

Chapter 4

Foundations of Probability

Probability theory is one of the foundation stones of statistics and econometrics. We'll start at the basics but move relatively quickly.

4.1 Probabilistic Models

Probability is a technically demanding field. The main reason is that we want to assign probabilities to all manner of events. The set of possible events can be very large, and we need ways to manage this complexity. In this and the next few sections, we'll review the basic machinery involved and why this machinery is necessary. At the same time, we'll skip several details that would take us too far away from econometric theory. This will cost little in terms of understanding.

One concept worth mentioning before we start is countable and uncountable sets. In essence, a nonempty set S is called **countable** if it is either finite or can be represented as a sequence, and **uncountable** if it is not countable. In probability theory this distinction is important. A quick review can be found in §15.3.

4.1.1 Sample Spaces and Events

In setting up a probabilistic model, we start with the notion of a **sample space**, which we think of as being a "list" of all possible outcomes in a given random experiment. In general, the sample space can be any nonempty set, and is usually denoted Ω . A typical element of Ω is denoted ω . The general idea is that a realization of uncertainty will lead to the selection of a particular $\omega \in \Omega$.

Example 4.1.1 In a random experiment that involves rolling a die once, the set of possible outcomes is naturally represented by $\Omega := \{1, \dots, 6\}$.

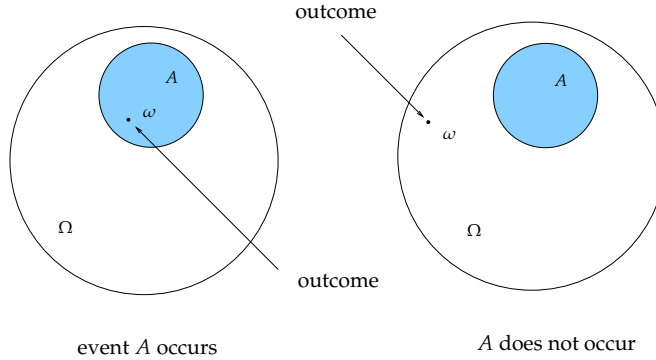


Figure 4.1 Outcomes and events

Example 4.1.2 Imagine that Burton Malkiel’s blindfolded monkey throws darts at a dartboard of radius 1. To record the outcome we impose ordinary Cartesian coordinates with origin at the center of the board. Let (h, v) be a typical location measured on the horizontal and vertical coordinates respectively. A natural sample space is $\Omega := \{(h, v) \in \mathbb{R}^2 : \|(h, v)\| \leq 1\}$. This set is also called the **unit disk** in $\mathbb{R}^2$.

Example 4.1.3 Consider an experiment where we roll a die twice. A suitable sample space is the set of integer pairs (i, j) with $1 \leq i, j \leq 6$. The first element i represents the outcome of the first roll, while the second element j represents the outcome of the second roll. Thus $\Omega := \{(i, j) : i, j \in \{1, \dots, 6\}\} = \{1, \dots, 6\} \times \{1, \dots, 6\}$.

Once we have a sample space in place we can start to think about events. In the language of probability theory, an **event** is just a subset of Ω (with some caveats that will be addressed below). An event A occurs whenever the individual $\omega \in \Omega$ selected in the random experiment happens to lie in A . Figure 4.1 illustrates.

4.1.1.1 Probabilities and Events

The next thing we need to do is to assign probabilities to outcomes. One idea is to start by assigning appropriate probabilities to every ω in Ω . We could then derive the probability of events by looking at the probabilities attached to points contained in those events. It turns out, however, that this is not the best approach, at least if we want to consider uncountable Ω . Rather, the standard approach is to *directly assign probabilities to subsets of Ω instead*.

To see why the first approach is problematic, consider the dart throwing model in example 4.1.2, where Ω is the unit disk in $\mathbb{R}^2$. Imagine that hitting any point in Ω is

"equally likely." We can model this by assuming that the probability the dart lands in some region A of the dartboard (some $A \subset \Omega$) is simply proportional to the area of A .

Consider now the probability of hitting some individual point $\omega \in \Omega$. Intuitively, this should be less than the probability of hitting any region A containing ω . By choosing A to be sufficiently small in area, we can show that the probability of hitting ω is less than ϵ for any given $\epsilon > 0$. We are forced to conclude that the probability of hitting the individual point ω is zero.¹

This example shows that in many sensible models the probability assigned to individual points in the sample space must be zero. If you reflect on this for even a moment you will agree that there is no way to build up a useful probabilistic model with such information alone. Hence we must assign probabilities directly to the events themselves.

4.1.1.2 Choosing the Set of Events

Should we assign probabilities to every possible event—that is, to every possible subset of Ω ? In the dart model above, we could say that the probability the dart lands in any region $A \subset \Omega$ is equal to the area of A divided by the area of Ω itself, which is π . In symbols,

$$\mathbb{P}(A) = \frac{\lambda(A)}{\pi} \quad \text{where } \lambda(A) := \text{the area of set } A \quad (4.1)$$

This seems sensible enough. For example, if $A = \Omega$, then $\mathbb{P}(A) = \pi/\pi = 1$, as it should be. If A is the top half of the dartboard then $\mathbb{P}(A) = 0.5$, and so on.

Unfortunately, if we think carefully about this approach, some cracks start to appear. What exactly do we mean by "area of A "? If A is a circle, then its area is π times the square of its radius. If A is a rectangle, then its area is the product of its two sides. But how about the area of more general subsets of Ω ?

It turns out that assigning area to all subsets of Ω in a consistent way is problematic. The reason is that Ω contains an awful lot of subsets, and some of them can be manipulated to exhibit very strange phenomena. (If you are curious, look up the Banach–Tarski paradox for a hint of the problems that can occur when we try to assign notions like area or volume to arbitrary sets.) We can dream up all sorts of strange events that confound our ability to construct a sensible probabilistic model.

Because of this, when Ω is uncountable, we usually resist taking the set of events to be all subsets of Ω . Instead, we take them to be some family of "well-behaved" subsets of Ω , denoted below by $\mathcal{F}$, and only assign probabilities to elements of $\mathcal{F}$.

There is of course a trade-off here. If we restrict $\mathcal{F}$ to be a small fraction of the subsets of Ω , then assigning probabilities becomes easy and we can eliminate strange

1. If p is a nonnegative number satisfying $p \geq 0$ and $p \leq \epsilon$ for all $\epsilon > 0$, then $p = 0$ must hold.

behavior. But if we make $\mathcal{F}$ too small, then we might not be able to assign probabilities to events that we care about. For example, in the dartboard model, if we restrict $\mathcal{F}$ to be the set of rectangles, then we can easily measure area and hence assign probabilities via (4.1). But this excludes interesting events such as the dart landing in the top half of the board.

How can we ensure that $\mathcal{F}$ is large enough? To think about this in a systematic way, suppose that $A \in \mathcal{F}$, and that $\mathbb{P}(A)$ is well-defined and represents the “probability of event A .” Now, given that we can assign a probability to the event A , it would be unfortunate if we couldn’t assign a probability to the event “not A ”, which corresponds to A^c . So normally we require that if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$. When this is true, we say that $\mathcal{F}$ is “closed under the taking of complements.”

Further, suppose that A and B are both in $\mathcal{F}$, and we can assign probabilities to these events. In this case, it would be natural to think about the probability of the event “ A or B ”, which corresponds to $A \cup B$. So we also require that if A and B are in $\mathcal{F}$, then $A \cup B$ is likewise in $\mathcal{F}$. We say that $\mathcal{F}$ is “closed under the taking of unions.” In fact for technical reasons it turns out that we prefer $\mathcal{F}$ to be closed under the taking of countable unions, which means that if $A_1, A_2, \dots$ is a sequence of sets in $\mathcal{F}$, then its union is likewise in $\mathcal{F}$.

A set $\mathcal{F}$ of subsets of a given set Ω that satisfies all these properties is called a σ -algebra. Formally, $\mathcal{F}$ is a σ -algebra on Ω if

- (i) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$,
- (ii) $A_1, A_2, \dots \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$, and
- (iii) $\Omega \in \mathcal{F}$.

Note that (i) and (iii) together imply that $\emptyset \in \mathcal{F}$. In this context, the empty set $\emptyset$ is called the **impossible event**. In contrast, Ω is called the **certain event** because it always occurs (regardless of which outcome ω is selected, $\omega \in \Omega$ is true by definition).

The σ -algebra of events $\mathcal{F}$ we choose to pair with a given Ω can vary from problem to problem. There is however one particular σ -algebra that you should know at least something about, and that is the standard σ -algebra we use to pair with ordinary vector space $\mathbb{R}^N$. This σ -algebra is called the Borel sets.

To understand the idea, consider the following: The property of $\mathcal{F}$ being a σ -algebra seems to imply that $\mathcal{F}$ is a relatively rich class of sets. After all, we can take complements and unions (and intersections too—see ex. 4.4.1) without ever leaving $\mathcal{F}$. However, there are trivial examples of σ -algebras, such as $\mathcal{F} := \{\emptyset, \Omega\}$. So if we want to pin down a large and useful class of events, the best way to proceed is to assume not only that $\mathcal{F}$ is a σ -algebra, but also that $\mathcal{F}$ contains some fundamental sets that we are likely to want to attach probabilities to, such as intervals and rectangles.

In this connection, the **Borel subsets of $\mathbb{R}^N$** are defined as the smallest σ -algebra

of subset of $\mathbb{R}^N$ that contains all rectangles in $\mathbb{R}^N$. (See §15.1 if you wish to learn more about rectangles.) This definition turns out to be highly successful: Restricting attention to Borel sets does a great job of excluding the kind of strange behavior found in the Banach-Tarski paradox, and, moreover, every subset of $\mathbb{R}^N$ we need to deal with in day to day analysis is a Borel set. These include planes and hyperplanes, circles, spheres, polygons, finite sets, and sequences of points. In what follows we use $\mathcal{B}(\mathbb{R}^N)$ to represent the Borel sets of $\mathbb{R}^N$.

4.1.2 Probabilities

Now let's take a σ -algebra of events as given and consider how to assign probabilities to elements of $\mathcal{F}$. For given event B , the symbol $\mathbb{P}(B)$ will be interpreted as representing "the probability that event B occurs." The way you should think about it is this:

$\mathbb{P}(B)$ represents the probability that when uncertainty is resolved and some $\omega \in \Omega$ is selected by "nature," the statement $\omega \in B$ is true.

In order to make sure our model of probability is well behaved, we need to put certain restrictions on $\mathbb{P}$. For example, it wouldn't do to have a B with $\mathbb{P}(B) = -93$, since negative probabilities don't make sense. The standard restrictions are imposed in the next definition:

Let Ω be a nonempty set and let $\mathcal{F}$ be a σ -algebra of subsets of Ω . A **probability** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function from $\mathcal{F}$ to $[0, 1]$ that satisfies

- (i) $\mathbb{P}(\Omega) = 1$ and
- (ii) $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ for any disjoint sequence of sets $A_1, A_2, \dots \in \mathcal{F}$.

In this context $\mathbb{P}$ is also called a **probability measure**. Together, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

In the definition, axiom (i) is straightforward. We require $\mathbb{P}(\Omega) = 1$ because, by construction, every possible ω lies in the set Ω . Axiom (ii) is called **countable additivity**. The concept of disjointness in the statement of the axiom is pairwise: Any distinct pair A_i, A_j share no points in common. By taking some of the sets as empty, countable additivity can be shown to imply finite **additivity**, which requires that

$$\mathbb{P}(A_1 \cup \dots \cup A_k) = \mathbb{P}(A_1) + \dots + \mathbb{P}(A_k) \quad (4.2)$$

whenever $A_1, \dots, A_k$ are disjoint.

The general notion of additivity is a natural requirement for probabilistic models. It says that to find the probability of a given event, we can determine all the different (think "disjoint") ways that the event could occur, and then sum their probabilities.

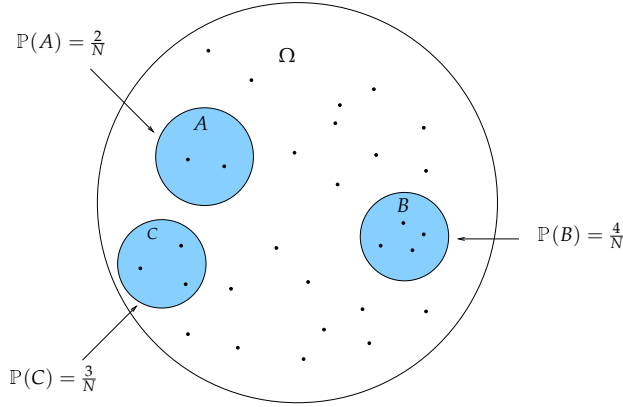


Figure 4.2 Each of the N dots occurs with probability $1/N$

Probabilists boost this up to countable additivity because it helps certain limiting arguments work smoothly.

Figure 4.2 gives an illustration of additivity. Suppose that the N dots in the figure are the only possible outcomes. Suppose that each of these occurs with probability $1/N$, and hence the probability of an event is the number of dots it contains divided by N . Hence $\mathbb{P}(A) = 2/N$, $\mathbb{P}(B) = 4/N$ and $\mathbb{P}(C) = 3/N$. In addition there are 9 dots in $A \cup B \cup C$, so

$$\mathbb{P}(A \cup B \cup C) = \frac{9}{N} = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$$

Notice how disjointness is important in making additivity work here: If the sets were not disjoint we would run the risk of double counting on the right hand side of this sum, breaking the equality.

Example 4.1.4 Let $\Omega := \{1, \dots, 6\}$ represent the six different faces of a die, as in example 4.1.1. Since Ω is finite, we can and do take $\mathcal{F}$ to be the set of all subsets of Ω . We define a probability $\mathbb{P}: \mathcal{F} \rightarrow [0, 1]$ by

$$\mathbb{P}(A) := \frac{|A|}{6} \quad \text{where } |A| := \text{number of elements in set } A \quad (4.3)$$

For example, $\mathbb{P}\{2, 4, 6\} = 3/6 = 1/2$. It's easy to see that $0 \leq \mathbb{P}(A) \leq 1$ for any $A \in \mathcal{F}$, and that $\mathbb{P}(\Omega) = 1$. Regarding additivity, suppose that A and B are two disjoint subsets of $\{1, \dots, 6\}$. Then $|A \cup B| = |A| + |B|$, since, by disjointness, the

number of elements in the union is just the number contributed by A plus the number contributed by B . Hence

$$\mathbb{P}(A \cup B) = \frac{|A \cup B|}{6} = \frac{|A| + |B|}{6} = \frac{|A|}{6} + \frac{|B|}{6} = \mathbb{P}(A) + \mathbb{P}(B)$$

This proves additivity for pairs of sets. An analogous argument confirms additivity for any finite collection, in the sense that (4.2) holds. Finite additivity is in this case equivalent to countable additivity, since the total number of distinct events is finite.

Example 4.1.5 Continuing example 4.1.4, if we roll the die, the probability of getting an even number should be equal to the probability of getting a 2 plus that of getting a 4 plus that of getting a 6. This works because of additivity. Formally,

$$\mathbb{P}\{2, 4, 6\} = \mathbb{P}[\{2\} \cup \{4\} \cup \{6\}] = \mathbb{P}\{2\} + \mathbb{P}\{4\} + \mathbb{P}\{6\} = 1/2$$

Example 4.1.6 Imagine flipping a coin N times and recording the outcome of each flip. One sample space for this experiment is

$$\Omega_0 := \{(b_1, \dots, b_N) : b_n \in \{\text{heads}, \text{tails}\} \text{ for each } n\}$$

Letting zero represent tails and one represent heads yields the more practical space

$$\Omega := \{0, 1\}^N := \{(b_1, \dots, b_N) : b_n \in \{0, 1\} \text{ for each } n\}$$

Thus Ω is the set of all binary sequences of length N . Since Ω is again finite, we take $\mathcal{F}$ to be the set of all subsets of Ω . As our probability, we define

$$\mathbb{P}(A) := 2^{-N}|A|$$

where $|A|$ is again the number of distinct elements in A . This is a kind of uniform probability, where all outcomes are equally likely. Indeed, if $(b_1, \dots, b_N)$ is any complete sequence of flips and $A \in \mathcal{F}$ is the singleton set containing just this draw, then $\mathbb{P}(A) = 2^{-N}$. Hence all such sequences have the same probability under $\mathbb{P}$.

Regarding the axioms on page 83, it is not hard to check that $0 \leq \mathbb{P}(A) \leq 1$ for all $A \subset \Omega$ and $\mathbb{P}(\Omega) = 1$ (note that the number of distinct binary sequences of length N is 2^N). In addition $\mathbb{P}$ is additive. Exercise 4.4.8 asks you to confirm this.

Example 4.1.7 Consider again the dartboard model from example 4.1.2, where Ω is the unit disk in $\mathbb{R}^2$. For the event space, the set of all subsets of Ω is problematic (see §4.1.1), so we take instead $\mathcal{F}$ to be the set of Borel subsets of $\mathbb{R}^2$ that lie inside Ω . For $\mathbb{P}$ we follow the “uniform” probability assignment given in (4.1). That is, $\mathbb{P}(B) = \lambda(B)/\pi$ for every $B \in \mathcal{F}$. The function λ that assigns area to Borel sets

is known to be countably additive, in that $\lambda(\cup_n A_n) = \sum_{n=1}^{\infty} \lambda(A_n)$ whenever these sets are disjoint. From this it follows that $\mathbb{P}$ is likewise countably additive. Evidently $\mathbb{P}(\Omega) = 1$.

The function λ from example 4.1.7 that maps Borel sets to their “area” is formally known as the **Lebesgue measure**. §15.3.1 provides a brief introduction to this concept.

4.1.2.1 Implications of the Axioms

Now let’s go back to the general case, where $(\Omega, \mathcal{F}, \mathbb{P})$ is an arbitrary probability space. From the axioms on page 83, we can derive a surprising number of properties. Let’s list the key ones, starting with the next fact.

Fact 4.1.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A, B \in \mathcal{F}$. If $A \subset B$, then

- (i) $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$,
- (ii) $\mathbb{P}(A) \leq \mathbb{P}(B)$,
- (iii) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, and
- (iv) $\mathbb{P}(\emptyset) = 0$.

(Here $A^c = \Omega \setminus A$. See §15.1 for a review of set-theoretic operations.)

To prove fact 4.1.1, note that when $A \subset B$, we have $B = (B \setminus A) \cup A$. Since $B \setminus A$ and A are disjoint, additivity of $\mathbb{P}$ now gives

$$\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A) \quad (\text{whenever } A \subset B)$$

This equality implies (i)–(iv) of fact 4.1.1. Rearranging gives (i), while nonnegativity of $\mathbb{P}$ gives (ii). Specializing to $B = \Omega$ gives (iii), and setting $B = A$ gives (iv).

The property $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$ is called **monotonicity**, and is both fundamental and intuitive: If $A \subset B$, then we know that B occurs whenever A occurs, since ω lands in B whenever it lands in A . Hence the probability of B should be no smaller. Many results in probability turn on this idea.

Fact 4.1.2 If A and B are any (not necessarily disjoint) events, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Exercise 4.4.2 asks you to prove this. It follows from fact 4.1.2 that for any $A, B \in \mathcal{F}$, we have $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. This inequality is called **subadditivity**. Thus probabilities are subadditive over arbitrary pairs of events and additive over disjoint pairs.

4.1.2.2 Dependence and Independence

If A and B are events, then the **conditional probability of A given B** is

$$\mathbb{P}(A | B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (4.4)$$

It represents the probability that A will occur, given the information that B has occurred. For the definition to make sense, we need $\mathbb{P}(B) > 0$. Events A and B are called independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. More generally, events $A_1, \dots, A_N \in \mathcal{F}$ are called **independent** if

$$\mathbb{P}\left(\bigcap_{n=1}^N A_n\right) = \prod_{n=1}^N \mathbb{P}(A_n) \quad (4.5)$$

Note that if A and B are independent, then the conditional probability of A given B is just the probability of A .

Example 4.1.8 Recall example 4.1.3, where we roll a die twice. The sample space is $\Omega := \{(i, j) : i, j \in \{1, \dots, 6\}\}$. Since Ω is finite we can again take $\mathcal{F}$ to be the set of all subsets of Ω . For our probability, we define $\mathbb{P}(E) := |E|/36$. (Here elements of E are pairs, so $|E|$ is the number of pairs in E .) Now consider the events

$$A := \{(i, j) \in \Omega : i \text{ is even}\} \quad \text{and} \quad B := \{(i, j) \in \Omega : j \text{ is even}\}$$

In this case we have

$$A \cap B = \{(i, j) \in \Omega : i \text{ and } j \text{ are even}\}$$

The events A and B are independent under $\mathbb{P}$. To check this, recall the basic principle for counting ordered tuples: The total number of possible tuples is the product of the number of possibilities for each element. For example, the number of distinct tuples

$$(i, j, k) \text{ where } i \in I, j \in J \text{ and } k \in K$$

is $|I| \times |J| \times |K|$. Hence, the number of elements in A is $3 \times 6 = 18$, the number of elements in B is $6 \times 3 = 18$, and the number of elements in $A \cap B$ is $3 \times 3 = 9$. As a result

$$\mathbb{P}(A \cap B) = 9/36 = 1/4 = (18/36) \times (18/36) = \mathbb{P}(A)\mathbb{P}(B)$$

This confirms that A and B are independent.

To state the next result, we need the concept of a **partition** of Ω , by which we mean a countable set of events $\{B_m\}_{m \geq 1}$ with the property that distinct pairs are disjoint and $\bigcup_{m \geq 1} B_m = \Omega$.

Example 4.1.9 Let $\Omega = (0, \infty)$ and $B_m = (m-1, m]$ for all $m \in \mathbb{N}$. The events $\{B_m\}_{m \geq 1}$ form a partition of Ω .

The following fact is known as the **law of total probability**.

Fact 4.1.3 Let $\{B_m\}_{m \geq 1}$ be a partition of Ω and let A be any event. If $\mathbb{P}(B_m) > 0$ for all m , then

$$\mathbb{P}(A) = \sum_{m \geq 1} \mathbb{P}(A \mid B_m) \cdot \mathbb{P}(B_m)$$

The proof is as follows: It is a basic result from set theory that $A \cap (\cup_m B_m) = \cup_m (A \cap B_m)$. From this equality, the definition of a partition and countable additivity of $\mathbb{P}$, we get

$$\mathbb{P}(A) = \mathbb{P}[A \cap (\cup_{m \geq 1} B_m)] = \mathbb{P}[\cup_{m \geq 1} (A \cap B_m)] = \sum_{m \geq 1} \mathbb{P}(A \cap B_m)$$

Applying (4.4) completes the proof.

Example 4.1.10 Consider a bank that lends to two types of customers, type A and type B. Types are not observable to the bank. Type A customers have a default probability of 5%. Type B customers have a default probability of 10%. Of potential customers, 90% are of type B and 10% are of type A. Given a randomly selected customer, we wish to calculate on behalf of the bank the default probability associated with this customer.

To this end, let D be the event that the customer defaults, let A be the event that the customer is of type A, and let B be the event that the customer is of type B. Since the customer must be one of these two types, we have, by the law of total probability,

$$\mathbb{P}(D) = \mathbb{P}(D \mid A)\mathbb{P}(A) + \mathbb{P}(D \mid B)\mathbb{P}(B)$$

The right-hand side evaluates to $0.05 \times 0.1 + 0.1 \times 0.9 = 0.005 + 0.09 = 0.095$.

Another useful result concerning conditional probability is **Bayes' law**, which states that for any events A and B with positive probability,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{4.6}$$

To prove (4.6), observe that, from the definition of conditional probability,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{and} \quad \mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Hence $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \mathbb{P}(B) = \mathbb{P}(B \mid A) \mathbb{P}(A)$. Rearranging yields (4.6).

Example 4.1.11 Banks use automated systems to detect fraudulent or illegal transactions. Consider a test that responds to each transaction with P or N , where P means “positive” (transaction flagged as fraudulent) and N means “negative” (transaction flagged as normal). Letting F mean fraudulent, suppose that

- $\mathbb{P}(P | F) = 0.99$ (the test flags 99% of fraudulent transactions),
- $\mathbb{P}(P | F^c) = 0.01$ (rate of false positives), and
- $\mathbb{P}(F) = 0.001$ (prevalence of fraud).

We want to know the probability of fraud given a positive test. To answer this question, we use Bayes’ law

$$\mathbb{P}(F | P) = \frac{\mathbb{P}(P | F)\mathbb{P}(F)}{\mathbb{P}(P)}$$

and the law of total probability $\mathbb{P}(P) = \mathbb{P}(P | F)\mathbb{P}(F) + \mathbb{P}(P | F^c)\mathbb{P}(F^c)$ to get

$$\mathbb{P}(F | P) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.01 \times 0.999} = \frac{11}{122} \approx \frac{1}{11}$$

Less than one in ten positives are actually fraudulent.

4.1.3 Random Variables

In some probability courses, a random variable is defined as a “value that changes randomly,” or something to that effect. This is a fairly useless definition. A better definition of a random variable is a function x from a sample space Ω into $\mathbb{R}$. Think of it this way:

- (i) “Nature” picks out an element ω in Ω according to some probability.
- (ii) The random variable x sends this ω into numerical value $x(\omega)$.

Thus a random variable is a map x that converts outcomes in sample space—which can be any kind of object—into *numerical outcomes*. This is valuable because numerical outcomes are easy to order, add, subtract and so forth. In other words, random variables “report” the outcome of an experiment in a format amenable to further analysis.

Example 4.1.12 Consider an experiment where we flip a coin until we get heads. Let 0 represent tails and 1 represent heads. As our sample space we take the set of infinite binary sequences

$$\Omega := \{(b_1, b_2, \dots) : b_n \in \{0, 1\} \text{ for each } n\}$$

Here's a random variable giving the number of flips until the first heads:

$$x(\omega) = x(b_1, b_2, \dots) = \min\{n \in \mathbb{N} : b_n = 1\}$$

Suppose that we also wish to know whether any heads occur in the first 10 flips. If so we can introduce a new random variable

$$y(\omega) = y(b_1, b_2, \dots) := \min \left\{ \sum_{n=1}^{10} b_n, 1 \right\} \quad (4.7)$$

that returns a 1 when this event occurs and zero otherwise. As per the definition, x and y are functions from Ω into $\mathbb{R}$.²

The random variable y in (4.7) is an example of a **binary** (or **Bernoulli**) random variable, which is a random variable taking values in the set $\{0, 1\}$. There is a generic way to create binary random variables that often comes in handy, using **indicator functions**. If Q is a statement, such as “on the planet Uranus, there exists a tribe of three-headed monkeys,” then $\mathbb{1}\{Q\}$ is considered as equal to 1 when the statement Q is true, and zero when the statement Q is false. Another common variation on the notation is, for arbitrary $A \in \mathcal{F}$,

$$\mathbb{1}_A(\omega) := \mathbb{1}\{\omega \in A\} := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

So defined, $\mathbb{1}_A$ is a binary random variable. In fact any binary random variable x can be expressed as $\mathbb{1}_A$ by setting $A := \{\omega \in \Omega : x(\omega) = 1\}$.

Fact 4.1.4 If $A_1, \dots, A_N$ are subsets of Ω , then

- (i) $\mathbb{1}_{\cap_{n=1}^N A_n} = \prod_{n=1}^N \mathbb{1}_{A_n}$ and
- (ii) $\mathbb{1}_{\cup_{n=1}^N A_n} = \sum_{n=1}^N \mathbb{1}_{A_n}$ whenever the sets are disjoint.

Here equality means when evaluated at any $\omega \in \Omega$. See exercise 4.4.5.

4.1.3.1 Notational Conventions

Given random variable x , you will often see notation such as

$$\{x \text{ has some property}\}$$

2. Is this true? If $\omega = \omega_0$ is an infinite sequence containing only zeros, then $\{n \in \mathbb{N} : b_n = 1\} = \emptyset$. What is $x(\omega_0)$ in this case? The convention is $\min \emptyset = \infty$, which is reasonable, but now x can take the value ∞ . However, in most applications this event has probability zero. We could set $x(\omega_0) = 0$ without changing anything significant. Now we're back to a well-defined function from Ω to $\mathbb{R}$.

This is a shorthand for the *event*

$$\{\omega \in \Omega : x(\omega) \text{ has some property}\}$$

Example 4.1.13 Consider the claim that, for any random variable x ,

$$\mathbb{P}\{x \leq a\} \leq \mathbb{P}\{x \leq b\} \quad \text{whenever} \quad a \leq b \quad (4.8)$$

The proof goes as follows: Pick any $a, b \in \mathbb{R}$ with $a \leq b$. The symbol $\{x \leq a\}$ is shorthand for the event $\{\omega \in \Omega : x(\omega) \leq a\}$. Likewise, $\{x \leq b\}$ is shorthand for $\{\omega \in \Omega : x(\omega) \leq b\}$. Since $a \leq b$, we have

$$\{\omega \in \Omega : x(\omega) \leq a\} \subset \{\omega \in \Omega : x(\omega) \leq b\}$$

(Any ω giving $x(\omega) \leq a$ must also give $x(\omega) \leq b$.) The result in (4.8) now follows from monotonicity of $\mathbb{P}$ (fact 4.1.1 on page 86).

Example 4.1.14 Recall example 4.1.6, with sample space $\Omega := \{0, 1\}^N$, events $\mathcal{F} :=$ all subsets of Ω and $\mathbb{P}(A) := 2^{-N}|A|$. Consider a random variable x on Ω that returns the first element of any given sequence. That is, $x(\omega) = x(b_1, \dots, b_N) = b_1$. The probability that $x = 1$ is $1/2$. Indeed

$$\mathbb{P}\{x = 1\} := \mathbb{P}\{\omega \in \Omega : x(\omega) = 1\} = \mathbb{P}\{(b_1, \dots, b_N) \in \Omega : b_1 = 1\}$$

Because the number of length N binary sequences with $b_1 = 1$ is 2^{N-1} , we have $\mathbb{P}\{x = 1\} = 2^{-N}2^{N-1} = 1/2$.

Another notational convention you should be aware of is that, for random variables, equalities, inequalities and arithmetic operations should be interpreted *point-wise*. For example,

- $x \leq y \iff x(\omega) \leq y(\omega) \text{ for all } \omega \in \Omega,$
- $x = y \iff x(\omega) = y(\omega) \text{ for all } \omega \in \Omega,$ and
- $z = \alpha x + \beta y \iff z(\omega) = \alpha x(\omega) + \beta y(\omega) \text{ for all } \omega \in \Omega.$

4.1.3.2 Random Variables Are Measurable Functions

We've omitted a technical detail in our construction of random variables. It arises because we cannot always take the set of events $\mathcal{F}$ to be the set of all subsets of Ω (since assigning probabilities to arbitrarily complex sets causes problems—see the discussion in §4.1.1).

To explain, let $(\Omega, \mathcal{F}, \mathbb{P})$ be any probability space. Let B be any subset of $\mathbb{R}$ and consider evaluating the probability

$$\mathbb{P}\{x \in B\} := \mathbb{P}\{\omega \in \Omega : x(\omega) \in B\}$$

Here x is some function from Ω to $\mathbb{R}$. Now observe that if we put no restrictions on x , we might end up in a mess. The reason is that we have no way of being sure that the set $\{\omega \in \Omega : x(\omega) \in B\}$ is in fact an element of $\mathcal{F}$. Since $\mathbb{P}$ is only defined over elements of $\mathcal{F}$ (i.e., for *bona fide* events), it is possible that $\mathbb{P}\{x \in B\}$ is undefined.

To prevent such embarrassment, it's standard to put some restrictions both on x and B that ensure $\{x \in B\} \in \mathcal{F}$. For B , we naturally restrict attention to $\mathcal{B}(\mathbb{R})$, the Borel subsets of $\mathbb{R}$, since we know these sets to be both relatively well behaved and common enough to satisfy our purposes (see §4.1.1). For x , we simply require that $\{x \in B\} \in \mathcal{F}$ whenever B is a Borel set. Thus, our formal definition is that a **random variable** on $(\Omega, \mathcal{F})$ is a function $x: \Omega \rightarrow \mathbb{R}$ satisfying

$$\{\omega \in \Omega : x(\omega) \in B\} \in \mathcal{F} \quad \text{for all } B \in \mathcal{B}(\mathbb{R}) \quad (4.9)$$

These kinds of functions are also sometimes called $\mathcal{F}$ -measurable functions. In other texts you will see (4.9) expressed using preimage notation. In the preimage notation, $x^{-1}(B)$ is all $\omega \in \Omega$ such that $x(\omega) \in B$. This is just the left-hand side of (4.9), so we can rewrite (4.9) as

$$x^{-1}(B) \in \mathcal{F} \quad \text{for all } B \in \mathcal{B}(\mathbb{R})$$

Thus x “pulls back” Borel sets to events.

4.1.3.3 Measurable Transformations

So far so good. If x is a random variable, then $\{x \in B\}$ is a well-defined event for every $B \in \mathcal{B}(\mathbb{R})$. But suppose now that we want to discuss some transformation of x . For example, we might be interested in the properties of $y := e^x$. Is y also a random variable? That is, is $\{y \in B\}$ always a well-defined event?

The answer to this question is affirmative provided that our transformation satisfies a regularity condition called Borel measurability. Formally, $f: \mathbb{R} \rightarrow \mathbb{R}$ is called **Borel measurable**, or **$\mathcal{B}$ -measurable**, if

$$f^{-1}(B) \in \mathcal{B}(\mathbb{R}) \quad \text{for all } B \in \mathcal{B}(\mathbb{R}) \quad (4.10)$$

Thus f “pulls back” Borel sets to Borel sets. Although we omit the details, the class of $\mathcal{B}$ -measurable functions is vast (any continuous function, any increasing function, etc., etc.), and all transformations we work with below are $\mathcal{B}$ -measurable.

So why is $\mathcal{B}$ -measurability of f the key to ensuring that $y = f(x)$ is a random

variable? Suppose that f is $\mathcal{B}$ -measurable and that x is a random variable. We claim that $\{y \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$. This holds because

$$\{y \in B\} = \{f(x) \in B\} = \{x \in f^{-1}(B)\} \quad (4.11)$$

for all $B \in \mathcal{B}(\mathbb{R})$. By (4.10) we know that $f^{-1}(B)$ is a Borel set. Since x is a random variable, it follows that $\{x \in f^{-1}(B)\}$ is an element of $\mathcal{F}$.

4.1.4 Expectations

Our next task is to define expectations of random variables and state their basic properties. Elementary probability courses define expectations by summing or integrating functions over densities. The view of a random variable as a function on a sample space plays no part in these definitions. However, when we move to more sophisticated arguments, particularly those involving a large number of random variables, we need a better definition. This section outlines the modern treatment of expectations found in probability and statistics. As we'll see later, the definition in this section is a generalization of the elementary definition. The two agree whenever the latter is well defined.

For finite random variables (i.e., random variables with finite range) the expectation is just the sum of all possible values of the variable, weighted by the corresponding probabilities. More formally, given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variable x taking only finitely many distinct values $s_1, \dots, s_J$, the **expectation** of x is defined as

$$\mathbb{E}x = \sum_{j=1}^J s_j \mathbb{P}\{x = s_j\} \quad (4.12)$$

Example 4.1.15 Let's apply this definition to the simplest possible case, which is a random variable x satisfying $x(\omega) = \alpha$ for all $\omega \in \Omega$, where α is some constant scalar value. In this case the sum in (4.12) has only one term, and

$$\mathbb{E}x = \alpha \mathbb{P}\{x = \alpha\} = \alpha \mathbb{P}\{\omega \in \Omega : x(\omega) = \alpha\} = \alpha \mathbb{P}(\Omega) = \alpha$$

Example 4.1.16 To evaluate the expectation of a binary random variable x , we apply (4.12) to obtain

$$\mathbb{E}x = 1 \times \mathbb{P}\{x = 1\} + 0 \times \mathbb{P}\{x = 0\} = \mathbb{P}\{x = 1\} \quad (4.13)$$

Example 4.1.17 Recall example 4.1.14, which models N flips of a fair coin. The sample space is $\Omega := \{0, 1\}^N$, the events are $\mathcal{F} :=$ all subsets of Ω , and $\mathbb{P}(A) := 2^{-N}|A|$ for all $A \in \mathcal{F}$. Let $x(\omega) = x(b_1, \dots, b_N) = \sum_{n=1}^N b_n$. This value corresponds to the

number of “heads” observed in one possible outcome $(b_1, \dots, b_N)$. The probability $\mathbb{P}$ we’ve assigned makes all draws equally likely. Hence a good guess is that half the flips will come up heads, or $\mathbb{E}x = N/2$. Let’s try to obtain this from (4.12).

Observe first that $0 \leq x \leq N$. Also, by the definition of $\mathbb{P}$, for any k we have $\mathbb{P}\{x = k\} = 2^{-N}|A_k|$, where

$$A_k := \{x = k\} = \left\{ (b_1, \dots, b_N) \in \Omega : \sum_{n=1}^N b_n = k \right\}$$

It is a standard result from combinatorics that $|A_k| = \binom{N}{k}$, where the right-hand side is the so-called **binomial coefficient** for N, k . One identity it satisfies is $\sum_{k=0}^N k \binom{N}{k} = N2^{N-1}$ for all N . Combining this identity with (4.12), the expectation of x is

$$\mathbb{E}x = \sum_{k=0}^N k 2^{-N} |A_k| = 2^{-N} \sum_{k=0}^N k \binom{N}{k} = \frac{N}{2}$$

4.1.4.1 Expectations for Arbitrary Random Variables

Definition (4.12) only applies for random variables with finite range. We still need to define expectation for general x . This is done by bootstrapping up from the finite case. The procedure is based on approximating arbitrary random variables with finite random variables. One finite approximation x_n to an arbitrary random variable x is shown in figure 4.3. The approximation can be improved without limit if we allow x_n to take a larger and larger number of distinct values. This process yields a sequence of finite random variables x_n converging to x (e.g., see Çinlar (2011), lemma 2.16). The expectation of x is then defined as

$$\mathbb{E}x := \lim_{n \rightarrow \infty} \mathbb{E}x_n \tag{4.14}$$

The value $\mathbb{E}x$ is also referred to as the **Lebesgue integral** of x with respect to $\mathbb{P}$, with the alternative notation $\mathbb{E}x = \int x(\omega) \mathbb{P}(d\omega)$.

Can we be sure that the limit in (4.14) makes sense? Suppose first that x is nonnegative (i.e., $x(\omega) \geq 0$ for all $\omega \in \Omega$). It can be shown that the sequence $\{x_n\}$ approximating x can then be chosen so that $\mathbb{E}x_n$ is monotone increasing and hence convergent, and that the limiting value does not depend on the particular approximating sequence $\{x_n\}$. See, for example, Dudley (2002), proposition 4.1.5. Hence $\mathbb{E}x$ is well defined, although we cannot rule out divergence of the sequence, which gives $\mathbb{E}x = \infty$.

We’ve assumed that x is nonnegative. If we drop this assumption, we can still treat x with similar methods by writing it as the difference between two nonnegative ran-

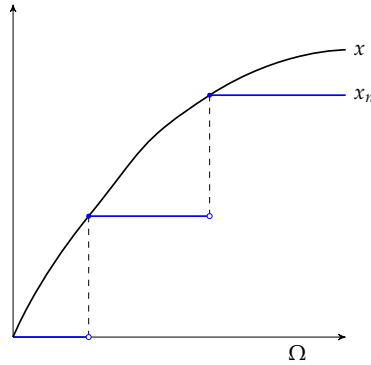


Figure 4.3 Finite approximation to a general random variable

dom variables. Typically this is done by writing $x = x^+ - x^-$, where $x^+ := \max\{x, 0\}$ and $x^- := -\min\{x, 0\}$. Both x^+ and x^- are nonnegative random variables. The expectation can then be defined as

$$\mathbb{E}x := \mathbb{E}x^+ - \mathbb{E}x^- \quad (4.15)$$

This is fine except that (4.15) might have the form $\infty - \infty$, which isn't well-defined. For this reason we often restrict attention to **integrable** random variables, which are all random variables x such that $\mathbb{E}|x| < \infty$. Since $x^+ \leq |x|$ and $x^- \leq |x|$, this is enough to ensure that (4.15) is well-defined.

The definition of $\mathbb{E}$ given in this section doesn't really tell us how to calculate $\mathbb{E}x$ in specific instances, but we'll see how to do so in §4.2.5.

4.1.4.2 Properties of Expectation

Let's summarize the main results concerning expectations and list some properties of this operation.

Fact 4.1.5 Given any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, there exists a uniquely defined function $\mathbb{E}$ that maps each integrable random variable x on $(\Omega, \mathcal{F}, \mathbb{P})$ into a value

$$\mathbb{E}x = \int x(\omega) \mathbb{P}(d\omega) \quad (4.16)$$

in $\mathbb{R}$, called the **expectation of x** under $\mathbb{P}$. The function has the following properties:

- (i) $\mathbb{E}\alpha = \alpha$ for all $\alpha \in \mathbb{R}$.

- (ii) $\mathbb{E} \mathbb{1}_A = \mathbb{P}(A)$ for all $A \in \mathcal{F}$.
- (iii) $x \leq y \implies \mathbb{E}x \leq \mathbb{E}y$.
- (iv) $\mathbb{E}[\alpha x + \beta y] = \alpha \mathbb{E}x + \beta \mathbb{E}y$ for all integrable x, y and constants α, β .

If we need to remind ourselves of the underlying probability measure $\mathbb{P}$ we will write $\mathbb{E}_{\mathbb{P}}x$ instead of $\mathbb{E}x$.

Since we only know how to take expectations of random variables, the expression $\mathbb{E}\alpha$ in part (i) should be understood as the expectation of a constant random variable everywhere equal to α . (We could write this object more precisely as $\alpha \mathbb{1}_{\Omega}$.) That the expectation equals α was already established in example 4.1.15. It also follows from (ii) and (iv), by taking $x = \mathbb{1}_{\Omega}$ and $\beta = 0$.

Part (ii) follows from example 4.1.16. It is restated here to emphasize this fundamental link between probabilities and expectations. Part (iii) is called **monotonicity of expectations**. The statement $x \leq y$ means that $x(\omega) \leq y(\omega)$ for all $\omega \in \Omega$. Exercise 4.4.23 asks you to prove this monotonicity in a special case. For a general proof see lemma 4.1.9 of Dudley (2002). Part (iv) is called **linearity of expectations**. Here $\alpha x + \beta y$ should be understood as the random variable $(\alpha x + \beta y)(\omega) := \alpha x(\omega) + \beta y(\omega)$. Exercises 4.4.13 and 4.4.14 ask you to check linearity in some special cases. For the general case see theorem 4.1.10 of Dudley (2002).

As an exercise, let's use these results to prove that if x is a finite random variable with range $\{s_j\}_{j=1}^J$ and h is any $\mathcal{B}$ -measurable function, then $h(x)$ has expectation

$$\mathbb{E}h(x) = \sum_{j=1}^J h(s_j) \mathbb{P}\{x = s_j\} \quad (4.17)$$

First observe that $\sum_{j=1}^J \mathbb{1}\{x = s_j\} = 1$, and hence we can write $h(x)$ as

$$h(x) = h(x) \sum_{j=1}^J \mathbb{1}\{x = s_j\} = \sum_{j=1}^J h(s_j) \mathbb{1}\{x = s_j\}$$

Using linearity of expectations gives $\mathbb{E}h(x) = \sum_{j=1}^J h(s_j) \mathbb{E} \mathbb{1}\{x = s_j\}$. Applying part (ii) of fact 4.1.5 leads to (4.17).

The next result is another useful link between probabilities and expectations, typically called **Chebyshev's inequality**.

Fact 4.1.6 For any nonnegative random variable x and any $\delta > 0$, we have

$$\mathbb{P}\{x \geq \delta\} \leq \frac{\mathbb{E}x}{\delta} \quad (4.18)$$

A common variation of Chebyshev's inequality is the bound

$$\mathbb{P}\{|x| \geq \delta\} \leq \frac{\mathbb{E}x^2}{\delta^2} \quad (4.19)$$

Exercise 4.4.29 asks you to check both (4.18) and (4.19).

4.1.5 Moments and Co-Moments

Let x be a random variable and let $k \in \mathbb{N}$. If x^k is integrable, then

- $\mathbb{E}[x^k]$ is called the **k th moment** of x .
- $\mathbb{E}[(x - \mathbb{E}x)^k]$ is called the **k th central moment** of x .

If $\mathbb{E}[|x|^k] = \infty$, then the k th moment is said not to exist. For some random variables even the first moment does not exist. For others every moment exists.

Fact 4.1.7 If the k th moment of x exists, then so does the j th for all $j \leq k$.

Exercise 4.4.24 asks you to confirm this. The next fact is called the **Cauchy–Schwarz inequality for random variables**. Its proof will be given later, once we've covered orthogonal projection (ex. 5.4.15).

Fact 4.1.8 If x and y are random variables with finite second moment, then

$$|\mathbb{E}[xy]| \leq \sqrt{\mathbb{E}[x^2]\mathbb{E}[y^2]} \quad (4.20)$$

The second central moment of x is called the **variance** of x :

$$\text{var } x := \mathbb{E}[(x - \mathbb{E}x)^2]$$

This gives a measure of the dispersion of x . The **standard deviation** of x is

$$\sigma_x := \sqrt{\text{var } x}$$

The **covariance** of random variables x and y is defined as

$$\text{cov}[x, y] := \mathbb{E}[(x - \mathbb{E}x)(y - \mathbb{E}y)]$$

Fact 4.1.9 If x and y have finite second moments, then

- (i) $\text{var } x$ and $\text{cov}[x, y]$ are finite,
- (ii) $\text{var } x = \mathbb{E}[x^2] - [\mathbb{E}x]^2$, and
- (iii) $\text{cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$.

Part (i) follows from (ii)–(iii), the Cauchy–Schwarz inequality and fact 4.1.7. Parts (ii)–(iii) follow from linearity of $\mathbb{E}$ and some simple manipulations.

Fact 4.1.10 If $x_1, \dots, x_N$ are random variables and $\alpha_0, \alpha_1, \dots, \alpha_N$ are constant scalars, then

$$\text{var} \left[\alpha_0 + \sum_{n=1}^N \alpha_n x_n \right] = \sum_{n=1}^N \alpha_n^2 \text{var}[x_n] + 2 \sum_{n < m} \alpha_n \alpha_m \text{cov}[x_n, x_m]$$

Some simple implications of this fact are that

- (i) $\text{var}[\alpha + \beta x] = \beta^2 \text{var}[x]$ and
- (ii) $\text{var}[\alpha x + \beta y] = \alpha^2 \text{var}[x] + \beta^2 \text{var}[y] + 2\alpha\beta \text{cov}[x, y]$.

The **correlation** of x and y is defined as

$$\text{corr}[x, y] := \frac{\text{cov}[x, y]}{\sigma_x \sigma_y}$$

If $\text{corr}[x, y] = 0$, we say that x and y are **uncorrelated**. Positive correlation means that $\text{corr}[x, y]$ is positive and negative correlation means that $\text{corr}[x, y]$ is negative. The first part of the next fact follows immediately from fact 4.1.8; the second is just algebra.

Fact 4.1.11 Given any two random variables x, y and positive constants α, β , we have

$$-1 \leq \text{corr}[x, y] \leq 1 \quad \text{and} \quad \text{corr}[\alpha x, \beta y] = \text{corr}[x, y]$$

4.1.5.1 Best Linear Predictors

As an application of covariance and a prelude to our discussion of prediction, let's consider the problem of predicting the value of a random variable y given knowledge of the value of a second random variable x (and also knowledge of the underlying probability distributions, which makes this a problem in probability rather than statistics). Thus we seek a function f such that $f(x)$ is close to y on average. To measure the latter, we will use **mean squared error**, which amounts in this case to

$$\mathbb{E}[(y - f(x))^2]$$

As we'll see in §5.2.5, the minimizer of the mean squared deviation over all functions of x is obtained by choosing $f(x) = \mathbb{E}[y | x]$, where the right-hand side is the conditional expectation of y given x . Here we'll consider the simpler problem of finding a good predictor of y within the class of “linear” functions

$$\mathcal{H}_\ell := \{ \text{all functions of the form } \ell(x) = \alpha + \beta x \}$$

(We defer to common usage and call such functions linear in this context, despite the fact that they are, in general, affine; see example 2.1.23.) Thus we consider the problem

$$\min_{\ell \in \mathcal{H}_\ell} \mathbb{E}[(y - \ell(x))^2] = \min_{\alpha, \beta \in \mathbb{R}} \mathbb{E}[(y - \alpha - \beta x)^2] \quad (4.21)$$

If α and β solve (4.21), then the function

$$\ell^*(x) := \alpha^* + \beta^* x \quad (4.22)$$

is called the **best linear predictor** of y given x .

Example 4.1.18 In finance, the relationship between returns on a given asset R_a and returns on a market benchmark R_m is called the **beta** of the asset. It measures exposure to systemic risk, as opposed to idiosyncratic risk specific to the asset. The beta of R_a is often defined as the coefficient β^* in the best linear prediction (4.22) when x is returns on the market benchmark and $y = R_a$.

To solve (4.21) we expand the square on the right-hand side and use linearity of $\mathbb{E}$ to write the objective function as

$$\psi(\alpha, \beta) := \mathbb{E}[y^2] - 2\alpha\mathbb{E}[y] - 2\beta\mathbb{E}[xy] + 2\alpha\beta\mathbb{E}[x] + \alpha^2 + \beta^2\mathbb{E}[x^2]$$

Computing the derivatives and solving the first-order conditions yields (ex. 4.4.27) the minimizers

$$\beta^* := \frac{\text{cov}[x, y]}{\text{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}[y] - \beta^*\mathbb{E}[x] \quad (4.23)$$

If you've studied simple linear least squares regression before, you might realize that α^* and β^* are the "population" counterparts of coefficient estimates in an empirical setting. In example 4.1.18, this allows us to take the coefficient β^* as the definition of an asset's beta, independent of any particular data set. If we wish to evaluate β^* , however, we need to use data, since the underlying probabilities, and hence the expectation operator $\mathbb{E}$, are not observable. It is at this stage that the population regression is replaced by simple linear least squares using data. More on this below.

4.2 Distributions

Any random variable x on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ assigns to each Borel set $B \subset \mathbb{R}$ a number $\mathbb{P}\{x \in B\}$ indicating the probability that x lies in B . While this is an important construction, as a means of specifying probabilities for outcomes in $\mathbb{R}$, it is somewhat indirect. When building statistical models, it's often more convenient to assign probabilities to subsets of $\mathbb{R}$ directly. This is typically done by specifying a

distribution over $\mathbb{R}$. In this section we look at how one might go about specifying such distributions.

4.2.1 Defining Distributions on $\mathbb{R}$

How can we specify probabilities over subsets of $\mathbb{R}$? In §4.1.1.2–§4.1.2 we learned how to construct a set of events $\mathcal{F}$ on an arbitrary sample space Ω and define a probability measure $\mathbb{P}$ on $\mathcal{F}$. If we want to describe probabilities over $\mathbb{R}$, then the natural thing to do is to repeat the process with Ω specialized to $\mathbb{R}$. For the set of events over $\mathbb{R}$ we take the Borel sets $\mathcal{B}(\mathbb{R})$, as discussed in §4.1.1.2. A probability measure defined over $\mathcal{B}(\mathbb{R})$ is called a **law** or **distribution**.

Distributions will be represented by symbols such as P and Q . Referring back to the definition on page 83, a distribution P is a map from $\mathcal{B}(\mathbb{R})$ to $[0, 1]$ such that

- (i) $P(\mathbb{R}) = 1$ and
- (ii) $P(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n)$ for any disjoint sequence $\{B_n\}$.

Being a probability measure, P satisfies all the properties in fact 4.1.1 on page 86. If there exists a Borel set S with $P(S) = 1$, then we say that P is **supported** on S .

Probability measures are both general and comprehensive. Each one pins down probabilistic information for every event in $\mathbb{R}$ that we might care about in econometrics or statistics. At the same time, this kind of generality can also be a hindrance. For example, it's not a trivial task to assign a specific value $P(B)$ to every single Borel set.

Fortunately, we can characterize distributions more succinctly. To begin, we recall the notion of a **cumulative distributed function**, or CDF, which is any function $F: \mathbb{R} \rightarrow [0, 1]$ satisfying

- (i) **monotonicity**: $s \leq s'$ implies $F(s) \leq F(s')$,
- (ii) **right-continuity**: $F(s_n) \downarrow F(s)$ whenever $s_n \downarrow s$, and
- (iii) $\lim_{s \rightarrow -\infty} F(s) = 0$ and $\lim_{s \rightarrow \infty} F(s) = 1$.

CDFs and distributions on $\mathbb{R}$ can be put in one-to-one correspondence as follows: First, it's true by definition that any two distributions Q and P satisfying $Q(B) = P(B)$ for all $B \in \mathcal{B}(\mathbb{R})$ are the same distribution, but it's also true that if $Q(B) = P(B)$ for all sets B of the form $(-\infty, s]$, then Q and P are again the same distribution. In other words, a distribution P is entirely characterized by the values of the function

$$F(s) := P((-\infty, s]) \quad (s \in \mathbb{R}) \quad (4.24)$$

Moreover this function turns out to have properties (i)–(iii) above. The next fact summarizes the relationship between F and P .

Fact 4.2.1 The following statements are true:

- (i) If P is any distribution on $\mathbb{R}$, then the function F in (4.24) is a CDF.
- (ii) Given any CDF F on $\mathbb{R}$, there exists exactly one distribution P satisfying (4.24).

For a full proof, see Williams (1991), lemma 1.6, or Dudley (2002), theorem 9.1.1. Here let's restrict ourselves to showing that the function F in (4.24) satisfies part (i) of the definition of a CDF. To see why (i) holds for this F , observe that $s \leq s'$ implies $(-\infty, s] \subset (-\infty, s']$. Part (ii) of fact 4.1.1 on page 86 then gives $P((-\infty, s]) \leq P((-\infty, s'])$. Hence $F(s) \leq F(s')$ as claimed.

CDFs are relatively easy to specify and work with. Here are some examples.

Example 4.2.1 The **univariate normal distributions** or **Gaussian distributions** refer to the class of distributions identified by CDFs of the form

$$F(s) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^s \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt \quad (s \in \mathbb{R})$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. We represent the distribution associated with (μ, σ) by $N(\mu, \sigma^2)$. The distribution $N(0, 1)$ is called the **standard normal distribution**. We use the symbol Φ for its CDF. Examples of normal CDFs are shown in figure 4.4.

Example 4.2.2 The **Pareto distributions** are the univariate distributions with CDFs of the form

$$F(s) = \begin{cases} 0 & \text{if } s < s_0 \\ 1 - \left(\frac{s_0}{s}\right)^\alpha & \text{if } s_0 \leq s \end{cases} \quad (s \in \mathbb{R}, s_0, \alpha > 0)$$

Pareto distributions are often used to model phenomena with heavy right-hand tails, such as the distribution of wealth or income.

Example 4.2.3 The class of **beta CDFs** is given by

$$F(s) = \begin{cases} 0 & \text{if } s \leq 0 \\ \frac{1}{B(\alpha, \beta)} \int_0^s u^{\alpha-1} (1-u)^{\beta-1} du & \text{if } 0 < s < 1 \\ 1 & \text{if } 1 \leq s \end{cases}$$

where $\alpha, \beta > 0$. See figure 4.5. In this example $B(\alpha, \beta)$ is the **beta function**

$$B(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad \text{where} \quad \Gamma(a) := \int_0^\infty u^{a-1} e^{-u} du$$

The function Γ is called the **gamma function**.

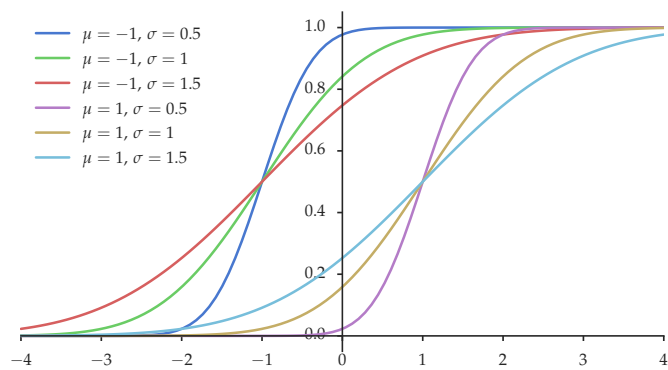


Figure 4.4 Normal CDFs

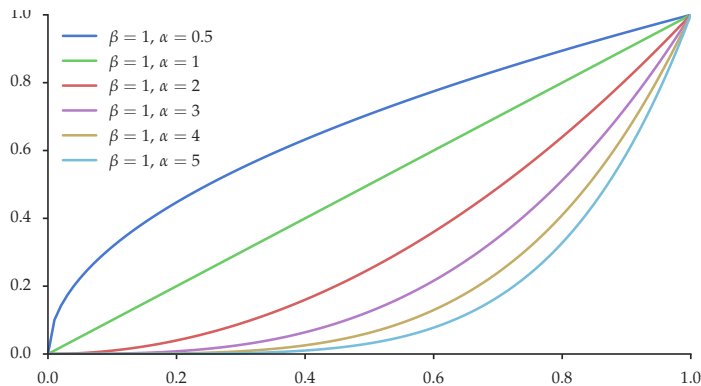


Figure 4.5 Beta CDFs

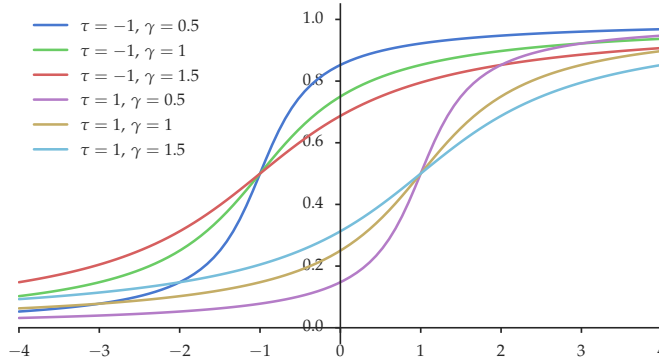


Figure 4.6 Cauchy CDFs

Example 4.2.4 The class of **Cauchy CDFs** is given by

$$F(s) = \frac{1}{\pi} \arctan \left(\frac{s - \tau}{\gamma} \right) + \frac{1}{2} \quad (s \in \mathbb{R})$$

The parameters $\tau \in \mathbb{R}$ and $\gamma > 0$ are the location and scale parameters respectively. If $\tau = 0$ and $\gamma = 1$, then F is called **standard Cauchy**. See figure 4.6.

Example 4.2.5 Given $a < b$, the **uniform CDF** on $[a, b]$ is the CDF

$$F(s) = \begin{cases} 0 & \text{if } s \leq a \\ \frac{s-a}{b-a} & \text{if } a < s < b \\ 1 & \text{if } b \leq s \end{cases}$$

We represent this distribution symbolically by $U[a, b]$.

4.2.2 Densities and PMFs

There are two important sub-cases for distributions: the discrete case and the absolutely continuous case. As we'll see, in these two sub-cases distributions have nice visual representations and expectations are easy to compute. Almost all "named" distributions fall into one of these two categories.

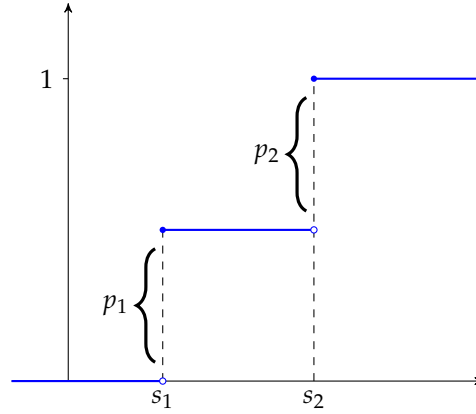


Figure 4.7 Discrete CDF

4.2.2.1 The Discrete Case

A distribution P is called **discrete** if it is supported on a countable set; that is, if there exists a countable set $\{s_j\}_{j \geq 1}$ with $P(\{s_j\}_{j \geq 1}) = 1$. For such a P let

$$p_j := P\{s_j\} := P(\{s_j\}) = \text{probability mass on the single point } s_j$$

Property (ii) of P on page 100 and $P(\{s_j\}_{j \geq 1}) = 1$ imply that $\sum_{j \geq 1} p_j = 1$ (ex. 4.4.15). Hence $\{p_j\}_{j \geq 1}$ is an example of a **probability mass function**, or **PMF**, which is any nonnegative sequence (finite or infinite) that sums to unity.

Together, $\{s_j\}_{j \geq 1}$ and $\{p_j\}_{j \geq 1}$ characterize the underlying distribution. For example, we can express the CDF corresponding to P via (4.24) as

$$F(s) = \sum_{j \geq 1} \mathbb{1}\{s_j \leq s\} p_j \quad (4.25)$$

This is intuitive: $F(s)$ represents the probability of an outcome in $(-\infty, s]$ under P . Since P puts all its probability mass on $\{s_j\}_{j \geq 1}$, this should equal the sum of the probabilities of hitting s_j over all $s_j \leq s$.

Figure 4.7 gives a visualization when $\{s_j\}_{j \geq 1} = \{s_1, s_2\}$. F is a step function, with a jump up of size p_j at point s_j . The CDF is right-continuous but not continuous.

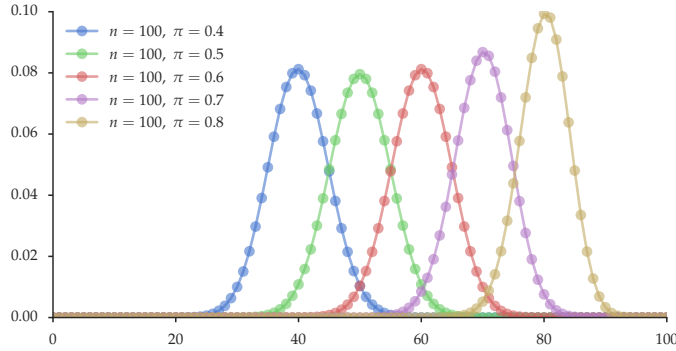


Figure 4.8 Binomial PMFs

Example 4.2.6 Given $N \in \mathbb{N}$ and $\pi \in (0, 1)$, the sequence $\{p_0, \dots, p_N\}$ defined by

$$p_j = \binom{N}{j} \pi^j (1 - \pi)^{N-j}$$

is called the **binomial PMF**. The value p_j represents the probability of j successes in N independent trials, each having success probability π .

4.2.2.2 The Absolutely Continuous Case

Another case of interest is what elementary books on probability call the “continuous” case, where distributions can be described by densities. A **density** is a nonnegative function p on $\mathbb{R}$ that integrates to 1. We say that distribution P is **represented by density p** (or **has density p**) if p is a density and

$$P(B) = \int_B p(s) \, ds \quad \text{for all } B \in \mathcal{B}(\mathbb{R}) \quad (4.26)$$

Here the integral over B can be understood as

$$\int_B p(s) \, ds := \int_{-\infty}^{\infty} \mathbb{1}_B(s) p(s) \, ds$$

Recall from our discussion in §4.2.1 that distributions are characterized by their values on simple sets, like half lines of the form $(-\infty, s]$. Thus, if p is a density and

$$P((-\infty, s]) = \int_{-\infty}^s p(t) dt \quad \text{for all } s \in \mathbb{R} \quad (4.27)$$

then (4.26) also holds.

Not every distribution has a density representation. An exact necessary and sufficient condition for existence is absolute continuity. A distribution P on the Borel subsets of $\mathbb{R}$ is called **absolutely continuous** if $P(B) = 0$ whenever B has Lebesgue measure zero. A definition of the latter property is given in §15.3.1, although for us it is enough to know that any countable subset of $\mathbb{R}$ has Lebesgue measure zero. This leads to the following fact:

Fact 4.2.2 If P is absolutely continuous, then $P(C) = 0$ whenever C is countable.

Compare with the discrete distributions, which put all mass on a countable set. One important implication of fact 4.2.2 is that individual points receive no probability mass under absolute continuity, and the corresponding CDF contains no jumps. In fact, when (4.27) holds, the fundamental theorem of calculus tells us that $F(s) = P((-\infty, s])$ is differentiable at all continuity points of p , and, in particular,

$$F'(s) = p(s) \quad \text{for all } s \in \mathbb{R} \text{ such that } p \text{ is continuous at } s$$

Example 4.2.7 The normal CDFs from example 4.2.1 are differentiable for all μ, σ , with density

$$p(s) = F'(s) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(s - \mu)^2}{2\sigma^2} \right\}$$

See figure 4.9. We reserve the symbol ϕ for the standard normal density.

Example 4.2.8 The Cauchy CDF from example 4.2.4 has density

$$p(s) = \frac{1}{\pi\gamma} \left[1 + \left(\frac{s - \tau}{\gamma} \right)^2 \right]^{-1} \quad (s \in \mathbb{R}, \gamma > 0, \tau \in \mathbb{R})$$

The Cauchy densities are more peaked around their modes and have greater mass in their tails than normal densities. See figure 4.10.

Example 4.2.9 The beta CDFs from example 4.2.3 have densities given by

$$p(s) = \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha, \beta)} \quad (\alpha, \beta > 0)$$

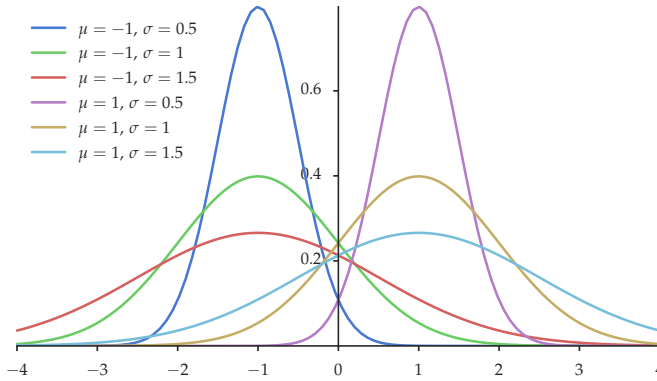


Figure 4.9 Normal densities

when $0 < s < 1$ and zero elsewhere.

Example 4.2.10 The $U[a, b]$ distribution (see page 103) is represented by the density

$$p(s) = \frac{1}{b-a} \mathbb{1}\{a \leq s \leq b\} \quad (s \in \mathbb{R}, a, b \in \mathbb{R}, a < b)$$

Example 4.2.11 The **gamma distribution** with shape parameter α and scale parameter β is the distribution with density

$$p(s) = \frac{s^{\alpha-1} e^{-s/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad (\alpha, \beta > 0)$$

when $0 < s < 1$ and zero elsewhere.

Example 4.2.12 The **chi-squared distribution with k degrees of freedom** is the distribution with density

$$p(s) := \frac{1}{2^{k/2} \Gamma(k/2)} s^{k/2-1} e^{-s/2} \quad (s > 0, k \in \mathbb{N})$$

This distribution is represented by the symbol $\chi^2(k)$. Figure 4.11 illustrates.

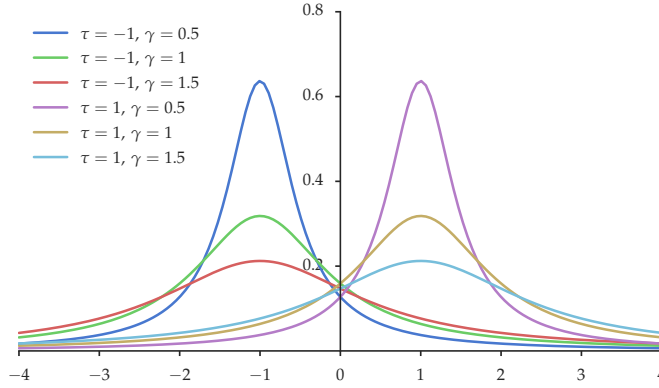


Figure 4.10 Cauchy densities

Example 4.2.13 **Student's t -distribution with k degrees of freedom**, or, more simply, the t -distribution with k degrees of freedom, is the distribution on $\mathbb{R}$ with density

$$p(s) := \frac{\Gamma(\frac{k+1}{2})}{(k\pi)^{1/2}\Gamma(\frac{k}{2})} \left(1 + \frac{s^2}{k}\right)^{-(k+1)/2} \quad (s \in \mathbb{R}, k > 0)$$

Example 4.2.14 The **F -distribution** with parameters k_1, k_2 is the distribution with the unlikely looking density

$$p(s) := \frac{\sqrt{(k_1 s)^{k_1} k_2^{k_2} / [k_1 s + k_2]^{k_1 + k_2}}}{s B(k_1/2, k_2/2)} \quad (s \geq 0, k_1, k_2 > 0)$$

The F -distribution arises in a number of hypothesis tests, as discussed below.

4.2.3 Integrating with Distributions

We all understand what it means to take the ordinary integral $\int_a^b h(s) ds$ of a well-behaved function h on some interval $[a, b]$. Sometimes we want to weight this integral, assigning more mass to different regions of $[a, b]$. This leads to integrals such as $\int_a^b h(s)p(s) ds$. For example, h could be a welfare function and p might record the density of agents of a certain type. In a probabilistic context, p might be a density indicating probabilities of outcomes, h might be a payoff function and $\int_a^b h(s)p(s) ds$ would be expected payoff.

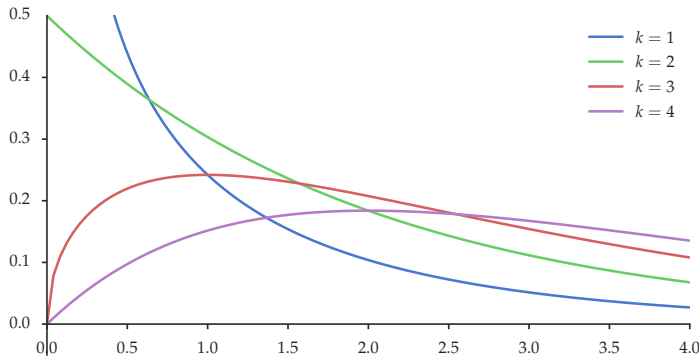


Figure 4.11 Chi-squared densities

We now wish to give meaning to expressions of the form $\int h(s)P(ds)$ that can be viewed as generalizations of these integrals. The nature of the generalization is that the probability P might not have a density, and yet we still want to compute a weighted integral using P . As motivation we'll continue with the previous example, where $h(s)$ is the payoff from outcome $s \in \mathbb{R}$, and the likelihood of outcomes is determined by probability measure P . To provide some basic regularity, we assume that h is a $\mathcal{B}$ -measurable function $\mathbb{R}$.

Actually our task is straightforward given the foundations already laid. In particular, we already have a general notion of integration suitable for understanding the expression $\int h(s)P(ds)$. In §4.1.4 we saw how to take an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and construct an expectation operator $\mathbb{E}$ that assigns to any integrable random variable a notion of expectation. Expectations can be thought of as integrals, as suggested by the alternative notation $\int x(\omega)\mathbb{P}(d\omega)$ for $\mathbb{E}x$ used in (4.16) on page 95.

So now take our distribution P on $\mathbb{R}$ and consider $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$ as a probability space. This new probability space has its own expectation operator $\mathbb{E}_P$. (See fact 4.1.5 if you wish to refresh your memory on this point.) Moreover we can use it to take the expectation of the payoff function h , which, being a $\mathcal{B}$ -measurable function on $\mathbb{R}$, is by definition a random variable on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$. This gives us a value $\mathbb{E}_P h$, which is exactly the integral we're after:

$$\mathbb{E}_P h := \int h(s)P(ds) := \text{the expectation of } h \text{ under } P$$

While the preceding discussion gives a neat definition of $\int h(s)P(ds)$, it would

be disappointing if this definition didn't lead us back to familiar expressions like $\int h(s)p(s) ds$ when P has a density. Fortunately, it does:

Fact 4.2.3 Let $h: \mathbb{R} \rightarrow \mathbb{R}$ be $\mathcal{B}$ -measurable and let P be a distribution on $\mathbb{R}$. If P is discrete, with PMF $\{p_j\}_{j \geq 1}$ and support $\{s_j\}_{j \geq 1}$, then

$$\int h(s)P(ds) = \sum_{j \geq 1} h(s_j)p_j \quad (4.28)$$

If P is absolutely continuous with density p , then

$$\int h(s)P(ds) = \int_{-\infty}^{\infty} h(s)p(s) ds \quad (4.29)$$

For a proof of fact 4.2.3, see §6.12 of Williams (1991).

4.2.4 Distributions of Random Variables

In §4.1.3 we defined random variables. Over the last few sections we've defined and discussed distributions over the Borel subsets of $\mathbb{R}$. Our next job is to connect them. We'll see that the connection runs in both directions: Every random variable defines a distribution on $\mathbb{R}$. Conversely, given any distribution, we can reverse engineer a random variable having that distribution.

To begin, let x be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By construction, the probability $\mathbb{P}\{x \in B\}$ is well-defined for every $B \in \mathcal{B}(\mathbb{R})$ (see (4.9)). The values obtained from this expression when x is held fixed and B is varied is called the **distribution of x** . In particular, the set function P defined by

$$P(B) = \mathbb{P}\{x \in B\} \quad (B \in \mathcal{B}(\mathbb{R})) \quad (4.30)$$

is both the distribution of x and also a distribution in the sense of §4.1.3 (i.e., a probability measure over the Borel sets). It records the probabilities of all outcomes of interest for the variable x when regarded as a single entity.³ In what follows we write $\mathcal{L}(x)$ for the distribution (or "law") of x . The relationship in (4.30) is summarized by writing $\mathcal{L}(x) = P$.

By (4.24), the CDF corresponding to the distribution P of x satisfies

$$F(s) = \mathbb{P}\{x \leq s\} \quad (s \in \mathbb{R}) \quad (4.31)$$

Sometimes F is called the **distribution function** or just distribution of x . Since F characterizes P , we also write $\mathcal{L}(x) = F$ to indicate that F represents the distribution of x .

3. When we want to combine x with other random variables we'll need their joint distribution. See §5.1.3.

Fact 4.2.4 If $\mathcal{L}(x) = F$, then $\mathbb{P}\{a < x \leq b\} = F(b) - F(a)$ for any $a \leq b$.

Indeed, if $a \leq b$, then $\{a < x \leq b\} = \{x \leq b\} \setminus \{x \leq a\}$ and $\{x \leq a\} \subset \{x \leq b\}$. Applying fact 4.1.1 on page 86 gives

$$\mathbb{P}\{a < x \leq b\} = \mathbb{P}\{x \leq b\} - \mathbb{P}\{x \leq a\} = F(b) - F(a)$$

From (4.31) we get for each random variable a uniquely defined CDF. It's also true that for every CDF F , there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $x: \Omega \rightarrow \mathbb{R}$ such that $\mathcal{L}(x) = F$; §7.4.1 outlines one construction.

In what follows, if $\mathcal{L}(x) = P$ and P has density p , we'll say that x **has density** p . If the distribution of x is discrete, we'll call x a **discrete random variable**. In the former case, where the distribution of x is absolutely continuous, no probability mass is placed on finite sets, leading to the next fact.

Fact 4.2.5 If x has a density, then $\mathbb{P}\{x = s\} = 0$ for all $s \in \mathbb{R}$, and for any $a < b$,

$$\mathbb{P}\{a < x < b\} = \mathbb{P}\{a < x \leq b\} = \mathbb{P}\{a \leq x < b\} = \mathbb{P}\{a \leq x \leq b\}$$

4.2.4.1 Distributions of Transformations

One often needs to obtain the distribution of the transform of a random variable. This is easy in the monotone case.

Fact 4.2.6 If $\mathcal{L}(x) = F$ and $y := \psi(x)$, where $\psi: \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing, then $\mathcal{L}(y) = G$ where $G(s) := F(\psi^{-1}(s))$.

To see why, observe that under these hypotheses ψ^{-1} exists and is (weakly) increasing. Hence

$$\mathbb{P}\{y \leq s\} = \mathbb{P}\{\psi(x) \leq s\} = \mathbb{P}\{x \leq \psi^{-1}(s)\} = F(\psi^{-1}(s))$$

Note how monotonicity is used in the second equality. (A related result can be derived for the strictly decreasing case, where the direction of the inequality flips.)

Example 4.2.15 If $\mathcal{L}(x) = F$ and $y := \exp(x)$, then the CDF of y is $G(s) := F(\ln(s))$.

An important related question is to determine when $y = \psi(x)$ has a density, and what the form of this density is.

Fact 4.2.7 If x has density p on $\mathbb{R}$ and $y := \psi(x)$ where ψ is a diffeomorphism on $\mathbb{R}$, then the distribution of y is absolutely continuous, with density

$$q(s) = p(\psi^{-1}(s)) \left| \frac{d\psi^{-1}(s)}{ds} \right| \quad (s \in \mathbb{R}) \quad (4.32)$$

Exercise 5.4.1 asks you to confirm (4.32). The term **diffeomorphism** just means that ψ is a bijection on $\mathbb{R}$ and both ψ and its inverse are differentiable.

Example 4.2.16 If x has density p on $\mathbb{R}$ and μ and σ are constants with $\sigma > 0$, then the density of $y := \mu + \sigma x$ is

$$q(s) = p\left(\frac{s - \mu}{\sigma}\right) \frac{1}{\sigma} \quad (s \in \mathbb{R})$$

When x is standard normal, this implies that $y = \mu + \sigma x$ is $N(\mu, \sigma^2)$, as can be seen by taking p to be the standard normal density ϕ and comparing with the expression for the $N(\mu, \sigma^2)$ density on page 106.

Fact 4.2.7 can be generalized. See, for example, theorem 15.5 of Schilling (2005).

4.2.5 Expectations from Distributions

Let x be a random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The distribution of x encodes all the information necessary to calculate expectations of x or of any $\mathcal{B}$ -measurable transformation $h(x)$. This is clearest when x is finite. For example, suppose that $\mathcal{L}(x) = P$, that $h: \mathbb{R} \rightarrow \mathbb{R}$ is any $\mathcal{B}$ -measurable function, and that P puts all mass on finite set $\{s_j\}_{j=1}^J$. Let $\{p_j\}$ be the PMF of P , so that $p_j = P\{s_j\}$. Applying the definition of expectations (see in particular (4.17) on page 96) and using $\mathbb{P}\{x = s_j\} = P\{s_j\}$ gives

$$\mathbb{E}h(x) = \sum_{j=1}^J h(s_j) \mathbb{P}\{x = s_j\} = \sum_{j=1}^J h(s_j) P\{s_j\} = \sum_{j=1}^J h(s_j) p_j \quad (4.33)$$

Comparing this with (4.28) on page 110, we can also write $\mathbb{E}h(x) = \int h(s) P(ds)$. Put differently, the expectation of $h(x)$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is equal to the expectation of h on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$. This is also true in the infinite case (see, e.g., theorem 5.2 of Çinlar (2011)), as stated in the next fact.

Fact 4.2.8 Let x be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{L}(x) = P$ and let h be a $\mathcal{B}$ -measurable function such that $h(x)$ is integrable. The expectation $\mathbb{E}h(x)$ is entirely determined by h and P . In particular,

$$\mathbb{E}h(x) = \int h(s) P(ds)$$

where $\int h(s) P(ds)$ is the expectation of h on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$.

Example 4.2.17 Let x be a random variable whose distribution P is the uniform distribution on $[a, b]$. Applying the definition of the uniform density in example 4.2.10, we obtain

$$\mathbb{E}x = \int sP(ds) = \int sp(s) ds = \int_{-\infty}^{\infty} \frac{s}{b-a} \mathbb{1}\{a \leq s \leq b\} ds$$

Solving the integral gives $\mathbb{E}x = \mu := (a + b)/2$. The variance is

$$\text{var}[x] = \int (s - \mu)^2 P(ds) = \int_a^b \left(s - \frac{a+b}{2}\right)^2 \frac{1}{b-a} ds = \frac{1}{12}(b-a)^2$$

Example 4.2.18 Suppose that $\mathcal{L}(x) = \mathcal{N}(\mu, \sigma)$. In the case where $\sigma > 0$, the mean can be computed via

$$\mathbb{E}x = \int_{-\infty}^{\infty} s \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\} ds = \mu$$

The variance is given by

$$\text{var}[x] = \int_{-\infty}^{\infty} (s - \mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\} ds = \sigma^2$$

For a discussion of how these integrals are solved, see §9.4 of Dudley (2002).

4.2.5.1 Moments from Distributions

In §4.1.5 we defined moments of random variables. For example, $\mathbb{E}x^k$ is the k th moment of x , assuming it exists. It is clear from fact 4.2.8 that any two random variables with the same distribution share the same moments. Hence moments are best thought of as a property of the distribution, not the random variable. Thus we define

- the **mean** of P as $\mu = \int sP(ds)$,
- the **k th moment** of P as $\int s^k P(ds)$,
- the **variance** of P as $\int (s - \mu)^2 P(ds)$,

and so on.

4.2.6 Quantile Functions

In this section it will be convenient to describe distributions in terms of CDFs. So let F be a CDF on $\mathbb{R}$, which, for the moment, we assume to be strictly increasing. Given $\tau \in (0, 1)$, the **τ th quantile** of F is the $\xi \in \mathbb{R}$ that solves $F(\xi) = \tau$. Under our

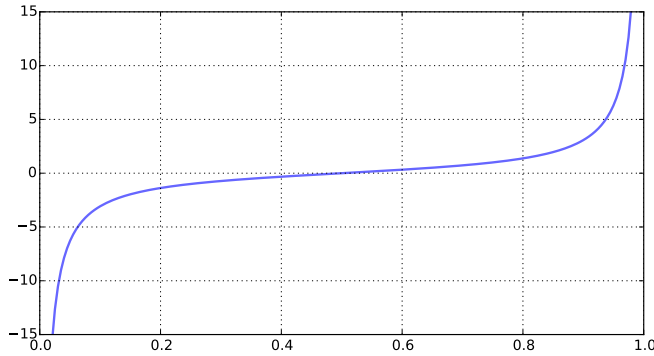


Figure 4.12 Cauchy quantile function

assumptions on F , such a ξ exists and is uniquely defined. The 0.5th quantile is called the **median** of F . It provides a measure of central tendency for the distribution.

We can of course express the τ th quantile of F using the inverse function:

$$F^{-1}(\tau) := \text{the unique } \xi \text{ such that } F(\xi) = \tau \quad (0 < \tau < 1) \quad (4.34)$$

Not surprisingly, the inverse of F is also called the **quantile function**.

Example 4.2.19 The quantile function associated with the standard Cauchy distribution (see page 103) is $F^{-1}(\tau) = \tan[\pi(\tau - 1/2)]$. Figure 4.12 illustrates. The horizontal axis is $\tau \in (0, 1)$.

If F is not strictly increasing, then there exist at least two distinct points s, s' such that $F(s) = F(s')$, and hence F^{-1} in (4.34) is not well-defined. This problem is negotiated by setting

$$F^{-1}(\tau) := \inf\{s \in \mathbb{R} : F(s) \geq \tau\} \quad (0 < \tau < 1) \quad (4.35)$$

(See §15.4.1 for a definition of $\inf$.) In the case where F is strictly increasing, (4.35) reduces to (4.34).

The quantile function is used in hypothesis testing to determine critical values. Let's outline a common scenario.

Fact 4.2.9 Let x be a random variable with density p . If p is symmetric, then the CDF

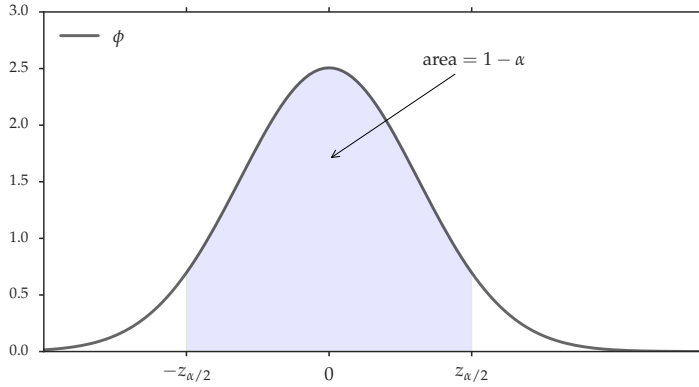


Figure 4.13 Critical values for the standard normal density

G of $y := |x|$ is given by

$$G(s) := \mathbb{P}\{y \leq s\} = \begin{cases} 2F(s) - 1 & \text{if } s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Here **symmetry** of p means that $p(s) = p(-s)$ for all $s \in \mathbb{R}$. This is equivalent to $F(s) = 1 - F(-s)$, as can be seen by differentiating or integrating. Exercise 4.4.18 asks you to prove fact 4.2.9.

Now take a random variable x with $\mathcal{L}(x) = F$ and constant $\alpha \in (0, 1)$ as given, and consider the c that solves $\mathbb{P}\{-c \leq x \leq c\} = 1 - \alpha$.

Fact 4.2.10 If $\mathcal{L}(x) = F$, x has a symmetric density and F is strictly increasing, then

$$c = F^{-1}(1 - \alpha/2) \implies \mathbb{P}\{-c \leq x \leq c\} = 1 - \alpha \quad (4.36)$$

To see this, fix $\alpha \in (0, 1)$ and let c be as in (4.36). From fact 4.2.9, we have

$$\mathbb{P}\{|x| \leq c\} = 2F(c) - 1 = 2F[F^{-1}(1 - \alpha/2)] - 1 = 1 - \alpha$$

In the case where F is the standard normal CDF Φ , this value c is usually denoted by $z_{\alpha/2}$. We will adopt the same notation:

$$z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2) \quad (4.37)$$

See figure 4.13.

4.3 Further Reading

The modern axiomatic treatment of probability based on measure theory was formulated in the 1930s by Andrei N. Kolmogorov. A discussion of his life and work can be found in Kendall et al. (1990). Nice introductory texts on measure-theoretic probability include Williams (1991) and Rosenthal (2006). Beautiful expositions at a slightly higher level can be found in Dudley (2002), Çinlar (2011), and Pollard (2002).

4.4 Exercises

Ex. 4.4.1 Show that if $\mathcal{F}$ is a σ -algebra and $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$ also holds.⁴

Ex. 4.4.2 Prove fact 4.1.2: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for any A, B .⁵

Ex. 4.4.3 Given sample space $\Omega := \{1, 2, 3\}$, let $A := \{1\}$, $B := \{2\}$, and $C := \{3\}$. Let $\mathbb{P}(A) = \mathbb{P}(B) = 1/3$. Compute $\mathbb{P}(C)$, $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B)$, $\mathbb{P}(A^c)$, $\mathbb{P}(A^c \cup B^c)$, and $\mathbb{P}(A | B)$. Are A and C independent?

Ex. 4.4.4 A die is designed so that the probability of getting face m is qm , where $m \in \{1, \dots, 6\}$ and q is a constant. Compute q .

Ex. 4.4.5 Prove the claims in fact 4.1.4 on page 90.

Ex. 4.4.6 Let Ω be a nonempty finite set, and let ω_0 be a fixed element of Ω . For each $A \subset \Omega$, define $\mathbb{P}(A) := \mathbb{1}\{\omega_0 \in A\}$. Is $\mathbb{P}$ a probability on Ω ? Why or why not?

Ex. 4.4.7 Let Ω be any sample space, and let $\mathbb{P}$ be a probability on the subsets $\mathcal{F}$. Let $A \in \mathcal{F}$. Show that if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, then A is independent of every other event in $\mathcal{F}$. Show that if A is independent of itself, then either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. Show that if A and B are independent, then A^c and B^c are also independent.

Ex. 4.4.8 Let $\mathbb{P}$ and Ω be defined as in example 4.1.6. Show that $\mathbb{P}$ is additive, in the sense that if A and B are disjoint events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Ex. 4.4.9 Let $\mathbb{P}$ and Ω be defined as in example 4.1.6. Let A be the event that the first flip is heads, and let B be the event that the second flip is heads. Show that A and B are independent under $\mathbb{P}$.

Ex. 4.4.10 Show that when Ω is finite, each random variable on Ω is finite.⁶

4. Hint: See the rules concerning set-theoretic operations in fact 15.1.1 on page 407.

5. Hint: Sketching the Venn diagram, convince yourself that $A = [(A \cup B) \setminus B] \cup (A \cap B)$. Finish the proof using the definition of a probability and fact 4.1.1 (page 86).

6. Hint: Review the definition of a function in §15.2.

Ex. 4.4.11 Show that any finite random variable can be expressed as

$$x = \sum_{j=1}^J s_j \mathbb{1}_{A_j} \quad (4.38)$$

where the scalars $\{s_j\}$ are distinct and the sets $\{A_j\}$ are events that partition Ω .

Ex. 4.4.12 For the generic finite random variable x defined in exercise 4.4.11, show that $\mathbb{E}x := \sum_{j=1}^J s_j \mathbb{P}(A_j)$.

Ex. 4.4.13 Fact 4.1.5 on page 95 states, among other things, that if x is a random variable and α is a constant, then $\mathbb{E}[\alpha x] = \alpha \mathbb{E}[x]$. Prove this for the case where x is a finite random variable using the definition (4.12).

Ex. 4.4.14 Fact 4.1.5 on page 95 states that, among other things, if x and y are two random variables, then $\mathbb{E}[x + y] = \mathbb{E}[x] + \mathbb{E}[y]$. Prove this for the binary random variables

$$x(\omega) = \mathbb{1}_A(\omega) \quad \text{and} \quad y(\omega) = \mathbb{1}_B(\omega) \quad (4.39)$$

using the definition (4.12).

Ex. 4.4.15 Let P be a discrete distribution with $P(\{s_j\}_{j \geq 1}) = 1$, and let $p_j := P\{s_j\}$. Show that $\sum_{j \geq 1} p_j = 1$.

Ex. 4.4.16 Let P, s_j and p_j be as in exercise 4.4.15. Show that if F is the CDF of P , then $F(s) = \sum_{j \geq 1} \mathbb{1}\{s_j \leq s\} p_j$ for all $s \in \mathbb{R}$.

Ex. 4.4.17 Let P, s_j and p_j be as in exercise 4.4.15. Show that $\lim_{s \rightarrow \infty} F(s) = 1$.

Ex. 4.4.18 Prove the claim in fact 4.2.9 on page 114.

Ex. 4.4.19 Let x be a discrete random variable taking values $s_1, \dots, s_J$, and let $p_j := \mathbb{P}\{x = s_j\}$. Show that $0 \leq p_j \leq 1$ for each j , and $\sum_{j=1}^J p_j = 1$.

Ex. 4.4.20 Let $\mathcal{L}(x) = F$ where F is the uniform CDF on $[0, 1]$. Give an expression for the CDF G of the random variable $y = x^2$.

Ex. 4.4.21 Let F be the CDF on $\mathbb{R}$ defined by $F(s) = e^s / (1 + e^s)$ for $s \in \mathbb{R}$.

(i) Obtain the quantile function corresponding to F .

(ii) Let $\mathcal{L}(x) = F$ and compute $\mathbb{P}\{0 \leq x \leq \ln 2\}$.

Ex. 4.4.22 Let $\mathcal{L}(y) = F$, where F is a CDF. Show that $F(s) = \mathbb{E} \mathbb{1}\{y \leq s\}$ for any s .

Ex. 4.4.23 Confirm monotonicity of expectations (see fact 4.1.5 on page 95) for the special case where x and y are the random variables in (4.39).

Ex. 4.4.24 Prove fact 4.1.7 (existence of k th moment implies existence of j th moment for all $j \leq k$).

Ex. 4.4.25 Confirm the expression for variance of linear combinations in fact 4.1.10.

Ex. 4.4.26 Let x and y be scalar random variables. With reference to fact 4.1.11 on page 98, is it true that $\text{corr}[\alpha x, \beta y] = \text{corr}[x, y]$ for *any* constant scalars α and β ? Why or why not?

Ex. 4.4.27 Confirm the solutions in (4.23).

Ex. 4.4.28 Consider the setting of §4.1.5. Let α^* , β^* and ℓ^* be as defined there. Let the prediction error u be defined as $u := y - \ell^*(x)$. Show that

- (i) $\mathbb{E} \ell^*(x) = \mathbb{E} y$
- (ii) $\text{var}[\ell^*(x)] = \text{corr}[x, y]^2 \text{var}[y]$
- (iii) $\text{cov}[\ell^*(x), u] = 0$
- (iv) $\text{var}[u] = (1 - \text{corr}[x, y]^2) \text{var}[y]$

Ex. 4.4.29 Prove the Chebyshev inequalities (4.18) and (4.19).

4.4.1 Solutions to Selected Exercises

Solution to Ex. 4.4.1. Let $\mathcal{F}$ be a σ -algebra and let $A, B \in \mathcal{F}$. The claim is that $A \cap B \in \mathcal{F}$. To see this, note that by the definition of a σ -algebra we must have $A^c \in \mathcal{F}$ and $B^c \in \mathcal{F}$. Since $\mathcal{F}$ is closed under the taking of unions, this means that $A^c \cup B^c \in \mathcal{F}$. Applying fact 15.1.1 on page 407 gives $(A \cap B)^c \in \mathcal{F}$. But then $A \cap B \in \mathcal{F}$, as was to be shown. $\square$

Solution to Ex. 4.4.2. Pick any sets $A, B \in \mathcal{F}$. To show that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

we start by decomposing A into the union of two disjoint sets: $A = [(A \cup B) \setminus B] \cup (A \cap B)$. Using additivity of $\mathbb{P}$, we then have

$$\mathbb{P}(A) = \mathbb{P}[(A \cup B) \setminus B] + \mathbb{P}(A \cap B)$$

Since $B \subset (A \cup B)$, we can apply part 1 of fact 4.1.1 (page 86) to obtain

$$\mathbb{P}(A) = \mathbb{P}(A \cup B) - \mathbb{P}(B) + \mathbb{P}(A \cap B)$$

Rearranging this expression gives the result that we are seeking. $\square$

Solution to Ex. 4.4.3. First $\mathbb{P}(C) = 1/3$ as $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) = 1/3 + 1/3 + \mathbb{P}(C)$, and hence $\mathbb{P}(C) = 1/3$. In addition $\mathbb{P}(A \cup B) = 2/3$, $\mathbb{P}(A \cap B) = 0$, $\mathbb{P}(A^c) = 2/3$, $\mathbb{P}(A^c \cup B^c) = \mathbb{P}((A \cap B)^c) = \mathbb{P}(\Omega) = 1$, and $\mathbb{P}(A \cap C) = 0 \neq 1/9 = \mathbb{P}(A)\mathbb{P}(C)$. Therefore A is not independent of C . $\square$

Solution to Ex. 4.4.4. When the die is rolled one face must come up, so the sum of the probabilities is one. More formally, letting $\Omega = \{1, \dots, 6\}$ be the sample space, we have

$$\mathbb{P}\{1, \dots, 6\} = \mathbb{P} \cup_{m=1}^6 \{m\} = \sum_{m=1}^6 \mathbb{P}\{m\} = \sum_{m=1}^6 qm = 1$$

Solving the last equality for q , we get $q = 1/21$. $\square$

Solution to Ex. 4.4.5. Let $A_1, \dots, A_N$ be subsets of Ω . Pick any $\omega \in \Omega$. Regarding part (i), suppose first that $\mathbb{1}_{\cap_{n=1}^N A_n}(\omega) = 1$. Then $\omega \in \cap_{n=1}^N A_n$, and hence $\omega \in A_n$ for all n . But then $\prod_{n=1}^N \mathbb{1}_{A_n}(\omega) = 1$. Conversely, if $\mathbb{1}_{\cap_{n=1}^N A_n}(\omega) = 0$, then ω is not in at least one A_n , so $\prod_{n=1}^N \mathbb{1}_{A_n}(\omega) = 0$. This verifies part (i) of fact 4.1.4. The proof of (ii) is similar. $\square$

Solution to Ex. 4.4.6. To show that $\mathbb{P}$ is a probability on Ω , we need to check that

- (i) $\mathbb{1}\{\omega_0 \in A\} \in [0, 1]$ for every $A \subset \Omega$.
- (ii) $\mathbb{1}\{\omega_0 \in \Omega\} = 1$.
- (iii) If $A \cap B = \emptyset$, then $\mathbb{1}\{\omega_0 \in A \cup B\} = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$.

Part (i) is immediate from the definition of an indicator function. Part (ii) holds because $\omega_0 \in \Omega$. Regarding (iii), pick any disjoint A and B . If $\omega_0 \in A$, then $\omega_0 \notin B$, and we have

$$\mathbb{1}\{\omega_0 \in A \cup B\} = 1 = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$$

If $\omega_0 \in B$, then $\omega_0 \notin A$, and once again we have

$$\mathbb{1}\{\omega_0 \in A \cup B\} = 1 = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$$

Finally, if ω_0 is in neither A nor B , then

$$\mathbb{1}\{\omega_0 \in A \cup B\} = 0 = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$$

We have shown that (i)–(iii) hold, and hence $\mathbb{P}$ is a probability on Ω . $\square$

Solution to Ex. 4.4.7. Suppose that $\mathbb{P}(A) = 0$ and that $B \in \mathcal{F}$. We claim that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, or, in this case, $\mathbb{P}(A \cap B) = 0$. Using nonnegativity and monotonic-

ity of $\mathbb{P}$ (fact 4.1.1), we obtain

$$0 \leq \mathbb{P}(A \cap B) \leq P(A) = 0$$

Therefore $\mathbb{P}(A \cap B) = 0$ as claimed.

Now suppose that $\mathbb{P}(A) = 1$. We claim that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, or, in this case, $\mathbb{P}(A \cap B) = \mathbb{P}(B)$. In view of fact 4.1.2 on page 86, we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$$

Since $\mathbb{P}(A) = 1$, it suffices to show that $\mathbb{P}(A \cup B) = 1$. This last equality is implied by monotonicity of $\mathbb{P}$, because $1 = \mathbb{P}(A) \leq \mathbb{P}(A \cup B) \leq 1$.

Next, suppose that A is independent of itself. In that case we have $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2$. If $a = a^2$, then either $a = 0$ or $a = 1$.

Finally, let A and B be independent. We have

$$\mathbb{P}(A^c \cap B^c) = \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B)$$

Applying fact 4.1.2 and independence, we can transform the right-hand side to obtain

$$\mathbb{P}(A^c \cap B^c) = (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) = \mathbb{P}(A^c)\mathbb{P}(B^c)$$

In other words, A^c and B^c are independent. □

Solution to Ex. 4.4.8. The proof is almost identical to the proof of additivity in example 4.1.5 (page 85). □

Solution to Ex. 4.4.9. The proof of independence is essentially the same as the proof of independence of A and B in example 4.1.8 (page 87). □

Solution to Ex. 4.4.11. Let x be finite-valued, with range $\{s_1, \dots, s_J\}$. Let $A_j := \{x = s_j\}$. By the definition of a function (see §15.2), each $\omega \in \Omega$ lies in exactly one A_j . Hence $A_1, \dots, A_J$ is a partition of Ω , and $x = \sum_{j=1}^J s_j \mathbb{1}_{A_j}$. □

Solution to Ex. 4.4.12. Clearly x is finite-valued, so, by (4.12), the expectation is given by $\mathbb{E}x = \sum_{j=1}^J s_j \mathbb{P}\{x = s_j\}$. Hence, to confirm $\mathbb{E}x := \sum_{j=1}^J s_j \mathbb{P}(A_j)$, we need only show that $\{x = s_j\} = A_j$. That the last statement is true follows from (4.38) and the fact that $A_1, \dots, A_J$ is a partition of Ω . □

Solution to Ex. 4.4.13. Let x be finite with range $\{s_1, \dots, s_J\}$ and let α be any constant. We claim that $\mathbb{E}[\alpha x] = \alpha \mathbb{E}[x]$. If $\alpha = 0$ the result is trivial, as both sides are clearly

zero. If not then $\alpha x = \alpha s_j$ if and only if $x = s_j$. Hence, applying (4.12),

$$\mathbb{E}[\alpha x] = \sum_{j=1}^J \alpha s_j \mathbb{P}\{x = s_j\} = \alpha \left[\sum_{j=1}^J s_j \mathbb{P}\{x = s_j\} \right] = \alpha \mathbb{E}[x] \quad \square$$

Solution to Ex. 4.4.14. Consider the sum $x + y$. By this, we mean the random variable $(x + y)(\omega) := x(\omega) + y(\omega)$. We claim that $\mathbb{E}[x + y] = \mathbb{E}[x] + \mathbb{E}[y]$. To see this, note first that

$$(x + y)(\omega) = \mathbb{1}_{A \setminus B}(\omega) + \mathbb{1}_{B \setminus A}(\omega) + 2\mathbb{1}_{A \cap B}(\omega)$$

(To check this, just go through the different cases for ω , and verify that the right-hand side of this expression agrees with $x(\omega) + y(\omega)$. Sketching a Venn diagram will help.) Therefore, by the definition of expectation,

$$\mathbb{E}[x + y] = \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + 2\mathbb{P}(A \cap B) \quad (4.40)$$

Now observe that $A = (A \setminus B) \cup (A \cap B)$, and hence, by disjointness,

$$\mathbb{E}[x] := \mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$$

Performing a similar calculation with y produces

$$\mathbb{E}[y] := \mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)$$

Summing these equations produces the value on the right-hand side of (4.40), thereby confirming $\mathbb{E}[x + y] = \mathbb{E}[x] + \mathbb{E}[y]$. $\square$

Solution to Ex. 4.4.15. Let $P(\{s_j\}_{j \geq 1}) = 1$ and $p_j := P\{s_j\}$ as stated in the exercise. The set $\{s_j\}_{j \geq 1}$ can be written as $\cup_{j=1}^{\infty} \{s_j\}$. Since P is a distribution it is countably additive (see page 100) and hence $1 = P(\cup_{j=1}^{\infty} \{s_j\}) = \sum_{j=1}^{\infty} P\{s_j\} = \sum_{j=1}^{\infty} p_j$. $\square$

Solution to Ex. 4.4.16. Using additivity and the fact that $P\{s_j\}_{j \geq 1} = 1$, we have

$$F(s) = P(-\infty, s] = P\left(\bigcup_{j \text{ s.t. } s_j \leq s} \{s_j\}\right) = \sum_{j \text{ s.t. } s_j \leq s} P\{s_j\} = \sum_{j=1}^J \mathbb{1}\{s_j \leq s\} p_j \quad \square$$

Solution to Ex. 4.4.17. We are assuming that x has finite range, and hence takes only finitely many different values. Let m be the largest such value. For this m , we have

$$\lim_{s \rightarrow \infty} F_x(s) \geq F_x(m) = \mathbb{P}\{\omega \in \Omega : x(\omega) \leq m\} = \mathbb{P}(\Omega) = 1$$

(The inequality is due to the fact that F_x is increasing.) On the other hand,

$$\lim_{s \rightarrow \infty} F_x(s) = \lim_{s \rightarrow \infty} \mathbb{P}\{x \leq s\} \leq \lim_{s \rightarrow \infty} \mathbb{P}(\Omega) = 1$$

From these two inequalities we get $1 \leq \lim_{s \rightarrow \infty} F_x(s) \leq 1$, which is equivalent to $\lim_{s \rightarrow \infty} F_x(s) = 1$. $\square$

Solution to Ex. 4.4.18. Fix $s \geq 0$. Using additivity over disjoint sets, we have

$$F_{|x|}(s) := \mathbb{P}\{|x| \leq s\} = \mathbb{P}\{-s \leq x \leq s\} = \mathbb{P}\{x = -s\} + \mathbb{P}\{-s < x \leq s\}$$

By assumption, $\mathbb{P}\{x = -s\} = 0$. Applying fact 4.2.4 on page 111 then yields

$$F_{|x|}(s) = \mathbb{P}\{-s < x \leq s\} = F(s) - F(-s)$$

As discussed after fact 4.2.9, symmetry is equivalent to $F(-s) = 1 - F(s)$. Plugging this in to the right-hand side of the previous expression gives $F_{|x|}(s) = 2F(s) - 1$. $\square$

Solution to Ex. 4.4.19. That $0 \leq p_j \leq 1$ for each j follows immediately from the definition of $\mathbb{P}$. In addition, using additivity of $\mathbb{P}$, we have

$$\sum_{j=1}^J p_j = \sum_{j=1}^J \mathbb{P}\{x = s_j\} = \mathbb{P} \bigcup_{j=1}^J \{x = s_j\} = \mathbb{P}(\Omega) = 1 \quad (4.41)$$

(We are using the fact that the sets $\{x = s_j\}$ are disjoint. Why is this always true? Look at the definition of a function given in §15.2.) $\square$

Solution to Ex. 4.4.20. Evidently $G(s) = 0$ when $s < 0$. For $s \geq 0$ we have

$$\mathbb{P}\{x^2 \leq s\} = \mathbb{P}\{|x| \leq \sqrt{s}\} = \mathbb{P}\{x \leq \sqrt{s}\} = F(\sqrt{s})$$

Thus, $G(s) = F(\sqrt{s})\mathbb{1}\{s \geq 0\}$. $\square$

Solution to Ex. 4.4.23. If $x(\omega) := \mathbb{1}\{\omega \in A\} \leq \mathbb{1}\{\omega \in B\} =: y(\omega)$ for any $\omega \in \Omega$, then $A \subset B$. (If $\omega \in A$, then $x(\omega) = 1$. Since $x(\omega) \leq y(\omega) \leq 1$, we have $y(\omega) = 1$, and hence $\omega \in B$.) Using fact 4.1.5 and monotonicity of $\mathbb{P}$, we then have

$$\mathbb{E}x = \mathbb{E}\mathbb{1}\{\omega \in A\} = \mathbb{P}(A) \leq \mathbb{P}(B) = \mathbb{E}\mathbb{1}\{\omega \in B\} = \mathbb{E}y \quad \square$$

Solution to Ex. 4.4.24. Let a be any nonnegative number, and let $j \leq k$. If $a \geq 1$, then $a^j \leq a^k$. If $a < 1$, then $a^j \leq 1$. Thus, for any $a \geq 0$, we have $a^j \leq a^k + 1$, and for any random variable x we have $|x|^j \leq |x|^k + 1$. Using linearity and monotonicity of

expectations and $\mathbb{E}1 = 1$, we then have $\mathbb{E}|x|^j \leq \mathbb{E}|x|^k + 1$. Hence the j th moment exists whenever the k th moment exists. $\square$

Solution to Ex. 4.4.25. We have

$$\begin{aligned}
 \text{var} \left[\sum_{n=1}^N \alpha_n x_n \right] &= \mathbb{E} \left[\left(\sum_{n=1}^N \alpha_n x_n - \mathbb{E} \left[\sum_{n=1}^N \alpha_n x_n \right] \right)^2 \right] \\
 &= \mathbb{E} \left[\left(\sum_{n=1}^N \alpha_n (x_n - \mathbb{E}[x_n]) \right)^2 \right] \\
 &= \mathbb{E} \left[\sum_{n=1}^N \alpha_n^2 (x_n - \mathbb{E}[x_n])^2 + 2 \sum_{n < m} \alpha_n \alpha_m (x_n - \mathbb{E}[x_n]) (x_m - \mathbb{E}[x_m]) \right] \\
 &= \sum_{n=1}^N \alpha_n^2 \text{var}[x_n] + \sum_{n < m} \alpha_n \alpha_m \text{cov}[x_n, x_m]
 \end{aligned}$$

as required. $\square$

Solution to Ex. 4.4.29. Pick any nonnegative random variable x and $\delta > 0$. By considering what happens at an arbitrary $\omega \in \Omega$, you should be able to convince yourself that

$$x = \mathbb{1}\{x \geq \delta\}x + \mathbb{1}\{x < \delta\}x \geq \mathbb{1}\{x \geq \delta\}\delta$$

Using the results in fact 4.1.5 (page 95) and rearranging gives (4.18). Regarding (4.19), observe that

$$x^2 = \mathbb{1}\{|x| \geq \delta\}x^2 + \mathbb{1}\{|x| < \delta\}x^2 \geq \mathbb{1}\{|x| \geq \delta\}\delta^2$$

Proceeding as before leads to (4.19). $\square$