is likely to be observed in a particular experiment is still to be discussed.

Example 1.2.5 (Defining probabilities–I) Consider the simple experiment of tossing a fair coin, so $S = \{\mathrm{H}, \mathrm{T}\}$. By a "fair" coin we mean a balanced coin that is equally as likely to land heads up as tails up, and hence the reasonable probability function is the one that assigns equal probabilities to heads and tails, that is,

$$P(\{\mathrm{H}\}) = P(\{\mathrm{T}\}). \tag{1.2.2}$$

Note that (1.2.2) does not follow from the Axioms of Probability but rather is outside of the axioms. We have used a symmetry interpretation of probability (or just intuition) to impose the requirement that heads and tails be equally probable. Since $S = \{\mathrm{H}\} \cup \{\mathrm{T}\}$, we have, from Axiom 1, $P(\{\mathrm{H}\} \cup \{\mathrm{T}\}) = 1$. Also, $\{\mathrm{H}\}$ and $\{\mathrm{T}\}$ are disjoint, so $P(\{\mathrm{H}\} \cup \{\mathrm{T}\}) = P(\{\mathrm{H}\}) + P(\{\mathrm{T}\})$ and

$$P(\{\mathrm{H}\}) + P(\{\mathrm{T}\}) = 1. \tag{1.2.3}$$

Simultaneously solving (1.2.2) and (1.2.3) shows that $P(\{\mathrm{H}\}) = P(\{\mathrm{T}\}) = \frac{1}{2}$.

Since (1.2.2) is based on our knowledge of the particular experiment, not the axioms, any nonnegative values for $P(\{\mathrm{H}\})$ and $P(\{\mathrm{T}\})$ that satisfy (1.2.3) define a legitimate probability function. For example, we might choose $P(\{\mathrm{H}\}) = \frac{1}{9}$ and $P(\{\mathrm{T}\}) = \frac{8}{9}$. ||

We need general methods of defining probability functions that we know will always satisfy Kolmogorov's Axioms. We do not want to have to check the Axioms for each new probability function, like we did in Example 1.2.5. The following gives a common method of defining a legitimate probability function.

Theorem 1.2.6 *Let $S = \{s_1, \ldots, s_n\}$ be a finite set. Let $\mathcal{B}$ be any sigma algebra of subsets of S. Let $p_1, \ldots, p_n$ be nonnegative numbers that sum to 1. For any $A \in \mathcal{B}$, define $P(A)$ by*

$$P(A) = \sum_{\{i: s_i \in A\}} p_i.$$

(The sum over an empty set is defined to be 0.) Then P is a probability function on $\mathcal{B}$. This remains true if $S = \{s_1, s_2, \ldots\}$ is a countable set.

Proof: We will give the proof for finite S. For any $A \in \mathcal{B}$, $P(A) = \sum_{\{i:s_i \in A\}} p_i \geq 0$, because every $p_i \geq 0$. Thus, Axiom 1 is true. Now,

$$P(S) = \sum_{\{i:s_i \in S\}} p_i = \sum_{i=1}^{n} p_i = 1.$$

Thus, Axiom 2 is true. Let $A_1, \ldots, A_k$ denote pairwise disjoint events. ($\mathcal{B}$ contains only a finite number of sets, so we need consider only finite disjoint unions.) Then,

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{\{j:s_j \in \cup_{i=1}^{k} A_i\}} p_j = \sum_{i=1}^{k} \sum_{\{j:s_j \in A_i\}} p_j = \sum_{i=1}^{k} P(A_i).$$

The first and third equalities are true by the definition of $P(A)$. The disjointedness of the A_is ensures that the second equality is true, because the same p_js appear exactly once on each side of the equality. Thus, Axiom 3 is true and Kolmogorov's Axioms are satisfied. □

The physical reality of the experiment might dictate the probability assignment, as the next example illustrates.

Example 1.2.7 (Defining probabilities–II) The game of darts is played by throwing a dart at a board and receiving a score corresponding to the number assigned to the region in which the dart lands. For a novice player, it seems reasonable to assume that the probability of the dart hitting a particular region is proportional to the area of the region. Thus, a bigger region has a higher probability of being hit.

Referring to Figure 1.2.1, we see that the dart board has radius r and the distance between rings is $r/5$. If we make the assumption that the board is always hit (see Exercise 1.7 for a variation on this), then we have

$$P(\text{scoring } i \text{ points}) = \frac{\text{Area of region } i}{\text{Area of dart board}}.$$

For example

$$P(\text{scoring 1 point}) = \frac{\pi r^2 - \pi(4r/5)^2}{\pi r^2} = 1 - \left(\frac{4}{5}\right)^2.$$

It is easy to derive the general formula, and we find that

$$P(\text{scoring } i \text{ points}) = \frac{(6-i)^2 - (5-i)^2}{5^2}, \quad i = 1, \ldots, 5,$$

independent of π and r. The sum of the areas of the disjoint regions equals the area of the dart board. Thus, the probabilities that have been assigned to the five outcomes sum to 1, and, by Theorem 1.2.6, this is a probability function (see Exercise 1.8). ||

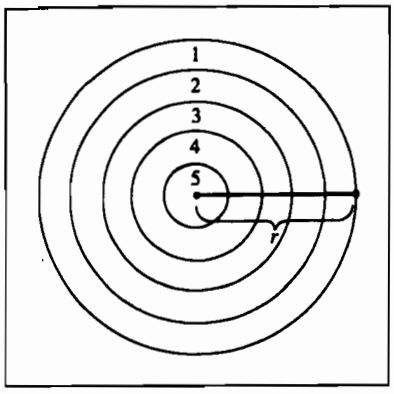


Figure 1.2.1. *Dart board for Example 1.2.7*

Before we leave the axiomatic development of probability, there is one further point to consider. Axiom 3 of Definition 1.2.4, which is commonly known as the Axiom of Countable Additivity, is not universally accepted among statisticians. Indeed, it can be argued that axioms should be simple, self-evident statements. Comparing Axiom 3 to the other axioms, which are simple and self-evident, may lead us to doubt whether it is reasonable to assume the truth of Axiom 3.

The Axiom of Countable Additivity is rejected by a school of statisticians led by deFinetti (1972), who chooses to replace this axiom with the Axiom of Finite Additivity.

Axiom of Finite Additivity: If $A \in \mathcal{B}$ and $B \in \mathcal{B}$ are disjoint, then

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

While this axiom may not be entirely self-evident, it is certainly simpler than the Axiom of Countable Additivity (and is implied by it – see Exercise 1.12).

Assuming only finite additivity, while perhaps more plausible, can lead to unexpected complications in statistical theory – complications that, at this level, do not necessarily enhance understanding of the subject. We therefore proceed under the assumption that the Axiom of Countable Additivity holds.

1.2.2 The Calculus of Probabilities

From the Axioms of Probability we can build up many properties of the probability function, properties that are quite helpful in the calculation of more complicated probabilities. Some of these manipulations will be discussed in detail in this section; others will be left as exercises.

We start with some (fairly self-evident) properties of the probability function when applied to a single event.

Theorem 1.2.8 *If P is a probability function and A is any set in $\mathcal{B}$, then*

a. $P(\emptyset) = 0$, *where $\emptyset$ is the empty set;*

b. $P(A) \leq 1$;

c. $P(A^c) = 1 - P(A)$.

Proof: It is easiest to prove (c) first. The sets A and A^c form a partition of the sample space, that is, $S = A \cup A^c$. Therefore,

$$P(A \cup A^c) = P(S) = 1 \tag{1.2.4}$$

by the second axiom. Also, A and A^c are disjoint, so by the third axiom,

$$P(A \cup A^c) = P(A) + P(A^c). \tag{1.2.5}$$

Combining (1.2.4) and (1.2.5) gives (c).

Since $P(A^c) \geq 0$, (b) is immediately implied by (c). To prove (a), we use a similar argument on $S = S \cup \emptyset$. (Recall that both S and $\emptyset$ are always in $\mathcal{B}$.) Since S and $\emptyset$ are disjoint, we have

$$1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset),$$

and thus $P(\emptyset) = 0$. □

Theorem 1.2.8 contains properties that are so basic that they also have the flavor of axioms, although we have formally proved them using only the original three Kolmogorov Axioms. The next theorem, which is similar in spirit to Theorem 1.2.8, contains statements that are not so self-evident.

Theorem 1.2.9 *If P is a probability function and A and B are any sets in $\mathcal{B}$, then*

a. $P(B \cap A^c) = P(B) - P(A \cap B)$;

b. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

c. *If $A \subset B$, then* $P(A) \leq P(B)$.

Proof: To establish (a) note that for any sets A and B we have

$$B = \{B \cap A\} \cup \{B \cap A^c\},$$

and therefore

$$P(B) = P(\{B \cap A\} \cup \{B \cap A^c\}) = P(B \cap A) + P(B \cap A^c), \tag{1.2.6}$$

where the last equality in (1.2.6) follows from the fact that $B \cap A$ and $B \cap A^c$ are disjoint. Rearranging (1.2.6) gives (a).

To establish (b), we use the identity

$$A \cup B = A \cup \{B \cap A^c\}. \tag{1.2.7}$$

A Venn diagram will show why (1.2.7) holds, although a formal proof is not difficult (see Exercise 1.2). Using (1.2.7) and the fact that A and $B \cap A^c$ are disjoint (since A and A^c are), we have

$$P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B) - P(A \cap B) \tag{1.2.8}$$

from (a).

If $A \subset B$, then $A \cap B = A$. Therefore, using (a) we have

$$0 \le P(B \cap A^c) = P(B) - P(A),$$

establishing (c). □

Formula (b) of Theorem 1.2.9 gives a useful inequality for the probability of an intersection. Since $P(A \cup B) \le 1$, we have from (1.2.8), after some rearranging,

$$P(A \cap B) \ge P(A) + P(B) - 1. \tag{1.2.9}$$

This inequality is a special case of what is known as *Bonferroni's Inequality* (Miller 1981 is a good reference). Bonferroni's Inequality allows us to bound the probability of a simultaneous event (the intersection) in terms of the probabilities of the individual events.

Example 1.2.10 (Bonferroni's Inequality) Bonferroni's Inequality is particularly useful when it is difficult (or even impossible) to calculate the intersection probability, but some idea of the size of this probability is desired. Suppose A and B are two events and each has probability .95. Then the probability that both will occur is bounded below by

$$P(A \cap B) \ge P(A) + P(B) - 1 = .95 + .95 - 1 = .90.$$

Note that unless the probabilities of the individual events are sufficiently large, the Bonferroni bound is a useless (but correct!) negative number. ‖

We close this section with a theorem that gives some useful results for dealing with a collection of sets.

Theorem 1.2.11 *If P is a probability function, then*

a. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ *for any partition* $C_1, C_2, \ldots$;

b. $P(\cup_{i=1}^{\infty} A_i) \le \sum_{i=1}^{\infty} P(A_i)$ *for any sets* $A_1, A_2, \ldots$. (*Boole's Inequality*)

Proof: Since $C_1, C_2, \ldots$ form a partition, we have that $C_i \cap C_j = \emptyset$ for all $i \neq j$, and $S = \cup_{i=1}^{\infty} C_i$. Hence,

$$A = A \cap S = A \cap \left(\bigcup_{i=1}^{\infty} C_i \right) = \bigcup_{i=1}^{\infty} (A \cap C_i),$$

where the last equality follows from the Distributive Law (Theorem 1.1.4). We therefore have

$$P(A) = P\left(\bigcup_{i=1}^{\infty}(A \cap C_i)\right).$$

Now, since the C_i are disjoint, the sets $A \cap C_i$ are also disjoint, and from the properties of a probability function we have

$$P\left(\bigcup_{i=1}^{\infty}(A \cap C_i)\right) = \sum_{i=1}^{\infty} P(A \cap C_i),$$

establishing (a).

To establish (b) we first construct a disjoint collection $A_1^*, A_2^*, \ldots$, with the property that $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$. We define A_i^* by

$$A_1^* = A_1, \quad A_i^* = A_i \backslash \left(\bigcup_{j=1}^{i-1} A_j\right), \quad i = 2, 3, \ldots,$$

where the notation $A \backslash B$ denotes the part of A that does not intersect with B. In more familiar symbols, $A \backslash B = A \cap B^c$. It should be easy to see that $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$, and we therefore have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i^*\right) = \sum_{i=1}^{\infty} P(A_i^*),$$

where the last equality follows since the A_i^* are disjoint. To see this, we write

$$\begin{aligned} A_i^* \cap A_k^* &= \left\{A_i \backslash \left(\bigcup_{j=1}^{i-1} A_j\right)\right\} \cap \left\{A_k \backslash \left(\bigcup_{j=1}^{k-1} A_j\right)\right\} && \text{(definition of } A_i^*) \\ &= \left\{A_i \cap \left(\bigcup_{j=1}^{i-1} A_j\right)^c\right\} \cap \left\{A_k \cap \left(\bigcup_{j=1}^{k-1} A_j\right)^c\right\} && \text{(definition of "\")} \\ &= \left\{A_i \cap \bigcap_{j=1}^{i-1} A_j^c\right\} \cap \left\{A_k \cap \bigcap_{j=1}^{k-1} A_j^c\right\} && \text{(DeMorgan's Laws)} \end{aligned}$$

Now if $i > k$, the first intersection above will be contained in the set A_k^c, which will have an empty intersection with A_k. If $k > i$, the argument is similar. Further, by construction $A_i^* \subset A_i$, so $P(A_i^*) \leq P(A_i)$ and we have

$$\sum_{i=1}^{\infty} P(A_i^*) \leq \sum_{i=1}^{\infty} P(A_i),$$

establishing (b). □

There is a similarity between Boole's Inequality and Bonferroni's Inequality. In fact, they are essentially the same thing. We could have used Boole's Inequality to derive (1.2.9). If we apply Boole's Inequality to A^c, we have

$$P\left(\bigcup_{i=1}^{n} A_i^c\right) \leq \sum_{i=1}^{n} P(A_i^c),$$

and using the facts that $\cup A_i^c = (\cap A_i)^c$ and $P(A_i^c) = 1 - P(A_i)$, we obtain

$$1 - P\left(\bigcap_{i=1}^{n} A_i\right) \leq n - \sum_{i=1}^{n} P(A_i).$$

This becomes, on rearranging terms,

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1), \tag{1.2.10}$$

which is a more general version of the Bonferroni Inequality of (1.2.9).

1.2.3 Counting

The elementary process of counting can become quite sophisticated when placed in the hands of a statistician. Most often, methods of counting are used in order to construct probability assignments on finite sample spaces, although they can be used to answer other questions also.

Example 1.2.12 (Lottery–I) For a number of years the New York state lottery operated according to the following scheme. From the numbers 1, 2, ..., 44, a person may pick any six for her ticket. The winning number is then decided by randomly selecting six numbers from the forty-four. To be able to calculate the probability of winning we first must count how many different groups of six numbers can be chosen from the forty-four. ||

Example 1.2.13 (Tournament) In a single-elimination tournament, such as the U.S. Open tennis tournament, players advance only if they win (in contrast to double-elimination or round-robin tournaments). If we have 16 entrants, we might be interested in the number of paths a particular player can take to victory, where a path is taken to mean a sequence of opponents. ||

Counting problems, in general, sound complicated, and often we must do our counting subject to many restrictions. The way to solve such problems is to break them down into a series of simple tasks that are easy to count, and employ known rules of combining tasks. The following theorem is a first step in such a process and is sometimes known as the Fundamental Theorem of Counting.

Theorem 1.2.14 *If a job consists of k separate tasks, the ith of which can be done in n_i ways, $i = 1, \ldots, k$, then the entire job can be done in $n_1 \times n_2 \times \cdots \times n_k$ ways.*

Proof: It suffices to prove the theorem for $k = 2$ (see Exercise 1.15). The proof is just a matter of careful counting. The first task can be done in n_1 ways, and for each of these ways we have n_2 choices for the second task. Thus, we can do the job in

$$\underbrace{(1 \times n_2) + (1 \times n_2) + \cdots + (1 \times n_2)}_{n_1 \text{ terms}} = n_1 \times n_2$$

ways, establishing the theorem for $k = 2$. □

Example 1.2.15 (Lottery–II) Although the Fundamental Theorem of Counting is a reasonable place to start, in applications there are usually more aspects of a problem to consider. For example, in the New York state lottery the first number can be chosen in 44 ways, and the second number in 43 ways, making a total of $44 \times 43 = 1{,}892$ ways of choosing the first two numbers. However, if a person is allowed to choose the same number twice, then the first two numbers can be chosen in $44 \times 44 = 1{,}936$ ways. ‖

The distinction being made in Example 1.2.15 is between counting *with replacement* and counting *without replacement*. There is a second crucial element in any counting problem, whether or not the ordering of the tasks is important. To illustrate with the lottery example, suppose the winning numbers are selected in the order 12, 37, 35, 9, 13, 22. Does a person who selected 9, 12, 13, 22, 35, 37 qualify as a winner? In other words, does the order in which the task is performed actually matter? Taking all of these considerations into account, we can construct a 2×2 table of possibilities:

Possible methods of counting

	Without replacement	With replacement
Ordered		
Unordered		

Before we begin to count, the following definition gives us some extremely helpful notation.

Definition 1.2.16 For a positive integer n, $n!$ (read n factorial) is the product of all of the positive integers less than or equal to n. That is,

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1.$$

Furthermore, we define $0! = 1$.

Let us now consider counting all of the possible lottery tickets under each of these four cases.

1. *Ordered, without replacement* From the Fundamental Theorem of Counting, the first number can be selected in 44 ways, the second in 43 ways, etc. So there are

$$44 \times 43 \times 42 \times 41 \times 40 \times 39 = \frac{44!}{38!} = 5{,}082{,}517{,}440$$

possible tickets.

2. *Ordered, with replacement* Since each number can now be selected in 44 ways (because the chosen number is replaced), there are

$$44 \times 44 \times 44 \times 44 \times 44 \times 44 = 44^6 = 7{,}256{,}313{,}856$$

possible tickets.

3. *Unordered, without replacement* We know the number of possible tickets when the ordering must be accounted for, so what we must do is divide out the redundant orderings. Again from the Fundamental Theorem, six numbers can be arranged in $6 \times 5 \times 4 \times 3 \times 2 \times 1$ ways, so the total number of unordered tickets is

$$\frac{44 \times 43 \times 42 \times 41 \times 40 \times 39}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{44!}{6!\,38!} = 7{,}059{,}052.$$

This form of counting plays a central role in much of statistics—so much, in fact, that it has earned its own notation.

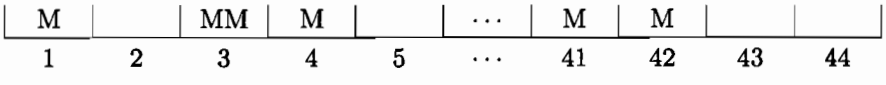
Definition 1.2.17 For nonnegative integers n and r, where $n \geq r$, we define the symbol $\binom{n}{r}$, read n *choose* r, as

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

In our lottery example, the number of possible tickets (unordered, without replacement) is $\binom{44}{6}$. These numbers are also referred to as *binomial coefficients*, for reasons that will become clear in Chapter 3.

4. *Unordered, with replacement* This is the most difficult case to count. You might first guess that the answer is $44^6/(6 \times 5 \times 4 \times 3 \times 2 \times 1)$, but this is not correct (it is too small).

To count in this case, it is easiest to think of placing 6 markers on the 44 numbers. In fact, we can think of the 44 numbers defining bins in which we can place the six markers, M, as shown, for example, in this figure.



The number of possible tickets is then equal to the number of ways that we can put the 6 markers into the 44 bins. But this can be further reduced by noting that all we need to keep track of is the arrangement of the markers and the walls of the bins. Note further that the two outermost walls play no part. Thus, we have to count all of the arrangements of 43 walls (44 bins yield 45 walls, but we disregard the two end walls) and 6 markers. We therefore have $43 + 6 = 49$ objects, which can be arranged in 49! ways. However, to eliminate the redundant orderings we must divide by both 6! and 43!, so the total number of arrangements is

$$\frac{49!}{6!\,43!} = 13{,}983{,}816.$$

Although all of the preceding derivations were done in terms of an example, it should be easy to see that they hold in general. For completeness, we can summarize these situations in Table 1.2.1.

Table 1.2.1. *Number of possible arrangements of size r from n objects*

	Without replacement	With replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

1.2.4 Enumerating Outcomes

The counting techniques of the previous section are useful when the sample space S is a finite set and all the outcomes in S are equally likely. Then probabilities of events can be calculated by simply counting the number of outcomes in the event. To see this, suppose that $S = \{s_1, \ldots, s_N\}$ is a finite sample space. Saying that all the outcomes are equally likely means that $P(\{s_i\}) = 1/N$ for every outcome s_i. Then, using Axiom 3 from Definition 1.2.4, we have, for any event A,

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{N} = \frac{\#\text{ of elements in } A}{\#\text{ of elements in } S}.$$

For large sample spaces, the counting techniques might be used to calculate both the numerator and denominator of this expression.

Example 1.2.18 (Poker) Consider choosing a five-card poker hand from a standard deck of 52 playing cards. Obviously, we are sampling without replacement from the deck. But to specify the possible outcomes (possible hands), we must decide whether we think of the hand as being dealt sequentially (ordered) or all at once (unordered). If we wish to calculate probabilities for events that depend on the order, such as the probability of an ace in the first two cards, then we must use the ordered outcomes. But if our events do not depend on the order, we can use the unordered outcomes. For this example we use the unordered outcomes, so the sample space consists of all the five-card hands that can be chosen from the 52-card deck. There are $\binom{52}{5} = 2{,}598{,}960$ possible hands. If the deck is well shuffled and the cards are randomly dealt, it is reasonable to assign probability $1/2{,}598{,}960$ to each possible hand.

We now calculate some probabilities by counting outcomes in events. What is the probability of having four aces? How many different hands are there with four aces? If we specify that four of the cards are aces, then there are 48 different ways of specifying the fifth card. Thus,

$$P(\text{four aces}) = \frac{48}{2{,}598{,}960},$$

less than 1 chance in 50,000. Only slightly more complicated counting, using Theorem 1.2.14, allows us to calculate the probability of having four of a kind. There are 13

ways to specify which denomination there will be four of. After we specify these four cards, there are 48 ways of specifying the fifth. Thus, the total number of hands with four of a kind is (13)(48) and

$$P(\text{four of a kind}) = \frac{(13)(48)}{2{,}598{,}960} = \frac{624}{2{,}598{,}960}.$$

To calculate the probability of exactly one pair (not two pairs, no three of a kind, etc.) we combine some of the counting techniques. The number of hands with exactly one pair is

$$13\binom{4}{2}\binom{12}{3}4^3 = 1{,}098{,}240. \tag{1.2.11}$$

Expression (1.2.11) comes from Theorem 1.2.14 because

$13 = \#$ of ways to specify the denomination for the pair,

$\binom{4}{2} = \#$ of ways to specify the two cards from that denomination,

$\binom{12}{3} = \#$ of ways of specifying the other three denominations,

$4^3 = \#$ of ways of specifying the other three cards from those denominations.

Thus,

$$P(\text{exactly one pair}) = \frac{1{,}098{,}240}{2{,}598{,}960}. \qquad \|$$

When sampling without replacement, as in Example 1.2.18, if we want to calculate the probability of an event that does not depend on the order, we can use either the ordered or unordered sample space. Each outcome in the unordered sample space corresponds to $r!$ outcomes in the ordered sample space. So, when counting outcomes in the ordered sample space, we use a factor of $r!$ in both the numerator and denominator that will cancel to give the same probability as if we counted in the unordered sample space.

The situation is different if we sample with replacement. Each outcome in the unordered sample space corresponds to some outcomes in the ordered sample space, but the number of outcomes differs.

Example 1.2.19 (Sampling with replacement) Consider sampling $r = 2$ items from $n = 3$ items, with replacement. The outcomes in the ordered and unordered sample spaces are these.

Unordered	$\{1,1\}$	$\{2,2\}$	$\{3,3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
Ordered	$(1,1)$	$(2,2)$	$(3,3)$	$(1,2),(2,1)$	$(1,3),(3,1)$	$(2,3),(3,2)$
Probability	1/9	1/9	1/9	2/9	2/9	2/9

The probabilities come from considering the nine outcomes in the ordered sample space to be equally likely. This corresponds to the common interpretation of "sampling with replacement"; namely, one of the three items is chosen, each with probability $1/3$; the item is noted and replaced; the items are mixed and again one of the three items is chosen, each with probability $1/3$. It is seen that the six outcomes in the unordered sample space are not equally likely under this kind of sampling. The formula for the number of outcomes in the unordered sample space is useful for enumerating the outcomes, but ordered outcomes must be counted to correctly calculate probabilities. ||

Some authors argue that it is appropriate to assign equal probabilities to the unordered outcomes when "randomly distributing r indistinguishable balls into n distinguishable urns." That is, an urn is chosen at random and a ball placed in it, and this is repeated r times. The order in which the balls are placed is not recorded so, in the end, an outcome such as $\{1,3\}$ means one ball is in urn 1 and one ball is in urn 3.

But here is the problem with this interpretation. Suppose two people observe this process, and Observer 1 records the order in which the balls are placed but Observer 2 does not. Observer 1 will assign probability $2/9$ to the event $\{1,3\}$. Observer 2, who is observing exactly the same process, should also assign probability $2/9$ to this event. But if the six unordered outcomes are written on identical pieces of paper and one is randomly chosen to determine the placement of the balls, then the unordered outcomes each have probability $1/6$. So Observer 2 will assign probability $1/6$ to the event $\{1,3\}$.

The confusion arises because the phrase "with replacement" will typically be interpreted with the sequential kind of sampling we described above, leading to assigning a probability $2/9$ to the event $\{1,3\}$. This is the correct way to proceed, as probabilities should be determined by the sampling mechanism, not whether the balls are distinguishable or indistinguishable.

Example 1.2.20 (Calculating an average) As an illustration of the distinguishable/indistinguishable approach, suppose that we are going to calculate all possible averages of four numbers selected from

$$2, 4, 9, 12$$

where we draw the numbers with replacement. For example, possible draws are $\{2,4,4,9\}$ with average 4.75 and $\{4,4,9,9\}$ with average 6.5. If we are only interested in the average of the sampled numbers, the ordering is unimportant, and thus the total number of distinct samples is obtained by counting according to unordered, with-replacement sampling.

The total number of distinct samples is $\binom{n+n-1}{n}$. But now, to calculate the probability distribution of the sampled averages, we must count the different ways that a particular average can occur.

The value 4.75 can occur only if the sample contains one 2, two 4s, and one 9. The number of possible samples that have this configuration is given in the following table:

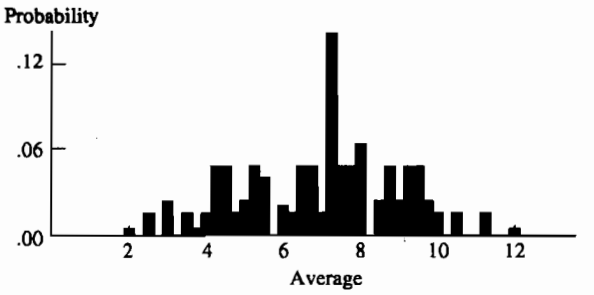


Figure 1.2.2. *Histogram of averages of samples with replacement from the four numbers* $\{2, 4, 4, 9\}$

Unordered	Ordered
$\{2, 4, 4, 9\}$	$(2,4,4,9), (2,4,9,4), (2,9,4,4), (4,2,4,9),$ $(4,2,9,4), (4,4,2,9), (4,4,9,2), (4,9,2,4),$ $(4,9,4,2), (9,2,4,4), (9,4,2,4), (9,4,4,2)$

The total number of ordered samples is $n^n = 4^4 = 256$, so the probability of drawing the unordered sample $\{2, 4, 4, 9\}$ is $12/256$. Compare this to the probability that we would have obtained if we regarded the unordered samples as equally likely – we would have assigned probability $1/\binom{n+n-1}{n} = 1/\binom{7}{4} = 1/35$ to $\{2, 4, 4, 9\}$ and to every other unordered sample.

To count the number of ordered samples that would result in $\{2, 4, 4, 9\}$, we argue as follows. We need to enumerate the possible orders of the four numbers $\{2, 4, 4, 9\}$, so we are essentially using counting method 1 of Section 1.2.3. We can order the sample in $4 \times 3 \times 2 \times 1 = 24$ ways. But there is a bit of double counting here, since we cannot count distinct arrangements of the two 4s. For example, the 24 ways would count $\{9, 4, 2, 4\}$ twice (which would be OK if the 4s were different). To correct this, we divide by 2! (there are 2! ways to arrange the two 4s) and obtain $24/2 = 12$ ordered samples. In general, if there are k places and we have m different numbers repeated $k_1, k_2, \ldots, k_m$ times, then the number of ordered samples is $\dfrac{k!}{k_1!k_2!\cdots k_m!}$. This type of counting is related to the *multinomial distribution*, which we will see in Section 4.6. Figure 1.2.2 is a histogram of the probability distribution of the sample averages, reflecting the multinomial counting of the samples.

There is also one further refinement that is reflected in Figure 1.2.2. It is possible that two different unordered samples will result in the same mean. For example, the unordered samples $\{4, 4, 12, 12\}$ and $\{2, 9, 9, 12\}$ both result in an average value of 8. The first sample has probability $3/128$ and the second has probability $3/64$, giving the value 8 a probability of $9/128 = .07$. See Example A.0.1 in Appendix A for details on constructing such a histogram. The calculation that we have done in this example is an elementary version of a very important statistical technique known as the *bootstrap* (Efron and Tibshirani 1993). We will return to the bootstrap in Section 10.1.4. ||

1.3 Conditional Probability and Independence

All of the probabilities that we have dealt with thus far have been unconditional probabilities. A sample space was defined and all probabilities were calculated with respect to that sample space. In many instances, however, we are in a position to update the sample space based on new information. In such cases, we want to be able to update probability calculations or to calculate *conditional probabilities.*

Example 1.3.1 (Four aces) Four cards are dealt from the top of a well-shuffled deck. What is the probability that they are the four aces? We can calculate this probability by the methods of the previous section. The number of distinct groups of four cards is

$$\binom{52}{4} = 270{,}725.$$

Only one of these groups consists of the four aces and every group is equally likely, so the probability of being dealt all four aces is $1/270{,}725$.

We can also calculate this probability by an "updating" argument, as follows. The probability that the first card is an ace is $4/52$. *Given that the first card is an ace,* the probability that the second card is an ace is $3/51$ (there are 3 aces and 51 cards left). Continuing this argument, we get the desired probability as

$$\frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{1}{270{,}725}. \qquad \|$$

In our second method of solving the problem, we updated the sample space after each draw of a card; we calculated conditional probabilities.

Definition 1.3.2 If A and B are events in S, and $P(B) > 0$, then the *conditional probability of A given B*, written $P(A|B)$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.3.1}$$

Note that what happens in the conditional probability calculation is that B becomes the sample space: $P(B|B) = 1$. The intuition is that our original sample space, S, has been updated to B. All further occurrences are then calibrated with respect to their relation to B. In particular, note what happens to conditional probabilities of disjoint sets. Suppose A and B are disjoint, so $P(A \cap B) = 0$. It then follows that $P(A|B) = P(B|A) = 0$.

Example 1.3.3 (Continuation of Example 1.3.1) Although the probability of getting all four aces is quite small, let us see how the conditional probabilities change given that some aces have already been drawn. Four cards will again be dealt from a well-shuffled deck, and we now calculate

$$P(\text{4 aces in 4 cards} \,|\, i \text{ aces in } i \text{ cards}), \quad i = 1, 2, 3.$$

The event {4 aces in 4 cards} is a subset of the event {i aces in i cards}. Thus, from the definition of conditional probability, (1.3.1), we know that

$$\begin{aligned} &P(\text{4 aces in 4 cards} \,|\, i \text{ aces in } i \text{ cards}) \\ &\quad = \frac{P(\{\text{4 aces in 4 cards}\} \cap \{i \text{ aces in } i \text{ cards}\})}{P(i \text{ aces in } i \text{ cards})} \\ &\quad = \frac{P(\text{4 aces in 4 cards})}{P(i \text{ aces in } i \text{ cards})}. \end{aligned}$$

The numerator has already been calculated, and the denominator can be calculated with a similar argument. The number of distinct groups of i cards is $\binom{52}{i}$, and

$$P(i \text{ aces in } i \text{ cards}) = \frac{\binom{4}{i}}{\binom{52}{i}}.$$

Therefore, the conditional probability is given by

$$P(\text{4 aces in 4 cards} \,|\, i \text{ aces in } i \text{ cards}) = \frac{\binom{52}{i}}{\binom{52}{4}\binom{4}{i}} = \frac{(4-i)!48!}{(52-i)!} = \frac{1}{\binom{52-i}{4-i}}.$$

For $i = 1$, 2, and 3, the conditional probabilities are .00005, .00082, and .02041, respectively. ||

For any B for which $P(B) > 0$, it is straightforward to verify that the probability function $P(\cdot|B)$ satisfies Kolmogorov's Axioms (see Exercise 1.35). You may suspect that requiring $P(B) > 0$ is redundant. Who would want to condition on an event of probability 0? Interestingly, sometimes this is a particularly useful way of thinking of things. However, we will defer these considerations until Chapter 4.

Conditional probabilities can be particularly slippery entities and sometimes require careful thought. Consider the following often-told tale.

Example 1.3.4 (Three prisoners) Three prisoners, A, B, and C, are on death row. The governor decides to pardon one of the three and chooses at random the prisoner to pardon. He informs the warden of his choice but requests that the name be kept secret for a few days.

The next day, A tries to get the warden to tell him who had been pardoned. The warden refuses. A then asks which of B or C will be executed. The warden thinks for a while, then tells A that B is to be executed.

Warden's reasoning: Each prisoner has a $\frac{1}{3}$ chance of being pardoned. Clearly, either B or C must be executed, so I have given A no information about whether A will be pardoned.

A's reasoning: Given that B will be executed, then either A or C will be pardoned. My chance of being pardoned has risen to $\frac{1}{2}$.

It should be clear that the warden's reasoning is correct, but let us see why. Let A, B, and C denote the events that A, B, or C is pardoned, respectively. We know

that $P(A) = P(B) = P(C) = \frac{1}{3}$. Let $\mathcal{W}$ denote the event that the warden says B will die. Using (1.3.1), A can update his probability of being pardoned to

$$P(A|\mathcal{W}) = \frac{P(A \cap \mathcal{W})}{P(\mathcal{W})}.$$

What is happening can be summarized in this table:

Prisoner pardoned	Warden tells A	
A	B dies	each with equal probability
A	C dies	
B	C dies	
C	B dies	

Using this table, we can calculate

$$\begin{aligned} P(\mathcal{W}) &= P(\text{warden says B dies}) \\ &= P(\text{warden says B dies and A pardoned}) \\ &\quad + P(\text{warden says B dies and C pardoned}) \\ &\quad + P(\text{warden says B dies and B pardoned}) \\ &= \frac{1}{6} + \frac{1}{3} + 0 \;=\; \frac{1}{2}. \end{aligned}$$

Thus, using the warden's reasoning, we have

$$\begin{aligned} P(A|\mathcal{W}) &= \frac{P(A \cap \mathcal{W})}{P(\mathcal{W})} \\ &= \frac{P(\text{warden says B dies and A pardoned})}{P(\text{warden says B dies})} \;=\; \frac{1/6}{1/2} \;=\; \frac{1}{3}. \end{aligned} \tag{1.3.2}$$

However, A falsely interprets the event $\mathcal{W}$ as equal to the event B^c and calculates

$$P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{1/3}{2/3} = \frac{1}{2}.$$

We see that conditional probabilities can be quite slippery and require careful interpretation. For some other variations of this problem, see Exercise 1.37. ||

Re-expressing (1.3.1) gives a useful form for calculating intersection probabilities,

$$P(A \cap B) = P(A|B)P(B), \tag{1.3.3}$$

which is essentially the formula that was used in Example 1.3.1. We can take advantage of the symmetry of (1.3.3) and also write

$$P(A \cap B) = P(B|A)P(A). \tag{1.3.4}$$

When faced with seemingly difficult calculations, we can break up our calculations according to (1.3.3) or (1.3.4), whichever is easier. Furthermore, we can equate the right-hand sides of these equations to obtain (after rearrangement)

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}, \tag{1.3.5}$$

which gives us a formula for "turning around" conditional probabilities. Equation (1.3.5) is often called Bayes' Rule for its discoverer, Sir Thomas Bayes (although see Stigler 1983).

Bayes' Rule has a more general form than (1.3.5), one that applies to partitions of a sample space. We therefore take the following as the definition of Bayes' Rule.

Theorem 1.3.5 (Bayes' Rule) *Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots$,*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}.$$

Example 1.3.6 (Coding) When coded messages are sent, there are sometimes errors in transmission. In particular, Morse code uses "dots" and "dashes," which are known to occur in the proportion of 3:4. This means that for any given symbol,

$$P(\text{dot sent}) = \frac{3}{7} \quad \text{and} \quad P(\text{dash sent}) = \frac{4}{7}.$$

Suppose there is interference on the transmission line, and with probability $\frac{1}{8}$ a dot is mistakenly received as a dash, and vice versa. If we receive a dot, can we be sure that a dot was sent? Using Bayes' Rule, we can write

$$P(\text{dot sent} \,|\, \text{dot received}) = P(\text{dot received} \,|\, \text{dot sent})\frac{P(\text{dot sent})}{P(\text{dot received})}.$$

Now, from the information given, we know that $P(\text{dot sent}) = \frac{3}{7}$ and $P(\text{dot received}|\ \text{dot sent}) = \frac{7}{8}$. Furthermore, we can also write

$$\begin{aligned} P(\text{dot received}) &= P(\text{dot received} \cap \text{dot sent}) + P(\text{dot received} \cap \text{dash sent}) \\ &= P(\text{dot received} \,|\, \text{dot sent})P(\text{dot sent}) \\ &\quad + P(\text{dot received} \,|\, \text{dash sent})P(\text{dash sent}) \\ &= \frac{7}{8} \times \frac{3}{7} + \frac{1}{8} \times \frac{4}{7} \;=\; \frac{25}{56}. \end{aligned}$$

Combining these results, we have that the probability of correctly receiving a dot is

$$P(\text{dot sent} \,|\, \text{dot received}) = \frac{(7/8) \times (3/7)}{25/56} = \frac{21}{25}. \qquad \|$$

In some cases it may happen that the occurrence of a particular event, B, has no effect on the probability of another event, A. Symbolically, we are saying that

$$P(A|B) = P(A). \tag{1.3.6}$$

If this holds, then by Bayes' Rule (1.3.5) and using (1.3.6) we have

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)} = P(A)\frac{P(B)}{P(A)} = P(B), \tag{1.3.7}$$

so the occurrence of A has no effect on B. Moreover, since $P(B|A)P(A) = P(A \cap B)$, it then follows that

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

which we take as the definition of statistical independence.

Definition 1.3.7 Two events, A and B, are *statistically independent* if

$$P(A \cap B) = P(A)P(B). \tag{1.3.8}$$

Note that independence could have been equivalently defined by either (1.3.6) or (1.3.7) (as long as either $P(A) > 0$ or $P(B) > 0$). The advantage of (1.3.8) is that it treats the events symmetrically and will be easier to generalize to more than two events.

Many gambling games provide models of independent events. The spins of a roulette wheel and the tosses of a pair of dice are both series of independent events.

Example 1.3.8 (Chevalier de Meré) The gambler introduced at the start of the chapter, the Chevalier de Meré, was particularly interested in the event that he could throw at least 1 six in 4 rolls of a die. We have

$$\begin{aligned} P(\text{at least 1 six in 4 rolls}) &= 1 - P(\text{no six in 4 rolls}) \\ &= 1 - \prod_{i=1}^{4} P(\text{no six on roll } i), \end{aligned}$$

where the last equality follows by independence of the rolls. On any roll, the probability of *not* rolling a six is $\frac{5}{6}$, so

$$P(\text{at least 1 six in 4 rolls}) = 1 - \left(\frac{5}{6}\right)^4 = .518. \qquad \|$$

Independence of A and B implies independence of the complements also. In fact, we have the following theorem.

Theorem 1.3.9 *If A and B are independent events, then the following pairs are also independent:*

a. *A and B^c,*

b. *A^c and B,*

c. *A^c and B^c.*

Proof: We will prove only (a), leaving the rest as Exercise 1.40. To prove (a) we must show that $P(A \cap B^c) = P(A)P(B^c)$. From Theorem 1.2.9a we have

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \qquad (A \text{ and } B \text{ are independent}) \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c). \end{aligned}$$

□

Independence of more than two events can be defined in a manner similar to (1.3.8), but we must be careful. For example, we might think that we could say A, B, and C are independent if $P(A \cap B \cap C) = P(A)P(B)P(C)$. However, this is not the correct condition.

Example 1.3.10 (Tossing two dice) Let an experiment consist of tossing two dice. For this experiment the sample space is

$$S = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(2,6),\ldots,(6,1),\ldots,(6,6)\};$$

that is, S consists of the 36 ordered pairs formed from the numbers 1 to 6. Define the following events:

$$\begin{aligned} A &= \{\text{doubles appear}\} = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}, \\ B &= \{\text{the sum is between 7 and 10}\}, \\ C &= \{\text{the sum is 2 or 7 or 8}\}. \end{aligned}$$

The probabilities can be calculated by counting among the 36 possible outcomes. We have

$$P(A) = \frac{1}{6}, \quad P(B) = \frac{1}{2}, \quad \text{and} \quad P(C) = \frac{1}{3}.$$

Furthermore,

$$\begin{aligned} P(A \cap B \cap C) &= P(\text{the sum is 8, composed of double 4s}) \\ &= \frac{1}{36} \\ &= \frac{1}{6} \times \frac{1}{2} \times \frac{1}{3} \\ &= P(A)P(B)P(C). \end{aligned}$$

However,

$$P(B \cap C) = P(\text{sum equals 7 or 8}) = \frac{11}{36} \neq P(B)P(C).$$

Similarly, it can be shown that $P(A \cap B) \neq P(A)P(B)$; therefore, the requirement $P(A \cap B \cap C) = P(A)P(B)P(C)$ is not a strong enough condition to guarantee pairwise independence. ||

A second attempt at a general definition of independence, in light of the previous example, might be to define A, B, and C to be independent if all the pairs are independent. Alas, this condition also fails.

Example 1.3.11 (Letters) Let the sample space S consist of the 3! permutations of the letters a, b, and c along with the three triples of each letter. Thus,

$$S = \begin{Bmatrix} \text{aaa} & \text{bbb} & \text{ccc} \\ \text{abc} & \text{bca} & \text{cba} \\ \text{acb} & \text{bac} & \text{cab} \end{Bmatrix}.$$

Furthermore, let each element of S have probability $\frac{1}{9}$. Define

$$A_i = \{i\text{th place in the triple is occupied by a}\}.$$

It is then easy to count that

$$P(A_i) = \frac{1}{3}, \quad i = 1, 2, 3,$$

and

$$P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{9},$$

so the A_is are pairwise independent. But

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{9} \neq P(A_1)P(A_2)P(A_3),$$

so the A_is do not satisfy the probability requirement. ||

The preceding two examples show that simultaneous (or mutual) independence of a collection of events requires an extremely strong definition. The following definition works.

Definition 1.3.12 A collection of events $A_1, \ldots, A_n$ are *mutually independent* if for any subcollection $A_{i_1}, \ldots, A_{i_k}$, we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P\left(A_{i_j}\right).$$

Example 1.3.13 (Three coin tosses–I) Consider the experiment of tossing a coin three times. A sample point for this experiment must indicate the result of each toss. For example, HHT could indicate that two heads and then a tail were observed. The sample space for this experiment has eight points, namely,

$$\{\text{HHH}, \text{HHT}, \text{HTH}, \text{THH}, \text{TTH}, \text{THT}, \text{HTT}, \text{TTT}\}.$$

Let H_i, $i = 1, 2, 3$, denote the event that the ith toss is a head. For example,

$$H_1 = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}\}. \tag{1.3.9}$$

If we assign probability $\frac{1}{8}$ to each sample point, then using enumerations such as (1.3.9), we see that $P(H_1) = P(H_2) = P(H_3) = \frac{1}{2}$. This says that the coin is fair and has an equal probability of landing heads or tails on each toss.

Under this probability model, the events H_1, H_2, and H_3 are also mutually independent. To verify this we note that

$$P(H_1 \cap H_2 \cap H_3) = P(\{\text{HHH}\}) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_2)P(H_3).$$

To verify the condition in Definition 1.3.12, we also must check each pair. For example,

$$P(H_1 \cap H_2) = P(\{\text{HHH}, \text{HHT}\}) = \frac{2}{8} = \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_2).$$

The equality is also true for the other two pairs. Thus, H_1, H_2, and H_3 are mutually independent. That is, the occurrence of a head on any toss has no effect on any of the other tosses.

It can be verified that the assignment of probability $\frac{1}{8}$ to each sample point is the only probability model that has $P(H_1) = P(H_2) = P(H_3) = \frac{1}{2}$ and H_1, H_2, and H_3 mutually independent. ‖

1.4 Random Variables

In many experiments it is easier to deal with a summary variable than with the original probability structure. For example, in an opinion poll, we might decide to ask 50 people whether they agree or disagree with a certain issue. If we record a "1" for agree and "0" for disagree, the sample space for this experiment has 2^{50} elements, each an ordered string of 1s and 0s of length 50. We should be able to reduce this to a reasonable size! It may be that the only quantity of interest is the number of people who agree (equivalently, disagree) out of 50 and, if we define a variable X = number of 1s recorded out of 50, we have captured the essence of the problem. Note that the sample space for X is the set of integers $\{0, 1, 2, \ldots, 50\}$ and is much easier to deal with than the original sample space.

In defining the quantity X, we have defined a mapping (a function) from the original sample space to a new sample space, usually a set of real numbers. In general, we have the following definition.

Definition 1.4.1 A *random variable* is a function from a sample space S into the real numbers.

Example 1.4.2 (Random variables) In some experiments random variables are implicitly used; some examples are these.

Examples of random variables

Experiment	Random variable
Toss two dice	X = sum of the numbers
Toss a coin 25 times	X = number of heads in 25 tosses
Apply different amounts of fertilizer to corn plants	X = yield/acre

In defining a random variable, we have also defined a new sample space (the range of the random variable). We must now check formally that our probability function, which is defined on the original sample space, can be used for the random variable.

Suppose we have a sample space

$$S = \{s_1, \ldots, s_n\}$$

with a probability function P and we define a random variable X with range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We can define a probability function P_X on $\mathcal{X}$ in the following way. We will observe $X = x_i$ if and only if the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_i$. Thus,

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}). \tag{1.4.1}$$

Note that the left-hand side of (1.4.1), the function P_X, is an *induced* probability function on $\mathcal{X}$, defined in terms of the original function P. Equation (1.4.1) formally defines a probability function, P_X, for the random variable X. Of course, we have to verify that P_X satisfies the Kolmogorov Axioms, but that is not a very difficult job (see Exercise 1.45). Because of the equivalence in (1.4.1), we will simply write $P(X = x_i)$ rather than $P_X(X = x_i)$.

A note on notation: Random variables will always be denoted with uppercase letters and the realized values of the variable (or its range) will be denoted by the corresponding lowercase letters. Thus, the random variable X can take the value x.

Example 1.4.3 (Three coin tosses–II) Consider again the experiment of tossing a fair coin three times from Example 1.3.13. Define the random variable X to be the number of heads obtained in the three tosses. A complete enumeration of the value of X for each point in the sample space is

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(s)$	3	2	2	2	1	1	1	0

The range for the random variable X is $\mathcal{X} = \{0, 1, 2, 3\}$. Assuming that all eight points in S have probability $\frac{1}{8}$, by simply counting in the above display we see that

the induced probability function on $\mathcal{X}$ is given by

x	0	1	2	3
$P_X(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

For example, $P_X(X = 1) = P(\{\text{HTT}, \text{THT}, \text{TTH}\}) = \frac{3}{8}$. ‖

Example 1.4.4 (Distribution of a random variable) It may be possible to determine P_X even if a complete listing, as in Example 1.4.3, is not possible. Let S be the 2^{50} strings of 50 0s and 1s, X = number of 1s, and $\mathcal{X} = \{0, 1, 2, \ldots, 50\}$, as mentioned at the beginning of this section. Suppose that each of the 2^{50} strings is equally likely. The probability that $X = 27$ can be obtained by counting all of the strings with 27 1s in the original sample space. Since each string is equally likely, it follows that

$$P_X(X = 27) = \frac{\#\text{ strings with 27 1s}}{\#\text{ strings}} = \frac{\binom{50}{27}}{2^{50}}.$$

In general, for any $i \in \mathcal{X}$,

$$P_X(X = i) = \frac{\binom{50}{i}}{2^{50}}.$$ ‖

The previous illustrations had both a finite S and finite $\mathcal{X}$, and the definition of P_X was straightforward. Such is also the case if $\mathcal{X}$ is countable. If $\mathcal{X}$ is uncountable, we define the induced probability function, P_X, in a manner similar to (1.4.1). For any set $A \subset \mathcal{X}$,

$$P_X(X \in A) = P(\{s \in S : X(s) \in A\}). \tag{1.4.2}$$

This does define a legitimate probability function for which the Kolmogorov Axioms can be verified. (To be precise, we use (1.4.2) to define probabilities only for a certain sigma algebra of subsets of $\mathcal{X}$. But we will not concern ourselves with these technicalities.)

1.5 Distribution Functions

With every random variable X, we associate a function called the cumulative distribution function of X.

Definition 1.5.1 The *cumulative distribution function* or *cdf* of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x.$$

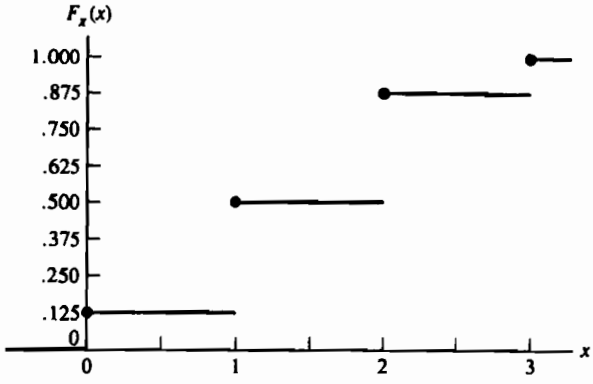


Figure 1.5.1. *Cdf of Example 1.5.2*

Example 1.5.2 (Tossing three coins) Consider the experiment of tossing three fair coins, and let X = number of heads observed. The cdf of X is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{8} & \text{if } 0 \le x < 1 \\ \frac{1}{2} & \text{if } 1 \le x < 2 \\ \frac{7}{8} & \text{if } 2 \le x < 3 \\ 1 & \text{if } 3 \le x < \infty. \end{cases} \tag{1.5.1}$$

The step function $F_X(x)$ is graphed in Figure 1.5.1. There are several points to note from Figure 1.5.1. F_X is defined for all values of x, not just those in $\mathcal{X} = \{0, 1, 2, 3\}$. Thus, for example,

$$F_X(2.5) = P(X \le 2.5) = P(X = 0, 1, \text{or } 2) = \frac{7}{8}.$$

Note that F_X has jumps at the values of $x_i \in \mathcal{X}$ and the size of the jump at x_i is equal to $P(X = x_i)$. Also, $F_X(x) = 0$ for $x < 0$ since X cannot be negative, and $F_X(x) = 1$ for $x \ge 3$ since x is certain to be less than or equal to such a value. ||

As is apparent from Figure 1.5.1, F_X can be discontinuous, with jumps at certain values of x. By the way in which F_X is defined, however, at the jump points F_X takes the value at the top of the jump. (Note the different inequalities in (1.5.1).) This is known as *right-continuity*—the function is continuous when a point is approached from the right. The property of right-continuity is a consequence of the definition of the cdf. In contrast, if we had defined $F_X(x) = P_X(X < x)$ (note strict inequality), F_X would then be *left-continuous*. The size of the jump at any point x is equal to $P(X = x)$.

Every cdf satisfies certain properties, some of which are obvious when we think of the definition of $F_X(x)$ in terms of probabilities.

Theorem 1.5.3 *The function $F(x)$ is a cdf if and only if the following three conditions hold:*

a. $\lim_{x\to-\infty}F(x)=0$ *and* $\lim_{x\to\infty}F(x)=1$.

b. $F(x)$ *is a nondecreasing function of* x.

c. $F(x)$ *is right-continuous; that is, for every number* x_0, $\lim_{x\downarrow x_0}F(x)=F(x_0)$.

Outline of proof: To prove necessity, the three properties can be verified by writing F in terms of the probability function (see Exercise 1.48). To prove sufficiency, that if a function F satisfies the three conditions of the theorem then it is a cdf for some random variable, is much harder. It must be established that there exists a sample space S, a probability function P on S, and a random variable X defined on S such that F is the cdf of X. □

Example 1.5.4 (Tossing for a head) Suppose we do an experiment that consists of tossing a coin until a head appears. Let p = probability of a head on any given toss, and define a random variable X = number of tosses required to get a head. Then, for any $x=1,2,\ldots,$

$$P(X=x)=(1-p)^{x-1}p, \tag{1.5.2}$$

since we must get $x-1$ tails followed by a head for the event to occur and all trials are independent. From (1.5.2) we calculate, for any positive integer x,

$$P(X\le x)=\sum_{i=1}^{x}P(X=i)=\sum_{i=1}^{x}(1-p)^{i-1}p. \tag{1.5.3}$$

The partial sum of the geometric series is

$$\sum_{k=1}^{n}t^{k-1}=\frac{1-t^n}{1-t},\quad t\neq 1, \tag{1.5.4}$$

a fact that can be established by induction (see Exercise 1.50). Applying (1.5.4) to our probability, we find that the cdf of the random variable X is

$$\begin{aligned}F_X(x)&=P(X\le x)\\&=\frac{1-(1-p)^x}{1-(1-p)}p\\&=1-(1-p)^x,\quad x=1,2,\ldots\ .\end{aligned}$$

The cdf $F_X(x)$ is flat between the nonnegative integers, as in Example 1.5.2.

It is easy to show that if $0<p<1$, then $F_X(x)$ satisfies the conditions of Theorem 1.5.3. First,

$$\lim_{x\to-\infty}F_X(x)=0$$

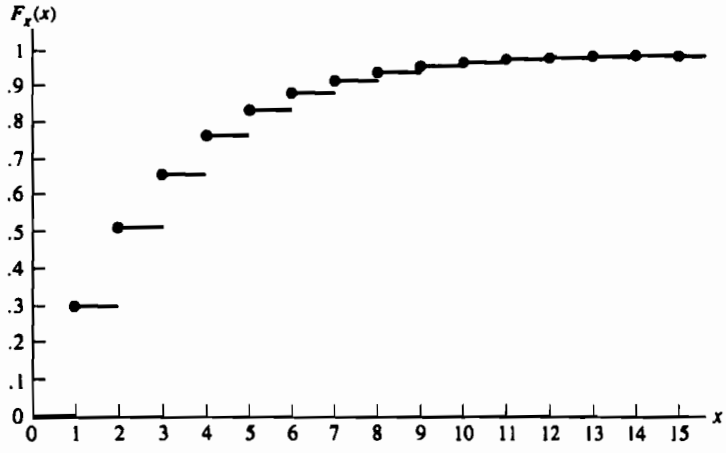


Figure 1.5.2. *Geometric cdf, $p = .3$*

since $F_X(x) = 0$ for all $x < 0$, and

$$\lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} 1 - (1-p)^x = 1,$$

where x goes through only integer values when this limit is taken. To verify property (b), we simply note that the sum in (1.5.3) contains more *positive* terms as x increases. Finally, to verify (c), note that, for any x, $F_X(x+\epsilon) = F_X(x)$ if $\epsilon > 0$ is sufficiently small. Hence,

$$\lim_{\epsilon\downarrow 0} F_X(x+\epsilon) = F_X(x),$$

so $F_X(x)$ is right-continuous. $F_X(x)$ is the cdf of a distribution called the *geometric distribution* (after the series) and is pictured in Figure 1.5.2. ||

Example 1.5.5 (Continuous cdf) An example of a continuous cdf is the function

$$F_X(x) = \frac{1}{1+e^{-x}}, \tag{1.5.5}$$

which satisfies the conditions of Theorem 1.5.3. For example,

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \text{since} \quad \lim_{x\to-\infty} e^{-x} = \infty$$

and

$$\lim_{x\to\infty} F_X(x) = 1 \quad \text{since} \quad \lim_{x\to\infty} e^{-x} = 0.$$

Differentiating $F_X(x)$ gives

$$\frac{d}{dx} F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2} > 0,$$

showing that $F_X(x)$ is increasing. F_X is not only right-continuous, but also continuous. This is a special case of the *logistic distribution.* ||

Example 1.5.6 (Cdf with jumps) If F_X is not a continuous function of x, it is possible for it to be a mixture of continuous pieces and jumps. For example, if we modify $F_X(x)$ of (1.5.5) to be, for some $\epsilon, 1 > \epsilon > 0$,

$$F_Y(y) = \begin{cases} \dfrac{1-\epsilon}{1+e^{-y}} & \text{if } y < 0 \\ \epsilon + \dfrac{(1-\epsilon)}{1+e^{-y}} & \text{if } y \geq 0, \end{cases} \tag{1.5.6}$$

then $F_Y(y)$ is the cdf of a random variable Y (see Exercise 1.47). The function F_Y has a jump of height ϵ at $y = 0$ and otherwise is continuous. This model might be appropriate if we were observing the reading from a gauge, a reading that could (theoretically) be anywhere between $-\infty$ and ∞. This particular gauge, however, sometimes sticks at 0. We could then model our observations with F_Y, where ϵ is the probability that the gauge sticks. ||

Whether a cdf is continuous or has jumps corresponds to the associated random variable being continuous or not. In fact, the association is such that it is convenient to define continuous random variables in this way.

Definition 1.5.7 A random variable X is *continuous* if $F_X(x)$ is a continuous function of x. A random variable X is *discrete* if $F_X(x)$ is a step function of x.

We close this section with a theorem formally stating that F_X completely determines the probability distribution of a random variable X. This is true if $P(X \in A)$ is defined only for events A in $\mathcal{B}^1$, the smallest sigma algebra containing all the intervals of real numbers of the form (a, b), $[a, b)$, $(a, b]$, and $[a, b]$. If probabilities are defined for a larger class of events, it is possible for two random variables to have the same distribution function but not the same probability for every event (see Chung 1974, page 27). In this book, as in most statistical applications, we are concerned only with events that are intervals, countable unions or intersections of intervals, etc. So we do not consider such pathological cases. We first need the notion of two random variables being identically distributed.

Definition 1.5.8 The random variables X and Y are *identically distributed* if, for every set $A \in \mathcal{B}^1$, $P(X \in A) = P(Y \in A)$.

Note that two random variables that are identically distributed are not necessarily equal. That is, Definition 1.5.8 does not say that $X = Y$.

Example 1.5.9 (Identically distributed random variables) Consider the experiment of tossing a fair coin three times as in Example 1.4.3. Define the random variables X and Y by

$$X = \text{number of heads observed} \quad \text{and} \quad Y = \text{number of tails observed.}$$

The distribution of X is given in Example 1.4.3, and it is easily verified that the distribution of Y is exactly the same. That is, for each $k = 0, 1, 2, 3$, we have $P(X = k) = P(Y = k)$. So X and Y are identically distributed. However, for no sample points do we have $X(s) = Y(s)$. ‖

Theorem 1.5.10 *The following two statements are equivalent:*

a. *The random variables X and Y are identically distributed.*
b. $F_X(x) = F_Y(x)$ *for every x.*

Proof: To show equivalence we must show that each statement implies the other. We first show that (a) $\Rightarrow$ (b).

Because X and Y are identically distributed, for any set $A \in \mathcal{B}^1$, $P(X \in A) = P(Y \in A)$. In particular, for every x, the set $(-\infty, x]$ is in $\mathcal{B}^1$, and

$$F_X(x) = P(X \in (-\infty, x]) = P(Y \in (-\infty, x]) = F_Y(x).$$

The converse implication, that (b) $\Rightarrow$ (a), is much more difficult to prove. The above argument showed that if the X and Y probabilities agreed on all sets, then they agreed on intervals. We now must prove the opposite; that is, if the X and Y probabilities agree on all intervals, then they agree on all sets. To show this requires heavy use of sigma algebras; we will not go into these details here. Suffice it to say that it is necessary to prove only that the two probability functions agree on all intervals (Chung 1974, Section 2.2). □

1.6 Density and Mass Functions

Associated with a random variable X and its cdf F_X is another function, called either the probability density function (pdf) or probability mass function (pmf). The terms pdf and pmf refer, respectively, to the continuous and discrete cases. Both pdfs and pmfs are concerned with "point probabilities" of random variables.

Definition 1.6.1 The *probability mass function (pmf)* of a discrete random variable X is given by

$$f_X(x) = P(X = x) \quad \text{for all } x.$$

Example 1.6.2 (Geometric probabilities) For the geometric distribution of Example 1.5.4, we have the pmf

$$f_X(x) = P(X = x) = \begin{cases} (1-p)^{x-1}p & \text{for } x = 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $P(X = x)$ or, equivalently, $f_X(x)$ is the size of the jump in the cdf at x. We can use the pmf to calculate probabilities. Since we can now measure the probability of a single point, we need only sum over all of the points in the appropriate event. Hence, for positive integers a and b, with $a \le b$, we have

$$P(a \le X \le b) = \sum_{k=a}^{b} f_X(k) = \sum_{k=a}^{b} (1-p)^{k-1}p.$$

As a special case of this we get

$$P(X \le b) = \sum_{k=1}^{b} f_X(k) = F_X(b). \tag{1.6.1}$$ ||

A widely accepted convention, which we will adopt, is to use an uppercase letter for the cdf and the corresponding lowercase letter for the pmf or pdf.

We must be a little more careful in our definition of a pdf in the continuous case. If we naively try to calculate $P(X = x)$ for a continuous random variable, we get the following. Since $\{X = x\} \subset \{x - \epsilon < X \le x\}$ for any $\epsilon > 0$, we have from Theorem 1.2.9(c) that

$$P(X = x) \le P(x - \epsilon < X \le x) = F_X(x) - F_X(x - \epsilon)$$

for any $\epsilon > 0$. Therefore,

$$0 \le P(X = x) \le \lim_{\epsilon \downarrow 0} [F_X(x) - F_X(x - \epsilon)] = 0$$

by the continuity of F_X. However, if we understand the purpose of the pdf, its definition will become clear.

From Example 1.6.2, we see that a pmf gives us "point probabilities." In the discrete case, we can sum over values of the pmf to get the cdf (as in (1.6.1)). The analogous procedure in the continuous case is to substitute integrals for sums, and we get

$$P(X \le x) = F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

Using the Fundamental Theorem of Calculus, if $f_X(x)$ is continuous, we have the further relationship

$$\frac{d}{dx} F_X(x) = f_X(x). \tag{1.6.2}$$

Note that the analogy with the discrete case is almost exact. We "add up" the "point probabilities" $f_X(x)$ to obtain interval probabilities.

Definition 1.6.3 The *probability density function* or *pdf*, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \quad \text{for all } x. \tag{1.6.3}$$

A note on notation: The expression "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$," where we read the symbol "$\sim$" as "is distributed as." We can similarly write $X \sim f_X(x)$ or, if X and Y have the same distribution, $X \sim Y$.

In the continuous case we can be somewhat cavalier about the specification of interval probabilities. Since $P(X = x) = 0$ if X is a continuous random variable,

$$P(a < X < b) = P(a < X \le b) = P(a \le X < b) = P(a \le X \le b).$$

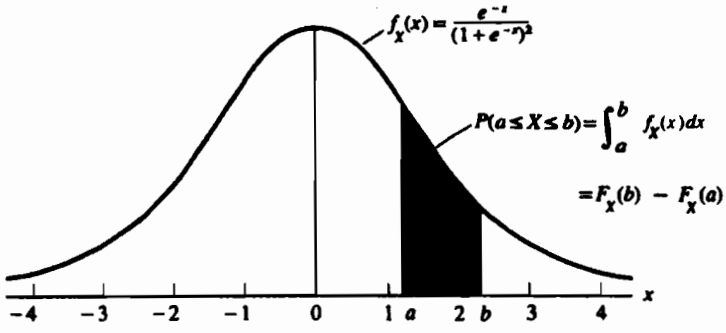


Figure 1.6.1. *Area under logistic curve*

It should be clear that the pdf (or pmf) contains the same information as the cdf. This being the case, we can use either one to solve problems and should try to choose the simpler one.

Example 1.6.4 (Logistic probabilities) For the logistic distribution of Example 1.5.5 we have

$$F_X(x) = \frac{1}{1+e^{-x}}$$

and, hence,

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

The area under the curve $f_X(x)$ gives us interval probabilities (see Figure 1.6.1):

$$\begin{aligned} P(a < X < b) &= F_X(b) - F_X(a) \\ &= \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)\,dx \\ &= \int_{a}^{b} f_X(x)\,dx. \end{aligned}$$

||

There are really only two requirements for a pdf (or pmf), both of which are immediate consequences of the definition.

Theorem 1.6.5 *A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if*

a. $f_X(x) \ge 0$ *for all* x.

b. $\sum_x f_X(x) = 1$ *(pmf)* or $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ *(pdf).*

Proof: If $f_X(x)$ is a pdf (or pmf), then the two properties are immediate from the definitions. In particular, for a pdf, using (1.6.3) and Theorem 1.5.3, we have that

$$1 = \lim_{x\to\infty} F_X(x) = \int_{-\infty}^{\infty} f_X(t)\,dt.$$

The converse implication is equally easy to prove. Once we have $f_X(x)$, we can define $F_X(x)$ and appeal to Theorem 1.5.3. □

From a purely mathematical viewpoint, any nonnegative function with a finite positive integral (or sum) can be turned into a pdf or pmf. For example, if $h(x)$ is any nonnegative function that is positive on a set A, 0 elsewhere, and

$$\int_{\{x\in A\}} h(x)\,dx = K < \infty$$

for some constant $K > 0$, then the function $f_X(x) = h(x)/K$ is a pdf of a random variable X taking values in A.

Actually, the relationship (1.6.3) does not always hold because $F_X(x)$ may be continuous but not differentiable. In fact, there exist continuous random variables for which the integral relationship does not exist for *any* $f_X(x)$. These cases are rather pathological and we will ignore them. Thus, in this text, we will assume that (1.6.3) holds for any continuous random variable. In more advanced texts (for example, Billingsley 1995, Section 31) a random variable is called *absolutely continuous* if (1.6.3) holds.

1.7 Exercises

1.1 For each of the following experiments, describe the sample space.

(a) Toss a coin four times.

(b) Count the number of insect-damaged leaves on a plant.

(c) Measure the lifetime (in hours) of a particular brand of light bulb.

(d) Record the weights of 10-day-old rats.

(e) Observe the proportion of defectives in a shipment of electronic components.

1.2 Verify the following identities.

(a) $A\backslash B = A\backslash(A\cap B) = A\cap B^c$

(b) $B = (B\cap A)\cup(B\cap A^c)$

(c) $B\backslash A = B\cap A^c$

(d) $A\cup B = A\cup(B\cap A^c)$

1.3 Finish the proof of Theorem 1.1.4. For any events A, B, and C defined on a sample space S, show that

(a) $A \cup B = B \cup A$ and $A \cap B = B \cap A$. (commutativity)
(b) $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$. (associativity)
(c) $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$. (DeMorgan's Laws)

1.4 For events A and B, find formulas for the probabilities of the following events in terms of the quantities $P(A)$, $P(B)$, and $P(A \cap B)$.

(a) either A or B or both
(b) either A or B but not both
(c) at least one of A or B
(d) at most one of A or B

1.5 Approximately one-third of all human twins are identical (one-egg) and two-thirds are fraternal (two-egg) twins. Identical twins are necessarily the same sex, with male and female being equally likely. Among fraternal twins, approximately one-fourth are both female, one-fourth are both male, and half are one male and one female. Finally, among all U.S. births, approximately 1 in 90 is a twin birth. Define the following events:

$$A = \{\text{a U.S. birth results in twin females}\}$$

$$B = \{\text{a U.S. birth results in identical twins}\}$$

$$C = \{\text{a U.S. birth results in twins}\}$$

(a) State, in words, the event $A \cap B \cap C$.
(b) Find $P(A \cap B \cap C)$.

1.6 Two pennies, one with $P(\text{head}) = u$ and one with $P(\text{head}) = w$, are to be tossed together independently. Define

$$p_0 = P(0 \text{ heads occur}),$$

$$p_1 = P(1 \text{ head occurs}),$$

$$p_2 = P(2 \text{ heads occur}).$$

Can u and w be chosen such that $p_0 = p_1 = p_2$? Prove your answer.

1.7 Refer to the dart game of Example 1.2.7. Suppose we do not assume that the probability of hitting the dart board is 1, but rather is proportional to the area of the dart board. Assume that the dart board is mounted on a wall that is hit with probability 1, and the wall has area A.

(a) Using the fact that the probability of hitting a region is proportional to area, construct a probability function for $P(\text{scoring } i \text{ points})$, $i = 0, \ldots, 5$. (No points are scored if the dart board is not hit.)
(b) Show that the conditional probability distribution $P(\text{scoring } i \text{ points} | \text{board is hit})$ is exactly the probability distribution of Example 1.2.7.

1.8 Again refer to the game of darts explained in Example 1.2.7.

(a) Derive the general formula for the probability of scoring i points.
(b) Show that $P(\text{scoring } i \text{ points})$ is a decreasing function of i, that is, as the points increase, the probability of scoring them decreases.
(c) Show that $P(\text{scoring } i \text{ points})$ is a probability function according to the Kolmogorov Axioms.

1.9 Prove the general version of DeMorgan's Laws. Let $\{A_\alpha: \alpha \in \Gamma\}$ be a (possibly uncountable) collection of sets. Prove that

(a) $(\cup_\alpha A_\alpha)^c = \cap_\alpha A_\alpha^c$. (b) $(\cap_\alpha A_\alpha)^c = \cup_\alpha A_\alpha^c$.

1.10 Formulate and prove a version of DeMorgan's Laws that applies to a finite collection of sets $A_1, \ldots, A_n$.

1.11 Let S be a sample space.

(a) Show that the collection $\mathcal{B} = \{\emptyset, S\}$ is a sigma algebra.

(b) Let $\mathcal{B} = \{$all subsets of S, including S itself$\}$. Show that $\mathcal{B}$ is a sigma algebra.

(c) Show that the intersection of two sigma algebras is a sigma algebra.

1.12 It was noted in Section 1.2.1 that statisticians who follow the deFinetti school do not accept the Axiom of Countable Additivity, instead adhering to the Axiom of Finite Additivity.

(a) Show that the Axiom of Countable Additivity implies Finite Additivity.

(b) Although, by itself, the Axiom of Finite Additivity does not imply Countable Additivity, suppose we supplement it with the following. Let $A_1 \supset A_2 \supset \cdots \supset A_n \supset \cdots$ be an infinite sequence of nested sets whose limit is the empty set, which we denote by $A_n \downarrow \emptyset$. Consider the following:

Axiom of Continuity: If $A_n \downarrow \emptyset$, then $P(A_n) \to 0$.

Prove that the Axiom of Continuity and the Axiom of Finite Additivity imply Countable Additivity.

1.13 If $P(A) = \frac{1}{3}$ and $P(B^c) = \frac{1}{4}$, can A and B be disjoint? Explain.

1.14 Suppose that a sample space S has n elements. Prove that the number of subsets that can be formed from the elements of S is 2^n.

1.15 Finish the proof of Theorem 1.2.14. Use the result established for $k = 2$ as the basis of an induction argument.

1.16 How many different sets of initials can be formed if every person has one surname and

(a) exactly two given names? (b) either one or two given names?

(b) either one or two or three given names?

(Answers: (a) 26^3 (b) $26^3 + 26^2$ (c) $26^4 + 26^3 + 26^2$)

1.17 In the game of dominoes, each piece is marked with two numbers. The pieces are symmetrical so that the number pair is not ordered (so, for example, $(2, 6) = (6, 2)$). How many different pieces can be formed using the numbers $1, 2, \ldots, n$?
(Answer: $n(n+1)/2$)

1.18 If n balls are placed at random into n cells, find the probability that exactly one cell remains empty.
(Answer: $\binom{n}{2} n!/n^n$)

1.19 If a multivariate function has continuous partial derivatives, the order in which the derivatives are calculated does not matter. Thus, for example, the function $f(x, y)$ of two variables has equal third partials

$$\frac{\partial^3}{\partial x^2 \partial y} f(x, y) = \frac{\partial^3}{\partial y \partial x^2} f(x, y).$$

(a) How many fourth partial derivatives does a function of three variables have?

(b) Prove that a function of n variables has $\binom{n+r-1}{r}$ rth partial derivatives.

1.20 My telephone rings 12 times each week, the calls being randomly distributed among the 7 days. What is the probability that I get at least one call each day?
(Answer: .2285)

1.21 A closet contains n pairs of shoes. If $2r$ shoes are chosen at random ($2r < n$), what is the probability that there will be no matching pair in the sample?
(Answer: $\binom{n}{2r}2^{2r}/\binom{2n}{2r}$)

1.22 (a) In a draft lottery containing the 366 days of the year (including February 29), what is the probability that the first 180 days drawn (without replacement) are evenly distributed among the 12 months?
(b) What is the probability that the first 30 days drawn contain none from September?
(Answers: (*a*) $.167 \times 10^{-8}$ (*b*) $\binom{336}{30}/\binom{366}{30}$)

1.23 Two people each toss a fair coin n times. Find the probability that they will toss the same number of heads.
(Answer: $\left(\frac{1}{4}\right)^n \binom{2n}{n}$)

1.24 Two players, A and B, alternately and independently flip a coin and the first player to obtain a head wins. Assume player A flips first.
(a) If the coin is fair, what is the probability that A wins?
(b) Suppose that $P(\text{head}) = p$, not necessarily $\frac{1}{2}$. What is the probability that A wins?
(c) Show that for all $p, 0 < p < 1, P(\text{A wins}) > \frac{1}{2}$. (*Hint:* Try to write $P(\text{A wins})$ in terms of the events $E_1, E_2, \ldots$, where $E_i = \{\text{head first appears on } i\text{th toss}\}$.)
(Answers: (*a*) 2/3 (*b*) $\frac{p}{1-(1-p)^2}$)

1.25 The Smiths have two children. At least one of them is a boy. What is the probability that both children are boys? (See Gardner 1961 for a complete discussion of this problem.)

1.26 A fair die is cast until a 6 appears. What is the probability that it must be cast more than five times?

1.27 Verify the following identities for $n \geq 2$.
(a) $\sum_{k=0}^{n}(-1)^k\binom{n}{k} = 0$ (b) $\sum_{k=1}^{n}k\binom{n}{k} = n2^{n-1}$
(c) $\sum_{k=1}^{n}(-1)^{k+1}k\binom{n}{k} = 0$

1.28 A way of approximating large factorials is through the use of *Stirling's Formula*:

$$n! \approx \sqrt{2\pi}n^{n+(1/2)}e^{-n},$$

a complete derivation of which is difficult. Instead, prove the easier fact,

$$\lim_{n\to\infty}\frac{n!}{n^{n+(1/2)}e^{-n}} = \text{a constant}.$$

(*Hint:* Feller 1968 proceeds by using the monotonicity of the logarithm to establish that

$$\int_{k-1}^{k}\log x\,dx < \log k < \int_{k}^{k+1}\log x\,dx, \quad k = 1, \ldots, n,$$

and hence

$$\int_{0}^{n}\log x\,dx < \log n! < \int_{1}^{n+1}\log x\,dx.$$

Now compare $\log n!$ to the average of the two integrals. See Exercise 5.35 for another derivation.)

1.29 (a) For the situation of Example 1.2.20, enumerate the ordered samples that make up the unordered samples $\{4, 4, 12, 12\}$ and $\{2, 9, 9, 12\}$.
(b) Enumerate the ordered samples that make up the unordered samples $\{4, 4, 12, 12\}$ and $\{2, 9, 9, 12\}$.

(c) Suppose that we had a collection of six numbers, $\{1, 2, 7, 8, 14, 20\}$. What is the probability of drawing, with replacement, the unordered sample $\{2, 7, 7, 8, 14, 14\}$?

(d) Verify that an unordered sample of size k, from m different numbers repeated $k_1, k_2, \ldots, k_m$ times, has $\dfrac{k!}{k_1!k_2!\cdots k_m!}$ ordered components, where $k_1 + k_2 + \cdots + k_m = k$.

(e) Use the result of the previous part to establish the identity

$$\sum_{k_1,k_2,\ldots,k_m : k_1+k_2+\cdots+k_m=k} \frac{k!}{k_1!k_2!\cdots k_m!} = \binom{k+m-1}{k}.$$

1.30 For the collection of six numbers, $\{1, 2, 7, 8, 14, 20\}$, draw a histogram of the distribution of all possible sample averages calculated from samples drawn with replacement.

1.31 For the situation of Example 1.2.20, the average of the original set of numbers $\{2, 4, 9, 12\}$ is $\frac{29}{4}$, which has the highest probability.

(a) Prove that, in general, if we sample with replacement from the set $\{x_1, x_2, \ldots, x_n\}$, the outcome with average $(x_1+x_2+\cdots+x_n)/n$ is the most likely, having probability $\frac{n!}{n^n}$.

(b) Use Stirling's Formula (Exercise 1.28) to show that $n!/n^n \approx \sqrt{2n\pi}/e^n$ (Hall 1992, Appendix I).

(c) Show that the probability that a particular x_i is missing from an outcome is $(1 - \frac{1}{n})^n \to e^{-1}$ as $n \to \infty$.

1.32 An employer is about to hire one new employee from a group of N candidates, whose future potential can be rated on a scale from 1 to N. The employer proceeds according to the following rules:

(a) Each candidate is seen in succession (in random order) and a decision is made whether to hire the candidate.

(b) Having rejected $m-1$ candidates ($m > 1$), the employer can hire the mth candidate only if the mth candidate is better than the previous $m - 1$.

Suppose a candidate is hired on the ith trial. What is the probability that the best candidate was hired?

1.33 Suppose that 5% of men and .25% of women are color-blind. A person is chosen at random and that person is color-blind. What is the probability that the person is male? (Assume males and females to be in equal numbers.)

1.34 Two litters of a particular rodent species have been born, one with two brown-haired and one gray-haired (litter 1), and the other with three brown-haired and two gray-haired (litter 2). We select a litter at random and then select an offspring at random from the selected litter.

(a) What is the probability that the animal chosen is brown-haired?

(b) Given that a brown-haired offspring was selected, what is the probability that the sampling was from litter 1?

1.35 Prove that if $P(\cdot)$ is a legitimate probability function and B is a set with $P(B) > 0$, then $P(\cdot|B)$ also satisfies Kolmogorov's Axioms.

1.36 If the probability of hitting a target is $\frac{1}{5}$, and ten shots are fired independently, what is the probability of the target being hit at least twice? What is the conditional probability that the target is hit at least twice, given that it is hit at least once?

1.37 Here we look at some variations of Example 1.3.4.

(a) In the warden's calculation of Example 1.3.4 it was assumed that if A were to be pardoned, then with equal probability the warden would tell A that either B or C would die. However, this need not be the case. The warden can assign probabilities γ and $1-\gamma$ to these events, as shown here:

Prisoner pardoned	Warden tells A	
A	B dies	with probability γ
A	C dies	with probability $1-\gamma$
B	C dies	
C	B dies	

Calculate $P(A|W)$ as a function of γ. For what values of γ is $P(A|W)$ less than, equal to, or greater than $\frac{1}{3}$?

(b) Suppose again that $\gamma = \frac{1}{2}$, as in the example. After the warden tells A that B will die, A thinks for a while and realizes that his original calculation was false. However, A then gets a bright idea. A asks the warden if he can swap fates with C. The warden, thinking that no information has been passed, agrees to this. Prove that A's reasoning is now correct and that his probability of survival has jumped to $\frac{2}{3}$!

A similar, but somewhat more complicated, problem, the "Monte Hall problem" is discussed by Selvin (1975). The problem in this guise gained a fair amount of notoriety when it appeared in a Sunday magazine (vos Savant 1990) along with a correct answer but with questionable explanation. The ensuing debate was even reported on the front page of the Sunday *New York Times* (Tierney 1991). A complete and somewhat amusing treatment is given by Morgan *et al.* (1991) [see also the response by vos Savant 1991]. Chun (1999) pretty much exhausts the problem with a very thorough analysis.

1.38 Prove each of the following statements. (Assume that any conditioning event has positive probability.)

(a) If $P(B) = 1$, then $P(A|B) = P(A)$ for any A.

(b) If $A \subset B$, then $P(B|A) = 1$ and $P(A|B) = P(A)/P(B)$.

(c) If A and B are mutually exclusive, then

$$P(A|A \cup B) = \frac{P(A)}{P(A) + P(B)}.$$

(d) $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$.

1.39 A pair of events A and B cannot be simultaneously *mutually exclusive* and *independent*. Prove that if $P(A) > 0$ and $P(B) > 0$, then:

(a) If A and B are mutually exclusive, they cannot be independent.

(b) If A and B are independent, they cannot be mutually exclusive.

1.40 Finish the proof of Theorem 1.3.9 by proving parts (b) and (c).

1.41 As in Example 1.3.6, consider telegraph signals "dot" and "dash" sent in the proportion 3:4, where erratic transmissions cause a dot to become a dash with probability $\frac{1}{4}$ and a dash to become a dot with probability $\frac{1}{3}$.

(a) If a dash is received, what is the probability that a dash has been sent?

(b) Assuming independence between signals, if the message dot–dot was received, what is the probability distribution of the four possible messages that could have been sent?

1.42 The *inclusion-exclusion identity* of Miscellanea 1.8.1 gets it name from the fact that it is proved by the method of inclusion and exclusion (Feller 1968, Section IV.1). Here we go into the details. The probability $P(\cup_{i=1}^n A_i)$ is the sum of the probabilities of all the sample points that are contained in at least one of the A_is. The method of inclusion and exclusion is a recipe for counting these points.

(a) Let E_k denote the set of all sample points that are contained in exactly k of the events $A_1, A_2, \ldots, A_n$. Show that $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(E_i)$.

(b) If E_1 is not empty, show that $P(E_1) = \sum_{i=1}^n P(A_i)$.

(c) Without loss of generality, assume that E_k is contained in $A_1, A_2, \ldots, A_k$. Show that $P(E_k)$ appears k times in the sum P_1, $\binom{k}{2}$ times in the sum P_2, $\binom{k}{3}$ times in the sum P_3, etc.

(d) Show that

$$k - \binom{k}{2} + \binom{k}{3} - \cdots \pm \binom{k}{k} = 1.$$

(See Exercise 1.27.)

(e) Show that parts $(a) - (c)$ imply $\sum_{i=1}^n P(E_i) = P_1 - P_2 = \cdots \pm P_n$, establishing the inclusion-exclusion identity.

1.43 For the *inclusion-exclusion identity* of Miscellanea 1.8.1:

(a) Derive both Boole's and Bonferroni's Inequality from the inclusion-exclusion identity.

(b) Show that the P_i satisfy $P_i \geq P_j$ if $i \geq j$ and that the sequence of bounds in Miscellanea 1.8.1 improves as the number of terms increases.

(c) Typically as the number of terms in the bound increases, the bound becomes more useful. However, Schwager (1984) cautions that there are some cases where there is not much improvement, in particular if the A_is are highly correlated. Examine what happens to the sequence of bounds in the extreme case when $A_i = A$ for every i. (See Worsley 1982 and the correspondence of Worsley 1985 and Schwager 1985.)

1.44 Standardized tests provide an interesting application of probability theory. Suppose first that a test consists of 20 multiple-choice questions, each with 4 possible answers. If the student guesses on each question, then the taking of the exam can be modeled as a sequence of 20 independent events. Find the probability that the student gets at least 10 questions correct, given that he is guessing.

1.45 Show that the induced probability function defined in (1.4.1) defines a legitimate probability function in that it satisfies the Kolmogorov Axioms.

1.46 Seven balls are distributed randomly into seven cells. Let X_i = the number of cells containing exactly i balls. What is the probability distribution of X_3? (That is, find $P(X_3 = x)$ for every possible x.)

1.47 Prove that the following functions are cdfs.

(a) $\frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$, $x \in (-\infty, \infty)$

(b) $(1 + e^{-x})^{-1}$, $x \in (-\infty, \infty)$

(c) $e^{-e^{-x}}$, $x \in (-\infty, \infty)$

(d) $1 - e^{-x}$, $x \in (0, \infty)$

(e) the function defined in (1.5.6)

1.48 Prove the necessity part of Theorem 1.5.3.

1.49 A cdf F_X is *stochastically greater* than a cdf F_Y if $F_X(t) \leq F_Y(t)$ for all t and $F_X(t) < F_Y(t)$ for some t. Prove that if $X \sim F_X$ and $Y \sim F_Y$, then

$$P(X > t) \geq P(Y > t) \quad \text{for every } t$$

and

$$P(X > t) > P(Y > t) \quad \text{for some } t,$$

that is, X tends to be bigger than Y.

1.50 Verify formula (1.5.4), the formula for the partial sum of the geometric series.

1.51 An appliance store receives a shipment of 30 microwave ovens, 5 of which are (unknown to the manager) defective. The store manager selects 4 ovens at random, without replacement, and tests to see if they are defective. Let X = number of defectives found. Calculate the pmf and cdf of X and plot the cdf.

1.52 Let X be a continuous random variable with pdf $f(x)$ and cdf $F(x)$. For a fixed number x_0, define the function

$$g(x) = \begin{cases} f(x)/[1 - F(x_0)] & x \geq x_0 \\ 0 & x < x_0. \end{cases}$$

Prove that $g(x)$ is a pdf. (Assume that $F(x_0) < 1$.)

1.53 A certain river floods every year. Suppose that the low-water mark is set at 1 and the high-water mark Y has distribution function

$$F_Y(y) = P(Y \leq y) = 1 - \frac{1}{y^2}, \quad 1 \leq y < \infty.$$

(a) Verify that $F_Y(y)$ is a cdf.
(b) Find $f_Y(y)$, the pdf of Y.
(c) If the low-water mark is reset at 0 and we use a unit of measurement that is $\frac{1}{10}$ of that given previously, the high-water mark becomes $Z = 10(Y - 1)$. Find $F_Z(z)$.

1.54 For each of the following, determine the value of c that makes $f(x)$ a pdf.

(a) $f(x) = c \sin x$, $0 < x < \pi/2$ (b) $f(x) = ce^{-|x|}$, $-\infty < x < \infty$

1.55 An electronic device has lifetime denoted by T. The device has value $V = 5$ if it fails before time $t = 3$; otherwise, it has value $V = 2T$. Find the cdf of V, if T has pdf

$$f_T(t) = \frac{1}{1.5} e^{-t/(1.5)}, \quad t > 0.$$

1.8 Miscellanea

1.8.1 Bonferroni and Beyond

The Bonferroni bound of (1.2.10), or Boole's Inequality (Theorem 1.2.11), provides simple bounds on the probability of an intersection or union. These bounds can be made more and more precise with the following expansion.

For sets $A_1, A_2, \ldots A_n$, we create a new set of nested intersections as follows. Let

$$P_1 = \sum_{i=1}^{n} P(A_i)$$

$$P_2 = \sum_{1 \le i < j \le n} P(A_i \cap A_j)$$

$$P_3 = \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k)$$

$$\vdots$$

$$P_n = P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Then the *inclusion-exclusion identity* says that

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P_1 - P_2 + P_3 - P_4 + \cdots \pm P_n.$$

Moreover, the P_i are ordered in that $P_i \ge P_j$ if $i \le j$, and we have the sequence of upper and lower bounds

$$\begin{aligned} P_1 \ge P(\cup_{i=1}^n A_i) &\ge P_1 - P_2 \\ P_1 - P_2 + P_3 \ge P(\cup_{i=1}^n A_i) &\ge P_1 - P_2 + P_3 - P_4 \\ &\vdots \end{aligned}$$

See Exercises 1.42 and 1.43 for details.

These bounds become increasingly tighter as the number of terms increases, and they provide a refinement of the original Bonferroni bounds. Applications of these bounds include approximating probabilities of runs (Karlin and Ost 1988) and multiple comparisons procedures (Naiman and Wynn 1992).