Introduction to Probability

Dimitri P. Bertsekas and John N. Tsitsiklis

These class notes are the currently used textbook for ``Probabilistic Systems Analysis," an introductory probability course at the Massachusetts Institute of Technology. The text of the notes is quite polished and complete, but the problems are less so.

The course is attended by a large number of undergraduate and graduate students with diverse backgrounds. Acccordingly, we have tried to strike a balance between simplicity in exposition and sophistication in analytical reasoning. Some of the more mathematically rigorous analysis has been just sketched or intuitively explained in the text, so that complex proofs do not stand in the way of an otherwise simple exposition. At the same time, some of this analysis and the necessary mathematical results are developed (at the level of advanced calculus) in theoretical problems, which are included at the end of the corresponding chapter. The theoretical problems (marked by *) constitute an important component of the text, and ensure that the mathematically oriented reader will find here a smooth development without major gaps.

We give solutions to all the problems, aiming to enhance the utility of the notes for self-study. We have additional problems, suitable for homework assignment (with solutions), which we make available to instructors.

Our intent is to gradually improve and eventually publish the notes as a textbook, and your comments will be appreciated

Dimitri P. Bertsekas (bertsekas@lids.mit.edu)

John N. Tsitsiklis (jnt@mit.edu)

LECTURE NOTES

Course 6.041-6.431

M.I.T.

FALL 2000

Introduction to Probability

Dimitri P. Bertsekas and John N. Tsitsiklis

Professors of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Contents

Preface

These class notes are the currently used textbook for “Probabilistic Systems Analysis,” an introductory probability course at the Massachusetts Institute of Technology. The text of the notes is quite polished and complete, but the problems are less so.

The course is attended by a large number of undergraduate and graduate students with diverse backgrounds. Acccordingly, we have tried to strike a balance between simplicity in exposition and sophistication in analytical reasoning. Some of the more mathematically rigorous analysis has been just sketched or intuitively explained in the text, so that complex proofs do not stand in the way of an otherwise simple exposition. At the same time, some of this analysis and the necessary mathematical results are developed (at the level of advanced calculus) in theoretical problems, which are included at the end of the corresponding chapter. The theoretical problems (marked by *) constitute an important component of the text, and ensure that the mathematically oriented reader will find here a smooth development without major gaps.

We give solutions to all the problems, aiming to enhance the utility of the notes for self-study. We have additional problems, suitable for homework assignment (with solutions), which we make available to instructors.

Our intent is to gradually improve and eventually publish the notes as a textbook, and your comments will be appreciated

Dimitri P. Bertsekas
bertsekas@lids.mit.edu

John N. Tsitsiklis
jnt@mit.edu

1

Sample Space and Probability

Contents

"Probability" is a very useful concept, but can be interpreted in a number of ways. As an illustration, consider the following.

> A patient is admitted to the hospital and a potentially life-saving drug is administered. The following dialog takes place between the nurse and a concerned relative.
>
> RELATIVE: Nurse, what is the probability that the drug will work?
> NURSE: I hope it works, we'll know tomorrow.
> RELATIVE: Yes, but what is the probability that it will?
> NURSE: Each case is different, we have to wait.
> RELATIVE: But let's see, out of a hundred patients that are treated under similar conditions, how many times would you expect it to work?
> NURSE (somewhat annoyed): I told you, every person is different, for some it works, for some it doesn't.
> RELATIVE (insisting): Then tell me, if you had to bet whether it will work or not, which side of the bet would you take?
> NURSE (cheering up for a moment): I'd bet it will work.
> RELATIVE (somewhat relieved): OK, now, would you be willing to lose two dollars if it doesn't work, and gain one dollar if it does?
> NURSE (exasperated): What a sick thought! You are wasting my time!

In this conversation, the relative attempts to use the concept of probability to discuss an **uncertain** situation. The nurse's initial response indicates that the meaning of "probability" is not uniformly shared or understood, and the relative tries to make it more concrete. The first approach is to define probability in terms of **frequency of occurrence**, as a percentage of successes in a moderately large number of similar situations. Such an interpretation is often natural. For example, when we say that a perfectly manufactured coin lands on heads "with probability 50%," we typically mean "roughly half of the time." But the nurse may not be entirely wrong in refusing to discuss in such terms. What if this was an experimental drug that was administered for the very first time in this hospital or in the nurse's experience?

While there are many situations involving uncertainty in which the frequency interpretation is appropriate, there are other situations in which it is not. Consider, for example, a scholar who asserts that the Iliad and the Odyssey were composed by the same person, with probability 90%. Such an assertion conveys some information, but not in terms of frequencies, since the subject is a one-time event. Rather, it is an expression of the scholar's **subjective belief**. One might think that subjective beliefs are not interesting, at least from a mathematical or scientific point of view. On the other hand, people often have to make choices in the presence of uncertainty, and a systematic way of making use of their beliefs is a prerequisite for successful, or at least consistent, decision making.

In fact, the choices and actions of a rational person, can reveal a lot about the inner-held subjective probabilities, even if the person does not make conscious use of probabilistic reasoning. Indeed, the last part of the earlier dialog was an attempt to infer the nurse's beliefs in an indirect manner. Since the nurse was willing to accept a one-for-one bet that the drug would work, we may infer that the probability of success was judged to be at least 50%. And had the nurse accepted the last proposed bet (two-for-one), that would have indicated a success probability of at least 2/3.

Rather than dwelling further into philosophical issues about the appropriateness of probabilistic reasoning, we will simply take it as a given that the theory of probability is useful in a broad variety of contexts, including some where the assumed probabilities only reflect subjective beliefs. There is a large body of successful applications in science, engineering, medicine, management, etc., and on the basis of this empirical evidence, probability theory is an extremely useful tool.

Our main objective in this book is to develop the art of describing uncertainty in terms of probabilistic models, as well as the skill of probabilistic reasoning. The first step, which is the subject of this chapter, is to describe the generic structure of such models, and their basic properties. The models we consider assign probabilities to collections (sets) of possible outcomes. For this reason, we must begin with a short review of set theory.

1.1 SETS

Probability makes extensive use of set operations, so let us introduce at the outset the relevant notation and terminology.

A **set** is a collection of objects, which are the **elements** of the set. If S is a set and x is an element of S, we write $x \in S$. If x is not an element of S, we write $x \notin S$. A set can have no elements, in which case it is called the **empty set**, denoted by $\emptyset$.

Sets can be specified in a variety of ways. If S contains a finite number of elements, say $x_1, x_2, \ldots, x_n$, we write it as a list of the elements, in braces:

$$S = \{x_1, x_2, \ldots, x_n\}.$$

For example, the set of possible outcomes of a die roll is $\{1, 2, 3, 4, 5, 6\}$, and the set of possible outcomes of a coin toss is $\{H, T\}$, where H stands for "heads" and T stands for "tails."

If S contains infinitely many elements $x_1, x_2, \ldots$, which can be enumerated in a list (so that there are as many elements as there are positive integers) we write

$$S = \{x_1, x_2, \ldots\},$$

and we say that S is **countably infinite**. For example, the set of even integers can be written as $\{0, 2, -2, 4, -4, \ldots\}$, and is countably infinite.

Alternatively, we can consider the set of all x that have a certain property P, and denote it by

$$\{x \,|\, x \text{ satisfies } P\}.$$

(The symbol "|" is to be read as "such that.") For example the set of even integers can be written as $\{k \,|\, k/2 \text{ is integer}\}$. Similarly, the set of all scalars x in the interval $[0, 1]$ can be written as $\{x \,|\, 0 \leq x \leq 1\}$. Note that the elements x of the latter set take a continuous range of values, and cannot be written down in a list (a proof is sketched in the end-of-chapter problems); such a set is said to be **uncountable**.

If every element of a set S is also an element of a set T, we say that S is a **subset** of T, and we write $S \subset T$ or $T \supset S$. If $S \subset T$ and $T \subset S$, the two sets are **equal**, and we write $S = T$. It is also expedient to introduce a **universal set**, denoted by Ω, which contains all objects that could conceivably be of interest in a particular context. Having specified the context in terms of a universal set Ω, we only consider sets S that are subsets of Ω.

Set Operations

The **complement** of a set S, with respect to the universe Ω, is the set $\{x \in \Omega \,|\, x \notin S\}$ of all elements of Ω that do not belong to S, and is denoted by S^c. Note that $\Omega^c = \emptyset$.

The **union** of two sets S and T is the set of all elements that belong to S or T (or both), and is denoted by $S \cup T$. The **intersection** of two sets S and T is the set of all elements that belong to both S and T, and is denoted by $S \cap T$. Thus,

$$S \cup T = \{x \,|\, x \in S \text{ or } x \in T\},$$
$$S \cap T = \{x \,|\, x \in S \text{ and } x \in T\}.$$

In some cases, we will have to consider the union or the intersection of several, even infinitely many sets, defined in the obvious way. For example, if for every positive integer n, we are given a set S_n, then

$$\bigcup_{n=1}^{\infty} S_n = S_1 \cup S_2 \cup \cdots = \{x \,|\, x \in S_n \text{ for some } n\},$$

and

$$\bigcap_{n=1}^{\infty} S_n = S_1 \cap S_2 \cap \cdots = \{x \,|\, x \in S_n \text{ for all } n\}.$$

Two sets are said to be disjoint if their intersection is empty. More generally, several sets are said to be **disjoint** if no two of them have a common element. A collection of sets is said to be a **partition** of a set S if the sets in the collection are disjoint and their union is S.

If x and y are two objects, we use (x, y) to denote the **ordered pair** of x and y. The set of scalars (real numbers) is denoted by $\Re$; the set of pairs (or triplets) of scalars, i.e., the two-dimensional plane (or three-dimensional space, respectively) is denoted by $\Re^2$ (or $\Re^3$, respectively).

Sets and the associated operations are easy to visualize in terms of **Venn diagrams**, as illustrated in Fig. 1.1.

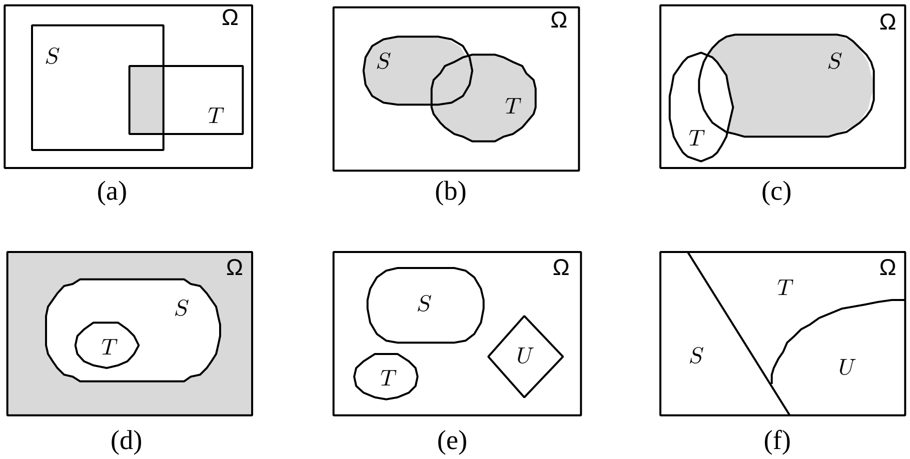


Figure 1.1: Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subset S$. The shaded region is the complement of S. (e) The sets S, T, and U are disjoint. (f) The sets S, T, and U form a partition of the set Ω.

The Algebra of Sets

Set operations have several properties, which are elementary consequences of the definitions. Some examples are:

$$\begin{array}{rlrl} S \cup T &= T \cup S, & S \cup (T \cup U) &= (S \cup T) \cup U, \\ S \cap (T \cup U) &= (S \cap T) \cup (S \cap U), & S \cup (T \cap U) &= (S \cup T) \cap (S \cup U), \\ (S^c)^c &= S, & S \cap S^c &= \emptyset, \\ S \cup \Omega &= \Omega, & S \cap \Omega &= S. \end{array}$$

Two particularly useful properties are given by **de Morgan's laws** which state that

$$\left(\bigcup_n S_n\right)^c = \bigcap_n S_n^c, \qquad \left(\bigcap_n S_n\right)^c = \bigcup_n S_n^c.$$

To establish the first law, suppose that $x \in (\cup_n S_n)^c$. Then, $x \notin \cup_n S_n$, which implies that for every n, we have $x \notin S_n$. Thus, x belongs to the complement of every S_n, and $x_n \in \cap_n S_n^c$. This shows that $(\cup_n S_n)^c \subset \cap_n S_n^c$. The converse inclusion is established by reversing the above argument, and the first law follows. The argument for the second law is similar.

1.2 PROBABILISTIC MODELS

A probabilistic model is a mathematical description of an uncertain situation. It must be in accordance with a fundamental framework that we discuss in this section. Its two main ingredients are listed below and are visualized in Fig. 1.2.

Elements of a Probabilistic Model

- The **sample space** Ω, which is the set of all possible outcomes of an experiment.
- The **probability law**, which assigns to a set A of possible outcomes (also called an **event**) a nonnegative number $\mathbf{P}(A)$ (called the **probability** of A) that encodes our knowledge or belief about the collective "likelihood" of the elements of A. The probability law must satisfy certain properties to be introduced shortly.

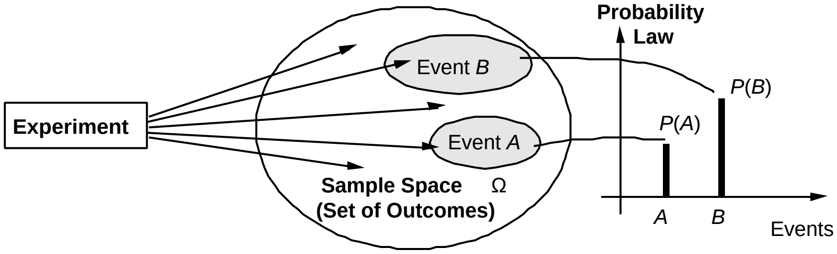


Figure 1.2: The main ingredients of a probabilistic model.

Sample Spaces and Events

Every probabilistic model involves an underlying process, called the **experiment**, that will produce exactly one out of several possible **outcomes**. The set of all possible outcomes is called the **sample space** of the experiment, and is denoted by Ω. A subset of the sample space, that is, a collection of possible outcomes, is called an **event**.[†] There is no restriction on what constitutes an experiment. For example, it could be a single toss of a coin, or three tosses, or an infinite sequence of tosses. However, it is important to note that in our formulation of a probabilistic model, there is only one experiment. So, three tosses of a coin constitute a single experiment, rather than three experiments.

The sample space of an experiment may consist of a finite or an infinite number of possible outcomes. Finite sample spaces are conceptually and mathematically simpler. Still, sample spaces with an infinite number of elements are quite common. For an example, consider throwing a dart on a square target and viewing the point of impact as the outcome.

Choosing an Appropriate Sample Space

Regardless of their number, different elements of the sample space should be distinct and **mutually exclusive** so that when the experiment is carried out, there is a unique outcome. For example, the sample space associated with the roll of a die cannot contain "1 or 3" as a possible outcome and also "1 or 4" as another possible outcome. When the roll is a 1, the outcome of the experiment would not be unique.

A given physical situation may be modeled in several different ways, depending on the kind of questions that we are interested in. Generally, the sample space chosen for a probabilistic model must be **collectively exhaustive**, in the sense that no matter what happens in the experiment, we always obtain an outcome that has been included in the sample space. In addition, the sample space should have enough detail to distinguish between all outcomes of interest to the modeler, while avoiding irrelevant details.

Example 1.1. Consider two alternative games, both involving ten successive coin tosses:

> *Game 1:* We receive $1 each time a head comes up.
>
> *Game 2:* We receive $1 for every coin toss, up to and including the first time a head comes up. Then, we receive $2 for every coin toss, up to the second time a head comes up. More generally, the dollar amount per toss is doubled each time a head comes up.

[†] Any collection of possible outcomes, including the entire sample space Ω and its complement, the empty set $\emptyset$, may qualify as an event. Strictly speaking, however, some sets have to be excluded. In particular, when dealing with probabilistic models involving an uncountably infinite sample space, there are certain unusual subsets for which one cannot associate meaningful probabilities. This is an intricate technical issue, involving the mathematics of measure theory. Fortunately, such pathological subsets do not arise in the problems considered in this text or in practice, and the issue can be safely ignored.

In game 1, it is only the total number of heads in the ten-toss sequence that matters, while in game 2, the order of heads and tails is also important. Thus, in a probabilistic model for game 1, we can work with a sample space consisting of eleven possible outcomes, namely, $0, 1, \ldots, 10$. In game 2, a finer grain description of the experiment is called for, and it is more appropriate to let the sample space consist of every possible ten-long sequence of heads and tails.

Sequential Models

Many experiments have an inherently sequential character, such as for example tossing a coin three times, or observing the value of a stock on five successive days, or receiving eight successive digits at a communication receiver. It is then often useful to describe the experiment and the associated sample space by means of a **tree-based sequential description**, as in Fig. 1.3.

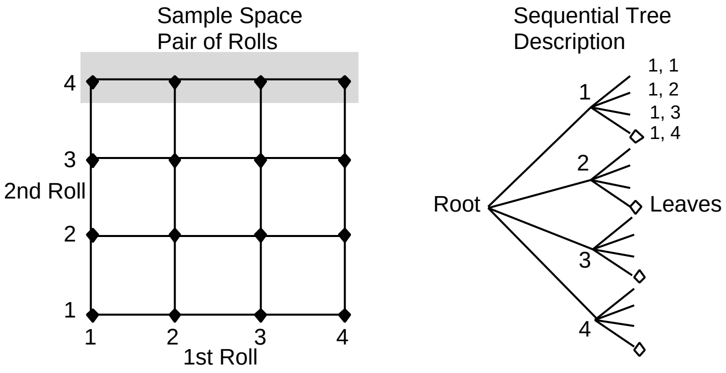


Figure 1.3: Two equivalent descriptions of the sample space of an experiment involving two rolls of a 4-sided die. The possible outcomes are all the ordered pairs of the form (i, j), where i is the result of the first roll, and j is the result of the second. These outcomes can be arranged in a 2-dimensional grid as in the figure on the left, or they can be described by the tree on the right, which reflects the sequential character of the experiment. Here, each possible outcome corresponds to a leaf of the tree and is associated with the unique path from the root to that leaf. The shaded area on the left is the event $\{(1, 4),\ (2, 4),\ (3, 4),\ (4, 4)\}$ that the result of the second roll is 4. That same event can be described as a set of leaves, as shown on the right. Note also that every node of the tree can be identified with an event, namely, the set of all leaves downstream from that node. For example, the node labeled by a 1 can be identified with the event $\{(1, 1), (1, 2), (1, 3), (1, 4)\}$ that the result of the first roll is 1.

Probability Laws

Suppose we have settled on the sample space Ω associated with an experiment.

Then, to complete the probabilistic model, we must introduce a **probability law**. Intuitively, this specifies the "likelihood" of any outcome, or of any set of possible outcomes (an event, as we have called it earlier). More precisely, the probability law assigns to every event A, a number $\mathbf{P}(A)$, called the **probability** of A, satisfying the following axioms.

Probability Axioms

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$, for every event A.
2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

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

Furthermore, if the sample space has an infinite number of elements and $A_1, A_2, \ldots$ is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, that is, $\mathbf{P}(\Omega) = 1$.

In order to visualize a probability law, consider a unit of mass which is to be "spread" over the sample space. Then, $\mathbf{P}(A)$ is simply the total mass that was assigned collectively to the elements of A. In terms of this analogy, the additivity axiom becomes quite intuitive: the total mass in a sequence of disjoint events is the sum of their individual masses.

A more concrete interpretation of probabilities is in terms of relative frequencies: a statement such as $\mathbf{P}(A) = 2/3$ often represents a belief that event A will materialize in about two thirds out of a large number of repetitions of the experiment. Such an interpretation, though not always appropriate, can sometimes facilitate our intuitive understanding. It will be revisited in Chapter 7, in our study of limit theorems.

There are many natural properties of a probability law which have not been included in the above axioms for the simple reason that they can be **derived** from them. For example, note that the normalization and additivity axioms imply that

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(\Omega \cup \emptyset) = \mathbf{P}(\Omega) + \mathbf{P}(\emptyset) = 1 + \mathbf{P}(\emptyset),$$

and this shows that the probability of the empty event is 0:

$$\mathbf{P}(\emptyset) = 0.$$

As another example, consider three disjoint events A_1, A_2, and A_3. We can use the additivity axiom for two disjoint events repeatedly, to obtain

$$\begin{aligned}\mathbf{P}(A_1 \cup A_2 \cup A_3) &= \mathbf{P}\big(A_1 \cup (A_2 \cup A_2)\big)\\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup A_3)\\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3).\end{aligned}$$

Proceeding similarly, we obtain that the probability of the union of finitely many disjoint events is always equal to the sum of the probabilities of these events. More such properties will be considered shortly.

Discrete Models

Here is an illustration of how to construct a probability law starting from some common sense assumptions about a model.

Example 1.2. Coin tosses. Consider an experiment involving a single coin toss. There are two possible outcomes, heads (H) and tails (T). The sample space is $\Omega = \{H, T\}$, and the events are

$$\{H, T\},\ \{H\},\ \{T\},\ \emptyset.$$

If the coin is fair, i.e., if we believe that heads and tails are "equally likely," we should assign equal probabilities to the two possible outcomes and specify that $\mathbf{P}\big(\{H\}\big) = \mathbf{P}\big(\{T\}\big) = 0.5$. The additivity axiom implies that

$$\mathbf{P}\big(\{H, T\}\big) = \mathbf{P}\big(\{H\}\big) + \mathbf{P}\big(\{T\}\big) = 1,$$

which is consistent with the normalization axiom. Thus, the probability law is given by

$$\mathbf{P}\big(\{H, T\}\big) = 1, \qquad \mathbf{P}\big(\{H\}\big) = 0.5, \qquad \mathbf{P}\big(\{T\}\big) = 0.5, \qquad \mathbf{P}(\emptyset) = 0,$$

and satisfies all three axioms.

Consider another experiment involving three coin tosses. The outcome will now be a 3-long string of heads or tails. The sample space is

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

We assume that each possible outcome has the same probability of 1/8. Let us construct a probability law that satisfies the three axioms. Consider, as an example, the event

$$A = \{\text{exactly 2 heads occur}\} = \{HHT,\ HTH,\ THH\}.$$

Using additivity, the probability of A is the sum of the probabilities of its elements:

$$\begin{aligned}\mathbf{P}\big(\{HHT, HTH, THH\}\big) &= \mathbf{P}\big(\{HHT\}\big) + \mathbf{P}\big(\{HTH\}\big) + \mathbf{P}\big(\{THH\}\big)\\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\\ &= \frac{3}{8}.\end{aligned}$$

Similarly, the probability of any event is equal to 1/8 times the number of possible outcomes contained in the event. This defines a probability law that satisfies the three axioms.

By using the additivity axiom and by generalizing the reasoning in the preceding example, we reach the following conclusion.

Discrete Probability Law

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event $\{s_1, s_2, \ldots, s_n\}$ is the sum of the probabilities of its elements:

$$\mathbf{P}\big(\{s_1, s_2, \ldots, s_n\}\big) = \mathbf{P}\big(\{s_1\}\big) + \mathbf{P}\big(\{s_2\}\big) + \cdots + \mathbf{P}\big(\{s_n\}\big).$$

In the special case where the probabilities $\mathbf{P}\big(\{s_1\}), \ldots, \mathbf{P}(\{s_n\}\big)$ are all the same (by necessity equal to $1/n$, in view of the normalization axiom), we obtain the following.

Discrete Uniform Probability Law

If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{Number of elements of } A}{n}.$$

Let us provide a few more examples of sample spaces and probability laws.

Example 1.3. Dice. Consider the experiment of rolling a pair of 4-sided dice (cf. Fig. 1.4). We assume the dice are fair, and we interpret this assumption to mean that each of the sixteen possible outcomes [ordered pairs (i, j), with $i, j = 1, 2, 3, 4$], has the same probability of 1/16. To calculate the probability of an event, we must count the number of elements of event and divide by 16 (the total number of possible outcomes). Here are some event probabilities calculated in this way:

$$\begin{aligned}
\mathbf{P}\big(\{\text{the sum of the rolls is even}\}\big) &= 8/16 = 1/2, \\
\mathbf{P}\big(\{\text{the sum of the rolls is odd}\}\big) &= 8/16 = 1/2, \\
\mathbf{P}\big(\{\text{the first roll is equal to the second}\}\big) &= 4/16 = 1/4, \\
\mathbf{P}\big(\{\text{the first roll is larger than the second}\}\big) &= 6/16 = 3/8, \\
\mathbf{P}\big(\{\text{at least one roll is equal to 4}\}\big) &= 7/16.
\end{aligned}$$

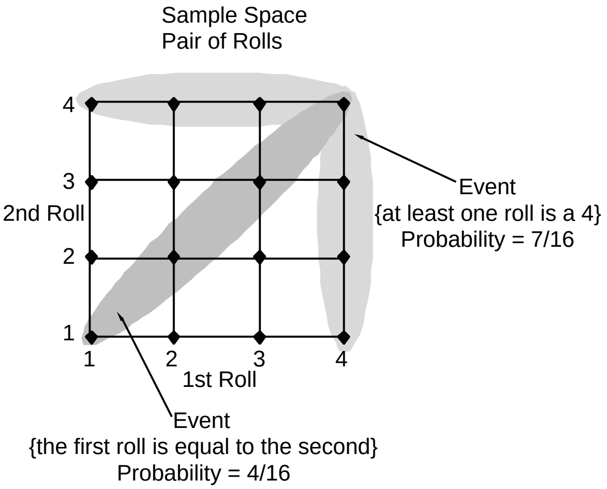


Figure 1.4: Various events in the experiment of rolling a pair of 4-sided dice, and their probabilities, calculated according to the discrete uniform law.

Continuous Models

Probabilistic models with continuous sample spaces differ from their discrete counterparts in that the probabilities of the single-element events may not be sufficient to characterize the probability law. This is illustrated in the following examples, which also illustrate how to generalize the uniform probability law to the case of a continuous sample space.

Example 1.4. A wheel of fortune is continuously calibrated from 0 to 1, so the possible outcomes of an experiment consisting of a single spin are the numbers in the interval $\Omega = [0, 1]$. Assuming a fair wheel, it is appropriate to consider all outcomes equally likely, but what is the probability of the event consisting of a single element? It cannot be positive, because then, using the additivity axiom, it would follow that events with a sufficiently large number of elements would have probability larger than 1. Therefore, the probability of any event that consists of a single element must be 0.

In this example, it makes sense to assign probability $b - a$ to any subinterval $[a, b]$ of $[0, 1]$, and to calculate the probability of a more complicated set by evaluating its "length."† This assignment satisfies the three probability axioms and qualifies as a legitimate probability law.

Example 1.5. Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being equally likely. The first to arrive will wait for 15 minutes and will leave if the other has not yet arrived. What is the probability that they will meet?

Let us use as sample space the square $\Omega = [0, 1] \times [0, 1]$, whose elements are the possible pairs of delays for the two of them. Our interpretation of "equally likely" pairs of delays is to let the probability of a subset of Ω be equal to its area. This probability law satisfies the three probability axioms. The event that Romeo and Juliet will meet is the shaded region in Fig. 1.5, and its probability is calculated to be 7/16.

Properties of Probability Laws

Probability laws have a number of properties, which can be deduced from the axioms. Some of them are summarized below.

Some Properties of Probability Laws

Consider a probability law, and let A, B, and C be events.

(a) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.

(b) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

(c) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.

(d) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

† The "length" of a subset S of $[0, 1]$ is the integral $\int_S dt$, which is defined, for "nice" sets S, in the usual calculus sense. For unusual sets, this integral may not be well defined mathematically, but such issues belong to a more advanced treatment of the subject.

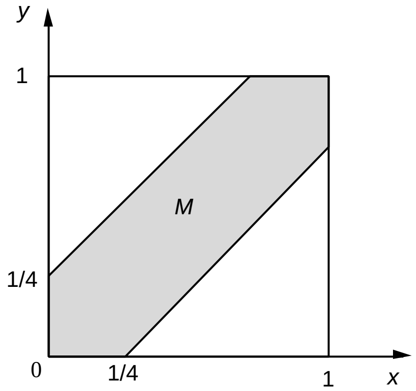


Figure 1.5: The event M that Romeo and Juliet will arrive within 15 minutes of each other (cf. Example 1.5) is

$$M = \Big\{(x, y) \;\Big|\; |x - y| \le 1/4,\ 0 \le x \le 1,\ 0 \le y \le 1\Big\},$$

and is shaded in the figure. The area of M is 1 minus the area of the two unshaded triangles, or $1 - (3/4) \cdot (3/4) = 7/16$. Thus, the probability of meeting is 7/16.

These properties, and other similar ones, can be visualized and verified graphically using Venn diagrams, as in Fig. 1.6. For a further example, note that we can apply property (c) repeatedly and obtain the inequality

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \le \sum_{i=1}^{n} \mathbf{P}(A_i).$$

In more detail, let us apply property (c) to the sets A_1 and $A_2 \cup \cdots \cup A_n$, to obtain

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \le \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup \cdots \cup A_n).$$

We also apply property (c) to the sets A_2 and $A_3 \cup \cdots \cup A_n$ to obtain

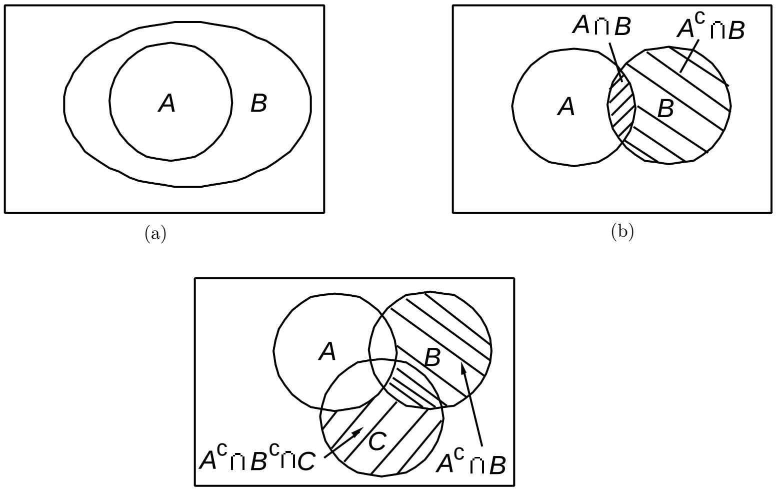
$$\mathbf{P}(A_2 \cup \cdots \cup A_n) \le \mathbf{P}(A_2) + \mathbf{P}(A_3 \cup \cdots \cup A_n),$$

continue similarly, and finally add.

Models and Reality

Using the framework of probability theory to analyze a physical but uncertain situation, involves two distinct stages.

(a) In the first stage, we construct a probabilistic model, by specifying a probability law on a suitably defined sample space. There are no hard rules to



(a)

(b)

(c)

Figure 1.6: Visualization and verification of various properties of probability laws using Venn diagrams. If $A \subset B$, then B is the union of the two disjoint events A and $A^c \cap B$; see diagram (a). Therefore, by the additivity axiom, we have

$$\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) \geq \mathbf{P}(A),$$

where the inequality follows from the nonnegativity axiom, and verifies property (a).

From diagram (b), we can express the events $A \cup B$ and B as unions of disjoint events:

$$A \cup B = A \cup (A^c \cap B), \qquad B = (A \cap B) \cup (A^c \cap B).$$

The additivity axiom yields

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B), \qquad \mathbf{P}(B) = \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B).$$

Subtracting the second equality from the first and rearranging terms, we obtain $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, verifying property (b). Using also the fact $\mathbf{P}(A \cap B) \geq 0$ (the nonnegativity axiom), we obtain $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$, verifying property (c)

From diagram (c), we see that the event $A \cup B \cup C$ can be expressed as a union of three disjoint events:

$$A \cup B \cup C = A \cup (A^c \cap B) \cup (A^c \cap B^c \cap C),$$

so property (d) follows as a consequence of the additivity axiom.

guide this step, other than the requirement that the probability law conform to the three axioms. Reasonable people may disagree on which model best represents reality. In many cases, one may even want to use a somewhat "incorrect" model, if it is simpler than the "correct" one or allows for tractable calculations. This is consistent with common practice in science and engineering, where the choice of a model often involves a tradeoff between accuracy, simplicity, and tractability. Sometimes, a model is chosen on the basis of historical data or past outcomes of similar experiments. Systematic methods for doing so belong to the field of **statistics**, a topic that we will touch upon in the last chapter of this book.

(b) In the second stage, we work within a fully specified probabilistic model and derive the probabilities of certain events, or deduce some interesting properties. While the first stage entails the often open-ended task of connecting the real world with mathematics, the second one is tightly regulated by the rules of ordinary logic and the axioms of probability. Difficulties may arise in the latter if some required calculations are complex, or if a probability law is specified in an indirect fashion. Even so, there is no room for ambiguity: all conceivable questions have precise answers and it is only a matter of developing the skill to arrive at them.

Probability theory is full of "paradoxes" in which different calculation methods seem to give different answers to the same question. Invariably though, these apparent inconsistencies turn out to reflect poorly specified or ambiguous probabilistic models.

1.3 CONDITIONAL PROBABILITY

Conditional probability provides us with a way to reason about the outcome of an experiment, based on **partial information**. Here are some examples of situations we have in mind:

(a) In an experiment involving two successive rolls of a die, you are told that the sum of the two rolls is 9. How likely is it that the first roll was a 6?

(b) In a word guessing game, the first letter of the word is a "t". What is the likelihood that the second letter is an "h"?

(c) How likely is it that a person has a disease given that a medical test was negative?

(d) A spot shows up on a radar screen. How likely is it that it corresponds to an aircraft?

In more precise terms, given an experiment, a corresponding sample space, and a probability law, suppose that we know that the outcome is within some given event B. We wish to quantify the likelihood that the outcome also belongs to some other given event A. We thus seek to construct a new probability law, which takes into account this knowledge and which, for any event A, gives us the **conditional probability of A given B**, denoted by $\mathbf{P}(A \,|\, B)$.

We would like the conditional probabilities $\mathbf{P}(A \,|\, B)$ of different events A to constitute a legitimate probability law, that satisfies the probability axioms. They should also be consistent with our intuition in important special cases, e.g., when all possible outcomes of the experiment are equally likely. For example, suppose that all six possible outcomes of a fair die roll are equally likely. If we are told that the outcome is even, we are left with only three possible outcomes, namely, 2, 4, and 6. These three outcomes were equally likely to start with, and so they should remain equally likely given the additional knowledge that the outcome was even. Thus, it is reasonable to let

$$\mathbf{P}(\text{the outcome is 6} \,|\, \text{the outcome is even}) = \frac{1}{3}.$$

This argument suggests that an appropriate definition of conditional probability when all outcomes are equally likely, is given by

$$\mathbf{P}(A \,|\, B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

Generalizing the argument, we introduce the following definition of conditional probability:

$$\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

where we assume that $\mathbf{P}(B) > 0$; the conditional probability is undefined if the conditioning event has zero probability. In words, out of the total probability of the elements of B, $\mathbf{P}(A \,|\, B)$ is the fraction that is assigned to possible outcomes that also belong to A.

Conditional Probabilities Specify a Probability Law

For a fixed event B, it can be verified that the conditional probabilities $\mathbf{P}(A \,|\, B)$ form a legitimate probability law that satisfies the three axioms. Indeed, nonnegativity is clear. Furthermore,

$$\mathbf{P}(\Omega \,|\, B) = \frac{\mathbf{P}(\Omega \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1,$$

and the normalization axiom is also satisfied. In fact, since we have $\mathbf{P}(B \,|\, B) = \mathbf{P}(B)/\mathbf{P}(B) = 1$, all of the conditional probability is concentrated on B. Thus, we might as well discard all possible outcomes outside B and treat the conditional probabilities as a probability law defined on the new universe B.

To verify the additivity axiom, we write for any two disjoint events A_1 and A_2,

$$\begin{aligned}\mathbf{P}(A_1 \cup A_2 \,|\, B) &= \frac{\mathbf{P}\big((A_1 \cup A_2) \cap B\big)}{\mathbf{P}(B)}\\ &= \frac{\mathbf{P}((A_1 \cap B) \cup (A_2 \cap B))}{\mathbf{P}(B)}\\ &= \frac{\mathbf{P}(A_1 \cap B) + \mathbf{P}(A_2 \cap B)}{\mathbf{P}(B)}\\ &= \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} + \frac{\mathbf{P}(A_2 \cap B)}{\mathbf{P}(B)}\\ &= \mathbf{P}(A_1 \,|\, B) + \mathbf{P}(A_2 \,|\, B),\end{aligned}$$

where for the second equality, we used the fact that $A_1 \cap B$ and $A_2 \cap B$ are disjoint sets, and for the third equality we used the additivity axiom for the (unconditional) probability law. The argument for a countable collection of disjoint sets is similar.

Since conditional probabilities constitute a legitimate probability law, all general properties of probability laws remain valid. For example, a fact such as $\mathbf{P}(A \cup C) \le \mathbf{P}(A) + \mathbf{P}(C)$ translates to the new fact

$$\mathbf{P}(A \cup C \,|\, B) \le \mathbf{P}(A \,|\, B) + \mathbf{P}(C \,|\, B).$$

Let us summarize the conclusions reached so far.

Properties of Conditional Probability

- The conditional probability of an event A, given an event B with $\mathbf{P}(B) > 0$, is defined by

$$\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

  and specifies a new (conditional) probability law on the same sample space Ω. In particular, all known properties of probability laws remain valid for conditional probability laws.

- Conditional probabilities can also be viewed as a probability law on a new universe B, because all of the conditional probability is concentrated on B.

- In the case where the possible outcomes are finitely many and equally likely, we have

$$\mathbf{P}(A \,|\, B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

Example 1.6. We toss a fair coin three successive times. We wish to find the conditional probability $\mathbf{P}(A \,|\, B)$ when A and B are the events

$$A = \{\text{more heads than tails come up}\}, \qquad B = \{\text{1st toss is a head}\}.$$

The sample space consists of eight sequences,

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

which we assume to be equally likely. The event B consists of the four elements HHH, HHT, HTH, HTT, so its probability is

$$\mathbf{P}(B) = \frac{4}{8}.$$

The event $A \cap B$ consists of the three elements outcomes HHH, HHT, HTH, so its probability is

$$\mathbf{P}(A \cap B) = \frac{3}{8}.$$

Thus, the conditional probability $\mathbf{P}(A \,|\, B)$ is

$$\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{3/8}{4/8} = \frac{3}{4}.$$

Because all possible outcomes are equally likely here, we can also compute $\mathbf{P}(A \,|\, B)$ using a shortcut. We can bypass the calculation of $\mathbf{P}(B)$ and $\mathbf{P}(A \cap B)$, and simply divide the number of elements shared by A and B (which is 3) with the number of elements of B (which is 4), to obtain the same result 3/4.

Example 1.7. A fair 4-sided die is rolled twice and we assume that all sixteen possible outcomes are equally likely. Let X and Y be the result of the 1st and the 2nd roll, respectively. We wish to determine the conditional probability $\mathbf{P}(A \,|\, B)$ where

$$A = \big\{\max(X, Y) = m\big\}, \qquad B = \big\{\min(X, Y) = 2\big\},$$

and m takes each of the values 1, 2, 3, 4.

As in the preceding example, we can first determine the probabilities $\mathbf{P}(A \cap B)$ and $\mathbf{P}(B)$ by counting the number of elements of $A \cap B$ and B, respectively, and dividing by 16. Alternatively, we can directly divide the number of elements of $A \cap B$ with the number of elements of B; see Fig. 1.7.

Example 1.8. A conservative design team, call it C, and an innovative design team, call it N, are asked to separately design a new product within a month. From past experience we know that:

(a) The probability that team C is successful is 2/3.

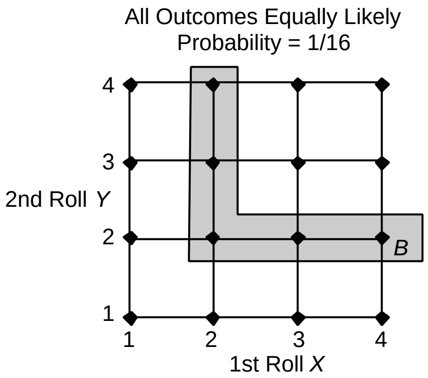


Figure 1.7: Sample space of an experiment involving two rolls of a 4-sided die. (cf. Example 1.7). The conditioning event $B = \{\min(X, Y) = 2\}$ consists of the 5-element shaded set. The set $A = \{\max(X, Y) = m\}$ shares with B two elements if $m = 3$ or $m = 4$, one element if $m = 2$, and no element if $m = 1$. Thus, we have

$$\mathbf{P}\big(\{\max(X,Y)=m\}\,|\,B\big) = \begin{cases} 2/5 & \text{if } m=3 \text{ or } m=4, \\ 1/5 & \text{if } m=2, \\ 0 & \text{if } m=1. \end{cases}$$

(b) The probability that team N is successful is 1/2.

(c) The probability that at least one team is successful is 3/4.

If both teams are successful, the design of team N is adopted. Assuming that exactly one successful design is produced, what is the probability that it was designed by team N?

There are four possible outcomes here, corresponding to the four combinations of success and failure of the two teams:

SS: both succeed,	FF: both fail,
SF: C succeeds, N fails,	FS: C fails, N succeeds.

We are given that the probabilities of these outcomes satisfy

$$\mathbf{P}(SS) + \mathbf{P}(SF) = \frac{2}{3}, \quad \mathbf{P}(SS) + \mathbf{P}(FS) = \frac{1}{2}, \quad \mathbf{P}(SS) + \mathbf{P}(SF) + \mathbf{P}(FS) = \frac{3}{4}.$$

From these relations, together with the normalization equation $\mathbf{P}(SS) + \mathbf{P}(SF) + \mathbf{P}(FS) + \mathbf{P}(FF) = 1$, we can obtain the probabilities of all the outcomes:

$$\mathbf{P}(SS) = \frac{5}{12}, \qquad \mathbf{P}(SF) = \frac{1}{4}, \qquad \mathbf{P}(FS) = \frac{1}{12}, \qquad \mathbf{P}(FF) = \frac{1}{4}.$$

The desired conditional probability is

$$\mathbf{P}\big(\{FS\}\,|\,\{SF, FS\}\big) = \frac{\frac{1}{12}}{\frac{1}{4} + \frac{1}{12}} = \frac{1}{4}.$$

Using Conditional Probability for Modeling

When constructing probabilistic models for experiments that have a sequential character, it is often natural and convenient to first specify conditional probabilities and then use them to determine unconditional probabilities. The rule $\mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A \mid B)$, which is a restatement of the definition of conditional probability, is often helpful in this process.

Example 1.9. Radar detection. If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of false alarm (a false indication of aircraft presence), and the probability of missed detection (nothing registers, even though an aircraft is present)?

A sequential representation of the sample space is appropriate here, as shown in Fig. 1.8. Let A and B be the events

$$A = \{\text{an aircraft is present}\},$$
$$B = \{\text{the radar registers an aircraft presence}\},$$

and consider also their complements

$$A^c = \{\text{an aircraft is not present}\},$$
$$B^c = \{\text{the radar does not register an aircraft presence}\}.$$

The given probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.8. Each event of interest corresponds to a leaf of the tree and its probability is equal to the product of the probabilities associated with the branches in a path from the root to the corresponding leaf. The desired probabilities of false alarm and missed detection are

$$\mathbf{P}(\text{false alarm}) = \mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B \mid A^c) = 0.95 \cdot 0.10 = 0.095,$$
$$\mathbf{P}(\text{missed detection}) = \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c \mid A) = 0.05 \cdot 0.01 = 0.0005.$$

Extending the preceding example, we have a general rule for calculating various probabilities in conjunction with a tree-based sequential description of an experiment. In particular:

(a) We set up the tree so that an event of interest is associated with a leaf. We view the occurrence of the event as a sequence of steps, namely, the traversals of the branches along the path from the root to the leaf.

(b) We record the conditional probabilities associated with the branches of the tree.

(c) We obtain the probability of a leaf by multiplying the probabilities recorded along the corresponding path of the tree.

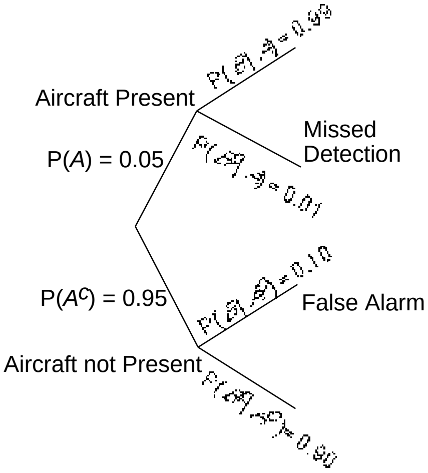


Figure 1.8: Sequential description of the sample space for the radar detection problem in Example 1.9

In mathematical terms, we are dealing with an event A which occurs if and only if each one of several events $A_1, \ldots, A_n$ has occurred, i.e., $A = A_1 \cap A_2 \cap \cdots \cap A_n$. The occurrence of A is viewed as an occurrence of A_1, followed by the occurrence of A_2, then of A_3, etc, and it is visualized as a path on the tree with n branches, corresponding to the events $A_1, \ldots, A_n$. The probability of A is given by the following rule (see also Fig. 1.9).

Multiplication Rule

Assuming that all of the conditioning events have positive probability, we have

$$\mathbf{P}\big(\cap_{i=1}^n A_i\big) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2) \cdots \mathbf{P}\big(A_n \,|\, \cap_{i=1}^{n-1} A_i\big).$$

The multiplication rule can be verified by writing

$$\mathbf{P}\big(\cap_{i=1}^n A_i\big) = \mathbf{P}(A_1)\frac{\mathbf{P}(A_1 \cap A_2)}{\mathbf{P}(A_1)}\frac{\mathbf{P}(A_1 \cap A_2 \cap A_3)}{\mathbf{P}(A_1 \cap A_2)} \cdots \frac{\mathbf{P}\big(\cap_{i=1}^n A_i\big)}{\mathbf{P}\big(\cap_{i=1}^{n-1} A_i\big)},$$

and by using the definition of conditional probability to rewrite the right-hand side above as

$$\mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2) \cdots \mathbf{P}\big(A_n \,|\, \cap_{i=1}^{n-1} A_i\big).$$

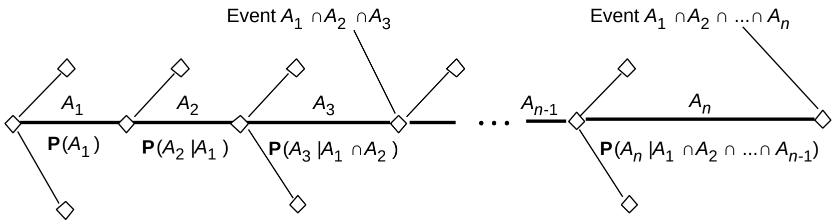

Event $A_1 \cap A_2 \cap A_3$
Event $A_1 \cap A_2 \cap ... \cap A_n$
A_1
A_2
A_3
A_{n-1}
A_n
$\mathbf{P}(A_1)$
$\mathbf{P}(A_2 | A_1)$
$\mathbf{P}(A_3 | A_1 \cap A_2)$
$\mathbf{P}(A_n | A_1 \cap A_2 \cap ... \cap A_{n-1})$

Figure 1.9: Visualization of the total probability theorem. The intersection event $A = A_1 \cap A_2 \cap \cdots \cap A_n$ is associated with a path on the tree of a sequential description of the experiment. We associate the branches of this path with the events $A_1, \ldots, A_n$, and we record next to the branches the corresponding conditional probabilities.

The final node of the path corresponds to the intersection event A, and its probability is obtained by multiplying the conditional probabilities recorded along the branches of the path

$$\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1) \cdots \mathbf{P}(A_n \,|\, A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Note that any intermediate node along the path also corresponds to some intersection event and its probability is obtained by multiplying the corresponding conditional probabilities up to that node. For example, the event $A_1 \cap A_2 \cap A_3$ corresponds to the node shown in the figure, and its probability is

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2).$$

For the case of just two events, A_1 and A_2, the multiplication rule is simply the definition of conditional probability.

Example 1.10. Three cards are drawn from an ordinary 52-card deck without replacement (drawn cards are not placed back in the deck). We wish to find the probability that none of the three cards is a heart. We assume that at each step, each one of the remaining cards is equally likely to be picked. By symmetry, this implies that every triplet of cards is equally likely to be drawn. A cumbersome approach, that we will not use, is to count the number of all card triplets that do not include a heart, and divide it with the number of all possible card triplets. Instead, we use a sequential description of the sample space in conjunction with the multiplication rule (cf. Fig. 1.10).

Define the events

$$A_i = \{\text{the } i\text{th card is not a heart}\}, \qquad i = 1, 2, 3.$$

We will calculate $\mathbf{P}(A_1 \cap A_2 \cap A_3)$, the probability that none of the three cards is a heart, using the multiplication rule,

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2).$$

We have

$$\mathbf{P}(A_1) = \frac{39}{52},$$

since there are 39 cards that are not hearts in the 52-card deck. Given that the first card is not a heart, we are left with 51 cards, 38 of which are not hearts, and

$$\mathbf{P}(A_2 \,|\, A_1) = \frac{38}{51}.$$

Finally, given that the first two cards drawn are not hearts, there are 37 cards which are not hearts in the remaining 50-card deck, and

$$\mathbf{P}(A_3 \,|\, A_1 \cap A_2) = \frac{37}{50}.$$

These probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.10. The desired probability is now obtained by multiplying the probabilities recorded along the corresponding path of the tree:

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50}.$$

Note that once the probabilities are recorded along the tree, the probability of several other events can be similarly calculated. For example,

$$\mathbf{P}(\text{1st is not a heart and 2nd is a heart}) = \frac{39}{52} \cdot \frac{13}{51},$$

$$\mathbf{P}(\text{1st two are not hearts and 3rd is a heart}) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{13}{50}.$$

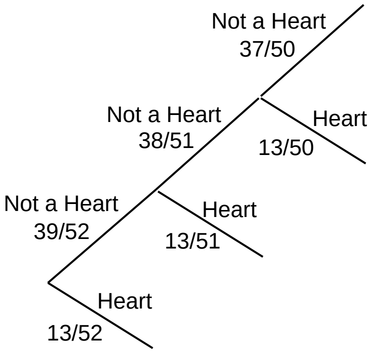


Figure 1.10: Sequential description of the sample space of the 3-card selection problem in Example 1.10.

Example 1.11. A class consisting of 4 graduate and 12 undergraduate students is randomly divided into 4 groups of 4. What is the probability that each group includes a graduate student? We interpret randomly to mean that given the assignment of some students to certain slots, any of the remaining students is equally likely to be assigned to any of the remaining slots. We then calculate the desired probability using the multiplication rule, based on the sequential description shown in Fig. 1.11. Let us denote the four graduate students by 1, 2, 3, 4, and consider the events

$$\begin{aligned} A_1 &= \{\text{students 1 and 2 are in different groups}\}, \\ A_2 &= \{\text{students 1, 2, and 3 are in different groups}\}, \\ A_3 &= \{\text{students 1, 2, 3, and 4 are in different groups}\}. \end{aligned}$$

We will calculate $\mathbf{P}(A_3)$ using the multiplication rule:

$$\mathbf{P}(A_3) = \mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2).$$

We have

$$\mathbf{P}(A_1) = \frac{12}{15},$$

since there are 12 student slots in groups other than the one of student 1, and there are 15 student slots overall, excluding student 1. Similarly,

$$\mathbf{P}(A_2 \,|\, A_1) = \frac{8}{14},$$

since there are 8 student slots in groups other than the one of students 1 and 2, and there are 14 student slots, excluding students 1 and 2. Also,

$$\mathbf{P}(A_3 \,|\, A_1 \cap A_2) = \frac{4}{13},$$

since there are 4 student slots in groups other than the one of students 1, 2, and 3, and there are 13 student slots, excluding students 1, 2, and 3. Thus, the desired probability is

$$\frac{12}{15} \cdot \frac{8}{14} \cdot \frac{4}{13},$$

and is obtained by multiplying the conditional probabilities along the corresponding path of the tree of Fig. 1.11.

1.4 TOTAL PROBABILITY THEOREM AND BAYES' RULE

In this section, we explore some applications of conditional probability. We start with the following theorem, which is often useful for computing the probabilities of various events, using a "divide-and-conquer" approach.

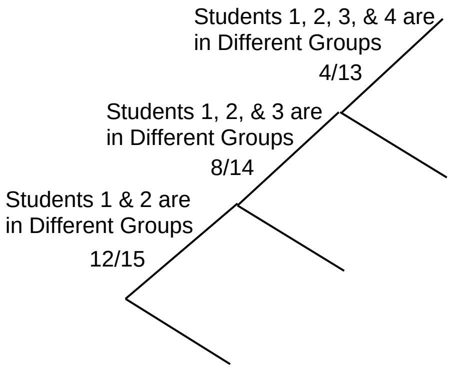


Figure 1.11: Sequential description of the sample space of the student problem in Example 1.11.

Total Probability Theorem

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in one and only one of the events $A_1, \ldots, A_n$) and assume that $\mathbf{P}(A_i) > 0$, for all $i = 1, \ldots, n$. Then, for any event B, we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \,|\, A_n).\end{aligned}$$

The theorem is visualized and proved in Fig. 1.12. Intuitively, we are partitioning the sample space into a number of scenarios (events) A_i. Then, the probability that B occurs is a weighted average of its conditional probability under each scenario, where each scenario is weighted according to its (unconditional) probability. One of the uses of the theorem is to compute the probability of various events B for which the conditional probabilities $\mathbf{P}(B \,|\, A_i)$ are known or easy to derive. The key is to choose appropriately the partition $A_1, \ldots, A_n$, and this choice is often suggested by the problem structure. Here are some examples.

Example 1.12. You enter a chess tournament where your probability of winning a game is 0.3 against half the players (call them type 1), 0.4 against a quarter of the players (call them type 2), and 0.5 against the remaining quarter of the players (call them type 3). You play a game against a randomly chosen opponent. What is the probability of winning?

Let A_i be the event of playing with an opponent of type i. We have

$$\mathbf{P}(A_1) = 0.5, \qquad \mathbf{P}(A_2) = 0.25, \qquad \mathbf{P}(A_3) = 0.25.$$

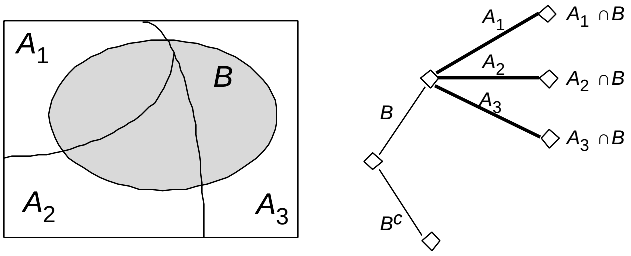


Figure 1.12: Visualization and verification of the total probability theorem. The events $A_1, \ldots, A_n$ form a partition of the sample space, so the event B can be decomposed into the disjoint union of its intersections $A_i \cap B$ with the sets A_i, i.e.,

$$B = (A_1 \cap B) \cup \cdots \cup (A_n \cap B).$$

Using the additivity axiom, it follows that

$$\mathbf{P}(B) = \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B).$$

Since, by the definition of conditional probability, we have

$$\mathbf{P}(A_i \cap B) = \mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i),$$

the preceding equality yields

$$\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \,|\, A_n).$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability of the leaf $A_i \cap B$ is the product $\mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i)$ of the probabilities along the path leading to that leaf. The event B consists of the three highlighted leaves and $\mathbf{P}(B)$ is obtained by adding their probabilities.

Let also B be the event of winning. We have

$$\mathbf{P}(B \,|\, A_1) = 0.3, \qquad \mathbf{P}(B \,|\, A_2) = 0.4, \qquad \mathbf{P}(B \,|\, A_3) = 0.5.$$

Thus, by the total probability theorem, the probability of winning is

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \mathbf{P}(A_2)\mathbf{P}(B \,|\, A_2) + \mathbf{P}(A_3)\mathbf{P}(B \,|\, A_3) \\ &= 0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5 \\ &= 0.375.\end{aligned}$$

Example 1.13. We roll a fair four-sided die. If the result is 1 or 2, we roll once more but otherwise, we stop. What is the probability that the sum total of our rolls is at least 4?

Let A_i be the event that the result of first roll is i, and note that $\mathbf{P}(A_i) = 1/4$ for each i. Let B be the event that the sum total is at least 4. Given the event A_1, the sum total will be at least 4 if the second roll results in 3 or 4, which happens with probability 1/2. Similarly, given the event A_2, the sum total will be at least 4 if the second roll results in 2, 3, or 4, which happens with probability 3/4. Also, given the event A_3, we stop and the sum total remains below 4. Therefore,

$$\mathbf{P}(B \,|\, A_1) = \frac{1}{2}, \qquad \mathbf{P}(B \,|\, A_2) = \frac{3}{4}, \qquad \mathbf{P}(B \,|\, A_3) = 0, \qquad \mathbf{P}(B \,|\, A_4) = 1.$$

By the total probability theorem,

$$\mathbf{P}(B) = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 = \frac{9}{16}.$$

The total probability theorem can be applied repeatedly to calculate probabilities in experiments that have a sequential character, as shown in the following example.

Example 1.14. Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). Alice is (by default) up-to-date when she starts the class. What is the probability that she is up-to-date after three weeks?

Let U_i and B_i be the events that Alice is up-to-date or behind, respectively, after i weeks. According to the total probability theorem, the desired probability $\mathbf{P}(U_3)$ is given by

$$\mathbf{P}(U_3) = \mathbf{P}(U_2)\mathbf{P}(U_3 \,|\, U_2) + \mathbf{P}(B_2)\mathbf{P}(U_3 \,|\, B_2) = \mathbf{P}(U_2) \cdot 0.8 + \mathbf{P}(B_2) \cdot 0.4.$$

The probabilities $\mathbf{P}(U_2)$ and $\mathbf{P}(B_2)$ can also be calculated using the total probability theorem:

$$\mathbf{P}(U_2) = \mathbf{P}(U_1)\mathbf{P}(U_2 \,|\, U_1) + \mathbf{P}(B_1)\mathbf{P}(U_2 \,|\, B_1) = \mathbf{P}(U_1) \cdot 0.8 + \mathbf{P}(B_1) \cdot 0.4,$$

$$\mathbf{P}(B_2) = \mathbf{P}(U_1)\mathbf{P}(B_2 \,|\, U_1) + \mathbf{P}(B_1)\mathbf{P}(B_2 \,|\, B_1) = \mathbf{P}(U_1) \cdot 0.2 + \mathbf{P}(B_1) \cdot 0.6.$$

Finally, since Alice starts her class up-to-date, we have

$$\mathbf{P}(U_1) = 0.8, \qquad \mathbf{P}(B_1) = 0.2.$$

We can now combine the preceding three equations to obtain

$$\mathbf{P}(U_2) = 0.8 \cdot 0.8 + 0.2 \cdot 0.4 = 0.72,$$

$$\mathbf{P}(B_2) = 0.8 \cdot 0.2 + 0.2 \cdot 0.6 = 0.28.$$

and by using the above probabilities in the formula for $\mathbf{P}(U_3)$:

$$\mathbf{P}(U_3) = 0.72 \cdot 0.8 + 0.28 \cdot 0.4 = 0.688.$$

Note that we could have calculated the desired probability $\mathbf{P}(U_3)$ by constructing a tree description of the experiment, by calculating the probability of every element of U_3 using the multiplication rule on the tree, and by adding. In experiments with a sequential character one may often choose between using the multiplication rule or the total probability theorem for calculation of various probabilities. However, there are cases where the calculation based on the total probability theorem is more convenient. For example, suppose we are interested in the probability $\mathbf{P}(U_{20})$ that Alice is up-to-date after 20 weeks. Calculating this probability using the multiplication rule is very cumbersome, because the tree representing the experiment is 20-stages deep and has 2^{20} leaves. On the other hand, with a computer, a sequential calculation using the total probability formulas

$$\mathbf{P}(U_{i+1}) = \mathbf{P}(U_i) \cdot 0.8 + \mathbf{P}(B_i) \cdot 0.4,$$

$$\mathbf{P}(B_{i+1}) = \mathbf{P}(U_i) \cdot 0.2 + \mathbf{P}(B_i) \cdot 0.6,$$

and the initial conditions $\mathbf{P}(U_1) = 0.8$, $\mathbf{P}(B_1) = 0.2$ is very simple.

The total probability theorem is often used in conjunction with the following celebrated theorem, which relates conditional probabilities of the form $\mathbf{P}(A \mid B)$ with conditional probabilities of the form $\mathbf{P}(B \mid A)$, in which the order of the conditioning is reversed.

Bayes' Rule

Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$, for all i. Then, for any event B such that $\mathbf{P}(B) > 0$, we have

$$\begin{aligned}\mathbf{P}(A_i \mid B) &= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)}.\end{aligned}$$

To verify Bayes' rule, note that $\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)$ and $\mathbf{P}(A_i \mid B)\mathbf{P}(B)$ are equal, because they are both equal to $\mathbf{P}(A_i \cap B)$. This yields the first equality. The second equality follows from the first by using the total probability theorem to rewrite $\mathbf{P}(B)$.

Bayes' rule is often used for **inference**. There are a number of "causes" that may result in a certain "effect." We observe the effect, and we wish to infer the cause. The events $A_1, \ldots, A_n$ are associated with the causes and the event B represents the effect. The probability $\mathbf{P}(B \mid A_i)$ that the effect will be observed when the cause A_i is present amounts to a probabilistic model of the cause-effect relation (cf. Fig. 1.13). Given that the effect B has been observed, we wish to evaluate the (conditional) probability $\mathbf{P}(A_i \mid B)$ that the cause A_i is present.

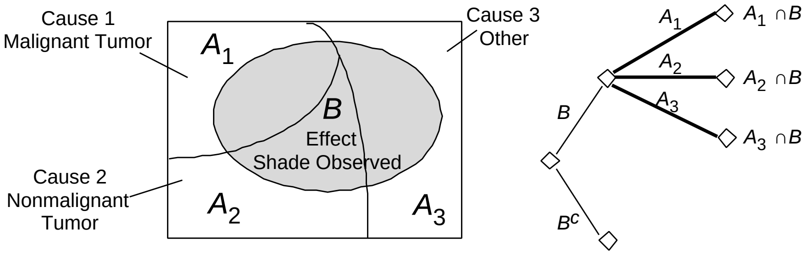


Figure 1.13: An example of the inference context that is implicit in Bayes' rule. We observe a shade in a person's X-ray (this is event B, the "effect") and we want to estimate the likelihood of three mutually exclusive and collectively exhaustive potential causes: cause 1 (event A_1) is that there is a malignant tumor, cause 2 (event A_2) is that there is a nonmalignant tumor, and cause 3 (event A_3) corresponds to reasons other than a tumor. We assume that we know the probabilities $\mathbf{P}(A_i)$ and $\mathbf{P}(B \mid A_i)$, $i = 1, 2, 3$. Given that we see a shade (event B occurs), Bayes' rule gives the conditional probabilities of the various causes as

$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \mathbf{P}(A_2)\mathbf{P}(B \mid A_2) + \mathbf{P}(A_3)\mathbf{P}(B \mid A_3)}, \quad i = 1, 2, 3.$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability $\mathbf{P}(A_1 \mid B)$ of a malignant tumor is the probability of the first highlighted leaf, which is $\mathbf{P}(A_1 \cap B)$, divided by the total probability of the highlighted leaves, which is $\mathbf{P}(B)$.

Example 1.15. Let us return to the radar detection problem of Example 1.9 and Fig. 1.8. Let

$$A = \{\text{an aircraft is present}\},$$
$$B = \{\text{the radar registers an aircraft presence}\}.$$

We are given that

$$\mathbf{P}(A) = 0.05, \qquad \mathbf{P}(B \mid A) = 0.99, \qquad \mathbf{P}(B \mid A^c) = 0.1.$$

Applying Bayes' rule, with $A_1 = A$ and $A_2 = A^c$, we obtain

$$\begin{aligned}\mathbf{P}(\text{aircraft present} \,|\, \text{radar registers}) &= \mathbf{P}(A \,|\, B)\\ &= \frac{\mathbf{P}(A)\mathbf{P}(B \,|\, A)}{\mathbf{P}(B)}\\ &= \frac{\mathbf{P}(A)\mathbf{P}(B \,|\, A)}{\mathbf{P}(A)\mathbf{P}(B \,|\, A) + \mathbf{P}(A^c)\mathbf{P}(B \,|\, A^c)}\\ &= \frac{0.05 \cdot 0.99}{0.05 \cdot 0.99 + 0.95 \cdot 0.1}\\ &\approx 0.3426.\end{aligned}$$

Example 1.16. Let us return to the chess problem of Example 1.12. Here A_i is the event of getting an opponent of type i, and

$$\mathbf{P}(A_1) = 0.5, \qquad \mathbf{P}(A_2) = 0.25, \qquad \mathbf{P}(A_3) = 0.25.$$

Also, B is the event of winning, and

$$\mathbf{P}(B \,|\, A_1) = 0.3, \qquad \mathbf{P}(B \,|\, A_2) = 0.4, \qquad \mathbf{P}(B \,|\, A_3) = 0.5.$$

Suppose that you win. What is the probability $\mathbf{P}(A_1 \,|\, B)$ that you had an opponent of type 1?

Using Bayes' rule, we have

$$\begin{aligned}\mathbf{P}(A_1 \,|\, B) &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \mathbf{P}(A_2)\mathbf{P}(B \,|\, A_2) + \mathbf{P}(A_3)\mathbf{P}(B \,|\, A_3)}\\ &= \frac{0.5 \cdot 0.3}{0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5}\\ &= 0.4.\end{aligned}$$

1.5 INDEPENDENCE

We have introduced the conditional probability $\mathbf{P}(A \,|\, B)$ to capture the partial information that event B provides about event A. An interesting and important special case arises when the occurrence of B provides no information and does not alter the probability that A has occurred, i.e.,

$$\mathbf{P}(A \,|\, B) = \mathbf{P}(A).$$

When the above equality holds, we say that A is **independent** of B. Note that by the definition $\mathbf{P}(A \mid B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$, this is equivalent to

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

We adopt this latter relation as the definition of independence because it can be used even if $\mathbf{P}(B) = 0$, in which case $\mathbf{P}(A \mid B)$ is undefined. The symmetry of this relation also implies that independence is a symmetric property; that is, if A is independent of B, then B is independent of A, and we can unambiguously say that A and B are **independent events**.

Independence is often easy to grasp intuitively. For example, if the occurrence of two events is governed by distinct and noninteracting physical processes, such events will turn out to be independent. On the other hand, independence is not easily visualized in terms of the sample space. A common first thought is that two events are independent if they are disjoint, but in fact the opposite is true: two disjoint events A and B with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$ are never independent, since their intersection $A \cap B$ is empty and has probability 0.

Example 1.17. Consider an experiment involving two successive rolls of a 4-sided die in which all 16 possible outcomes are equally likely and have probability 1/16.

(a) Are the events

$$A_i = \{\text{1st roll results in } i\}, \qquad B_j = \{\text{2nd roll results in } j\},$$

independent? We have

$$\begin{aligned}
\mathbf{P}(A \cap B) &= \mathbf{P}\big(\text{the result of the two rolls is } (i,j)\big) = \frac{1}{16},\\
\mathbf{P}(A_i) &= \frac{\text{number of elements of } A_i}{\text{total number of possible outcomes}} = \frac{4}{16},\\
\mathbf{P}(B_j) &= \frac{\text{number of elements of } B_j}{\text{total number of possible outcomes}} = \frac{4}{16}.
\end{aligned}$$

We observe that $\mathbf{P}(A_i \cap B_j) = \mathbf{P}(A_i)\mathbf{P}(B_j)$, and the independence of A_i and B_j is verified. Thus, our choice of the discrete uniform probability law (which might have seemed arbitrary) models the independence of the two rolls.

(b) Are the events

$$A = \{\text{1st roll is a 1}\}, \qquad B = \{\text{sum of the two rolls is a 5}\},$$

independent? The answer here is not quite obvious. We have

$$\mathbf{P}(A \cap B) = \mathbf{P}\big(\text{the result of the two rolls is (1,4)}\big) = \frac{1}{16},$$

and also

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

The event B consists of the outcomes (1,4), (2,3), (3,2), and (4,1), and

$$\mathbf{P}(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

Thus, we see that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, and the events A and B are independent.

(c) Are the events

$$A = \{\text{maximum of the two rolls is 2}\}, \quad B = \{\text{minimum of the two rolls is 2}\},$$

independent? Intuitively, the answer is "no" because the minimum of the two rolls tells us something about the maximum. For example, if the minimum is 2, the maximum cannot be 1. More precisely, to verify that A and B are not independent, we calculate

$$\mathbf{P}(A \cap B) = \mathbf{P}\big(\text{the result of the two rolls is (2,2)}\big) = \frac{1}{16},$$

and also

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{3}{16},$$

$$\mathbf{P}(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{5}{16}.$$

We have $\mathbf{P}(A)\mathbf{P}(B) = 15/(16)^2$, so that $\mathbf{P}(A \cap B) \neq \mathbf{P}(A)\mathbf{P}(B)$, and A and B are not independent.

Conditional Independence

We noted earlier that the conditional probabilities of events, conditioned on a particular event, form a legitimate probability law. We can thus talk about independence of various events with respect to this conditional law. In particular, given an event C, the events A and B are called **conditionally independent** if

$$\mathbf{P}(A \cap B \,|\, C) = \mathbf{P}(A \,|\, C)\mathbf{P}(B \,|\, C).$$

The definition of the conditional probability and the multiplication rule yield

$$\begin{aligned}\mathbf{P}(A \cap B \,|\, C) &= \frac{\mathbf{P}(A \cap B \cap C)}{\mathbf{P}(C)} \\ &= \frac{\mathbf{P}(C)\mathbf{P}(B \,|\, C)\mathbf{P}(A \,|\, B \cap C)}{\mathbf{P}(C)} \\ &= \mathbf{P}(B \,|\, C)\mathbf{P}(A \,|\, B \cap C).\end{aligned}$$

After canceling the factor $\mathbf{P}(B \,|\, C)$, assumed nonzero, we see that conditional independence is the same as the condition

$$\mathbf{P}(A \,|\, B \cap C) = \mathbf{P}(A \,|\, C).$$

In words, this relation states that if C is known to have occurred, the additional knowledge that B also occurred does not change the probability of A.

Interestingly, independence of two events A and B with respect to the unconditional probability law, does not imply conditional independence, and vice versa, as illustrated by the next two examples.

Example 1.18. Consider two independent fair coin tosses, in which all four possible outcomes are equally likely. Let

$$\begin{aligned} H_1 &= \{\text{1st toss is a head}\}, \\ H_2 &= \{\text{2nd toss is a head}\}, \\ D &= \{\text{the two tosses have different results}\}. \end{aligned}$$

The events H_1 and H_2 are (unconditionally) independent. But

$$\mathbf{P}(H_1 \,|\, D) = \frac{1}{2}, \qquad \mathbf{P}(H_2 \,|\, D) = \frac{1}{2}, \qquad \mathbf{P}(H_1 \cap H_2 \,|\, D) = 0,$$

so that $\mathbf{P}(H_1 \cap H_2 \,|\, D) \neq \mathbf{P}(H_1 \,|\, D)\mathbf{P}(H_2 \,|\, D)$, and H_1, H_2 are not conditionally independent.

Example 1.19. There are two coins, a blue and a red one. We choose one of the two at random, each being chosen with probability 1/2, and proceed with two independent tosses. The coins are biased: with the blue coin, the probability of heads in any given toss is 0.99, whereas for the red coin it is 0.01.

Let B be the event that the blue coin was selected. Let also H_i be the event that the ith toss resulted in heads. Given the choice of a coin, the events H_1 and H_2 are independent, because of our assumption of independent tosses. Thus,

$$\mathbf{P}(H_1 \cap H_2 \,|\, B) = \mathbf{P}(H_1 \,|\, B)\mathbf{P}(H_2 \,|\, B) = 0.99 \cdot 0.99.$$

On the other hand, the events H_1 and H_2 are not independent. Intuitively, if we are told that the first toss resulted in heads, this leads us to suspect that the blue coin was selected, in which case, we expect the second toss to also result in heads. Mathematically, we use the total probability theorem to obtain

$$\mathbf{P}(H_1) = \mathbf{P}(B)\mathbf{P}(H_1 \,|\, B) + \mathbf{P}(B^c)\mathbf{P}(H_1 \,|\, B^c) = \frac{1}{2} \cdot 0.99 + \frac{1}{2} \cdot 0.01 = \frac{1}{2},$$

as should be expected from symmetry considerations. Similarly, we have $\mathbf{P}(H_2) = 1/2$. Now notice that

$$\begin{aligned}\mathbf{P}(H_1 \cap H_2) &= \mathbf{P}(B)\mathbf{P}(H_1 \cap H_2 \,|\, B) + \mathbf{P}(B^c)\mathbf{P}(H_1 \cap H_2 \,|\, B^c) \\ &= \frac{1}{2} \cdot 0.99 \cdot 0.99 + \frac{1}{2} \cdot 0.01 \cdot 0.01 \approx \frac{1}{2}.\end{aligned}$$

Thus, $\mathbf{P}(H_1 \cap H_2) \neq \mathbf{P}(H_1)\mathbf{P}(H_2)$, and the events H_1 and H_2 are dependent, even though they are conditionally independent given B.

As mentioned earlier, if A and B are independent, the occurrence of B does not provide any new information on the probability of A occurring. It is then intuitive that the non-occurrence of B should also provide no information on the probability of A. Indeed, it can be verified that if A and B are independent, the same holds true for A and B^c (see the end-of-chapter problems).

We now summarize.

Independence

- Two events A and B are said to independent if

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

  If in addition, $\mathbf{P}(B) > 0$, independence is equivalent to the condition

  $$\mathbf{P}(A \,|\, B) = \mathbf{P}(A).$$

- If A and B are independent, so are A and B^c.
- Two events A and B are said to be conditionally independent, given another event C with $\mathbf{P}(C) > 0$, if

  $$\mathbf{P}(A \cap B \,|\, C) = \mathbf{P}(A \,|\, C)\mathbf{P}(B \,|\, C).$$

  If in addition, $\mathbf{P}(B \cap C) > 0$, conditional independence is equivalent to the condition

  $$\mathbf{P}(A \,|\, B \cap C) = \mathbf{P}(A \,|\, C).$$

- Independence does not imply conditional independence, and vice versa.

Independence of a Collection of Events

The definition of independence can be extended to multiple events.

Definition of Independence of Several Events

We say that the events $A_1, A_2, \ldots, A_n$ are **independent** if

$$\mathbf{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i), \qquad \text{for every subset } S \text{ of } \{1, 2, \ldots, n\}.$$

If we have a collection of three events, A_1, A_2, and A_3, independence amounts to satisfying the four conditions

$$\begin{aligned} \mathbf{P}(A_1 \cap A_2) &= \mathbf{P}(A_1)\,\mathbf{P}(A_2), \\ \mathbf{P}(A_1 \cap A_3) &= \mathbf{P}(A_1)\,\mathbf{P}(A_3), \\ \mathbf{P}(A_2 \cap A_3) &= \mathbf{P}(A_2)\,\mathbf{P}(A_3), \\ \mathbf{P}(A_1 \cap A_2 \cap A_3) &= \mathbf{P}(A_1)\,\mathbf{P}(A_2)\,\mathbf{P}(A_3). \end{aligned}$$

The first three conditions simply assert that any two events are independent, a property known as **pairwise independence**. But the fourth condition is also important and does not follow from the first three. Conversely, the fourth condition does not imply the first three; see the two examples that follow.

Example 1.20. Pairwise independence does not imply independence. Consider two independent fair coin tosses, and the following events:

$$\begin{aligned} H_1 &= \{\text{1st toss is a head}\}, \\ H_2 &= \{\text{2nd toss is a head}\}, \\ D &= \{\text{the two tosses have different results}\}. \end{aligned}$$

The events H_1 and H_2 are independent, by definition. To see that H_1 and D are independent, we note that

$$\mathbf{P}(D \,|\, H_1) = \frac{\mathbf{P}(H_1 \cap D)}{\mathbf{P}(H_1)} = \frac{1/4}{1/2} = \frac{1}{2} = \mathbf{P}(D).$$

Similarly, H_2 and D are independent. On the other hand, we have

$$\mathbf{P}(H_1 \cap H_2 \cap D) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbf{P}(H_1)\mathbf{P}(H_2)\mathbf{P}(D),$$

and these three events are not independent.

Example 1.21. The equality $\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\,\mathbf{P}(A_2)\,\mathbf{P}(A_3)$ is not enough for independence. Consider two independent rolls of a fair die, and the following events:

$$\begin{aligned}
A &= \{\text{1st roll is 1, 2, or 3}\},\\
B &= \{\text{1st roll is 3, 4, or 5}\},\\
C &= \{\text{the sum of the two rolls is 9}\}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathbf{P}(A\cap B) &= \frac{1}{6} \neq \frac{1}{2}\cdot\frac{1}{2} = \mathbf{P}(A)\mathbf{P}(B),\\
\mathbf{P}(A\cap C) &= \frac{1}{36} \neq \frac{1}{2}\cdot\frac{4}{36} = \mathbf{P}(A)\mathbf{P}(C),\\
\mathbf{P}(B\cap C) &= \frac{1}{12} \neq \frac{1}{2}\cdot\frac{4}{36} = \mathbf{P}(B)\mathbf{P}(C).
\end{aligned}$$

Thus the three events A, B, and C are not independent, and indeed no two of these events are independent. On the other hand, we have

$$\mathbf{P}(A\cap B\cap C) = \frac{1}{36} = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{4}{36} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

The intuition behind the independence of a collection of events is analogous to the case of two events. Independence means that the occurrence or non-occurrence of **any number** of the events from that collection carries no information on the remaining events or their complements. For example, if the events A_1, A_2, A_3, A_4 are independent, one obtains relations such as

$$\mathbf{P}(A_1\cup A_2\,|\,A_3\cap A_4) = \mathbf{P}(A_1\cup A_2)$$

or

$$\mathbf{P}(A_1\cup A_2^c\,|\,A_3^c\cap A_4) = \mathbf{P}(A_1\cup A_2^c);$$

see the end-of-chapter problems.

Reliability

In probabilistic models of complex systems involving several components, it is often convenient to assume that the components behave "independently" of each other. This typically simplifies the calculations and the analysis, as illustrated in the following example.

Example 1.22. Network connectivity. A computer network connects two nodes A and B through intermediate nodes C, D, E, F, as shown in Fig. 1.14(a). For every pair of directly connected nodes, say i and j, there is a given probability p_{ij} that the link from i to j is up. We assume that link failures are independent

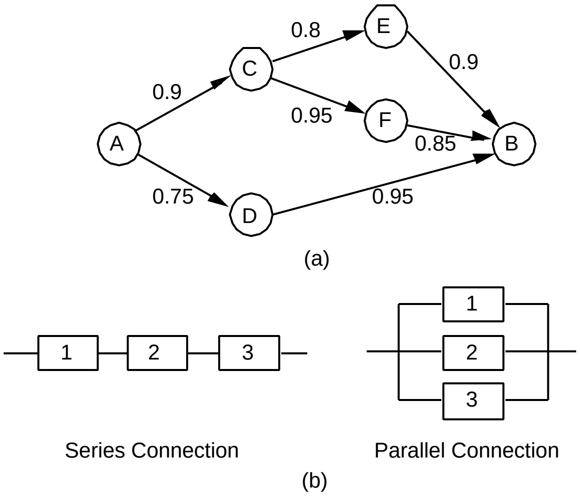


Figure 1.14: (a) Network for Example 1.22. The number next to each link (i, j) indicates the probability that the link is up. (b) Series and parallel connections of three components in a reliability problem.

of each other. What is the probability that there is a path connecting A and B in which all links are up?

This is a typical problem of assessing the reliability of a system consisting of components that can fail independently. Such a system can often be divided into subsystems, where each subsystem consists in turn of several components that are connected either in **series** or in **parallel**; see Fig. 1.14(b).

Let a subsystem consist of components $1, 2, \ldots, m$, and let p_i be the probability that component i is up ("succeeds"). Then, a series subsystem succeeds if **all** of its components are up, so its probability of success is the product of the probabilities of success of the corresponding components, i.e.,

$$\mathbf{P}(\text{series subsystem succeeds}) = p_1 p_2 \cdots p_m.$$

A parallel subsystem succeeds if **any one** of its components succeeds, so its probability of failure is the product of the probabilities of failure of the corresponding components, i.e.,

$$\begin{aligned}\mathbf{P}(\text{parallel subsystem succeeds}) &= 1 - \mathbf{P}(\text{parallel subsystem fails}) \\ &= 1 - (1 - p_1)(1 - p_2) \cdots (1 - p_m).\end{aligned}$$

Returning now to the network of Fig. 1.14(a), we can calculate the probability of success (a path from A to B is available) sequentially, using the preceding formulas, and starting from the end. Let us use the notation $X \rightarrow Y$ to denote the event that there is a (possibly indirect) connection from node X to node Y. Then,

$$\begin{aligned}\mathbf{P}(C \to B) &= 1 - \big(1 - \mathbf{P}(C \to E \text{ and } E \to B)\big)\big(1 - \mathbf{P}(C \to F \text{ and } F \to B)\big)\\ &= 1 - (1 - p_{CE}p_{EB})(1 - p_{CF}p_{FB})\\ &= 1 - (1 - 0.8 \cdot 0.9)(1 - 0.85 \cdot 0.95)\\ &= 0.946,\end{aligned}$$

$$\mathbf{P}(A \to C \text{ and } C \to B) = \mathbf{P}(A \to C)\mathbf{P}(C \to B) = 0.9 \cdot 0.946 = 0.851,$$

$$\mathbf{P}(A \to D \text{ and } D \to B) = \mathbf{P}(A \to D)\mathbf{P}(D \to B) = 0.75 \cdot 0.95 = 0.712,$$

and finally we obtain the desired probability

$$\begin{aligned}\mathbf{P}(A \to B) &= 1 - \big(1 - \mathbf{P}(A \to C \text{ and } C \to B)\big)\big(1 - \mathbf{P}(A \to D \text{ and } D \to B)\big)\\ &= 1 - (1 - 0.851)(1 - 0.712)\\ &= 0.957.\end{aligned}$$

Independent Trials and the Binomial Probabilities

If an experiment involves a sequence of independent but identical stages, we say that we have a sequence of **independent trials**. In the special case where there are only two possible results at each stage, we say that we have a sequence of independent **Bernoulli trials**. The two possible results can be anything, e.g., "it rains" or "it doesn't rain," but we will often think in terms of coin tosses and refer to the two results as "heads" (H) and "tails" (T).

Consider an experiment that consists of n independent tosses of a biased coin, in which the probability of "heads" is p, where p is some number between 0 and 1. In this context, independence means that the events $A_1, A_2, \ldots, A_n$ are independent, where $A_i = \{i\text{th toss is a head}\}$.

We can visualize independent Bernoulli trials by means of a sequential description, as shown in Fig. 1.15 for the case where $n = 3$. The conditional probability of any toss being a head, conditioned on the results of any preceding tosses is p, because of independence. Thus, by multiplying the conditional probabilities along the corresponding path of the tree, we see that any particular outcome (3-long sequence of heads and tails) that involves k heads and $3 - k$ tails has probability $p^k(1-p)^{3-k}$. This formula extends to the case of a general number n of tosses. We obtain that the probability of any particular n-long sequence that contains k heads and $n - k$ tails is $p^k(1-p)^{n-k}$, for all k from 0 to n.

Let us now consider the probability

$$p(k) = \mathbf{P}(k \text{ heads come up in an } n\text{-toss sequence}),$$

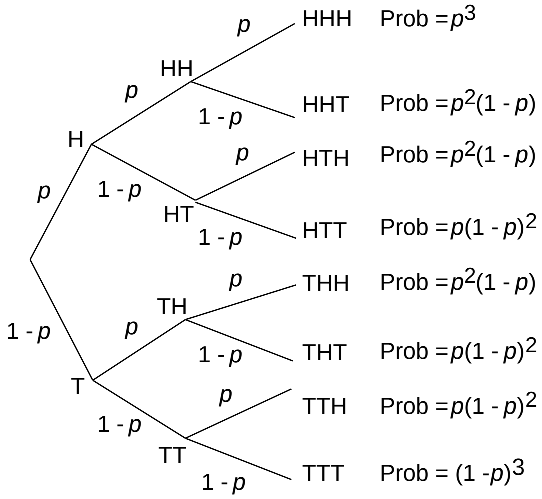


Figure 1.15: Sequential description of the sample space of an experiment involving three independent tosses of a biased coin. Along the branches of the tree, we record the corresponding conditional probabilities, and by the multiplication rule, the probability of obtaining a particular 3-toss sequence is calculated by multiplying the probabilities recorded along the corresponding path of the tree.

which will play an important role later. We showed above that the probability of any given sequence that contains k heads is $p^k(1-p)^{n-k}$, so we have

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \text{number of distinct } n\text{-toss sequences that contain } k \text{ heads.}$$

The numbers $\binom{n}{k}$ (called "n choose k") are known as the **binomial coefficients**, while the probabilities $p(k)$ are known as the **binomial probabilities**. Using a counting argument, to be given in Section 1.6, one finds that

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}, \qquad k = 0, 1, \ldots, n,$$

where for any positive integer i we have

$$i! = 1 \cdot 2 \cdots (i-1) \cdot i,$$

and, by convention, $0! = 1$. An alternative verification is sketched in the end-of-chapter problems. Note that the binomial probabilities $p(k)$ must add to 1, thus showing the **binomial formula**

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Example 1.23. Grade of service. An internet service provider has installed c modems to serve the needs of a population of n customers. It is estimated that at a given time, each customer will need a connection with probability p, independently of the others. What is the probability that there are more customers needing a connection than there are modems?

Here we are interested in the probability that more than c customers simultaneously need a connection. It is equal to

$$\sum_{k=c+1}^{n} p(k),$$

where

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

are the binomial probabilities.

This example is typical of problems of sizing the capacity of a facility to serve the needs of a homogeneous population, consisting of independently acting customers. The problem is to select the size c to achieve a certain threshold probability (sometimes called *grade of service*) that no user is left unserved.

1.6 COUNTING*

The calculation of probabilities often involves counting of the number of outcomes in various events. We have already seen two contexts where such counting arises.

(a) When the sample space Ω has a finite number of equally likely outcomes, so that the discrete uniform probability law applies. Then, the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{Number of elements of } A}{\text{Number of elements of } \Omega},$$

and involves counting the elements of A and of Ω.

(b) When we want to calculate the probability of an event A with a finite number of equally likely outcomes, each of which has an already known probability p. Then the probability of A is given by

$$\mathbf{P}(A) = p \cdot (\text{Number of elements of } A),$$

and involves counting of the number of elements of A. An example of this type is the calculation of the probability of k heads in n coin tosses (the binomial probabilities). We saw there that the probability of each distinct sequence involving k heads is easily obtained, but the calculation of the number of all such sequences is somewhat intricate, as will be seen shortly.

While counting is in principle straightforward, it is frequently challenging; the art of counting constitutes a large portion of a field known as **combinatorics**. In this section, we present the basic principle of counting and apply it to a number of situations that are often encountered in probabilistic models.

The Counting Principle

The counting principle is based on a divide-and-conquer approach, whereby the counting is broken down into stages through the use of a tree. For example, consider an experiment that consists of two consecutive stages. The possible results of the first stage are $a_1, a_2, \ldots, a_m$; the possible results of the second stage are $b_1, b_2, \ldots, b_n$. Then, the possible results of the two-stage experiment are all possible **ordered** pairs (a_i, b_j), $i = 1, \ldots, m$, $j = 1, \ldots, n$. Note that the number of such ordered pairs is equal to mn. This observation can be generalized as follows (see also Fig. 1.16).

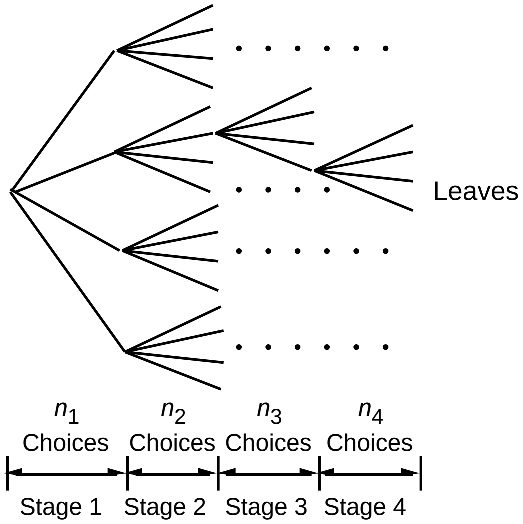


Figure 1.16: Illustration of the basic counting principle. The counting is carried out in r stages ($r = 4$ in the figure). The first stage has n_1 possible results. For every possible result of the first $i-1$ stages, there are n_i possible results at the ith stage. The number of leaves is $n_1 n_2 \cdots n_r$. This is the desired count.

The Counting Principle

Consider a process that consists of r stages. Suppose that:

(a) There are n_1 possible results for the first stage.

(b) For every possible result of the first stage, there are n_2 possible results at the second stage.

(c) More generally, for all possible results of the first $i-1$ stages, there are n_i possible results at the ith stage.

Then, the total number of possible results of the r-stage process is

$$n_1 \cdot n_2 \cdots n_r.$$

Example 1.24. The number of telephone numbers. A telephone number is a 7-digit sequence, but the first digit has to be different from 0 or 1. How many distinct telephone numbers are there? We can visualize the choice of a sequence as a sequential process, where we select one digit at a time. We have a total of 7 stages, and a choice of one out of 10 elements at each stage, except for the first stage where we only have 8 choices. Therefore, the answer is

$$8 \cdot \underbrace{10 \cdot 10 \cdots 10}_{6 \text{ times}} = 8 \cdot 10^6.$$

Example 1.25. The number of subsets of an n-element set. Consider an n-element set $\{s_1, s_2, \ldots, s_n\}$. How many subsets does it have (including itself and the empty set)? We can visualize the choice of a subset as a sequential process where we examine one element at a time and decide whether to include it in the set or not. We have a total of n stages, and a binary choice at each stage. Therefore the number of subsets is

$$\underbrace{2 \cdot 2 \cdots 2}_{n \text{ times}} = 2^n.$$

It should be noted that the Counting Principle remains valid even if each first-stage result leads to a different set of potential second-stage results, etc. The only requirement is that the number of possible second-stage results is constant, regardless of the first-stage result. This observation is used in the sequel.

In what follows, we will focus primarily on two types of counting arguments that involve the selection of k objects out of a collection of n objects. If the order of selection matters, the selection is called a **permutation**, and otherwise, it is called a **combination**. We will then discuss a more general type of counting, involving a **partition** of a collection of n objects into multiple subsets.

k-permutations

We start with n distinct objects, and let k be some positive integer, with $k \le n$. We wish to count the number of different ways that we can pick k out of these n objects and arrange them in a sequence, i.e., the number of distinct k-object sequences. We can choose any of the n objects to be the first one. Having chosen the first, there are only $n-1$ possible choices for the second; given the choice of the first two, there only remain $n-2$ available objects for the third stage, etc. When we are ready to select the last (the kth) object, we have already chosen $k-1$ objects, which leaves us with $n-(k-1)$ choices for the last one. By the Counting Principle, the number of possible sequences, called **k-permutations**, is

$$\begin{aligned} n(n-1)\cdots(n-k+1) &= \frac{n(n-1)\cdots(n-k+1)(n-k)\cdots 2\cdot 1}{(n-k)\cdots 2\cdot 1} \\ &= \frac{n!}{(n-k)!}. \end{aligned}$$

In the special case where $k=n$, the number of possible sequences, simply called **permutations**, is

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

(Let $k=n$ in the formula for the number of k-permutations, and recall the convention $0! = 1$.)

Example 1.26. Let us count the number of words that consist of four distinct letters. This is the problem of counting the number of 4-permutations of the 26 letters in the alphabet. The desired number is

$$\frac{n!}{(n-k)!} = \frac{26!}{22!} = 26\cdot 25\cdot 24\cdot 23 = 358,800.$$

The count for permutations can be combined with the Counting Principle to solve more complicated counting problems.

Example 1.27. You have n_1 classical music CDs, n_2 rock music CDs, and n_3 country music CDs. In how many different ways can you arrange them so that the CDs of the same type are contiguous?

We break down the problem in two stages, where we first select the order of the CD types, and then the order of the CDs of each type. There are 3! ordered sequences of the types of CDs (such as classical/rock/country, rock/country/classical, etc), and there are $n_1!$ (or $n_2!$, or $n_3!$) permutations of the classical (or rock, or country, respectively) CDs. Thus for each of the 3! CD type sequences, there are $n_1!\, n_2!\, n_3!$ arrangements of CDs, and the desired total number is $3!\, n_1!\, n_2!\, n_3!$.

Combinations

There are n people and we are interested in forming a committee of k. How many different committees are there? More abstractly, this is the same as the problem of counting the number of k-element subsets of a given n-element set. Notice that forming a combination is different than forming a k-permutation, because **in a combination there is no ordering of the selected elements**. Thus for example, whereas the 2-permutations of the letters A, B, C, and D are

AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC,

the combinations of two out of four of these letters are

AB, AC, AD, BC, BD, CD.

There is a close connection between the number of combinations and the binomial coefficient that was introduced in Section 1.5. To see this note that specifying an n-toss sequence with k heads is the same as picking k elements (those that correspond to heads) out of the n-element set of tosses. Thus, the number of combinations is the same as the binomial coefficient $\binom{n}{k}$ introduced in Section 1.5.

To count the number of combinations, note that selecting a k-permutation is the same as first selecting a combination of k items and then ordering them. Since there are $k!$ ways of ordering the k selected items, we see that the number of k-permutations is equal to the number of combinations times $k!$. Hence, the number of possible combinations, is given by

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

Example 1.28. The number of combinations of two out of the four letters A, B, C, and D is found by letting $n = 4$ and $k = 2$. It is

$$\binom{4}{2} = \frac{4!}{2!\,2!} = 6,$$

consistently with the listing given earlier.

It is worth observing that counting arguments sometimes lead to formulas that are rather difficult to derive algebraically. One example is the **binomial formula**

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$$

discussed in Section 1.5. Here is another example. Since $\binom{n}{k}$ is the number of k-element subsets of a given n-element subset, the sum over k of $\binom{n}{k}$ counts the number of subsets of all possible cardinalities. It is therefore equal to the number of all subsets of an n-element set, which is 2^n, and we obtain

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Partitions

Recall that a combination is a choice of k elements out of an n-element set without regard to order. This is the same as partitioning the set in two: one part contains k elements and the other contains the remaining $n-k$. We now generalize by considering partitions in more than two subsets.

We have n distinct objects and we are given nonnegative integers $n_1, n_2, \ldots, n_r$, whose sum is equal to n. The n items are to be divided into r disjoint groups, with the ith group containing exactly n_i items. Let us count in how many ways this can be done.

We form the groups one at a time. We have $\binom{n}{n_1}$ ways of forming the first group. Having formed the first group, we are left with $n - n_1$ objects. We need to choose n_2 of them in order to form the second group, and we have $\binom{n-n_1}{n_2}$ choices, etc. Using the Counting Principle for this r-stage process, the total number of choices is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r},$$

which is equal to

$$\frac{n!}{n_1!\,(n-n_1)!}\,\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}\cdots\frac{(n-n_1-\cdots-n_{r-1})!}{(n-n_1-\cdots-n_{r-1}-n_r)!\,n_r!}.$$

We note that several terms cancel and we are left with

$$\frac{n!}{n_1!\,n_2!\cdots n_r!}.$$

This is called the **multinomial coefficient** and is usually denoted by

$$\binom{n}{n_1, n_2, \ldots, n_r}.$$

Example 1.29. Anagrams. How many different letter sequences can be obtained by rearranging the letters in the word TATTOO? There are six positions to be filled by the available letters. Each rearrangement corresponds to a partition of the set of the six positions into a group of size 3 (the positions that get the letter T), a group of size 1 (the position that gets the letter A), and a group of size 2 (the positions that get the letter O). Thus, the desired number is

$$\frac{6!}{1!\,2!\,3!} = \frac{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6}{1\cdot 1\cdot 2\cdot 1\cdot 2\cdot 3} = 60.$$

It is instructive to rederive this answer using an alternative argument. (This argument can also be used to rederive the multinomial coefficient formula; see the end-of-chapter problems.) Let us rewrite TATTOO in the form $\mathrm{T_1AT_2T_3O_1O_2}$ pretending for a moment that we are dealing with 6 distinguishable objects. These 6 objects can be rearranged in 6! different ways. However, any of the 3! possible permutations of $\mathrm{T_1}$, $\mathrm{T_1}$, and $\mathrm{T_3}$, as well as any of the 2! possible permutations of $\mathrm{O_1}$ and $\mathrm{O_2}$, lead to the same word. Thus, when the subscripts are removed, there are only $6!/(3!\,2!)$ different words.

Example 1.30. A class consisting of 4 graduate and 12 undergraduate students is randomly divided into four groups of 4. What is the probability that each group includes a graduate student? This is the same as Example 1.11 in Section 1.3, but we will now obtain the answer using a counting argument.

We first determine the nature of the sample space. A typical outcome is a particular way of partitioning the 16 students into four groups of 4. We take the term "randomly" to mean that every possible partition is equally likely, so that the probability question can be reduced to one of counting.

According to our earlier discussion, there are

$$\binom{16}{4,4,4,4} = \frac{16!}{4!\,4!\,4!\,4!}$$

different partitions, and this is the size of the sample space.

Let us now focus on the event that each group contains a graduate student. Generating an outcome with this property can be accomplished in two stages:

(a) Take the four graduate students and distribute them to the four groups; there are four choices for the group of the first graduate student, three choices for the second, two for the third. Thus, there is a total of 4! choices for this stage.

(b) Take the remaining 12 undergraduate students and distribute them to the four groups (3 students in each). This can be done in

$$\binom{12}{3,3,3,3} = \frac{12!}{3!\,3!\,3!\,3!}$$

different ways.

By the Counting Principle, the event of interest can materialize in

$$\frac{4!\,12!}{3!\,3!\,3!\,3!}$$

different ways. The probability of this event is

$$\frac{\dfrac{4!\,12!}{3!\,3!\,3!\,3!}}{\dfrac{16!}{4!\,4!\,4!\,4!}}.$$

After some cancellations, we can see that this is the same as the answer $12 \cdot 8 \cdot 4/(15 \cdot 14 \cdot 13)$ obtained in Example 1.11.

Here is a summary of all the counting results we have developed.

Summary of Counting Results

- Permutations of n objects: $n!$
- k-permutations of n objects: $n!/(n-k)!$
- Combinations of k out of n objects: $\displaystyle\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Partitions of n objects into r groups with the ith group having n_i objects:

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!\; n_2! \cdots n_r!}.$$

1.7 SUMMARY AND DISCUSSION

A probability problem can usually be broken down into a few basic steps:

1. The description of the sample space, that is, the set of possible outcomes of a given experiment.
2. The (possibly indirect) specification of the probability law (the probability of each event).
3. The calculation of probabilities and conditional probabilities of various events of interest.

The probabilities of events must satisfy the nonnegativity, additivity, and normalization axioms. In the important special case where the set of possible outcomes is finite, one can just specify the probability of each outcome and obtain the probability of any event by adding the probabilities of the elements of the event.

Conditional probabilities can be viewed as probability laws on the same sample space. We can also view the conditioning event as a new universe, because only outcomes contained in the conditioning event can have positive conditional probability. Conditional probabilities are derived from the (unconditional) probability law using the definition $\mathbf{P}(A \mid B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$. However, the reverse process is often convenient, that is, first specify some conditional probabilities that are natural for the real situation that we wish to model, and then use them to derive the (unconditional) probability law. Two important tools in this context are the multiplication rule and the total probability theorem.

We have illustrated through examples three methods of specifying probability laws in probabilistic models:

(1) The **counting method**. This method applies to the case where the number of possible outcomes is finite, and all outcomes are equally likely. To calculate the probability of an event, we count the number of elements in the event and divide by the number of elements of the sample space.

(2) The **sequential method**. This method applies when the experiment has a sequential character, and suitable conditional probabilities are specified or calculated along the branches of the corresponding tree (perhaps using the counting method). The probabilities of various events are then obtained by multiplying conditional probabilities along the corresponding paths of the tree, using the multiplication rule.

(3) The **divide-and-conquer method**. Here, the probabilities $\mathbf{P}(B)$ of various events B are obtained from conditional probabilities $\mathbf{P}(B \mid A_i)$, where the A_i are suitable events that form a partition of the sample space and have known probabilities $\mathbf{P}(A_i)$. The probabilities $\mathbf{P}(B)$ are then obtained by using the total probability theorem.

Finally, we have focused on a few side topics that reinforce our main themes. We have discussed the use of Bayes' rule in inference, which is an important application context. We have also discussed some basic principles of counting and combinatorics, which are helpful in applying the counting method.

SOLVED PROBLEMS

SECTION 1.1. Sets

Problem 1. * Prove the identity $A \cup \Big(\cap_{n=1}^{\infty} B_n\Big) = \cap_{n=1}^{\infty}(A \cup B_n)$.

Solution. If x belongs to the set on the left, there are two possibilities. Either $x \in A$, in which case it belongs to all of the sets $A \cup B_n$, and therefore belongs to the set on the right. Alternatively, x belongs to all of the sets B_n in which case, it belongs to all of the sets $A \cup B_n$, and therefore again belongs to the set on the right.

Conversely, if x belongs to the set on the right, then it belongs to $A \cup B_n$ for all n. If x belongs to A, then it belongs to the set on the left. Otherwise, x must belong to every set B_n and again belongs to the set on the left.

Problem 2. * Cantor's diagonalization argument. Show that the set $[0, 1]$ is uncountable, that is, its elements cannot be arranged in a sequence.

Solution. Any number x in $[0, 1]$ can be represented in terms of its decimal expansion, e.g., $1/3 = 0.3333\cdots$. Note that most numbers have a unique decimal expansion, but there are a few exceptions. For example, $1/2$ can be represented as $0.5000\cdots$ or as $0.49999\cdots$. It can be shown that this is the only kind of exception, i.e., decimal expansions that end with an infinite string of zeroes or an infinite string of nines.

Suppose to derive a contradiction, that the elements of $[0, 1]$ can be arranged in a sequence $x_1, x_2, x_3, \ldots$, so that every element of $[0, 1]$ appears in the sequence. Consider the decimal expansion of x_n:

$$x_n = 0.a_n^1 a_n^2 a_n^3 \cdots,$$

where each digit a_n^i belongs to $\{0, 1, \ldots, 9\}$. Consider now a number y constructed as follows. The nth digit of y can be 1 or 2, and is chosen so that it is different from the nth digit of x_n. Note that y has a unique decimal expansion since it does not end with an infinite sequence of zeroes or nines. The number y differs from each x_n, since it has a different nth digit. Therefore, the sequence $x_1, x_2, \ldots$ does not exhaust the elements of $[0, 1]$, contrary to what was assumed. The contradiction establishes that the set $[0, 1]$ is uncountable.

SECTION 1.2. Probabilistic Models

Problem 3. Out of the students in a class, 60% are geniuses, 70% love chocolate, and 40% fall into both categories. Determine the probability that a randomly selected student is neither a genius nor a chocolate lover.

Solution. Let G and C be the events that the chosen student is a genius and a chocolate lover, respectively. We have $\mathbf{P}(G) = 0.6$, $\mathbf{P}(C) = 0.7$, and $\mathbf{P}(G \cap C) = 0.4$. We are interested in $\mathbf{P}(G^c \cap C^c)$, which is obtained with the following calculation:

$$\mathbf{P}(G^c \cap C^c) = 1 - \mathbf{P}(G \cup C) = 1 - \Big(\mathbf{P}(G) + \mathbf{P}(C) - \mathbf{P}(G \cap C)\Big) = 1 - (0.6 + 0.7 - 0.4) = 0.1.$$

Problem 4. A six-sided die is loaded in a way that each even face is twice as likely as each odd face. All even faces are equally likely, as are all odd faces. Construct a probabilistic model for a single roll of this die and find the probability that the outcome is less than 4.

Solution. We first determine the probabilities of the six possible outcomes. Let $a = \mathbf{P}(\{1\}) = \mathbf{P}(\{3\}) = \mathbf{P}(\{5\})$ and $b = \mathbf{P}(\{2\}) = \mathbf{P}(\{4\}) = \mathbf{P}(\{6\})$. We are given that $b = 2a$. By the additivity and normalization axioms, $1 = 3a + 3b = 3a + 6a = 9a$. Thus, $a = 1/9$, $b = 2/9$, and $\mathbf{P}(\{1, 2, 3\}) = 4/9$.

Problem 5. A four-sided die is rolled repeatedly, until the first time (if ever) that an even number is obtained. What is the sample space for this experiment?

Solution. The outcome of this experiment can be any finite sequence of the form $(a_1, a_2, \ldots, a_n)$, where n is an arbitrary positive integer, $a_1, a_2, \ldots, a_{n-1}$ belong to $\{1, 3\}$, and a_n belongs to $\{2, 4\}$. In addition, there are possible outcomes in which an even number is never obtained. Such outcomes are infinite sequences $(a_1, a_2, \ldots)$, with each element in the sequence belonging to $\{1, 3\}$. The sample space consists of all possible outcomes of the above two types.

Problem 6. * Bonferroni's inequality.

(a) Prove that for any two events A and B, we have

$$\mathbf{P}(A \cap B) \geq \mathbf{P}(A) + \mathbf{P}(B) - 1.$$

(b) Generalize to the case of n events $A_1, A_2, \ldots, A_n$, by showing that

$$\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_n) \geq \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots + \mathbf{P}(A_n) - (n - 1).$$

Solution. We have $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$ and $\mathbf{P}(A \cup B) \leq 1$, which implies part (a). For part (b), we use de Morgan's law to obtain

$$\begin{aligned} 1 - \mathbf{P}(A_1 \cap \cdots \cap A_n) &= \mathbf{P}\big((A_1 \cap \cdots \cap A_n)^c\big) \\ &= \mathbf{P}(A_1^c \cup \cdots \cup A_n^c) \\ &\leq \mathbf{P}(A_1^c) + \cdots + \mathbf{P}(A_n^c) \\ &= \big(1 - \mathbf{P}(A_1)\big) + \cdots + \big(1 - \mathbf{P}(A_n)\big) \\ &= n - \mathbf{P}(A_1) - \cdots - \mathbf{P}(A_n). \end{aligned}$$

Problem 7. * Continuity property of probabilities

(a) Let $A_1, A_2, \ldots$ be an infinite sequence of events, which is "monotonically increasing," meaning that $A_n \subset A_{n+1}$ for every n. Let $A = \cup_{n=1}^{\infty} A_n$. Show that $\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n)$.

(b) Suppose now that the events are "monotonically decreasing," i.e., $A_{n+1} \subset A_n$ for every n. Let $A = \cap_{n=1}^{\infty} A_n$. Show that $\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n)$.

(c) Consider an experiment whose sample space is the real line. Show that

$$\mathbf{P}\big([0,\infty)\big) = \lim_{n\to\infty} \mathbf{P}\big([0,n]\big) \qquad \text{and} \qquad \lim_{n\to\infty} \mathbf{P}\big([n,\infty)\big) = 0.$$

Solution. (a) Let $B_1 = A_1$ and, for $n \geq 2$, $B_n = A_n \cap A_{n-1}^c$. The events B_n are disjoint, $\cup_{k=1}^n B_k = A_n$, and $\cup_{k=1}^\infty B_k = A$. We have, by the additivity axiom,

$$\mathbf{P}(A) = \sum_{k=1}^{\infty} \mathbf{P}(B_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{P}(B_k) = \lim_{n\to\infty} \mathbf{P}(\cup_{k=1}^n B_k) = \lim_{n\to\infty} \mathbf{P}(A_n).$$

(b) Let $C_n = A_n^c$ and $C = A^c$. Since $A_{n+1} \subset A_n$, we obtain $C_n \subset C_{n+1}$, and the events C_n are increasing. Furthermore, $C = A^c = (\cap_{n=1}^\infty A_n)^c = \cup_{n=1}^\infty A_n^c = \cup_{n=1}^\infty C_n$. Using the result from part (a) for the sequence C_n, we obtain

$$1 - \mathbf{P}(A) = \mathbf{P}(A^c) = \mathbf{P}(C) = \lim_{n\to\infty} \mathbf{P}(C_n) = \lim_{n\to\infty} (1 - \mathbf{P}(A_n)),$$

from which we conclude that $\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n)$.

(c) For the first equality, use the result from part (a) with $A_n = [0, n]$ and $A = [0, \infty)$. For the second, use the result from part (b) with $A_n = [n, \infty)$ and $A = \cap_{n=1}^\infty A_n = \emptyset$.

SECTION 1.3. Conditional Probability

Problem 8. We roll two fair 6-sided dice. Each one of the 36 possible outcomes is assumed to be equally likely.

(a) Find the probability that doubles were rolled.

(b) Given that the roll resulted in a sum of 4 or less, find the conditional probability that doubles were rolled.

(c) Find the probability that at least one die is a 6.

(d) Given that the two dice land on different numbers, find the conditional probability that at least one die is a 6.

Solution. (a) There are 6 possible outcomes that are doubles, so the probability of doubles is $6/36 = 1/6$.

(b) The conditioning event (sum is 4 or less) consists of the 6 outcomes

$$\big\{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)\big\},$$

2 of which are doubles, so the conditional probability of doubles is $2/6 = 1/3$.

(c) There are 11 possible outcomes with at least one 6, namely, $(6,6)$, $(6,i)$, and $(i,6)$, for $i = 1, 2, \ldots, 5$. The probability that at least one die is a 6 is $11/36$.

(d) There are 30 possible outcomes where the dice land on different numbers. Out of these, there are 10 outcomes in which at least one of the rolls is a 6. Thus, the desired conditional probability is $10/30 = 1/3$.

Problem 9. A coin is tossed twice. Alice claims that the event of two heads is at least as likely if we know that the first toss is a head than if we know that at least one of the tosses is a head. Is she right? Does it make a difference if the coin is fair or unfair? How can we generalize Alice's reasoning?

Solution. Let A be the event that the first toss is a head and let B be the event that the second toss is a head. We must compare the conditional probabilities $\mathbf{P}(A \cap B \,|\, A)$ and $\mathbf{P}(A \cap B \,|\, A \cup B)$. We have

$$\mathbf{P}(A \cap B \,|\, A) = \frac{\mathbf{P}\big((A \cap B) \cap A\big)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)},$$

and

$$\mathbf{P}(A \cap B \,|\, A \cup B) = \frac{\mathbf{P}\big((A \cap B) \cap (A \cup B)\big)}{\mathbf{P}(A \cup B)} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A \cup B)}.$$

Since $\mathbf{P}(A \cup B) \geq \mathbf{P}(A)$, the first conditional probability above is at least as large, so Alice is right, regardless of whether the coin is fair or not. In the case where the coin is fair, that is, if all four outcomes HH, HT, TH, TT are equally likely, we have

$$\frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{1/4}{1/2} = \frac{1}{2}, \qquad \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A \cup B)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

A generalization of Alice's reasoning is that if A, B, and C are events such that $B \subset C$ and $A \cap B = A \cap C$ (for example, if $A \subset B \subset C$), then the event A is more likely if we know that B has occurred than if we know that C has occurred.

Problem 10. You are given three coins: one has heads in both faces, the second has tails in both faces, and the third has a head in one face and a tail in the other. You choose a coin at random, toss it, and it comes heads. What is the probability that the opposite face is tails?

Solution. In this problem, there is a tendency to reason that since the opposite face is either heads or tails, the desired probability is 1/2. This is, however, wrong, because given that heads came, it is more likely that the two-headed coin was chosen. The correct reasoning is to calculate the conditional probability

$$P = \mathbf{P}(\text{two-headed coin was chosen} \,|\, \text{heads came})$$

We have

$$\mathbf{P}(\text{two-headed coin was chosen and heads came}) = \frac{1}{3}$$

$$\mathbf{P}(\text{heads came}) = \frac{1}{2}$$

so by taking the ratio of the above two probabilities, we obtain $P = 2/3$. Thus the probability that the opposite face is tails is $1 - P = 1/3$.

Problem 11. Monty Hall problem. You are told that a prize is equally likely to be found behind one out of three closed doors. You point to one of the doors. Then, your friend opens for you one of the remaining two doors, after she makes sure that the prize is not behind it. At this point, you can stick to your initial choice, or switch to the other unopened door. You win the prize if it lies behind your final choice of a door.

(a) As far as the probability of winning is concerned, is there an advantage in switching doors? State precisely any modeling assumptions you are making.

(b) Consider the following strategy. You first point to door 1. If door 2 is opened, you do not switch. If door 3 is opened, you switch. What is the probability of winning under this strategy?

Solution. (a) Under the strategy of no switching, your initial choice will determine whether you win or not, and the probability of winning is 1/3. This is because the prize is equally likely to be behind each door, and we are assuming that you are given no additional information.

Let us now consider the strategy of switching. If the prize is behind the initially chosen door (probability 1/3), you do not win. If it is not (probability 2/3), and given that another door without a prize has been opened for you, you will get to the winning door once you switch. Thus, the probability of winning is now 2/3, and this is a better strategy.

(b) Under this strategy, there is insufficient information for determining the probability of winning. The answer depends on the way that your friend chooses which door to open. Let us consider two possibilities.

Suppose that if the prize is behind door 1, your friend always chooses to open door 2. (If the prize is behind door 2 or 3, your friend has no choice.) If the prize is behind door 1, your friend opens door 2, you do not switch and you win. If the prize is behind door 2, your friend opens door 3, you switch, and you win. If the prize is behind door 3, your friend opens door 2, you do not switch, and you lose. Thus, the probability of winning is 2/3.

Suppose now that if the prize is behind door 1, your friend is equally likely to open either door 2 or 3. If the prize is behind door 1 (probability 1/3), and if your friend opens door 2 (probability 1/2), you do not switch and you win (probability 1/6). But if your friend opens door 3, you switch and you lose. If the prize is behind door 2, your friend opens door 3, you switch, and you win (probability 1/3). If the prize is behind door 3, your friend opens door 2, you do not switch and you lose. Thus, the probability of winning is $\frac{1}{6}+\frac{1}{3}=\frac{1}{2}$.

Problem 12. * Show that $\mathbf{P}(A\cap B\,|\,B)=\mathbf{P}(A\,|\,B)$, assuming that $\mathbf{P}(B)>0$.

Solution. Using the definition of conditional probabilities, we have

$$\mathbf{P}(A\cap B\,|\,B)=\frac{\mathbf{P}(A\cap B\cap B)}{\mathbf{P}(B)}=\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}=\mathbf{P}(A\,|\,B).$$

SECTION 1.4. Total Probability Theorem and Bayes' Rule

Problem 13. Testing for a rare disease. A test for a certain rare disease has 90% accuracy: if a person has the disease, the test results are positive with probability 0.9, and if the person does not have the disease, the test results are negative with probability 0.9. A random person drawn from a certain population has probability 0.001 of having the disease. Given that the person just tested positive, what is the probability of having the disease?

Solution. Let A be the event that the person has the disease. Let B be the event that the test results are positive. The desired probability, $\mathbf{P}(A \,|\, B)$, is found by Bayes' rule:

$$\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A)\mathbf{P}(B \,|\, A)}{\mathbf{P}(A)\mathbf{P}(B \,|\, A) + \mathbf{P}(A^c)\mathbf{P}(B \,|\, A^c)} = \frac{0.001 \cdot 0.90}{0.001 \cdot 0.90 + 0.999 \cdot 0.10} = 0.089.$$

Note that even though the test was assumed to be fairly accurate, a person who has tested positive is still very unlikely to have the disease.

Problem 14. You have n drawers in your filing cabinet, and you left your term paper in drawer k with probability $p_k > 0$. The drawers are so messy that even if you correctly guess that the term paper is in drawer j, the probability that you find it is only d_j. You search for your paper in a particular drawer, say drawer i, but the search is unsuccessful. Conditional on this event, show that the probability that your paper is in drawer j, is given by

$$\frac{p_j}{1 - p_i d_i}, \qquad \text{if } j \neq i, \qquad\qquad \frac{p_i(1 - d_i)}{1 - p_i d_i}, \qquad \text{if } j = i.$$

Solution. Let i be the drawer that you search, and let A be the event that you find nothing. Since the paper is in drawer i with probability p_i, and your search is successful with probability d_i, the multiplication rule yields $\mathbf{P}(A^c) = p_i d_i$ and $\mathbf{P}(A) = 1 - p_i d_i$. Let B be the event that the paper is in drawer j. If $j \neq i$, then $A \cap B = B$, $\mathbf{P}(A \cap B) = \mathbf{P}(B)$, and we have

$$\mathbf{P}(B \,|\, A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(B)}{\mathbf{P}(A)} = \frac{p_j}{1 - p_i d_i},$$

as required.

Consider now the case $i = j$. Defining the events A and B as above, we obtain

$$\mathbf{P}(B \,|\, A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(B)\mathbf{P}(A \,|\, B)}{\mathbf{P}(A)} = \frac{p_i(1 - d_i)}{1 - p_i d_i}.$$

Problem 15. How an inferior player with a superior strategy can gain an advantage. Boris is about to play a two-game chess match with an opponent, and wants to find the strategy that maximizes his winning chances. Each game ends with either a win by one of the players, or a draw. If the score is tied at the end of the two games, the match goes into sudden-death mode, and the players continue to play until the first time one of them wins a game (and the match). Boris has two playing styles and he can choose one of the two at will in each game, no matter what style he chose in previous games. (1) *Timid play* with which he draws with probability $p_d > 0$, and he loses with probability $1 - p_d$. (2) *Bold play* with which he wins with probability p_w, and he loses with probability $1 - p_w$. Boris will always play bold during sudden death, but may switch style between games 1 and 2.

(a) Find the probability that Boris wins the match for each of the following strategies:

(i) Play bold in both games 1 and 2.

(ii) Play timid in both games 1 and 2.

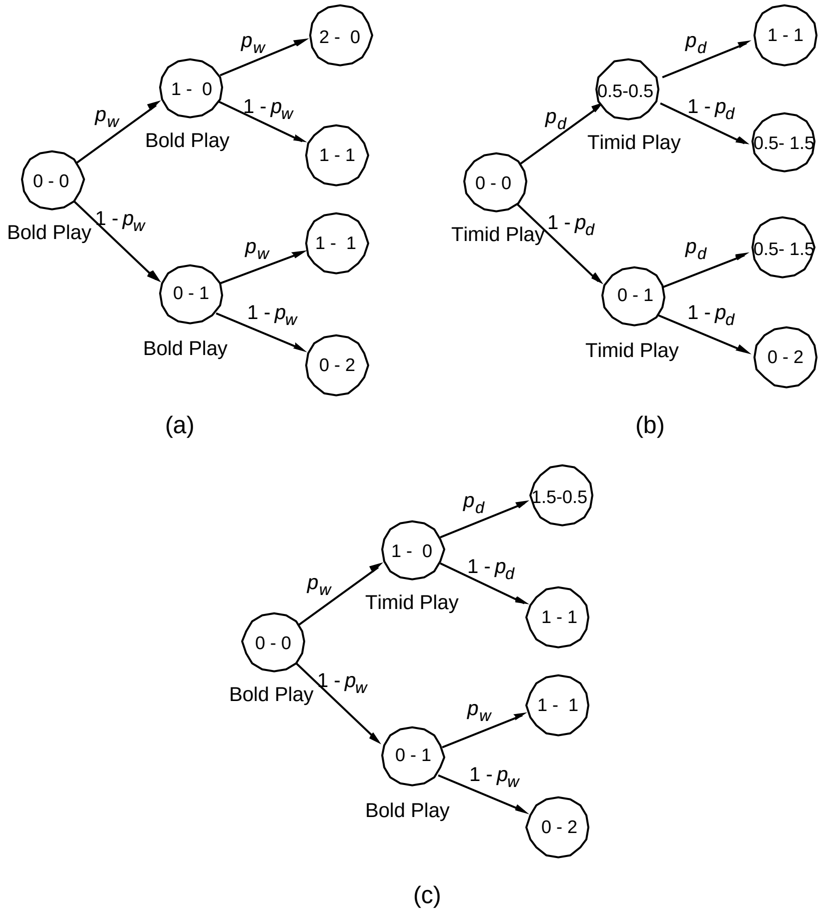


Figure 1.17: Sequential descriptions of the sample space of the chess match histories under strategies (i), (ii), and (iii).

(iii) Play timid whenever he is ahead in the score, and play bold otherwise.

(b) Assume that $p_w < 1/2$, so Boris is the worse player, regardless of the playing style he adopts. Show that with the strategy in (iii) above, and depending on the values of p_w and p_d, Boris may have a better than a 50-50 chance to win the match. How do you explain this advantage?

Solution. (a) Figure 1.17 provides a sequential description of the sample space for the three different strategies. Here we assume 1 point for a win, 0 for a loss, and 1/2 point for a draw. In the case of a tied 1-1 score, we go to sudden death in the next game, and Boris wins the match with probability p_w, or loses the match with probability $1 - p_w$.

(i) Using the total probability theorem and the sequential description of Fig. 1.17(a),

we have

$$\mathbf{P}(\text{Boris wins}) = p_w^2 + 2p_w(1-p_w)p_w.$$

The term p_w^2 corresponds to the win-win outcome, and the term $2p_w(1-p_w)p_w$ corresponds to the win-lose-win and the lose-win-win outcomes.

(ii) Using Fig. 1.17(b), we have

$$\mathbf{P}(\text{Boris wins}) = p_d^2 p_w,$$

corresponding to the draw-draw-win outcome.

(iii) Using Fig. 1.17(c), we have

$$\mathbf{P}(\text{Boris wins}) = p_w p_d + p_w(1-p_d)p_w + (1-p_w)p_w^2.$$

The term $p_w p_d$ corresponds to the win-draw outcome, the term $p_w(1-p_d)p_w$ corresponds to the win-lose-win outcome, and the term $(1-p_w)p_w^2$ corresponds to lose-win-win outcome.

(b) If $p_w < 1/2$, Boris has a greater probability of losing rather than winning any one game, regardless of the type of play he uses. Despite this, the probability of winning the match with strategy (iii) can be greater than 1/2, provided that p_w is close enough to 1/2 and p_d is close enough to 1. As an example, if $p_w = 0.45$ and $p_d = 0.9$, with strategy (iii) we have

$$\mathbf{P}(\text{Boris wins}) = 0.45 \cdot 0.9 + 0.45^2 \cdot (1-0.9) + (1-0.45) \cdot 0.45^2 \approx 0.54.$$

With strategies (i) and (ii), the corresponding probabilities of a win can be calculated to be approximately 0.43 and 0.36, respectively. What is happening here is that with strategy (iii), Boris is allowed to select a playing style *after* seeing the result of the first game, while his opponent is not. Thus, by being able to dictate the playing style in each game after receiving partial information about the match's outcome, Boris gains an advantage.

Problem 16. * Two players take turns removing a ball from a jar that initially contains m white and n black balls. The first player to remove a white ball wins.

(a) Derive a recursion for the probability $P(m,n)$ that the starting player wins.

(b) Fix n and let m increase. Use the recursion to show that $P(m,n)$ approaches 1 as $m \to \infty$.

(c) Fix m and let n increase. Use the recursion to show that $P(m,n)$ approaches 1/2 as $n \to \infty$.

Solution. (a) We have, using the total probability theorem,

$$P(m,n) = \frac{m}{m+n} + \frac{n}{m+n}\big(1 - P(m,n-1)\big) = 1 - \frac{n}{m+n}P(m,n-1).$$

The probabilities $P(m,1), P(m,2), \ldots$ can be calculated using this formula, starting with the initial condition $P(m,0) = 1$.

(b) From the recursion, we see that $P(m,n)$ is greater or equal to $m/(m+n)$, which approaches 1 as $m \to \infty$.

(c) For fixed m, define

$$\xi_n = P(m,n) - \frac{1}{2}.$$

Then the recursion of part (a) becomes

$$\xi_n = -\xi_{n-1} + \frac{m}{m+n},$$

from which

$$\xi_n = \xi_{n-2} + \frac{m}{m+n} - \frac{m}{m+n-1} = \xi_{n-2} - \frac{m}{(m+n)(m+n-1)}.$$

Therefore the sequence ξ_{2k}, $k = 0, 1, \ldots$ is monotonically decreasing with k, and since from the definition of ξ_k, it is bounded from below, it must converge to some number. From the above equation, we see that the only possible limit of ξ_{2k} is 0. From this and the equation $\xi_n = -\xi_{n-1} + m/(m+n)$, we see that ξ_{2k-1} also converges to 0 as $k \to \infty$. Therefore the entire sequence ξ_n converges to 1/2, or equivalently $P(m,n)$ tends to 1/2 as $n \to \infty$.

Problem 17. * Each of k jars contains m white and n black balls. A ball is randomly chosen from jar 1 and transferred to jar 2, then a ball is randomly chosen from jar 2 and transferred to jar 3, etc. Finally, a ball is randomly chosen from jar k. Show that the probability that the last ball is white is the same as the probability that the first ball is white, i.e., it is $m/(m+n)$.

Solution. We derive a recursion for the probability p_i that a white ball is chosen in the ith jar. We have, using the total probability theorem,

$$p_{i+1} = \frac{m+1}{m+n+1} p_i + \frac{m}{m+n+1}(1-p_i) = \frac{1}{m+n+1} p_i + \frac{m}{m+n+1},$$

starting with the initial condition $p_1 = m/(m+n)$. Thus, we have

$$p_2 = \frac{1}{m+n+1}\frac{m}{m+n} + \frac{m}{m+n+1} = \frac{m}{m+n}.$$

More generally, this calculation shows that if $p_{i-1} = m/(m+n)$, then $p_i = m/(m+n)$. Thus, we obtain $p_i = m/(m+n)$ for all i.

Problem 18. * We have two jars each containing initially n balls. We perform four successive ball exchanges. In each exchange, we pick simultaneously and at random a ball from each jar and move it to the other jar. What is the probability that at the end of the four exchanges all the balls will be in the jar where they started?

Solution. Let $p_{i,n-i}(k)$ denote the probability that after k exchanges, a jar will contain i balls that started in that jar and $n-i$ balls that started in the other jar. We want to find $p_{n,0}(4)$. We argue recursively, using the total probability theorem. We have

$$p_{n,0}(4) = \frac{1}{n} \cdot \frac{1}{n} \cdot p_{n-1,1}(3),$$

$$
\begin{aligned}
p_{n-1,1}(3) &= p_{n,0}(2) + 2\cdot\frac{n-1}{n}\cdot\frac{1}{n}\cdot p_{n-1,1}(2) + \frac{2}{n}\cdot\frac{2}{n}\cdot p_{n-2,2}(2),\\
p_{n,0}(2) &= \frac{1}{n}\cdot\frac{1}{n}\cdot p_{n-1,1}(1),\\
p_{n-1,1}(2) &= 2\cdot\frac{n-1}{n}\cdot\frac{1}{n}\cdot p_{n-1,1}(1)\\
p_{n-2,2}(2) &= \frac{n-1}{n}\cdot\frac{n-1}{n}\cdot p_{n-1,1}(1),\\
p_{n-1,1}(1) &= 1.
\end{aligned}
$$

Combining these equations, we obtain

$$
p_{n,0}(4) = \frac{1}{n^2}\left(\frac{1}{n^2} + \frac{4(n-1)^2}{n^4} + \frac{4(n-1)^2}{n^4}\right) = \frac{1}{n^2}\left(\frac{1}{n^2} + \frac{8(n-1)^2}{n^4}\right).
$$

Problem 19. * Conditional version of the total probability theorem. Show the identity

$$
\mathbf{P}(A\,|\,B) = \mathbf{P}(C\,|\,B)\mathbf{P}(A\,|\,B\cap C) + \mathbf{P}(C^c\,|\,B)\mathbf{P}(A\,|\,B\cap C^c),
$$

assuming all the conditioning events have positive probability.

Solution. Using the conditional probability definition and the additivity axiom on the disjoint sets $A\cap B\cap C$ and $A\cap B\cap C^c$, we obtain

$$
\begin{aligned}
\mathbf{P}(C\,|\,B)&\mathbf{P}(A\,|\,B\cap C) + \mathbf{P}(C^c\,|\,B)\mathbf{P}(A\,|\,B\cap C^c)\\
&= \frac{\mathbf{P}(B\cap C)}{\mathbf{P}(B)}\cdot\frac{\mathbf{P}(A\cap B\cap C)}{\mathbf{P}(B\cap C)} + \frac{\mathbf{P}(B\cap C^c)}{\mathbf{P}(B)}\cdot\frac{\mathbf{P}(A\cap B\cap C^c)}{\mathbf{P}(B\cap C^c)}\\
&= \frac{\mathbf{P}(A\cap B\cap C) + \mathbf{P}(A\cap B\cap C^c)}{\mathbf{P}(B)}\\
&= \frac{\mathbf{P}\big((A\cap B\cap C)\cup(A\cap B\cap C^c)\big)}{\mathbf{P}(B)}\\
&= \frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}\\
&= \mathbf{P}(A\,|\,B).
\end{aligned}
$$

Problem 20. * Let A and B be events with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$. We say that an event B *suggests* an event A if $\mathbf{P}(A\,|\,B) > \mathbf{P}(A)$, and *does not suggest* event A if $\mathbf{P}(A\,|\,B) < \mathbf{P}(A)$.

(a) Show that B suggests A if and only if A suggests B.

(b) Show that B suggests A if and only if B^c does not suggest A. Assume that $\mathbf{P}(B^c) > 0$.

(c) We know that a treasure is located in one of two places, with probabilities β and $1-\beta$, respectively, where $0 < \beta < 1$. We search the first place and if the treasure is there, we find it with probability $p > 0$. Show that the event of not finding the treasure in the first place suggests that the treasure is in the second place.

Solution. (a) We have $\mathbf{P}(A \,|\, B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$, so B suggests A if and only if $\mathbf{P}(A \cap B) > \mathbf{P}(A)\mathbf{P}(B)$, which is equivalent to A suggesting B, by symmetry.

(b) Arguing as in (a) above, we have that B^c does not suggest A if and only if

$$\mathbf{P}(A \cap B^c) < \mathbf{P}(A)\mathbf{P}(B^c).$$

We have

$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A) - \mathbf{P}(A \cap B), \qquad \mathbf{P}(B^c) = 1 - \mathbf{P}(B).$$

Substituting in the previous inequality, we see that B^c does not suggest A if and only if

$$\mathbf{P}(A) - \mathbf{P}(A \cap B) < \mathbf{P}(A)\big(1 - \mathbf{P}(B)\big),$$

or $\mathbf{P}(A \cap B) > \mathbf{P}(A)\mathbf{P}(B)$. By part (a), this is equivalent to B suggesting A.

(b) Since $\mathbf{P}(B) + \mathbf{P}(B^c) = 1$, we have

$$\mathbf{P}(B)\mathbf{P}(A) + \mathbf{P}(B^c)\mathbf{P}(A) = \mathbf{P}(A) = \mathbf{P}(B)\mathbf{P}(A \,|\, B) + \mathbf{P}(B^c)\mathbf{P}(A \,|\, B^c),$$

which implies that

$$\mathbf{P}(B)\big(\mathbf{P}(A \,|\, B) - \mathbf{P}(A)\big) = \mathbf{P}(B^c)\big(\mathbf{P}(A) - \mathbf{P}(A \,|\, B^c)\big).$$

Thus, $\mathbf{P}(A \,|\, B) > \mathbf{P}(A)$ (B suggests A) if and only if $\mathbf{P}(A) > \mathbf{P}(A \,|\, B^c)$ (B^c does not suggest A).

(c) Let A and B be the events

$$A = \{\text{the treasure is in the second place}\},$$
$$B = \{\text{we don't find the treasure in the first place}\}.$$

Using the total probability theorem, we have

$$\mathbf{P}(B) = \mathbf{P}(A^c)\mathbf{P}(B \,|\, A^c) + \mathbf{P}(A)\mathbf{P}(B \,|\, A) = \beta(1-p) + (1-\beta),$$

so

$$\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{1-\beta}{\beta(1-p) + (1-\beta)} = \frac{1-\beta}{1-\beta p} > 1 - \beta = \mathbf{P}(A).$$

It follows that event B suggests event A.

Problem 21. * The paradox of induction. Consider a statement whose truth is unknown. If we see many examples that are compatible with it, we are tempted to view the statement as more probable. Such reasoning is often referred to as *inductive inference* (in a philosophical, rather than mathematical sense). Consider now the statement that "all cows are white." An equivalent statement is that "everything that is not white is not a cow." We then observe several black crows. Our observations are clearly compatible with the statement, but do they make the hypothesis "all cows are white" more likely?

To analyze this situation, we consider a probabilistic model. Let us assume that there are two possible states of the world, which we model as complementary events:

$$A : \text{all cows are white},$$
$$A^c : 50\% \text{ of all cows are white}.$$

Let p be the prior probability $\mathbf{P}(A)$ that all cows are white. We make an observation of a cow or a crow, with probability q and $1-q$, respectively, independently of whether event A occurs or not. Assume that $0 < p < 1$, $0 < q < 1$, and that all crows are black.

(a) Given the event $B = \{\text{a black crow was observed}\}$, what is $\mathbf{P}(A \,|\, B)$?

(b) Given the event $C = \{\text{a white cow was observed}\}$, what is $\mathbf{P}(A \,|\, C)$?

Solution. (a) We use the formula

$$\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A)\mathbf{P}(B \,|\, A)}{\mathbf{P}(B)}.$$

Since all crows are black, we have $\mathbf{P}(B) = 1 - q$. Furthermore, $\mathbf{P}(A) = p$. Finally, $\mathbf{P}(B \,|\, A) = 1 - q = \mathbf{P}(B)$, since the probability of observing a (black) crow is not affected by the truth of our hypothesis. We conclude that $\mathbf{P}(A \,|\, B) = \mathbf{P}(A) = p$. Thus, the new evidence, while compatible with the hypothesis "all cows are white," does not change our beliefs about its truth.

(b) Once more,

$$\mathbf{P}(A \,|\, C) = \frac{\mathbf{P}(A \cap C)}{\mathbf{P}(C)} = \frac{\mathbf{P}(A)\mathbf{P}(C \,|\, A)}{\mathbf{P}(C)}.$$

Given the event A, a cow is observed with probability q, and it must be white. Thus, $\mathbf{P}(C \,|\, A) = q$. Giben the event A^c, a cow is observed with probability q, and it is white with probability $1/2$. Thus, $\mathbf{P}(C \,|\, A^c) = q/2$. Using the total probability theorem,

$$\mathbf{P}(C) = \mathbf{P}(A)\mathbf{P}(C \,|\, A) + \mathbf{P}(A^c)\mathbf{P}(C \,|\, A^c) = pq + (1-p)\frac{q}{2}.$$

Hence,

$$\mathbf{P}(A \,|\, C) = \frac{pq}{pq + (1-p)\frac{q}{2}} = \frac{2p}{1+p} > p.$$

Thus, the observation of a white cow makes the hypothesis "all cows are white" more likely to be true.

SECTION 1.5. Independence

Problem 22. A hunter has two hunting dogs. One day, on the trail of some animal, the hunter comes to a place where the road diverges into two paths. He knows that each dog, independently of the other, will choose the correct path with probability p. The hunter decides to let each dog choose a path, and if they agree, take that one, and if they disagree, to randomly pick a path. Is his strategy better than just letting one of the two dogs decide on a path?

Solution. Consider the sample space for the hunter's strategy. The outcomes that lead to the correct path are:

(1) Both dogs agree on the correct path (probability p^2, by independence).

(2) The dogs disagree, dog 1 chooses right path, and hunter follows dog 1 [probability $p(1-p)/2$].

(3) The dogs disagree, dog 2 chooses right path, and hunter follows dog 2 [probability $p(1-p)/2$].

The above events are mutually exclusive, so we can add the probabilities to find that under the hunter's strategy, the probability that he chooses the correct path is

$$p^2 + \frac{1}{2}p(1-p) + \frac{1}{2}p(1-p) = p.$$

On the other hand, if the hunter lets one dog choose the path, this dog will also choose the correct path with probability p. Thus, the two strategies are equally effective.

Problem 23. Communication through a noisy channel. A binary (0 or 1) message transmitted through a noisy communication channel is received incorrectly with probability ϵ_0 and ϵ_1, respectively (see Fig. 1.18). Errors in different symbol transmissions are independent.

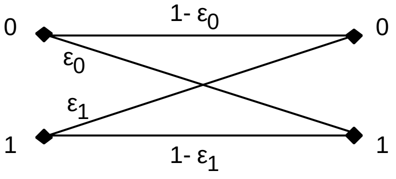


Figure 1.18: Error probabilities in a binary communication channel.

(a) Suppose that the channel source transmits a 0 with probability p and transmits a 1 with probability $1-p$. What is the probability that a randomly chosen symbol is received correctly?

(b) Suppose that the string of symbols 1011 is transmitted. What is the probability that all the symbols in the string are received correctly?

(c) In an effort to improve reliability, each symbol is transmitted three times and the received symbol is decoded by majority rule. In other words, a 0 (or 1) is transmitted as 000 (or 111, respectively), and it is decoded at the receiver as a 0 (or 1) if and only if the received three-symbol string contains at least two 0s (or 1s, respectively). What is the probability that a transmitted 0 is correctly decoded?

(d) Suppose that the channel source transmits a 0 with probability p and transmits a 1 with probability $1-p$, and that the scheme of part (c) is used. What is the probability that a 0 was transmitted given that the received string is 101?

Solution. (a) Let A be the event that a 0 is transmitted. Using the total probability theorem, the desired probability is

$$\mathbf{P}(A)(1-\epsilon_0)+\big(1-\mathbf{P}(A)\big)(1-\epsilon_1)=p(1-\epsilon_0)+(1-p)(1-\epsilon_1).$$

(b) By independence, the probability that the string 1011 is received correctly is

$$(1-\epsilon_0)(1-\epsilon_1)^3.$$

(c) In order for a 0 to be decoded correctly, the received string must be 000, 001, 010, or 100. Given that the string transmitted was 000, the probability of receiving 000 is $(1-\epsilon_0)^3$, and the probability of each of the strings 001, 010, and 100 is $\epsilon_0(1-\epsilon_0)^2$. Thus, the probability of correct decoding is

$$3\epsilon_0(1-\epsilon_0)^2+(1-\epsilon_0)^3.$$

(d) Using Bayes' rule, we have

$$\mathbf{P}(0\,|\,101)=\frac{\mathbf{P}(0)\mathbf{P}(101\,|\,0)}{\mathbf{P}(0)\mathbf{P}(101\,|\,0)+\mathbf{P}(1)\mathbf{P}(101\,|\,1)}.$$

The probabilities needed in the above formula are

$$\mathbf{P}(0)=p,\quad \mathbf{P}(1)=1-p,\quad \mathbf{P}(101\,|\,0)=\epsilon_0^2(1-\epsilon_0),\quad \mathbf{P}(101\,|\,1)=\epsilon_1(1-\epsilon_1)^2.$$

Problem 24. An electrical system consists of identical components that with probability p independently of other components. The components are connected as shown in Fig. 1.19. The system succeeds if there is a path that starts at point A, ends at point B, and consists of successful components. What is the probability of success?

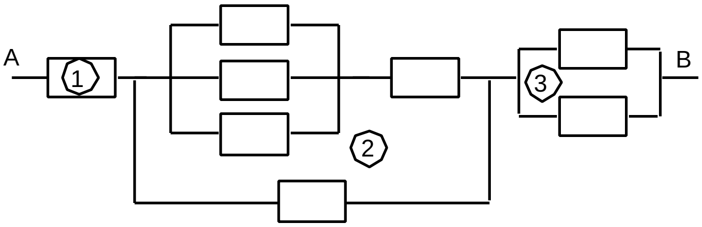


Figure 1.19: A system of identical components, which succeeds if there is a path that starts at point A, ends at point B, and consists of successful components.

Solution. The system may be viewed as a series connection of three subsystems, denoted 1, 2, and 3 in Fig. 1.19. The probability of success of the entire system is $p_1p_2p_3$, where

p_i is the probability of success of subsystem i. Using the formulas for the probability of success of a series or a parallel system given in Example 1.22, we have

$$p_1 = p, \qquad p_3 = 1 - (1-p)^2,$$

and

$$p_2 = 1 - (1-p)\bigl(1 - p\bigl(1 - (1-p)^3\bigr)\bigr).$$

Problem 25. Reliability of a k-out-of-n system. A system consists of n identical components that are reliable with probability p independently of other components. The system succeeds if at least k out of the n components work reliably. What is the probability of success?

Solution. Let A_i be the event that exactly i components work reliably. The probability of success is the union $\cup_{i=k}^n A_i$, and since the A_i are disjoint, it is equal to

$$\sum_{i=k}^{n} \mathbf{P}(A_i) = \sum_{i=k}^{n} p(i),$$

where $p(i)$ are the binomial probabilities. Thus, the probability of success is

$$\sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

Problem 26. A cellular phone system services a population of n_1 "voice users" (those that occasionally need a voice connection) and n_2 "data users" (those that occasionally need a data connection). We estimate that at a given time, each user will need to be connected to the system with probability p_1 (for voice users) or p_2 (for data users), independently of other users. The data rate for a voice user is r_1 bits/sec and for a data user is r_2 bits/sec. The cellular system has a total capacity of c bits/sec. What is the probability that more users want to use the system than the system can accommodate?

Solution. The probability that k_1 voice users and k_2 data users simultaneously need the facility is $p_1(k_1)p_2(k_2)$, where $p_1(k_1)$ and $p_2(k_2)$ are the corresponding binomial probabilities, given by

$$p_i(k_i) = \binom{n}{k_i} p_i^{k_i} (1-p_i)^{n-k_i}, \qquad i = 1, 2.$$

The probability that more users want to use the system than the system can accommodate is the sum of all products $p_1(k_1)p_2(k_2)$ as k_1 and k_2 range over all possible values whose total bit rate requirement $k_1r_1 + k_2r_2$ exceeds the capacity c of the system. Thus, the desired probability is

$$\sum_{\{(k_1,k_2)\,|\,k_1r_1+k_2r_2>c,\,k_1\le n_1,\,k_2\le n_2\}} p_1(k_1)p_2(k_2).$$

Problem 27. Imagine a drunk tightrope walker, who manages to keep his balance, and takes a forward step with probability p, or backward step with probability $(1-p)$, independently of what he did in previous steps.

(a) What is the probability that after two steps the tightrope walker will be at the same place on the rope?

(b) What is the probability that after three steps, he will be one step forward from where he began?

(c) Given that after three steps he has managed to move ahead one step, what is the probability that the first step he took was a step forward?

Solution. (a) In order to end up in the same place after two steps, the tightrope walker can make a forward step followed by a backward step, or vice versa. Therefore the required probability is $2p(1-p)$.

(b) The probability that after three steps he will be one step ahead of his starting point is the probability that there are exactly two forward steps out of a total of three steps. This corresponds to the event

$$A = \big\{(F,F,B),(F,B,F),(B,F,F)\big\}$$

where F and B stand for a forward or backward step, respectively. Each of the possible outcomes has probability $p^2(1-p)$, so

$$\mathbf{P}(A) = 3p^2(1-p).$$

(c) The conditioning event is the event A above. Two out of the three outcomes in A start with a forward step. Therefore, the desired conditional probability is $2/3$.

Problem 28. A power utility can supply electricity to a city from n different power plants. All n work quite efficiently together, but in fact any one by itself can produce enough electricity to supply the entire city. Now suppose that each one of these power plants fails independently of the others, and power plant i fails with probability p_i.

(a) What is the probability that the city will experience a black-out?

(b) Now suppose that the city increases its demand for electricity, so that two power plants are necessary to keep the city from a black-out. Find the probability that the city will experience a black-out.

Solution. (a) Let A denote the event that the city experiences a black-out. Since the power plants fail independently of each other, we know that

$$\mathbf{P}(A) = \prod_{i=1}^{n} p_i.$$

(b) If two power plants are necessary to keep the city from a black out, the city will black out if either all n or any $n-1$ power plants fail. These two events are mutually exclusive, so we can calculate the probability $\mathbf{P}(A)$ of a black-out by adding their probabilities:

$$\mathbf{P}(A) = \prod_{i=1}^{n} p_i + \sum_{i=1}^{n}\left((1-p_i)\prod_{j\neq i} p_j\right).$$

Here, $(1-p_i)\prod_{j\neq i} p_j$ is the probability that $n-1$ plants have failed and plant i is the one that has not failed.

Problem 29. **Independence/pairwise independence example.** Ina Rush is disappointed because she never gets all of the three traffic lights on her way to work to be green. Yet Ina is convinced that the three lights are independently red or green. Can Ina be right? If Ina is wrong, is it possible that the light colors she encounters are pairwise independent?

Solution. Let F, S, and T be the events that the first, second, and third lights are green, respectively. We know that $\mathbf{P}(F\cap S\cap T)=0$. If F, S, and T are independent, we must have $\mathbf{P}(F\cap S\cap T)=\mathbf{P}(F)\mathbf{P}(S)\mathbf{P}(T)$, so it follows that one of the three lights must always be red when Ina gets to it [$\mathbf{P}(F)=0$ or $\mathbf{P}(S)=0$ or $\mathbf{P}(T)=0$]. This may be possible. In particular, suppose that the first two lights stay green or red for 1 minute intervals, Ina takes half a minute to reach the second light from the first, and the second light becomes red half a minute after the first becomes green. For an example of a probability law under which the second light is always red but the three lights are independent, suppose that the light sequences where the second light is red,

$$(\text{R,R,R}),\quad (\text{G,R,G}),\quad (\text{R,R,G}),\quad (\text{G,R,R}),$$

all have probability 1/4.

Suppose that all the light sequences that Ina encounters involve either one or three red lights, and all these sequences are equally likely, i.e,

$$\mathbf{P}(\text{R,R,R})=\mathbf{P}(\text{R,G,G})=\mathbf{P}(\text{G,R,G})=\mathbf{P}(\text{G,G,R})=\frac{1}{4}.$$

Then it can be seen that F, S, and T have probability 1/2, and that the events $F\cap S$, $F\cap T$, $S\cap T$ have probabilities 1/4, so F, S, and T are pairwise independent. However, F, S, and T are not independent, since $\mathbf{P}(F\cap S\cap T)=0\neq 1/8=\mathbf{P}(F)\mathbf{P}(S)\mathbf{P}(T)$.

Problem 30. A company is interviewing potential employees. Suppose that each candidate is either qualified, or unqualified with given probabilities q and $1-q$, respectively. The company tries to determine this by asking 20 true or false questions. A candidate gets a C for each correct answer, and an I for each incorrect answer. A qualified candidate has probability p of answering the question correctly, while an unqualified candidate has a probability p of answering incorrectly. The answers to different questions are assumed to be independent. If the company considers anyone with at least 15 C's qualified, and everyone else unqualified, what is the probability that the 20 questions will correctly identify someone to be qualified or unqualified?

Solution. Let Q be the event that someone is qualified, Q^c the event that someone is unqualified, and A the event that the 20 questions correctly determine whether the candidate is qualified or not. Using the total probability theorem, we have

$$\begin{aligned}\mathbf{P}(A)&=\mathbf{P}(A\cap Q)+\mathbf{P}(A\cap Q^c)\\&=\mathbf{P}(Q)\cdot\mathbf{P}(A\,|\,Q)+\mathbf{P}(Q^c)\cdot\mathbf{P}(A\,|\,Q^c)\\&=q\sum_{i=15}^{20}\binom{20}{i}p^i(1-p)^{20-i}+(1-q)\sum_{i=6}^{20}\binom{20}{i}p^i(1-p)^{20-i}.\end{aligned}$$

Problem 31. The problem of points. Telis and Wendy play a round of golf (18 holes) for a \$10 stake, and their probabilities of winning on any one hole are p and $1 - p$, respectively, independently of their results in other holes. At the end of 10 holes, with the score 4 to 6 in favor of Wendy, Telis receives an urgent call and has to report back to work. They decide to split the stake in proportion to their probabilities of winning had they completed the round, as follows. If p_T and p_W are the conditional probabilities that Telis and Wendy, respectively, are ahead in the score after 18 holes given the 4-6 score after 10 holes, then Telis should get a fraction $p_T/(p_T + p_W)$ of the stake, and Wendy should get the remaining $p_W/(p_T + p_W)$. How much money should Telis get? *Note*: This is an example of the, so-called, problem of points, which played an important historical role in the development of probability theory. The problem was posed by Chevalier de Mere in the 17th century to Pascal, who introduced the idea that the stake of an interrupted game should be divided in proportion to the players' conditional probabilities of winning given the state of the game at the time of interruption. Pascal worked out some special cases and through a correspondence with Fermat, stimulated much thinking and several probability-related investigations.

Solution. We have

$$p_T = \mathbf{P}(\text{at least 6 out of the 8 remaining holes are won by Telis}),$$

$$p_W = \mathbf{P}(\text{at least 4 out of the 8 remaining holes are won by Wendy}).$$

Using the binomial formulas,

$$p_T = \sum_{k=6}^{8} \binom{8}{k} p^k (1-p)^{8-k} \qquad p_W = \sum_{k=4}^{8} \binom{8}{k} (1-p)^k p^{8-k}.$$

The amount of money that Telis should get is $10p_T/(p_T + p_W)$ dollars.

Problem 32. * Gambler's ruin. Two gamblers, G_1 and G_2, hold a sequence of independent betting rounds. In each round, G_1 wins with probability p, and G_2 wins with probability $q = 1 - p$. The winner of a round collects \$1 from the loser. Initially, G_1 has k dollars, and G_2 has $N - k$ dollars, and play continues until one of them has no money left. What is the probability that G_1 wins?

Solution. Let us denote by P_k the probability that G_1 wins all the money given that he starts with k and G_2 starts with $N - k$. Let us also denote by A the event that this happens, and by F the event that G_1 wins the first round. We apply the total probability theorem to obtain

$$\begin{aligned} P_k &= \mathbf{P}(A \,|\, F)\mathbf{P}(F) + \mathbf{P}(A \,|\, F^c)\mathbf{P}(F^c) \\ &= p\mathbf{P}(A \,|\, F) + q\mathbf{P}(A \,|\, F^c). \end{aligned}$$

By the independence of past and future events, $\mathbf{P}(A \,|\, F) = P_{k+1}$ and similarly $\mathbf{P}(A \,|\, F^c) = P_{k-1}$. Thus, we have $pP_k + qP_k = pP_{k+1} + qP_{k-1}$ for all k, or

$$P_{k+1} - P_k = r(P_k - P_{k-1}),$$

where $r = q/p$. We will solve for P_k in terms of p and q using iteration and the boundary values $P_0 = 0$ and $P_N = 1$. Since $P_0 = 0$, we obtain

$$P_k = \begin{cases} P_1 \dfrac{1-r^k}{1-r}, & \text{if } p \neq q, \\ kP_1, & \text{if } p = q. \end{cases}$$

Since $P_N = 1$, we can solve for P_1 and therefore for P_k:

$$P_1 = \begin{cases} \dfrac{1-r}{1-r^N}, & \text{if } p \neq q, \\ \dfrac{1}{N}, & \text{if } p = q, \end{cases}$$

so that

$$P_k = \begin{cases} \dfrac{1-r^k}{1-r^N}, & \text{if } p \neq q, \\ \dfrac{k}{N}, & \text{if } p = q. \end{cases}$$

Problem 33. * Consider a biased coin that comes up heads with probability p and tails with probability $1 - p$. Let q_n be the probability that after n independent tosses, there have been an even number of heads.

(a) Show that $q_n = p(1 - q_{n-1}) + (1 - p)q_{n-1}$.

(b) Show that $q_n = \big(1 + (1 - 2p)^n\big)/2$.

Solution. (a) Let A be the event that the first $n - 1$ tosses produce an even number of heads, and let E be the event that the nth toss is a head. We can obtain an even number of heads in n tosses in two distinct ways: 1) there is an even number of heads in the first $n - 1$ tosses, and the nth toss results in tails: this is the event $A \cap E^c$; 2) there is an odd number of heads in the first $n - 1$ tosses, and the nth toss results in heads: this is the event $A^c \cap E$. Using also the independence of A and E,

$$\begin{aligned} q_n &= \mathbf{P}\big((A \cap E^c) \cup (A^c \cap E)\big) \\ &= \mathbf{P}(A \cap E^c) + \mathbf{P}(A^c \cap E) \\ &= \mathbf{P}(A)\mathbf{P}(E^c) + \mathbf{P}(A^c)\mathbf{P}(E) \\ &= (1-p)q_{n-1} + p(1-q_{n-1}). \end{aligned}$$

(b) We will iterate the above expression, and use the property $q_0 = 1$ (with zero tosses, we have zero heads, which is even) to get the desired result:

$$\begin{aligned} q_n &= p(1-q_{n-1}) + (1-p)q_{n-1} \\ &= p + (1-2p)q_{n-1} \\ &= p + (1-2p)p + (1-2p)^2 q_{n-2} \\ &= \cdots \\ &= p\big(1 + (1-2p) + (1-2p)^2 + \cdots + (1-2p)^{n-1}\big) + (1-2p)^n. \end{aligned}$$

Assuming that $p \neq 0$, we can use the identity

$$1 + (1-2p) + (1-2p)^2 + (1-2p)^3 + \cdots + (1-2p)^{n-1} = \frac{1-(1-2p)^n}{1-(1-2p)}.$$

Therefore,

$$\begin{aligned} q_n &= p(1 + (1-2p) + (1-2p)^2 + \cdots + (1-2p)^{n-1}) + (1-2p)^n \\ &= p\frac{1-(1-2p)^n}{1-(1-2p)} + (1-2p)^n \\ &= \frac{1}{2} - \frac{(1-2p)^n}{2} + (1-2p)^n \\ &= \frac{1+(1-2p)^n}{2}. \end{aligned}$$

For the case $p = 0$, the number of heads will be zero, which is even, and $q_n = 1$, which agrees with the desired result.

Problem 34. * A particular class has had a history of low attendance. Fed up with the situation, the professor decides that she will not lecture unless at least k of the n students enrolled in the class are present. Each student will independently show up with probability p_g if the weather is good, and with probability p_b if the weather is bad. If the chance of bad weather tomorrow is $\mathbf{P}(B)$, what is the probability that the professor teaches her class?

Solution. Let the event A be the event that the professor teaches her class. We have

$$\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c) = \mathbf{P}(B)\mathbf{P}(A \,|\, B) + \mathbf{P}(B^c)\mathbf{P}(A \,|\, B^c).$$

We also have

$$\mathbf{P}(A \,|\, B) = \sum_{i=k}^{n} \binom{n}{i} \cdot p_b^i (1-p_b)^{n-i},$$

$$\mathbf{P}(A \,|\, B^c) = \sum_{i=k}^{n} \binom{n}{i} \cdot p_g^i (1-p_g)^{n-i}.$$

Therefore,

$$\mathbf{P}(A) = \mathbf{P}(B) \sum_{i=k}^{n} \binom{n}{i} \cdot p_b^i (1-p_b)^{n-i} + \big(1 - \mathbf{P}(B)\big) \sum_{i=k}^{n} \binom{n}{i} \cdot p_g^i (1-p_g)^{n-i}.$$

Problem 35. * Let A and B be independent events. Use the definition of independence to prove the following:

(a) A and B^c are independent.

(b) A^c and B^c are independent.

Solution. (a) The event A is the union of the disjoint events $A \cap B^c$ and $A \cap B$. Using the additivity axiom and the independence of A and B, we obtain

$$\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B) + \mathbf{P}(A \cap B^c).$$

It follows that $\mathbf{P}(A \cap B^c) = \mathbf{P}(A)\big(1 - \mathbf{P}(B)\big) = \mathbf{P}(A)\mathbf{P}(B^c)$.

(b) Apply the result of part (a) twice: first on A and B, then on B^c and A.

Problem 36. * Suppose that A, B, and C are independent. Use the mathematical definition of independence to show that A and $B \cup C$ are independent.

Solution. To prove independence, we need to show that $\mathbf{P}\big(A \cap (B \cup C)\big) = \mathbf{P}(A) \cdot \mathbf{P}(B \cup C)$. Using the identity $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, we obtain

$$\begin{aligned}\mathbf{P}\big(A \cap (B \cup C)\big) &= \mathbf{P}\big((A \cap B) \cup (A \cap C)\big)\\ &= \mathbf{P}(A \cap B) + \mathbf{P}(A \cap C) - \mathbf{P}\big((A \cap B) \cap (A \cap C)\big)\\ &= \mathbf{P}(A \cap B) + \mathbf{P}(A \cap C) - \mathbf{P}(A \cap B \cap C).\end{aligned}$$

The independence of A, B, and C implies that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, $\mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C)$, and $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$. It follows that

$$\begin{aligned}\mathbf{P}\big(A \cap (B \cup C)\big) &= \mathbf{P}(A)\mathbf{P}(B) + \mathbf{P}(A)\mathbf{P}(C) - \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)\\ &= \mathbf{P}(A)\big(\mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(B)\mathbf{P}(C)\big)\\ &= \mathbf{P}(A)\mathbf{P}(B \cup C).\end{aligned}$$

Problem 37. * Suppose that A, B, and C are independent. Prove formally that A and B are conditionally independent given C.

Solution. We need to show that $\mathbf{P}(A \cap B \,|\, C) = \mathbf{P}(A \,|\, C)\mathbf{P}(B \cap C)$. We have

$$\begin{aligned}\mathbf{P}(A \cap B \,|\, C) &= \frac{\mathbf{P}(A \cap B \cap C)}{\mathbf{P}(C)}\\ &= \frac{\mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)}{\mathbf{P}(C)}\\ &= \mathbf{P}(A)\mathbf{P}(B)\\ &= \mathbf{P}(A \,|\, C)\mathbf{P}(B \,|\, C).\end{aligned}$$

The first equality uses the definition of conditional probabilities; the second uses the assumed independence; the fourth uses the independence of A from C, and of B from C.

Problem 38. * Assume that the events A_1, A_2, A_3, A_4 are independent and that $\mathbf{P}(A_3 \cap A_4) > 0$. Show that

$$\mathbf{P}(A_1 \cup A_2 \,|\, A_3 \cap A_4) = \mathbf{P}(A_1 \cup A_2).$$

Solution. We have

$$\mathbf{P}(A_1 \,|\, A_3 \cap A_4) = \frac{\mathbf{P}(A_1 \cap A_3 \cap A_4)}{\mathbf{P}(A_3 \cap A_4)} = \frac{\mathbf{P}(A_1)\mathbf{P}(A_3)\mathbf{P}(A_4)}{\mathbf{P}(A_3)\mathbf{P}(A_4)} = \mathbf{P}(A_1).$$

One obtains similarly that $\mathbf{P}(A_2 \mid A_3 \cap A_4) = \mathbf{P}(A_2)$ and $\mathbf{P}(A_1 \cap A_2 \mid A_3 \cap A_4) = \mathbf{P}(A_1 \cap A_2)$, and, finally,

$$\begin{aligned}\mathbf{P}(A_1 \cup A_2 \mid A_3 \cap A_4) &= \mathbf{P}(A_1 \mid A_3 \cap A_4) + \mathbf{P}(A_2 \mid A_3 \cap A_4) \\ &\qquad - \mathbf{P}(A_1 \cap A_2 \mid A_3 \cap A_4) \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2) - \mathbf{P}(A_1 \cap A_2) \\ &= \mathbf{P}(A_1 \cup A_2).\end{aligned}$$

Problem 39. * Laplace's rule of succession. Consider $N+1$ boxes with the kth box containing k red balls and $N-k$ white balls, where k ranges from 0 to N. We choose a box at random (all boxes are equally likely) and then choose a ball at random from that box, n successive times (the ball drawn is replaced each time, and a new ball is selected independently). Suppose a red ball was drawn each of the n times. What is the probability that if we draw a ball one more time it will be red? Estimate this probability for large N.

Solution. We want to find the conditional probability $\mathbf{P}(E \mid R_n)$, where E is the event of a red ball drawn at time $n+1$, and R_n is the event of a red ball drawn each of the n preceding times. Intuitively, the consistent draw of a red ball indicates that a box with a high percentage of red balls was chosen, so we expect that $\mathbf{P}(E \mid R_n)$ is closer to 1 than to 0. In fact, Laplace used this example to calculate the probability that the sun will rise tomorrow given that it has risen for the preceding 5,000 years ($n = 1,826,213$ days). (It is not clear how serious Laplace was about this calculation, but the story is part of the folklore of probability theory.)

We have

$$\mathbf{P}(E \mid R_n) = \frac{\mathbf{P}(E \cap R_n)}{\mathbf{P}(R_n)},$$

where using the total probability theorem, we have

$$\begin{aligned}\mathbf{P}(R_n) &= \sum_{k=0}^{N} \mathbf{P}(k\text{th box chosen}) \left(\frac{k}{N}\right)^n = \frac{1}{N+1} \sum_{k=0}^{N} \left(\frac{k}{N}\right)^n, \\ \mathbf{P}(E \cap R_n) &= \mathbf{P}(R_{n+1}) = \frac{1}{N+1} \sum_{k=0}^{N} \left(\frac{k}{N}\right)^{n+1}.\end{aligned}$$

For large N, we can view $\mathbf{P}(R_n)$ as a piecewise constant approximation to an integral:

$$\mathbf{P}(R_n) = \frac{1}{N+1} \sum_{k=0}^{N} \left(\frac{k}{N}\right)^n \approx \frac{1}{(N+1)N^n} \int_0^N x^n dx = \frac{1}{(N+1)N^n} \cdot \frac{N^{n+1}}{n+1} \approx \frac{1}{n+1}.$$

Similarly,

$$\mathbf{P}(E \cap R_n) = \mathbf{P}(R_{n+1}) \approx \frac{1}{n+2},$$

so that

$$\mathbf{P}(E \mid R_n) \approx \frac{n+1}{n+2}.$$

Thus, for large N, drawing a red ball one more time is almost certain.

Problem 40. * Show the formula for the binomial coefficients

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

and also the binomial formula by using the, so called, **Pascal triangle**, given in Fig. 1.21.

Solution. Note that n-toss sequences that contain k heads can be obtained in two ways:

(1) By starting with an $n-1$-toss sequence that contains k heads and adding a tail at the end. There are $\binom{n-1}{k}$ different sequences of this type.

(2) By starting with an $n-1$-toss sequence that contains $k-1$ heads and adding a head at the end. There are $\binom{n-1}{k-1}$ different sequences of this type.

Thus,

$$\binom{n}{k} = \begin{cases} \binom{n-1}{k-1} + \binom{n-1}{k} & \text{if } k = 1, 2, \ldots, n-1, \\ 1 & \text{if } k = 0, n. \end{cases}$$

We now use the above relation to demonstrate the formula

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

by induction on n. Indeed, we have from the definition $\binom{1}{0} = \binom{1}{1} = 1$, so for $n = 1$ the above formula is seen to hold as long as we use the convention $0! = 1$. If the formula holds for each index up to $n-1$, we have for $k = 1, 2, \ldots, n-1$,

$$\begin{aligned} \binom{n}{k} &= \binom{n-1}{k-1} + \binom{n-1}{k} \\ &= \frac{(n-1)!}{(k-1)!(n-1-k+1)!} + \frac{(n-1)!}{k!(n-1-k)!} \\ &= \frac{k}{n} \cdot \frac{n!}{k!(n-k)!} + \frac{n-k}{n} \cdot \frac{n!}{k!(n-k)!} \\ &= \frac{n!}{k!(n-k)!}, \end{aligned}$$

and the induction is complete. Since the binomial probabilities $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ must add to 1, we have the binomial formula

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

$$\begin{array}{ccccccccc}
 & & & & \binom{0}{0} & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & \\
 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & \\
\binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} \\
 & & & & \cdots\cdots & & & &
\end{array}
\qquad\qquad
\begin{array}{ccccccccc}
 & & & & 1 & & & & \\
 & & & 1 & & 1 & & & \\
 & & 1 & & 2 & & 1 & & \\
 & 1 & & 3 & & 3 & & 1 & \\
1 & & 4 & & 6 & & 4 & & 1 \\
 & & & & \cdots\cdots & & & &
\end{array}$$

Figure 1.21: Sequential calculation method of the binomial coefficients using the Pascal triangle. Each term $\binom{n}{k}$ in the triangular array on the left is computed and placed in the triangular array on the right by adding its two neighbors in the row above it (except for the boundary terms with $k = 0$ or $k = n$, which are equal to 1).

SECTION 1.6. Counting

Problem 41. The birthday problem. Consider n people who are attending a party. We assume that every person has an equal probability of being born on any day during the year, independently of everyone else, and ignore the additional complication presented by leap years (i.e., nobody is born on February 29). What is the probability that each person has a distinct birthday?

Solution. The sample space consists of all possible choices for the birthday of each person. Since there are n persons, and each has 365 choices for their birthday, the sample space has 365^n elements. Let us now consider those choices of birthdays for which no two persons have the same birthday. Assuming that $n \leq 365$, there are 365 choices for the first person, 364 for the second, etc., for a total of $365 \cdot 364 \cdot (365 - n + 1)$. Thus,

$$\mathbf{P}(\text{no two birthdays coincide}) = \frac{365 \cdot 364 \cdot (365 - n + 1)}{365^n}.$$

It is interesting to note that for n as small as 23, the probability that there are two persons with the same birthday is larger than 1/2.

Problem 42. We draw the top 7 cards from a well-shuffled standard 52-card deck. Find the probability that:

(a) The 7 cards include exactly 3 aces.

(b) The 7 cards include exactly 2 kings.

(c) The probability that the 7 cards include exactly 3 aces or exactly 2 kings or both.

Solution. (a) The sample space consists of all ways of drawing 7 elements out of a 52-element set, and its cardinality is $\binom{52}{7}$. Let us count those outcomes that involve exactly 3 aces. We are free to select any 3 out of the 4 aces, and any 4 out of the 48 remaining cards, for a total of $\binom{4}{3}\binom{48}{4}$ choices. Thus,

$$\mathbf{P}(7 \text{ cards include exactly 3 aces}) = \frac{\binom{4}{3}\binom{48}{4}}{\binom{52}{7}}.$$

(b) Proceeding the same way as in part (a), we obtain

$$\mathbf{P}(7 \text{ cards include exactly 2 kings}) = \frac{\binom{4}{2}\binom{48}{5}}{\binom{52}{7}}.$$

(c) If A and B stand for the events in parts (a) and (b), respectively, we are looking for $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$. The event $A \cap B$ (having exactly 3 aces and exactly 2 kings) can materialize by choosing 3 out of the 4 available aces, 2 out of the 4 available kings, and 2 more cards out of the remaining 44. Thus, this event consists of $\binom{4}{3}\binom{4}{2}\binom{44}{2}$ distinct outcomes. Hence,

$$\mathbf{P}(7 \text{ cards include 3 aces and/or 2 kings}) = \frac{\binom{4}{3}\binom{48}{4} + \binom{4}{2}\binom{48}{5} - \binom{4}{3}\binom{4}{2}\binom{44}{2}}{\binom{52}{7}}.$$

Problem 43. A well-shuffled standard 52-card deck is dealt to 4 players. Find the probability that each of the players gets an ace.

Solution. The size of the sample space is the number of different ways that 52 objects can be divided in 4 groups of 13, and is given by the multinomial formula

$$\frac{52!}{13!\,13!\,13!\,13!}.$$

There are 4! different ways of distributing the 4 aces to the players, and there are

$$\frac{48!}{12!\,12!\,12!\,12!}$$

different ways of dividing the remaining 48 cards into 4 groups of 12. Thus, the desired probability is

$$\frac{4!\dfrac{48!}{12!\,12!\,12!\,12!}}{\dfrac{52!}{13!\,13!\,13!\,13!}}.$$

A second solution can be obtained by considering a different, but probabilistically equivalent of dealing the cards. Each player has 13 slots, each one of which is to receive one card. Instead of shuffling the deck, we place the 4 aces at the top, and start dealing the cards one at a time, with each free slot being equally likely to receive the next card. For the event of interest to materialize, the first ace can go anywhere; the second can go to any one of the 39 slots (out of the 51 available) that correspond to players that do not yet have an ace; the third can go to any one of the 26 slots (out of the 50 available) that correspond to the two players that do not yet have an ace; and finally, the fourth, can go to any one of the 13 slots (out of the 49 available) that correspond to the only player who does not yet have an ace. Thus, the desired probability is

$$\frac{39 \cdot 26 \cdot 13}{51 \cdot 50 \cdot 49}.$$

By simplifying our previous answer, it can be checked that it is the same as the one obtained here, thus corroborating the intuitive fact that the two different ways of dealing the cards are equivalent.

Problem 44. Twenty distinct cars park in the same parking lot every day. Ten of these cars are US-made, while the other ten are foreign-made. The parking lot has exactly twenty spaces, all in a row, so the cars park side by side. However, the drivers have varying schedules, so the position any car might take on a certain day is random.

(a) In how many different ways can the cars line up?

(b) What is the probability that on a given day, the cars will park in such a way that they alternate (no two US-made are adjacent and no two foreign-made are adjacent)?

Solution. (a) Since the cars are all distinct, there are 20! ways to line them up.

(b) To find the probability that the cars will be parked so that they alternate, we count the number of "favorable" outcomes, and divide by the total number of possible outcomes found in part (a). We count in the following manner. We first arrange the US cars in an ordered sequence (permutation). We can do this in 10! ways, since there are 10 distinct cars. Similarly, arrange the foreign cars in an ordered sequence, which can also be done in 10! ways. Finally, interleave the two sequences. This can be done in two different ways, since we can let the first car be either US-made or foreign. Thus, we have a total of $2 \cdot 10! \cdot 10!$ possibilities, and the desired probability is

$$\frac{2 \cdot 10! \cdot 10!}{20!}.$$

Note that we could have solved the second part of the problem by neglecting the fact that the cars are distinct. Suppose the foreign cars are indistinguishable, and also that the US cars are indistinguishable. Out of the 20 available spaces, we need to choose 10 spaces in which to place the US cars, and thus there are $\binom{20}{10}$ possible outcomes. Out of these outcomes, there are only two in which the cars alternate, depending on whether we start with a US or a foreign car. Thus, the desired probability is $2\binom{20}{10}$, which coincides with our earlier answer.

Problem 45. We deal from a well-shuffled 52-card deck.

(a) What is the probability that the 13th card dealt is a king?

(b) What is the probability that the 13th card is the first king to be dealt?

Solution. (a) Since we are given no information on the first 12 cards that are dealt, the probability that the 13th card dealt is a king is the same as the probability that the first card dealt is a king, and equals 4/52.

(b) The probability that the 13th card is the first king to be dealt is the probability that out of the first 13 cards to be dealt, exactly one was a king, and that the king was dealt last. Now, given that exactly one king was dealt in the first 13 cards, the probability that the King was dealt last is just $\frac{1}{13}$ since each place in line is equally likely. Thus, it remains to calculate the probability that there was exactly one King in the first 13 cards dealt. To calculate this probability we count the "favorable" outcomes and divide by the total number of possible outcomes. We first count the favorable outcomes, namely those with exactly one King in the first 13 cards dealt. We can choose a particular King in 4 ways, and we can choose the other 12 cards in $\binom{48}{12}$ ways, therefore there are $4 \cdot \binom{48}{12}$ favorable outcomes. There are $\binom{52}{13}$ total outcomes, so the desired probability is

$$\frac{1}{13} \cdot \frac{4 \cdot \binom{48}{12}}{\binom{52}{13}}.$$

For an alternative solution, we argue as in Example 1.10. The probability that the first card is not a king is 48/52. Given that, the probability that the second is not a king is 47/51. We continue similarly until the 12th card. The probability that the 12th card is not a king, given that none of the preceding 11 was a king is 37/41. (There are $52-11 = 41$ cards left, and $48-11 = 37$ of them are not kings.) Finally, the conditional probability that the 13th card is a king is 4/40. The desired probability is

$$\frac{48 \cdot 47 \cdots 37 \cdot 4}{52 \cdot 51 \cdots 41 \cdot 40}.$$

Problem 46. An urn contains m red and n white balls.

(a) We draw two balls simultaneously and at random. Describe the sample space and calculate the probability that the selected balls are of different color, by using two approaches: a counting approach based on the discrete uniform law, and a sequential approach based on the multiplication rule.

(b) A fair 3-sided die is rolled and if k comes up, where $k = 1, 2, 3$, we remove k balls from the urn at random and put them aside. Describe the sample space and calculate the probability that all of the balls drawn are red, using a divide-and-conquer approach and the total probability theorem.

Solution. (a) Although we pick the balls simultaneously, we can still reason as if we are picking them sequentially. One possible sample space is obtained by numbering the red balls from 1 to m, the white balls from $m + 1$ to $m + n$, and by letting the possible outcomes be all ordered pairs of integers (i, j) with $1 \leq i, j \leq m+n$ and $i \neq j$. All outcomes are equally likely, so we can use the counting method. The total number of possible outcomes is $(m + n)(m + n - 1)$. The number of outcomes corresponding to red-white selection, (i.e., $i \in \{1, \ldots, m\}$ and $j \in \{m + 1, \ldots, m + n\}$) is mn. The number of outcomes corresponding to white-red selection, (i.e., $i \in \{m + 1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$) is also mn. Thus, the desired probability that the balls are of different color is

$$\frac{2mn}{(m + n)(m + n - 1)}.$$

Another possible sample space consists of all the possible ordered color pairs, i.e., $\{RR, RW, WR, WW\}$. We then have to calculate the probability of the event $\{RW, WR\}$. We consider a sequential description of the experiment, i.e., we first select the first ball and then the second. In the first stage, the probability of a red ball is $m/(m+n)$. In the second stage, the probability of a red ball is either $m/(m+n-1)$ or $(m-1)/(m+n-1)$ depending on whether the first ball was blue or red, respectively. Therefore, using the multiplication rule, we have

$$\mathbf{P}(RR) = \frac{m}{m+n} \cdot \frac{m-1}{m-1+n}, \qquad \mathbf{P}(RW) = \frac{m}{m+n} \cdot \frac{n}{m-1+n},$$

$$\mathbf{P}(WR) = \frac{n}{m+n} \cdot \frac{m}{m+n-1}, \qquad \mathbf{P}(WW) = \frac{n}{m+n} \cdot \frac{n-1}{m+n-1}.$$

The desired probability is

$$\begin{aligned}\mathbf{P}\big(\{RW, WR\}\big) &= \mathbf{P}(RW) + \mathbf{P}(WR) \\ &= \frac{m}{m+n} \cdot \frac{n}{m-1+n} + \frac{n}{m+n} \cdot \frac{m}{m+n-1} \\ &= \frac{2mn}{(m+n)(m+n-1)}.\end{aligned}$$

(b) We calculate the conditional probability of all balls being red, given any of the possible values of k. We have $\mathbf{P}(R \,|\, k=1) = m/(m+n)$ and, as found in part (a), $\mathbf{P}(RR \,|\, k=2) = m(m-1)/(m+n)(m-1+n)$. Arguing sequentially as in part (a), we also have $\mathbf{P}(RRR \,|\, k=3) = m(m-1)(m-2)/(m+n)(m-1+n)(m-2+n)$. According to the total probability theorem, the desired answer is

$$\frac{1}{3}\left(\frac{m}{m+n} + \frac{m(m-1)}{(m+n)(m-1+n)} + \frac{m(m-1)(m-2)}{(m+n)(m-1+n)(m-2+n)}\right).$$

Problem 47. Eight rooks are randomly placed in distinct squares of a 8×8 chessboard. Find the probability that all the rooks are safe from one another, i.e., that there is no row or column with more than one rook.

Solution. We count the number of ways in which we can safely place 8 distinguishable rooks, and then divide this by the total number of possibilities. First we count the number of favorable positions for the rooks. We will place the rooks one by one on the 8×8 chessboard. For the first rook, there are no constraints, so we have 64 choices. Placing this rook, however, eliminates one row and one column. Thus, for our second rook, we can imagine that the illegal column and row have been removed, thus leaving us with a 7×7 chessboard, and thus with 49 choices. Similarly, for the third rook we have 36 choices, for the fourth 25, etc.

In the absence of any restrictions, there are $64 \cdot 63 \cdots 57$ ways we can place 8 rooks, so the desired probability is

$$\frac{64 \cdot 49 \cdot 36 \cdot 25 \cdot 16 \cdot 9 \cdot 4}{\frac{64!}{56!}}.$$

Problem 48. De Méré's Paradox. A six-sided die is rolled three times independently. Which is more likely: a sum of 11 or a sum of 12? (This question was posed by the French nobleman De Méré to his friend Pascal in the 17th century.)

Solution. A sum of 11 is obtained with the following 6 combinations:

$$(6,4,1)\ (6,3,2)\ (5,5,1)\ (5,4,2)\ (5,3,3)\ (4,4,3).$$

A sum of 12 is obtained with the following 6 combinations:

$$(6,5,1)\ (6,4,2)\ (6,3,3)\ (5,5,2)\ (5,4,3)\ (4,4,4).$$

Each combination of 3 distinct numbers corresponds to 6 permutations, while each combination of 3 numbers, two of which are equal corresponds to 3 permutations. Counting the number of permutations in the 6 combinations corresponding to a sum of 11, we obtain $6+6+3+6+3+3=27$ permutations. Counting the number of permutations in the 6 combinations corresponding to a sum of 12, we obtain $6+6+3+3+6+1=25$ permutations. Since all permutations are equally likely, a sum of 11 is more likely than a sum of 12.

Note also that the sample space has $6^3=216$ elements, so we have $\mathbf{P}(11)=27/216$, $\mathbf{P}(12)=25/216$

Problem 49. A club consists of a club leader and a number (possibly zero) of additional club members.

(a) Any subset of n given persons may decide to form a club. Explain why the number of possible clubs is $n2^{n-1}$.

(b) Find a different way of counting the number of possible clubs and establish that

$$\sum_{k=1}^{n} k\binom{n}{k} = n2^{n-1}.$$

Solution. (a) There are n choices for the club leader. Once the leader is chosen, we are left with a set of $n-1$ available persons, and we are free to choose any of the 2^{n-1} subsets.

(b) We can form a k-person club by first selecting k out of the n available persons (there are $\binom{n}{k}$ choices), and then selecting one of the members to be a leader (there are k choices). Thus, there is a total of $k\binom{n}{k}$ k-person clubs. We then sum over all k to obtain the number of possible clubs of any size.

Problem 50. * Correcting the number of permutations for indistinguishable objects. When permuting n objects some of which are indistinguishable, different permutations may lead to the same object sequence, so the number of distinguishable ordered sequences is less than $n!$. For example there are six permutations of the letters A, B, and C:

ABC, ACB, BAC, BCA, CAB, CBA,

but only three distinguishable permutations of the letters A, D, and D:

ADD, DAD, DDA.

(a) Suppose that k out of the n objects are indistinguishable. Show that the number of distinguishable ordered sequences is $n!/k!$.

(b) Suppose that we have r types of indistinguishable objects, and for each i, k_i objects of type i. Show that the number of distinguishable ordered sequences is

$$\frac{n!}{k_1!\,k_2!\cdots k_r!}.$$

Solution. (a) Each one of the $n!$ permutations corresponds to $k!$ duplicates which are obtained by permuting the k indistinguishable objects. Thus, the $n!$ permutations can be grouped into $n!/k!$ groups of $k!$ indistinguishable permutations each, and the number of distinguishable ordered sequences is $n!/k!$. For example the three letters A, D, and D give the $3! = 6$ permutations

ADD, ADD, DAD, DDA, DAD, DDA,

obtained by replacing B and C by D in the permutations of A, B, and C given earlier. However, these 6 permutations can be divided into the $n!/k! = 3!/2! = 3$ groups

$$\{ADD, ADD\},\ \{DAD, DAD\},\ \{DDA, DDA\},$$

each having $k! = 2! = 2$ indistinguishable permutations.

(b) One solution is to extend the argument in (a) above: for each object type i, there are $k_i!$ indistinguishable permutations of the k_i objects. Hence, for any given ordered sequence, there are $k_1!k_2!\cdots k_r!$ indistinguishable permutations.

An alternative argument goes as follows. Choosing an ordered sequence is the same as starting with n slots and for each i, choosing the k_i slots to be occupied by objects of type i. This is the same as partitioning the set $\{1, \ldots, n\}$ into groups of size $k_1, \ldots, k_r$ and the number of such partitions is given by the multinomial coefficient.

Problem 51. * Hypergeometric probabilities. An urn contains n balls, out of which m are red. We select k of the balls at random, without replacement (i.e., selected balls are not put back into the urn before the next selection). What is the probability that i of the selected balls are red? It is assumed that $i \le k \le n$.

Solution. The sample space consists of the $\binom{n}{k}$ different ways that we can select k out of the available balls. For the event of interest to materialize, we have to select i out of the m red balls, which can be done in $\binom{m}{i}$ ways, and also select $k - i$ out of the $n - m$ blue balls, which can be done in $\binom{n-m}{k-i}$ ways. Therefore, the desired probability is

$$\frac{\dbinom{m}{i}\dbinom{n-m}{k-i}}{\dbinom{n}{k}}.$$

2

Discrete Random Variables

Contents

2.1 BASIC CONCEPTS

In many probabilistic models, the outcomes are of a numerical nature, e.g., if they correspond to instrument readings or stock prices. In other experiments, the outcomes are not numerical, but they may be associated with some numerical values of interest. For example, if the experiment is the selection of students from a given population, we may wish to consider their grade point average. When dealing with such numerical values, it is often useful to assign probabilities to them. This is done through the notion of a **random variable**, the focus of the present chapter.

Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome; see Fig. 2.1. We refer to this number as the **numerical value** or the **experimental value** of the random variable. Mathematically, **a random variable is a real-valued function of the experimental outcome**.

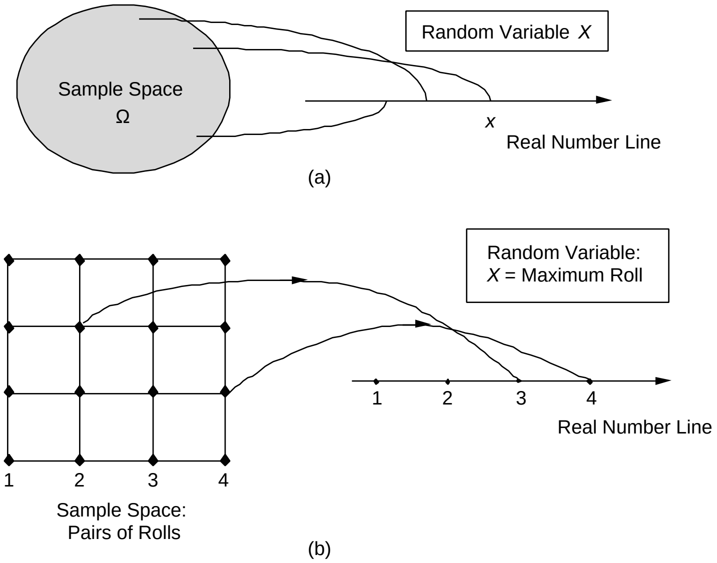


Figure 2.1: (a) Visualization of a random variable. It is a function that assigns a numerical value to each possible outcome of the experiment. (b) An example of a random variable. The experiment consists of two rolls of a 4-sided die, and the random variable is the maximum of the two rolls. If the outcome of the experiment is $(4, 2)$, the experimental value of this random variable is 4.

Here are some examples of random variables:

(a) In an experiment involving a sequence of 5 tosses of a coin, the number of heads in the sequence is a random variable. However, the 5-long sequence of heads and tails is not considered a random variable because it does not have an explicit numerical value.

(b) In an experiment involving two rolls of a die, the following are examples of random variables:

(1) The sum of the two rolls.

(2) The number of sixes in the two rolls.

(3) The second roll raised to the fifth power.

(c) In an experiment involving the transmission of a message, the time needed to transmit the message, the number of symbols received in error, and the delay with which the message is received are all random variables.

There are several basic concepts associated with random variables, which are summarized below.

Main Concepts Related to Random Variables

Starting with a probabilistic model of an experiment:

- A **random variable** is a real-valued function of the outcome of the experiment.
- A **function of a random variable** defines another random variable.
- We can associate with each random variable certain "averages" of interest, such the **mean** and the **variance**.
- A random variable can be **conditioned** on an event or on another random variable.
- There is a notion of **independence** of a random variable from an event or from another random variable.

A random variable is called **discrete** if its **range** (the set of values that it can take) is finite or at most countably infinite. For example, the random variables mentioned in (a) and (b) above can take at most a finite number of numerical values, and are therefore discrete.

A random variable that can take an uncountably infinite number of values is not discrete. For an example, consider the experiment of choosing a point a from the interval $[-1, 1]$. The random variable that associates the numerical value a^2 to the outcome a is not discrete. On the other hand, the random variable that associates with a the numerical value

$$\mathrm{sgn}(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a < 0, \end{cases}$$

is discrete.

In this chapter, we focus exclusively on discrete random variables, even though we will typically omit the qualifier "discrete."

Concepts Related to Discrete Random Variables

Starting with a probabilistic model of an experiment:

- A **discrete random variable** is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values.
- A (discrete) random variable has an associated **probability mass function** (PMF), which gives the probability of each numerical value that the random variable can take.
- A **function of a random variable** defines another random variable, whose PMF can be obtained from the PMF of the original random variable.

We will discuss each of the above concepts and the associated methodology in the following sections. In addition, we will provide examples of some important and frequently encountered random variables. In Chapter 3, we will discuss general (not necessarily discrete) random variables.

Even though this chapter may appear to be covering a lot of new ground, this is not really the case. The general line of development is to simply take the concepts from Chapter 1 (probabilities, conditioning, independence, etc.) and apply them to random variables rather than events, together with some appropriate new notation. The only genuinely new concepts relate to means and variances.

2.2 PROBABILITY MASS FUNCTIONS

The most important way to characterize a random variable is through the probabilities of the values that it can take. For a discrete random variable X, these are captured by the **probability mass function** (PMF for short) of X, denoted p_X. In particular, if x is any possible value of X, the **probability mass** of x, denoted $p_X(x)$, is the probability of the event $\{X = x\}$ consisting of all outcomes that give rise to a value of X equal to x:

$$p_X(x) = \mathbf{P}\big(\{X = x\}\big).$$

For example, let the experiment consist of two independent tosses of a fair coin, and let X be the number of heads obtained. Then the PMF of X is

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } x = 2, \\ 1/2 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

In what follows, we will often omit the braces from the event/set notation, when no ambiguity can arise. In particular, we will usually write $\mathbf{P}(X = x)$ in place of the more correct notation $\mathbf{P}\big(\{X = x\}\big)$. We will also adhere to the following convention throughout: **we will use upper case characters to denote random variables, and lower case characters to denote real numbers such as the numerical values of a random variable.**

Note that

$$\sum_x p_X(x) = 1,$$

where in the summation above, x ranges over all the possible numerical values of X. This follows from the additivity and normalization axioms, because the events $\{X = x\}$ are disjoint and form a partition of the sample space, as x ranges over all possible values of X. By a similar argument, for any set S of real numbers, we also have

$$\mathbf{P}(X \in S) = \sum_{x \in S} p_X(x).$$

For example, if X is the number of heads obtained in two independent tosses of a fair coin, as above, the probability of at least one head is

$$\mathbf{P}(X > 0) = \sum_{x>0} p_X(x) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

Calculating the PMF of X is conceptually straightforward, and is illustrated in Fig. 2.2.

Calculation of the PMF of a Random Variable X

For each possible value x of X:

1. Collect all the possible outcomes that give rise to the event $\{X = x\}$.
2. Add their probabilities to obtain $p_X(x)$.

The Bernoulli Random Variable

Consider the toss of a biased coin, which comes up a head with probability p, and a tail with probability $1 - p$. The **Bernoulli** random variable takes the two values 1 and 0, depending on whether the outcome is a head or a tail:

$$X = \begin{cases} 1 & \text{if a head,} \\ 0 & \text{if a tail.} \end{cases}$$

Its PMF is

$$p_X(x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

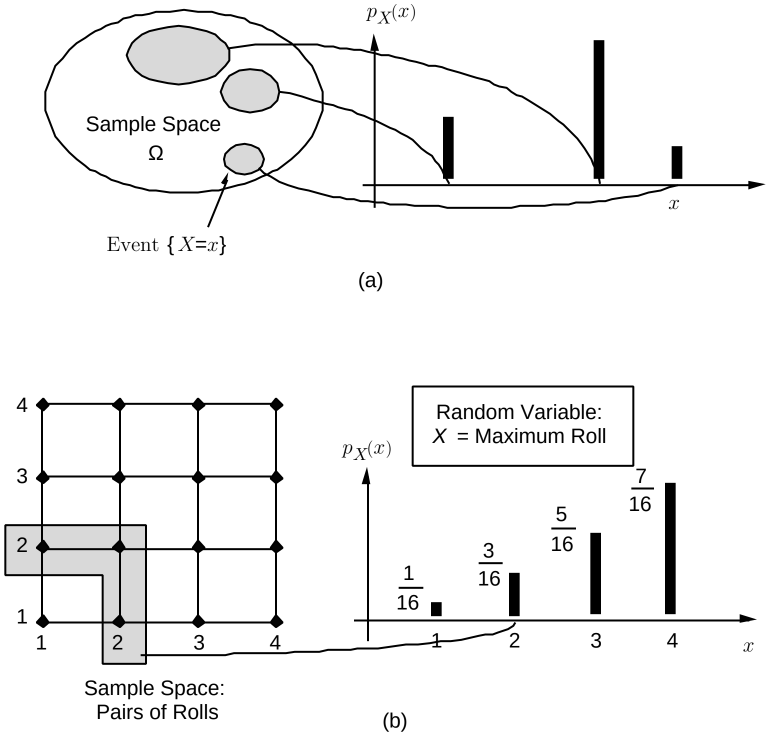


Figure 2.2: (a) Illustration of the method to calculate the PMF of a random variable X. For each possible value x, we collect all the outcomes that give rise to $X = x$ and add their probabilities to obtain $p_X(x)$. (b) Calculation of the PMF p_X of the random variable X = maximum roll in two independent rolls of a fair 4-sided die. There are four possible values x, namely, 1, 2, 3, 4. To calculate $p_X(x)$ for a given x, we add the probabilities of the outcomes that give rise to x. For example, there are three outcomes that give rise to $x = 2$, namely, $(1, 2), (2, 2), (2, 1)$. Each of these outcomes has probability 1/16, so $p_X(2) = 3/16$, as indicated in the figure.

For all its simplicity, the Bernoulli random variable is very important. In practice, it is used to model generic probabilistic situations with just two outcomes, such as:

(a) The state of a telephone at a given time that can be either free or busy.

(b) A person who can be either healthy or sick with a certain disease.

(c) The preference of a person who can be either for or against a certain political candidate.

Furthermore, by combining multiple Bernoulli random variables, one can construct more complicated random variables.

The Binomial Random Variable

A biased coin is tossed n times. At each toss, the coin comes up a head with probability p, and a tail with probability $1-p$, independently of prior tosses. Let X be the number of heads in the n-toss sequence. We refer to X as a **binomial** random variable **with parameters** n **and** p. The PMF of X consists of the binomial probabilities that were calculated in Section 1.4:

$$p_X(k) = \mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

(Note that here and elsewhere, we simplify notation and use k, instead of x, to denote the experimental values of integer-valued random variables.) The normalization property $\sum_x p_X(x) = 1$, specialized to the binomial random variable, is written as

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Some special cases of the binomial PMF are sketched in Fig. 2.3.

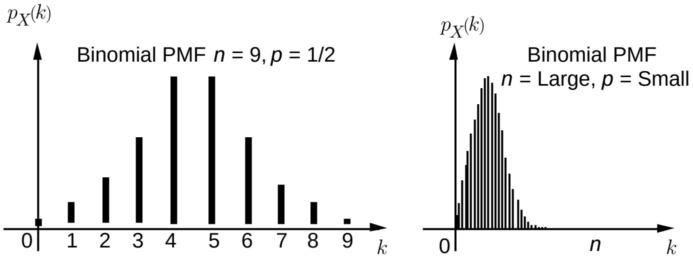


Figure 2.3: The PMF of a binomial random variable. If $p = 1/2$, the PMF is symmetric around $n/2$. Otherwise, the PMF is skewed towards 0 if $p < 1/2$, and towards n if $p > 1/2$.

The Geometric Random Variable

Suppose that we repeatedly and independently toss a biased coin with probability of a head p, where $0 < p < 1$. The **geometric** random variable is the number X of tosses needed for a head to come up for the first time. Its PMF is given by

$$p_X(k) = (1-p)^{k-1} p, \qquad k = 1, 2, \ldots,$$

since $(1-p)^{k-1}p$ is the probability of the sequence consisting of $k-1$ successive tails followed by a head; see Fig. 2.4. This is a legitimate PMF because

$$\sum_{k=1}^{\infty} p_X(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{k=0}^{\infty} (1-p)^k = p \cdot \frac{1}{1-(1-p)} = 1.$$

Naturally, the use of coin tosses here is just to provide insight. More generally, we can interpret the geometric random variable in terms of repeated independent trials until the first "success." Each trial has probability of success p and the number of trials until (and including) the first success is modeled by the geometric random variable.

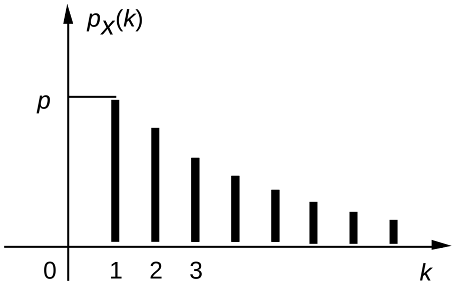


Figure 2.4: The PMF

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots,$$

of a geometric random variable. It decreases as a geometric progression with parameter $1 - p$.

The Poisson Random Variable

A Poisson random variable takes nonnegative integer values. Its PMF is given by

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots,$$

where λ is a positive parameter characterizing the PMF, see Fig. 2.5. It is a legitimate PMF because

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right) = e^{-\lambda}e^{\lambda} = 1.$$

To get a feel for the Poisson random variable, think of a binomial random variable with very small p and very large n. For example, consider the number of typos in a book with a total of n words, when the probability p that any one word is misspelled is very small (associate a word with a coin toss which comes a head when the word is misspelled), or the number of cars involved in accidents in a city on a given day (associate a car with a coin toss which comes a head when the car has an accident). Such a random variable can be well-modeled as a Poisson random variable.

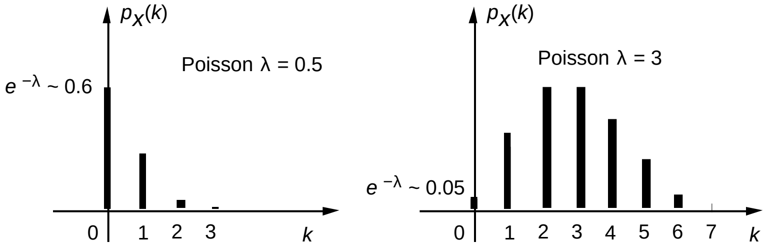


Figure 2.5: The PMF $e^{-\lambda}\frac{\lambda^k}{k!}$ of the Poisson random variable for different values of λ. Note that if $\lambda < 1$, then the PMF is monotonically decreasing, while if $\lambda > 1$, the PMF first increases and then decreases as the value of k increases (this is shown in the end-of-chapter problems).

More precisely, the Poisson PMF with parameter λ is a good approximation for a binomial PMF with parameters n and p, provided $\lambda = np$, n is very large, and p is very small, i.e.,

$$e^{-\lambda}\frac{\lambda^k}{k!} \approx \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

In this case, using the Poisson PMF may result in simpler models and calculations. For example, let $n = 100$ and $p = 0.01$. Then the probability of $k = 5$ successes in $n = 100$ trials is calculated using the binomial PMF as

$$\frac{100!}{95!\,5!}0.01^5(1-0.01)^{95} = 0.00290.$$

Using the Poisson PMF with $\lambda = np = 100 \cdot 0.01 = 1$, this probability is approximated by

$$e^{-1}\frac{1}{5!} = 0.00306.$$

We provide a formal justification of the Poisson approximation property in the end-of-chapter problems and also in Chapter 5, where we will further interpret it, extend it, and use it in the context of the Poisson process.

2.3 FUNCTIONS OF RANDOM VARIABLES

Consider a probability model of today's weather, let the random variable X be the temperature in degrees Celsius, and consider the transformation $Y = 1.8X + 32$, which gives the temperature in degrees Fahrenheit. In this example, Y is a **linear** function of X, of the form

$$Y = g(X) = aX + b,$$

where a and b are scalars. We may also consider nonlinear functions of the general form

$$Y = g(X).$$

For example, if we wish to display temperatures on a logarithmic scale, we would want to use the function $g(X) = \log X$.

If $Y = g(X)$ is a function of a random variable X, then Y is also a random variable, since it provides a numerical value for each possible outcome. This is because every outcome in the sample space defines a numerical value x for X and hence also the numerical value $y = g(x)$ for Y. If X is discrete with PMF p_X, then Y is also discrete, and its PMF p_Y can be calculated using the PMF of X. In particular, to obtain $p_Y(y)$ for any y, we add the probabilities of all values of x such that $g(x) = y$:

$$p_Y(y) = \sum_{\{x \,|\, g(x)=y\}} p_X(x).$$

Example 2.1. Let $Y = |X|$ and let us apply the preceding formula for the PMF p_Y to the case where

$$p_X(x) = \begin{cases} 1/9 & \text{if } x \text{ is an integer in the range } [-4, 4], \\ 0 & \text{otherwise.} \end{cases}$$

The possible values of Y are $y = 0, 1, 2, 3, 4$. To compute $p_Y(y)$ for some given value y from this range, we must add $p_X(x)$ over all values x such that $|x| = y$. In particular, there is only one value of X that corresponds to $y = 0$, namely $x = 0$. Thus,

$$p_Y(0) = p_X(0) = \frac{1}{9}.$$

Also, there are two values of X that correspond to each $y = 1, 2, 3, 4$, so for example,

$$p_Y(1) = p_X(-1) + p_X(1) = \frac{2}{9}.$$

Thus, the PMF of Y is

$$p_Y(y) = \begin{cases} 2/9 & \text{if } y = 1, 2, 3, 4, \\ 1/9 & \text{if } y = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For another related example, let $Z = X^2$. To obtain the PMF of Z, we can view it either as the square of the random variable X or as the square of the random variable Y. By applying the formula $p_Z(z) = \sum_{\{x \,|\, x^2=z\}} p_X(x)$ or the formula $p_Z(z) = \sum_{\{y \,|\, y^2=z\}} p_Y(y)$, we obtain

$$p_Z(z) = \begin{cases} 2/9 & \text{if } z = 1, 4, 9, 16, \\ 1/9 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

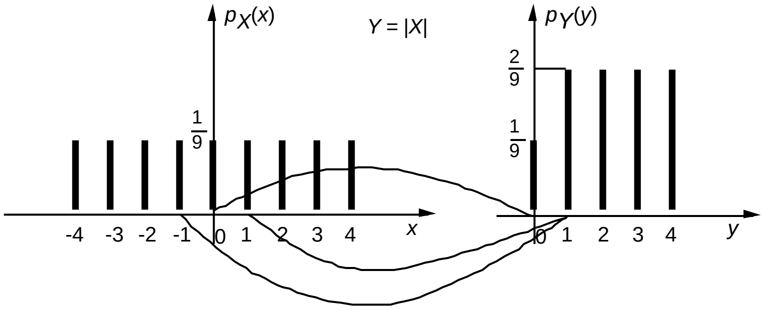


Figure 2.7: The PMFs of X and $Y = |X|$ in Example 2.1.

2.4 EXPECTATION, MEAN, AND VARIANCE

The PMF of a random variable X provides us with several numbers, the probabilities of all the possible values of X. It would be desirable to summarize this information in a single representative number. This is accomplished by the **expectation** of X, which is a weighted (in proportion to probabilities) average of the possible values of X.

As motivation, suppose you spin a wheel of fortune many times. At each spin, one of the numbers $m_1, m_2, \ldots, m_n$ comes up with corresponding probability $p_1, p_2, \ldots, p_n$, and this is your monetary reward from that spin. What is the amount of money that you "expect" to get "per spin"? The terms "expect" and "per spin" are a little ambiguous, but here is a reasonable interpretation.

Suppose that you spin the wheel k times, and that k_i is the number of times that the outcome is m_i. Then, the total amount received is $m_1k_1 + m_2k_2 + \cdots + m_nk_n$. The amount received per spin is

$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k}.$$

If the number of spins k is very large, and if we are willing to interpret probabilities as relative frequencies, it is reasonable to anticipate that m_i comes up a fraction of times that is roughly equal to p_i:

$$p_i \approx \frac{k_i}{k}, \qquad i = 1, \ldots, n.$$

Thus, the amount of money per spin that you "expect" to receive is

$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k} \approx m_1p_1 + m_2p_2 + \cdots + m_np_n.$$

Motivated by this example, we introduce an important definition.

Expectation

We define the **expected value** (also called the **expectation** or the **mean**) of a random variable X, with PMF $p_X(x)$, by†

$$\mathbf{E}[X] = \sum_x x p_X(x).$$

Example 2.2. Consider two independent coin tosses, each with a 3/4 probability of a head, and let X be the number of heads obtained. This is a binomial random variable with parameters $n = 2$ and $p = 3/4$. Its PMF is

$$p_X(k) = \begin{cases} (1/4)^2 & \text{if } k = 0, \\ 2 \cdot (1/4) \cdot (3/4) & \text{if } k = 1, \\ (3/4)^2 & \text{if } k = 2, \end{cases}$$

so the mean is

$$\mathbf{E}[X] = 0 \cdot \left(\frac{1}{4}\right)^2 + 1 \cdot \left(2 \cdot \frac{1}{4} \cdot \frac{3}{4}\right) + 2 \cdot \left(\frac{3}{4}\right)^2 = \frac{24}{16} = \frac{3}{2}.$$

It is useful to view the mean of X as a "representative" value of X, which lies somewhere in the middle of its range. We can make this statement more precise, by viewing the mean as the **center of gravity** of the PMF, in the sense explained in Fig. 2.8.

† When dealing with random variables that take a countably infinite number of values, one has to deal with the possibility that the infinite sum $\sum_x x p_X(x)$ is not well-defined. More concretely, we will say that the expectation is well-defined if $\sum_x |x| p_X(x) < \infty$. In that case, it is known that the infinite sum $\sum_x x p_X(x)$ converges to a finite value that is independent of the order in which the various terms are summed.

For an example where the expectation is not well-defined, consider a random variable X that takes the value 2^k with probability 2^{-k}, for $k = 1, 2, \ldots$. For a more subtle example, consider the random variable X that takes the values 2^k and -2^k with probability 2^{-k}, for $k = 2, 3, \ldots$. The expectation is again undefined, even though the PMF is symmetric around zero and one might be tempted to say that $\mathbf{E}[X]$ is zero.

Throughout this book, in lack of an indication to the contrary, we implicitly assume that the expected value of the random variables of interest is well-defined.

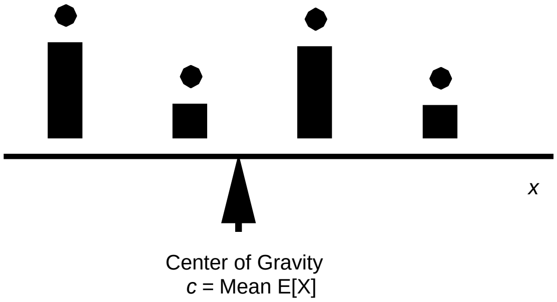


Figure 2.8: Interpretation of the mean as a center of gravity. Given a bar with a weight $p_X(x)$ placed at each point x with $p_X(x) > 0$, the center of gravity c is the point at which the sum of the torques from the weights to its left are equal to the sum of the torques from the weights to its right, that is,

$$\sum_x (x - c)p_X(x) = 0, \qquad \text{or} \quad c = \sum_x x p_X(x),$$

and the center of gravity is equal to the mean $\mathbf{E}[X]$.

There are many other quantities that can be associated with a random variable and its PMF. For example, we define the **2nd moment** of the random variable X as the expected value of the random variable X^2. More generally, we define the **nth moment** as $\mathbf{E}[X^n]$, the expected value of the random variable X^n. With this terminology, the 1st moment of X is just the mean.

The most important quantity associated with a random variable X, other than the mean, is its **variance**, which is denoted by $\text{var}(X)$ and is defined as the expected value of the random variable $\big(X - \mathbf{E}[X]\big)^2$, i.e.,

$$\text{var}(X) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big].$$

Since $\big(X - \mathbf{E}[X]\big)^2$ can only take nonnegative values, the variance is always nonnegative.

The variance provides a measure of dispersion of X around its mean. Another measure of dispersion is the **standard deviation** of X, which is defined as the square root of the variance and is denoted by σ_X:

$$\sigma_X = \sqrt{\text{var}(X)}.$$

The standard deviation is often easier to interpret, because it has the same units as X. For example, if X measures length in meters, the units of variance are square meters, while the units of the standard deviation are meters.

One way to calculate $\text{var}(X)$, is to use the definition of expected value, after calculating the PMF of the random variable $\big(X - \mathbf{E}[X]\big)^2$. This latter random variable is a function of X, and its PMF can be obtained in the manner discussed in the preceding section.

Example 2.3. Consider the random variable X of Example 2.1, which has the PMF

$$p_X(x) = \begin{cases} 1/9 & \text{if } x \text{ is an integer in the range } [-4,4], \\ 0 & \text{otherwise.} \end{cases}$$

The mean $\mathbf{E}[X]$ is equal to 0. This can be seen from the symmetry of the PMF of X around 0, and can also be verified from the definition:

$$\mathbf{E}[X] = \sum_x x p_X(x) = \frac{1}{9} \sum_{x=-4}^{4} x = 0.$$

Let $Z = \big(X - \mathbf{E}[X]\big)^2 = X^2$. As in Example 2.1, we obtain

$$p_Z(z) = \begin{cases} 2/9 & \text{if } z = 1, 4, 9, 16, \\ 1/9 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The variance of X is then obtained by

$$\text{var}(X) = \mathbf{E}[Z] = \sum_z z p_Z(z) = 0 \cdot \frac{1}{9} + 1 \cdot \frac{2}{9} + 4 \cdot \frac{2}{9} + 9 \cdot \frac{2}{9} + 16 \cdot \frac{2}{9} = \frac{60}{9}.$$

It turns out that there is an easier method to calculate var(X), which uses the PMF of X but *does not require the PMF of* $\big(X - \mathbf{E}[X]\big)^2$. This method is based on the following rule.

Expected Value Rule for Functions of Random Variables

Let X be a random variable with PMF $p_X(x)$, and let $g(X)$ be a real-valued function of X. Then, the expected value of the random variable $g(X)$ is given by

$$\mathbf{E}\big[g(X)\big] = \sum_x g(x) p_X(x).$$

To verify this rule, we use the formula $p_Y(y) = \sum_{\{x \,|\, g(x)=y\}} p_X(x)$ derived in the preceding section, we have

$$\begin{aligned}
\mathbf{E}\big[g(X)\big] &= \mathbf{E}[Y] \\
&= \sum_y y p_Y(y) \\
&= \sum_y y \sum_{\{x \,|\, g(x)=y\}} p_X(x) \\
&= \sum_y \sum_{\{x \,|\, g(x)=y\}} y p_X(x) \\
&= \sum_y \sum_{\{x \,|\, g(x)=y\}} g(x) p_X(x) \\
&= \sum_x g(x) p_X(x).
\end{aligned}$$

Using the expected value rule, we can write the variance of X as

$$\mathrm{var}(X) = \mathbf{E}\left[\big(X - \mathbf{E}[X]\big)^2\right] = \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x).$$

Similarly, the nth moment is given by

$$\mathbf{E}[X^n] = \sum_x x^n p_X(x),$$

and there is no need to calculate the PMF of X^n.

Example 2.3. (Continued) For the random variable X with PMF

$$p_X(x) = \begin{cases} 1/9 & \text{if } x \text{ is an integer in the range } [-4, 4], \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned}
\mathrm{var}(X) &= \mathbf{E}\left[\big(X - \mathbf{E}[X]\big)^2\right] \\
&= \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x) \\
&= \frac{1}{9} \sum_{x=-4}^{4} x^2 \qquad \text{since } \mathbf{E}[X] = 0 \\
&= \frac{1}{9}(16 + 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16) \\
&= \frac{60}{9},
\end{aligned}$$

which is consistent with the result obtained earlier.

As we have noted earlier, the variance is always nonnegative, but could it be zero? Since every term in the formula $\sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x)$ for the variance is nonnegative, the sum is zero if and only if $\big(x - \mathbf{E}[X]\big)^2 p_X(x) = 0$ for every x. This condition implies that for any x with $p_X(x) > 0$, we must have $x = \mathbf{E}[X]$ and the random variable X is not really "random": its experimental value is equal to the mean $\mathbf{E}[X]$, with probability 1.

Variance

The variance $\mathrm{var}(X)$ of a random variable X is defined by

$$\mathrm{var}(X) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big]$$

and can be calculated as

$$\mathrm{var}(X) = \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x).$$

It is always nonnegative. Its square root is denoted by σ_X and is called the **standard deviation**.

Let us now use the expected value rule for functions in order to derive some important properties of the mean and the variance. We start with a random variable X and define a new random variable Y, of the form

$$Y = aX + b,$$

where a and b are given scalars. Let us derive the mean and the variance of the linear function Y. We have

$$\mathbf{E}[Y] = \sum_x (ax + b)p_X(x) = a\sum_x x p_X(x) + b\sum_x p_X(x) = a\mathbf{E}[X] + b.$$

Furthermore,

$$\begin{aligned}\mathrm{var}(Y) &= \sum_x \big(ax + b - \mathbf{E}[aX + b]\big)^2 p_X(x)\\ &= \sum_x \big(ax + b - a\mathbf{E}[X] - b\big)^2 p_X(x)\\ &= a^2 \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x)\\ &= a^2 \mathrm{var}(X).\end{aligned}$$

Mean and Variance of a Linear Function of a Random Variable

Let X be a random variable and let

$$Y = aX + b,$$

where a and b are given scalars. Then,

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \qquad \text{var}(Y) = a^2\text{var}(X).$$

Let us also give a convenient formula for the variance of a random variable X with given PMF.

Variance in Terms of Moments Expression

$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.$$

This expression is verified as follows:

$$\begin{aligned}
\text{var}(X) &= \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x) \\
&= \sum_x \Big(x^2 - 2x\mathbf{E}[X] + \big(\mathbf{E}[X]\big)^2\Big) p_X(x) \\
&= \sum_x x^2 p_X(x) - 2\mathbf{E}[X] \sum_x x p_X(x) + \big(\mathbf{E}[X]\big)^2 \sum_x p_X(x) \\
&= \mathbf{E}[X^2] - 2\big(\mathbf{E}[X]\big)^2 + \big(\mathbf{E}[X]\big)^2 \\
&= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.
\end{aligned}$$

We will now derive the mean and the variance of a few important random variables.

Example 2.4. Mean and Variance of the Bernoulli. Consider the experiment of tossing a biased coin, which comes up a head with probability p and a tail with probability $1 - p$, and the Bernoulli random variable X with PMF

$$p_X(k) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0. \end{cases}$$

Its mean, second moment, and variance are given by the following calculations:

$$\begin{aligned}\mathbf{E}[X] &= 1\cdot p + 0\cdot(1-p) = p,\\ \mathbf{E}[X^2] &= 1^2\cdot p + 0\cdot(1-p) = p,\\ \mathrm{var}(X) &= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = p - p^2 = p(1-p).\end{aligned}$$

Example 2.5. Discrete Uniform Random Variable. What is the mean and variance of the roll of a fair six-sided die? If we view the result of the roll as a random variable X, its PMF is

$$p_X(k) = \begin{cases} 1/6 & \text{if } k = 1,2,3,4,5,6,\\ 0 & \text{otherwise.}\end{cases}$$

Since the PMF is symmetric around 3.5, we conclude that $\mathbf{E}[X] = 3.5$. Regarding the variance, we have

$$\begin{aligned}\mathrm{var}(X) &= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2\\ &= \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) - (3.5)^2,\end{aligned}$$

which yields $\mathrm{var}(X) = 35/12$.

The above random variable is a special case of a **discrete uniformly distributed** random variable (or **discrete uniform** for short), which by definition, takes one out of a range of contiguous integer values, with equal probability. More precisely, this random variable has a PMF of the form

$$p_X(k) = \begin{cases} \dfrac{1}{b-a+1} & \text{if } k = a, a+1, \ldots, b,\\ 0 & \text{otherwise,}\end{cases}$$

where a and b are two integers with $a < b$; see Fig. 2.9.

The mean is

$$\mathbf{E}[X] = \frac{a+b}{2},$$

as can be seen by inspection, since the PMF is symmetric around $(a+b)/2$. To calculate the variance of X, we first consider the simpler case where $a = 1$ and $b = n$. It can be verified by induction on n that

$$\mathbf{E}[X^2] = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{1}{6}(n+1)(2n+1).$$

We leave the verification of this as an exercise for the reader. The variance can now be obtained in terms of the first and second moments

$$\begin{aligned}\mathrm{var}(X) &= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2\\ &= \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2\\ &= \frac{1}{12}(n+1)(4n+2-3n-3)\\ &= \frac{n^2-1}{12}.\end{aligned}$$

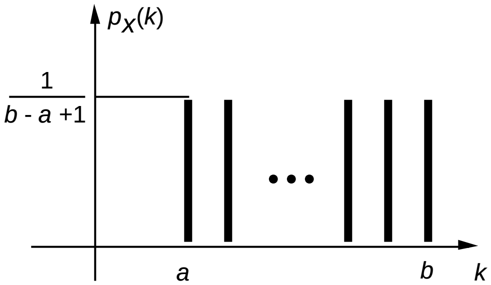


Figure 2.9: PMF of the discrete random variable that is uniformly distributed between two integers a and b. Its mean and variance are

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \mathrm{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

For the case of general integers a and b, we note that the uniformly distributed random variable over $[a, b]$ has the same variance as the uniformly distributed random variable over the interval $[1, b - a + 1]$, since these two random variables differ by the constant $a - 1$. Therefore, the desired variance is given by the above formula with $n = b - a + 1$, which yields

$$\mathrm{var}(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{(b-a)(b-a+2)}{12}.$$

Example 2.6. The Mean of the Poisson. The mean of the Poisson PMF

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots,$$

can be calculated is follows:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} && \text{the } k = 0 \text{ term is zero} \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^m}{m!} && \text{let } m = k - 1 \\ &= \lambda. \end{aligned}$$

The last equality is obtained by noting that $\sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^m}{m!} = \sum_{m=0}^{\infty} p_X(m) = 1$ is the normalization property for the Poisson PMF.

A similar calculation shows that the variance of a Poisson random variable is also λ (see the solved problems). We will have the occasion to derive this fact in a number of different ways in later chapters.

Expected values often provide a convenient vehicle for choosing optimally between several candidate decisions that result in different expected rewards. If we view the expected reward of a decision as its "average payoff over a large number of trials," it is reasonable to choose a decision with maximum expected reward. The following is an example.

Example 2.7. The Quiz Problem. This example, when generalized appropriately, is a prototypical model for optimal scheduling of a collection of tasks that have uncertain outcomes.

Consider a quiz game where a person is given two questions and must decide which question to answer first. Question 1 will be answered correctly with probability 0.8, and the person will then receive as prize \$100, while question 2 will be answered correctly with probability 0.5, and the person will then receive as prize \$200. If the first question attempted is answered incorrectly, the quiz terminates, i.e., the person is not allowed to attempt the second question. If the first question is answered correctly, the person is allowed to attempt the second question. Which question should be answered first to maximize the expected value of the total prize money received?

The answer is not obvious because there is a tradeoff: attempting first the more valuable but also more difficult question 2 carries the risk of never getting a chance to attempt the easier question 1. Let us view the total prize money received as a random variable X, and calculate the expected value $\mathbf{E}[X]$ under the two possible question orders (cf. Fig. 2.10):

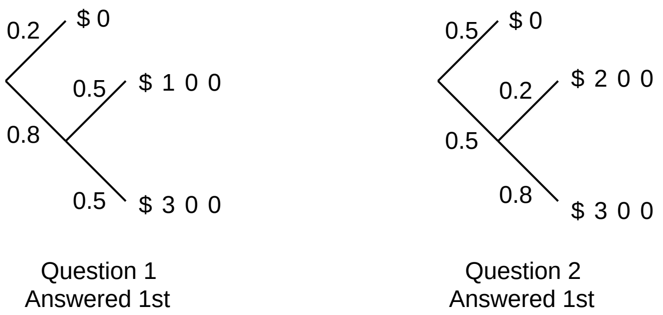


Figure 2.10: Sequential description of the sample space of the quiz problem for the two cases where we answer question 1 or question 2 first.

(a) *Answer question 1 first*: Then the PMF of X is (cf. the left side of Fig. 2.10)

$$p_X(0) = 0.2, \qquad p_X(100) = 0.8 \cdot 0.5, \qquad p_X(300) = 0.8 \cdot 0.5,$$

and we have

$$\mathbf{E}[X] = 0.8 \cdot 0.5 \cdot 100 + 0.8 \cdot 0.5 \cdot 300 = \$160.$$

(b) *Answer question 2 first*: Then the PMF of X is (cf. the right side of Fig. 2.10)

$$p_X(0) = 0.5, \qquad p_X(200) = 0.5 \cdot 0.2, \qquad p_X(300) = 0.5 \cdot 0.8,$$

and we have

$$\mathbf{E}[X] = 0.5 \cdot 0.2 \cdot 200 + 0.5 \cdot 0.8 \cdot 300 = \$140.$$

Thus, it is preferable to attempt the easier question 1 first.

Let us now generalize the analysis. Denote by p_1 and p_2 the probabilities of correctly answering questions 1 and 2, respectively, and by v_1 and v_2 the corresponding prizes. If question 1 is answered first, we have

$$\mathbf{E}[X] = p_1(1 - p_2)v_1 + p_1p_2(v_1 + v_2) = p_1v_1 + p_1p_2v_2,$$

while if question 2 is answered first, we have

$$\mathbf{E}[X] = p_2(1 - p_1)v_2 + p_2p_1(v_2 + v_1) = p_2v_2 + p_2p_1v_1.$$

It is thus optimal to answer question 1 first if and only if

$$p_1v_1 + p_1p_2v_2 \geq p_2v_2 + p_2p_1v_1,$$

or equivalently, if

$$\frac{p_1v_1}{1 - p_1} \geq \frac{p_2v_2}{1 - p_2}.$$

Thus, it is optimal to order the questions in decreasing value of the expression $pv/(1 - p)$, which provides a convenient index of quality for a question with probability of correct answer p and value v. Interestingly, this rule generalizes to the case of more than two questions (see the end-of-chapter problems).

We finally illustrate by example a common pitfall: unless $g(X)$ is a linear function, it is not generally true that $\mathbf{E}\big[g(X)\big]$ is equal to $g\big(\mathbf{E}[X]\big)$.

Example 2.8. Average Speed Versus Average Time. If the weather is good (which happens with probability 0.6), Alice walks the 2 miles to class at a speed of $V = 5$ miles per hour, and otherwise drives her motorcycle at a speed of $V = 30$ miles per hour. What is the mean of the time T to get to class?

The correct way to solve the problem is to first derive the PMF of T,

$$p_T(t) = \begin{cases} 0.6 & \text{if } t = 2/5 \text{ hours}, \\ 0.4 & \text{if } t = 2/30 \text{ hours}, \end{cases}$$

and then calculate its mean by

$$\mathbf{E}[T] = 0.6 \cdot \frac{2}{5} + 0.4 \cdot \frac{2}{30} = \frac{4}{15} \text{ hours.}$$

However, it is wrong to calculate the mean of the speed V,

$$\mathbf{E}[V] = 0.6 \cdot 5 + 0.4 \cdot 30 = 15 \text{ miles per hour,}$$

and then claim that the mean of the time T is

$$\frac{2}{\mathbf{E}[V]} = \frac{2}{15} \text{ hours.}$$

To summarize, in this example we have

$$T = \frac{2}{V}, \qquad \text{and } \mathbf{E}[T] = \mathbf{E}\left[\frac{2}{V}\right] \neq \frac{2}{\mathbf{E}[V]}.$$

2.5 JOINT PMFS OF MULTIPLE RANDOM VARIABLES

Probabilistic models often involve several random variables of interest. For example, in a medical diagnosis context, the results of several tests may be significant, or in a networking context, the workloads of several gateways may be of interest. All of these random variables are associated with the same experiment, sample space, and probability law, and their values may relate in interesting ways. This motivates us to consider probabilities involving simultaneously the numerical values of several random variables and to investigate their mutual couplings. In this section, we will extend the concepts of PMF and expectation developed so far to multiple random variables. Later on, we will also develop notions of conditioning and independence that closely parallel the ideas discussed in Chapter 1.

Consider two discrete random variables X and Y associated with the same experiment. The **joint** PMF of X and Y is defined by

$$p_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y)$$

for all pairs of numerical values (x, y) that X and Y can take. Here and elsewhere, we will use the abbreviated notation $\mathbf{P}(X = x, Y = y)$ instead of the more precise notations $\mathbf{P}(\{X = x\} \cap \{Y = y\})$ or $\mathbf{P}(X = x \text{ and } Y = x)$.

The joint PMF determines the probability of any event that can be specified in terms of the random variables X and Y. For example if A is the set of all pairs (x, y) that have a certain property, then

$$\mathbf{P}\big((X,Y) \in A\big) = \sum_{(x,y)\in A} p_{X,Y}(x,y).$$

In fact, we can calculate the PMFs of X and Y by using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

The formula for $p_X(x)$ can be verified using the calculation

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \sum_y \mathbf{P}(X = x, Y = y) \\ &= \sum_y p_{X,Y}(x,y), \end{aligned}$$

where the second equality follows by noting that the event $\{X = x\}$ is the union of the disjoint events $\{X = x, Y = y\}$ as y ranges over all the different values of Y. The formula for $p_Y(y)$ is verified similarly. We sometimes refer to p_X and p_Y as the **marginal** PMFs, to distinguish them from the joint PMF.

The example of Fig. 2.11 illustrates the calculation of the marginal PMFs from the joint PMF by using the **tabular method**. Here, the joint PMF of X and Y is arranged in a two-dimensional table, and **the marginal PMF of X or Y at a given value is obtained by adding the table entries along a corresponding column or row**, respectively.

Functions of Multiple Random Variables

When there are multiple random variables of interest, it is possible to generate new random variables by considering functions involving several of these random variables. In particular, a function $Z = g(X, Y)$ of the random variables X and Y defines another random variable. Its PMF can be calculated from the joint PMF $p_{X,Y}$ according to

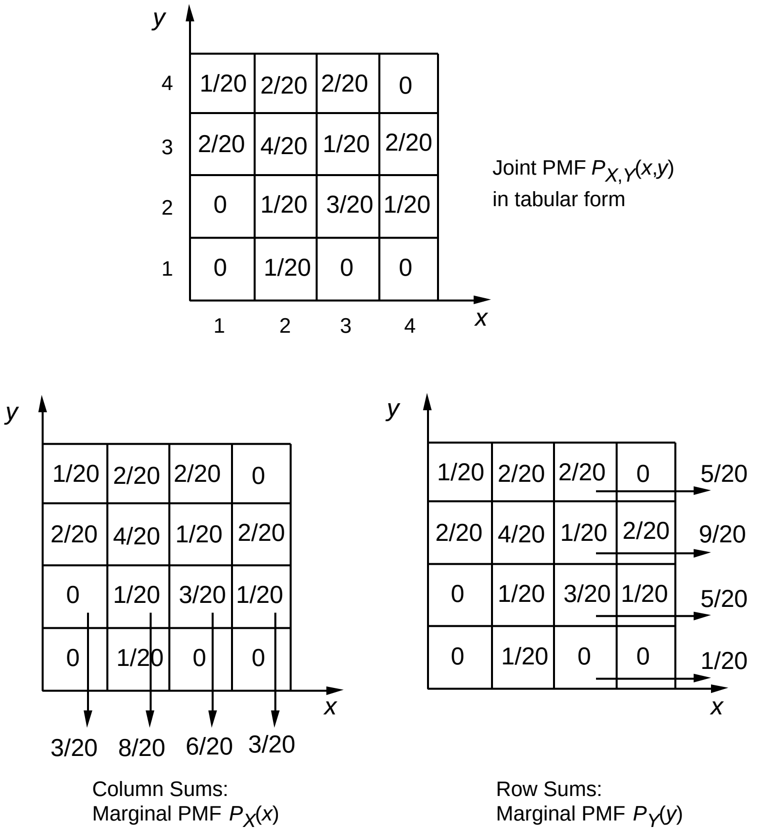
$$p_Z(z) = \sum_{\{(x,y)\,|\,g(x,y)=z\}} p_{X,Y}(x,y).$$

Furthermore, the expected value rule for functions naturally extends and takes the form

$$\mathbf{E}\big[g(X,Y)\big] = \sum_{x,y} g(x,y)p_{X,Y}(x,y).$$

The verification of this is very similar to the earlier case of a function of a single random variable. In the special case where g is linear and of the form $aX+bY+c$, where a, b, and c are given scalars, we have

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$



y \ x	1	2	3	4
4	1/20	2/20	2/20	0
3	2/20	4/20	1/20	2/20
2	0	1/20	3/20	1/20
1	0	1/20	0	0

Figure 2.11: Illustration of the tabular method for calculating marginal PMFs from joint PMFs. The joint PMF is represented by a table, where the number in each square (x, y) gives the value of $p_{X,Y}(x, y)$. To calculate the marginal PMF $p_X(x)$ for a given value of x, we add the numbers in the column corresponding to x. For example $p_X(2) = 8/20$. Similarly, to calculate the marginal PMF $p_Y(y)$ for a given value of y, we add the numbers in the row corresponding to y. For example $p_Y(2) = 5/20$.

More than Two Random Variables

The joint PMF of three random variables X, Y, and Z is defined in analogy with the above as

$$p_{X,Y,Z}(x, y, z) = \mathbf{P}(X = x, Y = y, Z = z),$$

for all possible triplets of numerical values (x, y, z). Corresponding marginal PMFs are analogously obtained by equations such as

$$p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, z),$$

and

$$p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, y, z).$$

The expected value rule for functions takes the form

$$\mathbf{E}\big[g(X, Y, Z)\big] = \sum_{x,y,z} g(x, y, z) p_{X,Y,Z}(x, y, z),$$

and if g is linear and of the form $aX + bY + cZ + d$, then

$$\mathbf{E}[aX + bY + cZ + d] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c\mathbf{E}[Z] + d.$$

Furthermore, there are obvious generalizations of the above to more than three random variables. For example, for any random variables $X_1, X_2, \ldots, X_n$ and any scalars $a_1, a_2, \ldots, a_n$, we have

$$\mathbf{E}[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1\mathbf{E}[X_1] + a_2\mathbf{E}[X_2] + \cdots + a_n\mathbf{E}[X_n].$$

Example 2.9. Mean of the Binomial. Your probability class has 300 students and each student has probability 1/3 of getting an A, independently of any other student. What is the mean of X, the number of students that get an A? Let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th student gets an A,} \\ 0 & \text{otherwise.} \end{cases}$$

Thus $X_1, X_2, \ldots, X_n$ are Bernoulli random variables with common mean $p = 1/3$ and variance $p(1 - p) = (1/3)(2/3) = 2/9$. Their sum

$$X = X_1 + X_2 + \cdots + X_n$$

is the number of students that get an A. Since X is the number of "successes" in n independent trials, it is a binomial random variable with parameters n and p.

Using the linearity of X as a function of the X_i, we have

$$\mathbf{E}[X] = \sum_{i=1}^{300} \mathbf{E}[X_i] = \sum_{i=1}^{300} \frac{1}{3} = 300 \cdot \frac{1}{3} = 100.$$

If we repeat this calculation for a general number of students n and probability of A equal to p, we obtain

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} p = np,$$

Example 2.10. The Hat Problem. Suppose that n people throw their hats in a box and then each picks up one hat at random. What is the expected value of X, the number of people that get back their own hat?

For the ith person, we introduce a random variable X_i that takes the value 1 if the person selects his/her own hat, and takes the value 0 otherwise. Since $\mathbf{P}(X_i = 1) = 1/n$ and $\mathbf{P}(X_i = 0) = 1 - 1/n$, the mean of X_i is

$$\mathbf{E}[X_i] = 1 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n}.$$

We now have

$$X = X_1 + X_2 + \cdots + X_n,$$

so that

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = n \cdot \frac{1}{n} = 1.$$

Summary of Facts About Joint PMFs

Let X and Y be random variables associated with the same experiment.

- The joint PMF of X and Y is defined by

$$p_{X,Y}(x,y) = \mathbf{P}(X = x, Y = y).$$

- The marginal PMFs of X and Y can be obtained from the joint PMF, using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

- A function $g(X,Y)$ of X and Y defines another random variable, and

$$\mathbf{E}\big[g(X,Y)\big] = \sum_{x,y} g(x,y) p_{X,Y}(x,y).$$

  If g is linear, of the form $aX + bY + c$, we have

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

- The above have natural extensions to the case where more than two random variables are involved.

2.6 CONDITIONING

If we have a probabilistic model and we are also told that a certain event A has occurred, we can capture this knowledge by employing the conditional instead of the original (unconditional) probabilities. As discussed in Chapter 1, conditional probabilities are like ordinary probabilities (satisfy the three axioms) except that they refer to a new universe in which event A is known to have occurred. In the same spirit, we can talk about conditional PMFs which provide the probabilities of the possible values of a random variable, conditioned on the occurrence of some event. This idea is developed in this section. In reality though, there is not much that is new, only an elaboration of concepts that are familiar from Chapter 1, together with a fair dose of new notation.

Conditioning a Random Variable on an Event

The **conditional PMF** of a random variable X, conditioned on a particular event A with $\mathbf{P}(A) > 0$, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x \,|\, A) = \frac{\mathbf{P}\big(\{X = x\} \cap A\big)}{\mathbf{P}(A)}.$$

Note that the events $\{X = x\} \cap A$ are disjoint for different values of x, their union is A, and, therefore,

$$\mathbf{P}(A) = \sum_x \mathbf{P}\big(\{X = x\} \cap A\big).$$

Combining the above two formulas, we see that

$$\sum_x p_{X|A}(x) = 1,$$

so $p_{X|A}$ is a legitimate PMF.

As an example, let X be the roll of a die and let A be the event that the roll is an even number. Then, by applying the preceding formula, we obtain

$$\begin{aligned} p_{X|A}(x) &= \mathbf{P}(X = x \,|\, \text{roll is even}) \\ &= \frac{\mathbf{P}(X = x \text{ and } X \text{ is even})}{\mathbf{P}(\text{roll is even})} \\ &= \begin{cases} 1/3 & \text{if } x = 2, 4, 6, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The conditional PMF is calculated similar to its unconditional counterpart: to obtain $p_{X|A}(x)$, we add the probabilities of the outcomes that give rise to $X = x$ **and** belong to the conditioning event A, and then normalize by dividing with $\mathbf{P}(A)$ (see Fig. 2.12).

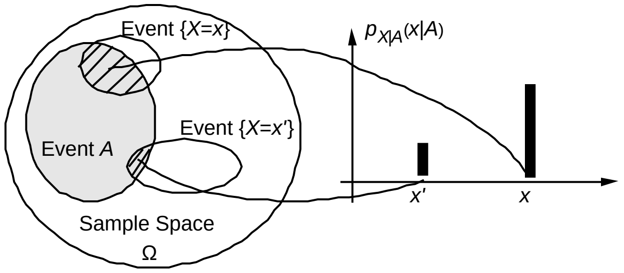


Figure 2.12: Visualization and calculation of the conditional PMF $p_{X|A}(x)$. For each x, we add the probabilities of the outcomes in the intersection $\{X = x\} \cap A$ and normalize by diving with $\mathbf{P}(A)$.

Conditioning one Random Variable on Another

Let X and Y be two random variables associated with the same experiment. If we know that the experimental value of Y is some particular y (with $p_Y(y) > 0$), this provides partial knowledge about the value of X. This knowledge is captured by the **conditional PMF** $p_{X|Y}$ of X given Y, which is defined by specializing the definition of $p_{X|A}$ to events A of the form $\{Y = y\}$:

$$p_{X|Y}(x \,|\, y) = \mathbf{P}(X = x \,|\, Y = y).$$

Using the definition of conditional probabilities, we have

$$p_{X|Y}(x \,|\, y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Let us fix some y, with $p_Y(y) > 0$ and consider $p_{X|Y}(x \,|\, y)$ as a function of x. This function is a valid PMF for X: it assigns nonnegative values to each possible x, and these values add to 1. Furthermore, this function of x, has the same shape as $p_{X,Y}(x, y)$ except that it is normalized by dividing with $p_Y(y)$, which enforces the normalization property

$$\sum_x p_{X|Y}(x \,|\, y) = 1.$$

Figure 2.13 provides a visualization of the conditional PMF.

The conditional PMF is often convenient for the calculation of the joint PMF, using a sequential approach and the formula

$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x \,|\, y),$$

or its counterpart

$$p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y \,|\, x).$$

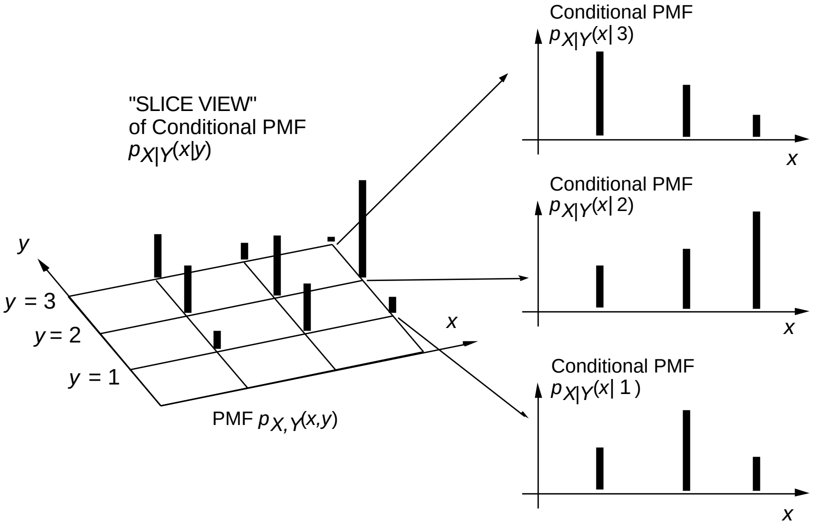


Figure 2.13: Visualization of the conditional PMF $p_{X|Y}(x\,|\,y)$. For each y, we view the joint PMF along the slice $Y = y$ and renormalize so that

$$\sum_x p_{X|Y}(x\,|\,y) = 1.$$

This method is entirely similar to the use of the multiplication rule from Chapter 1. The following examples provide an illustration.

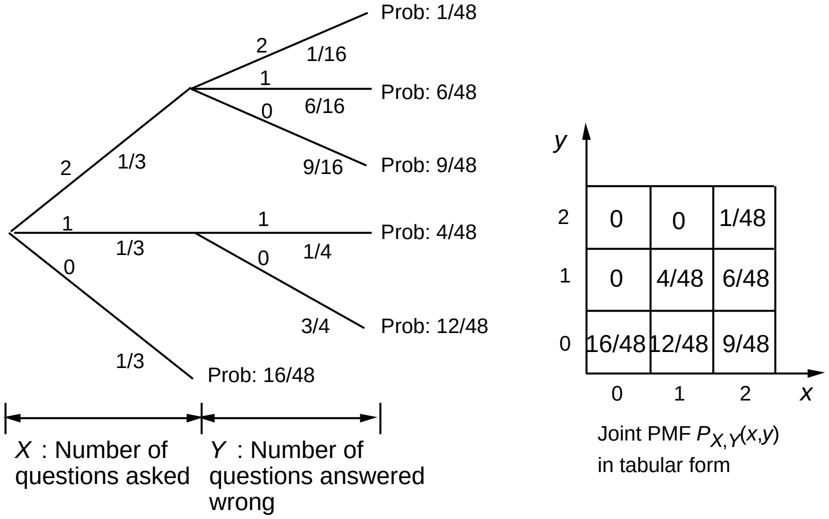
Example 2.11. Professor May B. Right often has her facts wrong, and answers each of her students' questions incorrectly with probability 1/4, independently of other questions. In each lecture May is asked 0, 1, or 2 questions with equal probability 1/3. Let X and Y be the number of questions May is asked and the number of questions she answers wrong in a given lecture, respectively. To construct the joint PMF $p_{X,Y}(x,y)$, we need to calculate all the probabilities $\mathbf{P}(X = x, Y = y)$ for all combinations of values of x and y. This can be done by using a sequential description of the experiment and the multiplication rule $p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x\,|\,y)$, as shown in Fig. 2.14. For example, for the case where one question is asked and is answered wrong, we have

$$p_{X,Y}(1,1) = p_X(x)p_{Y|X}(y\,|\,x) = \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12}.$$

The joint PMF can be represented by a two-dimensional table, as shown in Fig. 2.14. It can be used to calculate the probability of any event of interest. For instance, we have

$$\begin{aligned}\mathbf{P}(\text{at least one wrong answer}) &= p_{X,Y}(1,1) + p_{X,Y}(2,1) + p_{X,Y}(2,2)\\ &= \frac{4}{48} + \frac{6}{48} + \frac{1}{48}.\end{aligned}$$

.


Prob: 1/48
2
1/16
1
Prob: 6/48
0
6/16
2
1/3
9/16
Prob: 9/48
1
1
Prob: 4/48
1/3
0
1/4
0
3/4
Prob: 12/48
1/3
Prob: 16/48
X : Number of questions asked
Y : Number of questions answered wrong
y
x
Joint PMF $P_{X,Y}(x,y)$ in tabular form

2	0	0	1/48
1	0	4/48	6/48
0	16/48	12/48	9/48
	0	1	2

Figure 2.14: Calculation of the joint PMF $p_{X,Y}(x,y)$ in Example 2.11.

Example 2.12. Consider four independent rolls of a 6-sided die. Let X be the number of 1's and let Y be the number of 2's obtained. What is the joint PMF of X and Y?

The marginal PMF p_Y is given by the binomial formula

$$p_Y(y) = \binom{4}{y}\left(\frac{1}{6}\right)^y \left(\frac{5}{6}\right)^{4-y}, \qquad y = 0, 1, \ldots, 4.$$

To compute the conditional PMF $p_{X|Y}$, note that given that $Y = y$, X is the number of 1's in the remaining $4 - y$ rolls, each of which can take the 5 values $1, 3, 4, 5, 6$ with equal probability $1/5$. Thus, the conditional PMF $p_{X|Y}$ is binomial with parameters $4 - y$ and $p = 1/5$:

$$p_{X|Y}(x\,|\,y) = \binom{4-y}{x}\left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{4-y-x},$$

for all x and y such that $x,\ y = 0, 1, \ldots, 4$, and $0 \leq x + y \leq 4$. The joint PMF is now given by

$$\begin{aligned} p_{X,Y}(x,y) &= p_Y(y)p_{X|Y}(x\,|\,y) \\ &= \binom{4}{y}\left(\frac{1}{6}\right)^y\left(\frac{5}{6}\right)^{4-y}\binom{4-y}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{4-y-x}, \end{aligned}$$

for all nonnegative integers x and y such that $0 \leq x + y \leq 4$. For other values of x and y, we have $p_{X,Y}(x,y) = 0$.

The conditional PMF can also be used to calculate the marginal PMFs. In particular, we have by using the definitions,

$$p_X(x) = \sum_y p_{X,Y}(x,y) = \sum_y p_Y(y)p_{X|Y}(x\,|\,y).$$

This formula provides a divide-and-conquer method for calculating marginal PMFs. It is in essence identical to the total probability theorem given in Chapter 1, but cast in different notation. The following example provides an illustration.

Example 2.13. Consider a transmitter that is sending messages over a computer network. Let us define the following two random variables:

X : the travel time of a given message, Y : the length of the given message.

We know the PMF of the travel time of a message that has a given length, and we know the PMF of the message length. We want to find the (unconditional) PMF of the travel time of a message.

We assume that the length of a message can take two possible values: $y = 10^2$ bytes with probability 5/6, and $y = 10^4$ bytes with probability 1/6, so that

$$p_Y(y) = \begin{cases} 5/6 & \text{if } y = 10^2, \\ 1/6 & \text{if } y = 10^4. \end{cases}$$

We assume that the travel time X of the message depends on its length Y and the congestion level of the network at the time of transmission. In particular, the travel time is $10^{-4}Y$ secs with probability 1/2, $10^{-3}Y$ secs with probability 1/3, and $10^{-2}Y$ secs with probability 1/6. Thus, we have

$$p_{X|Y}(x\,|\,10^2) = \begin{cases} 1/2 & \text{if } x = 10^{-2}, \\ 1/3 & \text{if } x = 10^{-1}, \\ 1/6 & \text{if } x = 1, \end{cases} \qquad p_{X|Y}(x\,|\,10^4) = \begin{cases} 1/2 & \text{if } x = 1, \\ 1/3 & \text{if } x = 10, \\ 1/6 & \text{if } x = 100. \end{cases}$$

To find the PMF of X, we use the total probability formula

$$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x\,|\,y).$$

We obtain

$$p_X(10^{-2}) = \frac{5}{6} \cdot \frac{1}{2}, \qquad p_X(10^{-1}) = \frac{5}{6} \cdot \frac{1}{3}, \qquad p_X(1) = \frac{5}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{2},$$

$$p_X(10) = \frac{1}{6} \cdot \frac{1}{3}, \qquad p_X(100) = \frac{1}{6} \cdot \frac{1}{6}.$$

Note finally that one can define conditional PMFs involving more than two random variables, as in $p_{X,Y|Z}(x, y \,|\, z)$ or $p_{X|Y,Z}(x \,|\, y, z)$. The concepts and methods described above generalize easily (see the end-of-chapter problems).

Summary of Facts About Conditional PMFs

Let X and Y be random variables associated with the same experiment.

- Conditional PMFs are similar to ordinary PMFs, but refer to a universe where the conditioning event is known to have occurred.
- The conditional PMF of X given an event A with $\mathbf{P}(A) > 0$, is defined by
$$p_{X|A}(x) = \mathbf{P}(X = x \,|\, A)$$
and satisfies
$$\sum_x p_{X|A}(x) = 1.$$
- The conditional PMF of X given $Y = y$ is related to the joint PMF by
$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x \,|\, y).$$
This is analogous to the multiplication rule for calculating probabilities and can be used to calculate the joint PMF from the conditional PMF.
- The conditional PMF of X given Y can be used to calculate the marginal PMFs with the formula
$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x \,|\, y).$$
This is analogous to the divide-and-conquer approach for calculating probabilities using the total probability theorem.
- There are natural extensions to the above involving more than two random variables.

Conditional Expectation

A conditional PMF can be thought of as an ordinary PMF over a new universe determined by the conditioning event. In the same spirit, a conditional expectation is the same as an ordinary expectation, except that it refers to the new universe, and all probabilities and PMFs are replaced by their conditional counterparts. We list the main definitions and relevant facts below.

Summary of Facts About Conditional Expectations

Let X and Y be random variables associated with the same experiment.

- The conditional expectation of X given an event A with $\mathbf{P}(A) > 0$, is defined by
$$\mathbf{E}[X \mid A] = \sum_x x p_{X|A}(x \mid A).$$
For a function $g(X)$, it is given by
$$\mathbf{E}\big[g(X) \mid A\big] = \sum_x g(x) p_{X|A}(x \mid A).$$

- The conditional expectation of X given a value y of Y is defined by
$$\mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y).$$

- We have
$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X \mid Y = y].$$
This is the **total expectation theorem**.

- Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$ for all i. Then,
$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{P}(A_i) \mathbf{E}[X \mid A_i].$$

Let us verify the total expectation theorem, which basically says that "the unconditional average can be obtained by averaging the conditional averages." The theorem is derived using the total probability formula
$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x \mid y)$$

and the calculation

$$\begin{aligned}\mathbf{E}[X] &= \sum_x x p_X(x)\\ &= \sum_x x \sum_y p_Y(y) p_{X|Y}(x\,|\,y)\\ &= \sum_y p_Y(y) \sum_x x p_{X|Y}(x\,|\,y)\\ &= \sum_y p_Y(y)\mathbf{E}[X\,|\,Y=y].\end{aligned}$$

The relation $\mathbf{E}[X]=\sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}[X\,|\,A_i]$ can be verified by viewing it as a special case of the total expectation theorem. Let us introduce the random variable Y that takes the value i if and only if the event A_i occurs. Its PMF is given by

$$p_Y(i)=\begin{cases}\mathbf{P}(A_i) & \text{if } i=1,2,\ldots,n,\\ 0 & \text{otherwise.}\end{cases}$$

The total expectation theorem yields

$$\mathbf{E}[X]=\sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}[X\,|\,Y=i],$$

and since the event $\{Y=i\}$ is just A_i, we obtain the desired expression

$$\mathbf{E}[X]=\sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}[X\,|\,A_i].$$

The total expectation theorem is analogous to the total probability theorem. It can be used to calculate the unconditional expectation $\mathbf{E}[X]$ from the conditional PMF or expectation, using a divide-and-conquer approach.

Example 2.14. Messages transmitted by a computer in Boston through a data network are destined for New York with probability 0.5, for Chicago with probability 0.3, and for San Francisco with probability 0.2. The transit time X of a message is random. Its mean is 0.05 secs if it is destined for New York, 0.1 secs if it is destined for Chicago, and 0.3 secs if it is destined for San Francisco. Then, $\mathbf{E}[X]$ is easily calculated using the total expectation theorem as

$$\mathbf{E}[X]=0.5\cdot 0.05+0.3\cdot 0.1+0.2\cdot 0.3=0.115 \text{ secs.}$$

Example 2.15. Mean and Variance of the Geometric Random Variable. You write a software program over and over, and each time there is probability p

that it works correctly, independently from previous attempts. What is the mean and variance of X, the number of tries until the program works correctly?

We recognize X as a geometric random variable with PMF

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots .$$

The mean and variance of X are given by

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p, \qquad \mathrm{var}(X) = \sum_{k=1}^{\infty}(k - \mathbf{E}[X])^2(1-p)^{k-1}p,$$

but evaluating these infinite sums is somewhat tedious. As an alternative, we will apply the total expectation theorem, with $A_1 = \{X = 1\} = \{\text{first try is a success}\}$, $A_2 = \{X > 1\} = \{\text{first try is a failure}\}$, and end up with a much simpler calculation.

If the first try is successful, we have $X = 1$, and

$$\mathbf{E}[X \mid X = 1] = 1.$$

If the first try fails ($X > 1$), we have wasted one try, and we are back where we started. So, the expected number of remaining tries is $\mathbf{E}[X]$, and

$$\mathbf{E}[X \mid X > 1] = 1 + \mathbf{E}[X].$$

Thus,

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{P}(X = 1)\mathbf{E}[X \mid X = 1] + \mathbf{P}(X > 1)\mathbf{E}[X \mid X > 1] \\ &= p + (1-p)\big(1 + \mathbf{E}[X]\big),\end{aligned}$$

from which we obtain

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

With similar reasoning, we also have

$$\mathbf{E}[X^2 \mid X = 1] = 1, \qquad \mathbf{E}[X^2 \mid X > 1] = \mathbf{E}\big[(1+X)^2\big] = 1 + 2\mathbf{E}[X] + \mathbf{E}[X^2],$$

so that

$$\mathbf{E}[X^2] = p \cdot 1 + (1-p)\big(1 + 2\mathbf{E}[X] + \mathbf{E}[X^2]\big),$$

from which we obtain

$$\mathbf{E}[X^2] = \frac{1 + 2(1-p)\mathbf{E}[X]}{p},$$

and

$$\mathbf{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}.$$

We conclude that

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

2.7 INDEPENDENCE

We now discuss concepts of independence related to random variables. These concepts are analogous to the concepts of independence between events (cf. Chapter 1). They are developed by simply introducing suitable events involving the possible values of various random variables, and by considering their independence.

Independence of a Random Variable from an Event

The independence of a random variable from an event is similar to the independence of two events. The idea is that knowing the occurrence of the conditioning event tells us nothing about the value of the random variable. More formally, we say that the random variable X is **independent of the event** A if

$$\mathbf{P}(X = x \text{ and } A) = \mathbf{P}(X = x)\mathbf{P}(A) = p_X(x)\mathbf{P}(A), \qquad \text{for all } x,$$

which is the same as requiring that the two events $\{X = x\}$ and A be independent, for any choice x. As long as $\mathbf{P}(A) > 0$, and using the definition $p_{X|A}(x) = \mathbf{P}(X = x \text{ and } A)/\mathbf{P}(A)$ of the conditional PMF, we see that independence is the same as the condition

$$p_{X|A}(x) = p_X(x), \qquad \text{for all } x.$$

Example 2.16. Consider two independent tosses of a fair coin. Let X be the number of heads and let A be the event that the number of heads is even. The (unconditional) PMF of X is

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = 0, \\ 1/2 & \text{if } x = 1, \\ 1/4 & \text{if } x = 2, \end{cases}$$

and $\mathbf{P}(A) = 1/2$. The conditional PMF is obtained from the definition $p_{X|A}(x) = \mathbf{P}\big(X = x \text{ and } A\big)/\mathbf{P}(A)$:

$$p_{X|A}(x) = \begin{cases} 1/2 & \text{if } x = 0, \\ 0 & \text{if } x = 1, \\ 1/2 & \text{if } x = 2. \end{cases}$$

Clearly, X and A are not independent, since the PMFs p_X and $p_{X|A}$ are different. For an example of a random variable that is independent of A, consider the random variable that takes the value 0 if the first toss is a head, and the value 1 if the first toss is a tail. This is intuitively clear and can also be verified by using the definition of independence.

Independence of Random Variables

The notion of independence of two random variables is similar. We say that two **random variables** X and Y are **independent** if

$$p_{X,Y}(x,y) = p_X(x)\,p_Y(y), \qquad \text{for all } x, y.$$

This is the same as requiring that the two events $\{X = x\}$ and $\{Y = y\}$ be independent for every x and y. Finally, the formula $p_{X,Y}(x,y) = p_{X|Y}(x\,|\,y)p_Y(y)$ shows that independence is equivalent to the condition

$$p_{X|Y}(x\,|\,y) = p_X(x), \qquad \text{for all } y \text{ with } p_Y(y) > 0 \text{ and all } x.$$

Intuitively, independence means that the experimental value of Y tells us nothing about the value of X.

There is a similar notion of conditional independence of two random variables, given an event A with $\mathbf{P}(A > 0$. The conditioning event A defines a new universe and all probabilities (or PMFs) have to be replaced by their conditional counterparts. For example, X and Y are said to be **conditionally independent**, given a positive probability event A, if

$$\mathbf{P}(X = x, Y = y\,|\,A) = \mathbf{P}(X = x\,|\,A)\mathbf{P}(Y = y\,|\,A), \qquad \text{for all } x \text{ and } y,$$

or, in this chapter's notation,

$$p_{X,Y|A}(x,y) = p_{X|A}(x)p_{Y|A}(y), \qquad \text{for all } x \text{ and } y.$$

Once more, this is equivalent to

$$p_{X|Y,A}(x\,|\,y) = p_{X|A}(x) \qquad \text{for all } x \text{ and } y \text{ such that } p_{Y|A}(y) > 0.$$

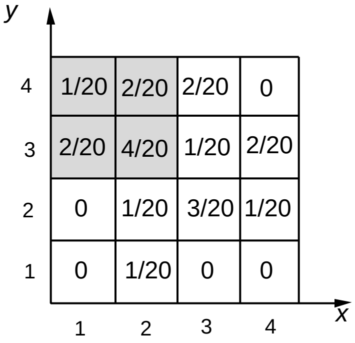
As in the case of events (Section 1.4), conditional independence may not imply unconditional independence and vice versa. This is illustrated by the example in Fig. 2.15.

If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y],$$

as shown by the following calculation:

$$\begin{aligned}
\mathbf{E}[XY] &= \sum_x \sum_y xyp_{X,Y}(x,y) \\
&= \sum_x \sum_y xyp_X(x)p_Y(y) \qquad \text{by independence} \\
&= \sum_x xp_X(x) \sum_y yp_Y(y) \\
&= \mathbf{E}[X]\,\mathbf{E}[Y].
\end{aligned}$$



y \ x	1	2	3	4
4	1/20	2/20	2/20	0
3	2/20	4/20	1/20	2/20
2	0	1/20	3/20	1/20
1	0	1/20	0	0

Figure 2.15: Example illustrating that conditional independence may not imply unconditional independence. For the PMF shown, the random variables X and Y are not independent. For example, we have

$$p_{X|Y}(1\,|\,1) = \mathbf{P}(X = 1\,|\,Y = 1) = 0 \neq \mathbf{P}(X = 1) = p_X(1).$$

On the other hand, conditional on the event $A = \{X \leq 2,\, Y \geq 3\}$ (the shaded set in the figure), the random variables X and Y can be seen to be independent. In particular, we have

$$p_{X|Y,A}(x\,|\,y) = \begin{cases} 1/3 & \text{if } x = 1, \\ 2/3 & \text{if } x = 2, \end{cases}$$

for both values $y = 3$ and $y = 4$.

A very similar calculation also shows that if X and Y are independent, then

$$\mathbf{E}\big[g(X)h(Y)\big] = \mathbf{E}\big[g(X)\big]\,\mathbf{E}\big[h(Y)\big],$$

for any functions g and h. In fact, this follows immediately once we realize that if X and Y are independent, then the same is true for $g(X)$ and $h(Y)$. This is intuitively clear and its formal verification is left as an end-of-chapter problem.

Consider now the sum $Z = X + Y$ of two independent random variables X and Y, and let us calculate the variance of Z. We have, using the relation $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$,

$$\begin{aligned}
\text{var}(Z) &= \mathbf{E}\Big[\big(X + Y - \mathbf{E}[X + Y]\big)^2\Big] \\
&= \mathbf{E}\Big[\big(X + Y - \mathbf{E}[X] - \mathbf{E}[Y]\big)^2\Big] \\
&= \mathbf{E}\Big[\Big(\big(X - \mathbf{E}[X]\big) + \big(Y - \mathbf{E}[Y]\big)\Big)^2\Big] \\
&= \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}\Big[\big(Y - \mathbf{E}[Y]\big)^2\Big] \\
&\quad + 2\mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\Big] \\
&= \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}\Big[\big(Y - \mathbf{E}[Y]\big)^2\Big].
\end{aligned}$$

To justify the last equality, note that the random variables $X-\mathbf{E}[X]$ and $Y-\mathbf{E}[Y]$ are independent (they are functions of the independent random variables X and Y, respectively) and

$$\mathbf{E}\big[\big(X-\mathbf{E}[X]\big)\big(Y-\mathbf{E}[Y]\big)\big]=\mathbf{E}\big[\big(X-\mathbf{E}[X]\big)\big]\,\mathbf{E}\big[\big(Y-\mathbf{E}[Y]\big)\big]=0.$$

We conclude that

$$\mathrm{var}(Z)=\mathrm{var}(X)+\mathrm{var}(Y).$$

Thus, the variance of the sum of two **independent** random variables is equal to the sum of their variances. As an interesting contrast, note that the mean of the sum of two random variables is always equal to the sum of their means, even if they are not independent.

Summary of Facts About Independent Random Variables

Let A be an event, with $\mathbf{P}(A)>0$, and let X and Y be random variables associated with the same experiment.

- X is independent of the event A if

  $$p_{X|A}(x)=p_X(x),\qquad \text{for all } x,$$

  that is, if for all x, the events $\{X=x\}$ and A are independent.

- X and Y are independent if for all possible pairs (x,y), the events $\{X=x\}$ and $\{Y=y\}$ are independent, or equivalently

  $$p_{X,Y}(x,y)=p_X(x)p_Y(y),\qquad \text{for all } x,y.$$

- If X and Y are independent random variables, then

  $$\mathbf{E}[XY]=\mathbf{E}[X]\mathbf{E}[Y].$$

  Furthermore, for any functions f and g, the random variables $g(X)$ and $h(Y)$ are independent, and we have

  $$\mathbf{E}\big[g(X)h(Y)\big]=\mathbf{E}\big[g(X)\big]\mathbf{E}\big[h(Y)\big].$$

- If X and Y are independent, then

  $$\mathrm{var}[X+Y]=\mathrm{var}(X)+\mathrm{var}(Y).$$

Independence of Several Random Variables

All of the above have natural extensions to the case of more than two random variables. For example, three random variables X, Y, and Z are said to be independent if

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_Y(y)p_Z(z), \qquad \text{for all } x,y,z.$$

If X, Y, and Z are independent random variables, then any three random variables of the form $f(X)$, $g(Y)$, and $h(Z)$, are also independent. Similarly, any two random variables of the form $g(X,Y)$ and $h(Z)$ are independent. On the other hand, two random variables of the form $g(X,Y)$ and $h(Y,Z)$ are usually not independent, because they are both affected by Y. Properties such as the above are intuitively clear if we interpret independence in terms of noninteracting (sub)experiments. They can be formally verified (see the end-of-chapter problems), but this is sometimes tedious. Fortunately, there is general agreement between intuition and what is mathematically correct. This is basically a testament that the definitions of independence we have been using adequately reflect the intended interpretation.

Another property that extends to multiple random variables is the following. If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$\text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n).$$

This can be verified by a calculation similar to the one for the case of two random variables and is left as an exercise for the reader.

Example 2.17. Variance of the Binomial. We consider n independent coin tosses, with each toss having probability p of coming up a head. For each i, we let X_i be the Bernoulli random variable which is equal to 1 if the ith toss comes up a head, and is 0 otherwise. Then, $X = X_1 + X_2 + \cdots + X_n$ is a binomial random variable. By the independence of the coin tosses, the random variables $X_1, \ldots, X_n$ are independent, and

$$\text{var}(X) = \sum_{i=1}^{n} \text{var}(X_i) = np(1-p).$$

The formulas for the mean and variance of a weighted sum of random variables form the basis for many statistical procedures that estimate the mean of a random variable by averaging many independent samples. A typical case is illustrated in the following example.

Example 2.18. Mean and Variance of the Sample Mean. We wish to estimate the approval rating of a president, to be called C. To this end, we ask n

persons drawn at random from the voter population, and we let X_i be a random variable that encodes the response of the ith person:

$$X_i = \begin{cases} 1 & \text{if the } i\text{th person approves C's performance,} \\ 0 & \text{if the } i\text{th person disapproves C's performance.} \end{cases}$$

We model $X_1, X_2, \ldots, X_n$ as independent Bernoulli random variables with common mean p and variance $p(1-p)$. Naturally, we view p as the true approval rating of C. We "average" the responses and compute the **sample mean** S_n, defined as

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Thus, S_n is the approval rating of C within our n-person sample.

We have, using the linearity of S_n as a function of the X_i,

$$\mathbf{E}[S_n] = \sum_{i=1}^{n} \frac{1}{n}\mathbf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} p = p,$$

and making use of the independence of $X_1, \ldots, X_n$,

$$\text{var}(S_n) = \sum_{i=1}^{n} \frac{1}{n^2}\text{var}(X_i) = \frac{p(1-p)}{n}.$$

The sample mean S_n can be viewed as a "good" estimate of the approval rating. This is because it has the correct expected value, which is the approval rating p, and its accuracy, as reflected by its variance, improves as the sample size n increases.

Note that even if the random variables X_i are not Bernoulli, the same calculation yields

$$\text{var}(S_n) = \frac{\text{var}(X)}{n},$$

as long as the X_i are independent, with common mean $\mathbf{E}[X]$ and variance $\text{var}(X)$. Thus, again, the sample mean becomes a very good estimate (in terms of variance) of the true mean $\mathbf{E}[X]$, as the sample size n increases. We will revisit the properties of the sample mean and discuss them in much greater detail in Chapter 7, when we discuss the laws of large numbers.

Example 2.19. Estimating Probabilities by Simulation. In many practical situations, the analytical calculation of the probability of some event of interest is very difficult. However, if we have a physical or computer model that can generate outcomes of a given experiment in accordance with their true probabilities, we can use simulation to calculate with high accuracy the probability of any given event A. In particular, we independently generate with our model n outcomes, we record the number m that belong to the event A of interest, and we approximate $\mathbf{P}(A)$ by m/n. For example, to calculate the probability $p = \mathbf{P}(\text{Heads})$ of a biased coin, we flip the coin n times, and we approximate p with the ratio (number of heads recorded)$/n$.

To see how accurate this process is, consider n independent Bernoulli random variables $X_1, \ldots, X_n$, each with PMF

$$p_{X_i}(x_i) = \begin{cases} \mathbf{P}(A) & \text{if } x_i = 1, \\ 0 & \text{if } x_i = 0. \end{cases}$$

In a simulation context, X_i corresponds to the ith outcome, and takes the value 1 if the ith outcome belongs to the event A. The value of the random variable

$$X = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is the estimate of $\mathbf{P}(A)$ provided by the simulation. According to Example 2.17, X has mean $\mathbf{P}(A)$ and variance $\mathbf{P}(A)\big(1 - \mathbf{P}(A)\big)/n$, so that for large n, it provides an accurate estimate of $\mathbf{P}(A)$.

2.8 SUMMARY AND DISCUSSION

Random variables provide the natural tools for dealing with probabilistic models in which the outcome determines certain numerical values of interest. In this chapter, we focused on discrete random variables, and developed the main concepts and some relevant tools. We also discussed several special random variables, and derived their PMF, mean, and variance, as summarized in the table that follows.

Summary of Results for Special Random Variables

Discrete Uniform over $[a, b]$**:**

$$p_X(k) = \begin{cases} \dfrac{1}{b - a + 1} & \text{if } k = a, a+1, \ldots, b, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \operatorname{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

Bernoulli with Parameter p**:** (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \qquad \operatorname{var}(X) = p(1-p).$$

Binomial with Parameters p and n: (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

$$\mathbf{E}[X] = np, \qquad \mathrm{var}(X) = np(1-p).$$

Geometric with Parameter p: (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \qquad \mathrm{var}(X) = \frac{1-p}{p^2}.$$

Poisson with Parameter λ: (Approximates the binomial PMF when n is large, p is small, and $\lambda = np$.)

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, \ldots,$$

$$\mathbf{E}[X] = \lambda, \qquad \mathrm{var}(X) = \lambda.$$

We also considered multiple random variables, and introduced their joint and conditional PMFs, and associated expected values. Conditional PMFs are often the starting point in probabilistic models and can be used to calculate other quantities of interest, such as marginal or joint PMFs and expectations, through a sequential or a divide-and-conquer approach. In particular, given the conditional PMF $p_{X|Y}(x \,|\, y)$:

(a) The joint PMF can be calculated by

$$p_{X,Y}(x, y) = p_Y(y)p_{X|Y}(x \,|\, y).$$

This can be extended to the case of three or more random variables, as in

$$p_{X,Y,Z}(x, y, z) = p_Y(y)p_{Y|Z}(y \,|\, z)p_{X|Y,Z}(x \,|\, y, z),$$

and is analogous to the sequential tree-based calculation method using the multiplication rule, discussed in Chapter 1.

(b) The marginal PMF can be calculated by

$$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x \,|\, y),$$

which generalizes the divide-and-conquer calculation method we discussed in Chapter 1.

(c) The divide-and-conquer calculation method in (b) above can be extended to compute expected values using the total expectation theorem:

$$\mathbf{E}[X] = \sum_y p_Y(y)\mathbf{E}[X \,|\, Y = y].$$

The concepts and methods of this chapter extend appropriately to general random variables (see the next chapter), and are fundamental for our subject.

SOLVED PROBLEMS

SECTION 2.2. Probability Mass Functions

Problem 1. The MIT soccer team has 2 games scheduled for one weekend. MIT has a 0.4 probability of not losing the first game, and a 0.6 probability of not losing the second game. If it does not lose, the team has a 50% chance of a win, and a 50% chance of a tie, independently of all other weekend events. MIT will receive 2 points for a win, 1 for a tie, and 0 for a loss. Let X be the number of points the MIT team earns over the weekend. Find the PMF of X.

*Solution.*This requires enumeration of the possibilities and straightforward computation:

$$\begin{aligned}
&\mathbf{P}(X=0) = 0.6 \cdot 0.4 = 0.24,\\
&\mathbf{P}(X=1) = 0.4 \cdot 0.5 \cdot 0.4 + 0.6 \cdot 0.5 \cdot 0.6 = 0.26,\\
&\mathbf{P}(X=2) = 0.4 \cdot 0.5 \cdot 0.4 + 0.6 \cdot 0.5 \cdot 0.6 + 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 = 0.32,\\
&\mathbf{P}(X=3) = 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 + 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 = 0.12,\\
&\mathbf{P}(X=4) = 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 = 0.06,\\
&\mathbf{P}(X>4) = 0.
\end{aligned}$$

Problem 2. The Celtics and the Lakers are set to play a playoff series of n basketball games, where n is odd. The Celtics have a probability p of winning any one game, independently of other games.

(a) Find values of p for which the Celtics would rather play $n = 5$ games rather than $n = 3$ games.

(b) For any $k > 0$, find the values for p for which $n = 2k + 1$ is better for the Celtics than $n = 2k - 1$.

Solution.(a) If $n = 5$, the Celtics need to win 3, 4, or 5 games (we assume here that all games are played even if the series is already decided). Thus the probability that the Celtics will win the series when $n = 5$ is the sum of the probabilities of 3, 4, or 5 wins:

$$\binom{5}{3} p^3(1-p)^2 + \binom{5}{4} p^4(1-p)^1 + \binom{5}{5} p^5(1-p)^0$$

Similarly, the probability that the Celtics will win the series when $n = 3$ is

$$\binom{3}{2} p^2(1-p)^1 + \binom{3}{3} p^3(1-p)^0.$$

We want to find p such that

$$10p^3(1-p)^2 + 5p^4(1-p) + p^5 > 3p^2(1-p) + p^3.$$

We see that for this to hold we must have $p > \frac{1}{2}$.

(b) It turns out that for the general case, we need $p > \frac{1}{2}$ as well. To prove this, let N be the number of Celtics wins in the first $2k-1$ games. If A denotes the event that the Celtics win with $n = 2k+1$, and B denotes the event that the Celtics win with $n = 2k-1$, then

$$\mathbf{P}(A) = \mathbf{P}(N \geq k+1) + \mathbf{P}(N = k) \cdot \big(1 - (1-p)^2\big) + \mathbf{P}(N = k-1) \cdot p^2,$$

$$\mathbf{P}(B) = \mathbf{P}(N \geq k) = \mathbf{P}(N = k) + \mathbf{P}(N \geq k+1),$$

and therefore

$$\begin{aligned}\mathbf{P}(A) - \mathbf{P}(B) &= \mathbf{P}(N = k-1) \cdot p^2 - \mathbf{P}(N = k) \cdot (1-p)^2 \\ &= \binom{2k-1}{k-1} p^{k-1}(1-p)^k p^2 - \binom{2k-1}{k} (1-p)^2 p^k (1-p)^{k-1}.\end{aligned}$$

By simplifying the above expression, it follows that for $\mathbf{P}(A) > \mathbf{P}(B)$, we need $p > \frac{1}{2}$.

Problem 3. You go to a party with 500 guests. What is the probability that exactly one other guest has the same birthday as you? Calculate this exactly and also approximately by using the Poisson PMF. (For simplicity, exclude birthdays on February 29.)

*Solution.*The number of guests that have the same birthday as you is binomial with $p = 1/365$ and $n = 499$. Thus the probability that exactly one other guest has the same birthday is

$$\binom{499}{1} \frac{1}{365} \left(\frac{364}{365}\right)^{498} \approx 0.3486.$$

Let $\lambda = np = 499/365 \approx 1.367$. The Poisson approximation is $e^{-\lambda}\lambda = e^{-1.367} \cdot 1.367 \approx 0.3483$, which closely agrees with the correct probability based on the binomial.

Problem 4. * The matchbox problem – inspired by Banach's smoking habits. A smoker mathematician carries one matchbox in his right pocket and one in his left pocket. Each time he wants to light a cigarette he selects a matchbox from the two pockets with probability $p = 1/2$, independently of earlier selections. The two matchboxes have initially N matches each. What is the PMF of the number of remaining matches at the moment when the mathematician reaches for a match and discovers that the corresponding matchbox is empty? How can we generalize for the case where the probabilities of a left and a right pocket selection are p and $1-p$, respectively?

*Solution.*Suppose that an empty box is first discovered in the left pocket. Let r_k be the conditional probability that the number of matches in the right pocket is k at that time, where $k = 0, 1, \ldots, N$. This number is k if and only if exactly $N-k$ right pocket selections precede the $(N+1)$st left pocket selection. Viewing a left and a right pocket selection as a "success" and a "failure," respectively, the probability of this happening is the probability of getting N successes in $2N-k$ trials. Thus

$$r_k = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k}.$$

The conditional probability of k matches in the left pocket given that an empty box is first discovered in the right pocket is also r_k. Thus, by the total probability theorem, the unconditional probability that the number of matches in the other pocket is k at the time when an empty box is first discovered is r_k, i.e.,

$$p_X(k) = \binom{2N-k}{N}\left(\frac{1}{2}\right)^{2N-k}, \qquad k = 0, 1, \ldots, N.$$

In the more general case, we must calculate separately the conditional probabilities of k matches in the other pocket, given that the left and the right pockets are first found empty. We must also calculate the probabilities that the left and the right pockets are first found empty. We then apply the total probability theorem.

Problem 5. * Recursive computation of the binomial PMF. Let X be the binomial random variable with parameters n and p. Show that the PMF can be computed, starting with $p_X(0) = (1-p)^n$, using the recursive formula

$$p_X(k+1) = \frac{p}{1-p} \cdot \frac{n-k}{k+1} \cdot p_X(k), \qquad k = 0, 1, \ldots, n-1.$$

*Solution.*For $k = 0, 1, \ldots, n-1$, we have

$$\frac{p_X(k+1)}{p_X(k)} = \frac{\binom{n}{k+1}p^{k+1}(1-p)^{n-k-1}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{p}{1-p} \cdot \frac{n-k}{k+1}.$$

Problem 6. * Form of the binomial PMF. Consider the binomial random variable X with parameters n and p. Show that, as k increases, the PMF $p_X(k)$ first increases monotonically and then decreases monotonically, with the maximum obtained when k is the largest integer that is less or equal to $(n+1)p$.

*Solution.*For $k = 1, \ldots, n$, we have

$$\frac{p_X(k)}{p_X(k-1)} = \frac{\binom{n}{k}p^k(1-p)^{n-k}}{\binom{n}{k-1}p^{k-1}(1-p)^{n-k+1}} = \frac{(n-k+1)p}{k(1-p)} = \frac{(n+1)p - pk}{k - pk}.$$

Thus if $k \leq (n+1)p$, or equivalently $k - pk \leq (n+1)p - pk$, the above ratio is greater than or equal to 1, and it follows that $p_X(k)$ is monotonically nondecreasing. Otherwise, the ratio is less than one, and $p_X(k)$ is monotonically decreasing, as required.

Problem 7. * Form of the Poisson PMF. Let X be the Poisson random variable with parameter λ. Show that the PMF $p_X(k)$ increases monotonically with k up to the point where k reaches the largest integer not exceeding λ, and after that decreases monotonically with k.

*Solution.*Using the expression for the Poisson PMF, we have

$$\frac{p_X(k)}{p_X(k-1)} = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \cdot \frac{(k-1)!}{\lambda^{k-1} \cdot e^{-\lambda}} = \frac{\lambda}{k}.$$

Thus if $k < \lambda$ the ratio is greater than 1, and it follows that $p_X(k)$ is monotonically increasing. Otherwise, the ratio is less than one, and $p_X(k)$ is monotonically decreasing, as required.

Problem 8. *   Justification of the Poisson approximation property. Consider the PMF of the binomial random variable with parameters n and p. Show that asymptotically, as

$$n \to \infty, \qquad p \to 0,$$

while np is fixed at a given scalar λ, this PMF approaches the PMF of the Poisson random variable with parameter λ.

*Solution.*Write the binomial PMF as

$$\begin{aligned} p_X(k) &= \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1-\frac{\lambda}{n}\right)^{n-k}. \end{aligned}$$

Fix k, and let $n \to \infty$ with $\lambda = np$ fixed. We have, for $j = 1, \ldots, k$,

$$\frac{n-k+j}{n} \to 1, \quad \left(1-\frac{\lambda}{n}\right)^k \to 1, \quad \left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}.$$

Thus, for each fixed k, as $n \to \infty$ we have

$$p_X(k) \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

SECTION 2.3. Functions of Random Variables

Problem 9. Let X be a random variable that takes values from 0 to 9 with equal probability 1/10.

(a) Find the PMF of the random variable $Y = X \bmod(3)$.

(b) Find the PMF of the random variable $Z = 5 \bmod(\mathrm{X}+1)$.

Solution.(a) Using the formula $p_Y(y) = \sum_{\{x \,|\, x \pmod 3 = y\}} p_X(x)$, we obtain

$$\begin{aligned} &p_Y(0) = p_X(0) + p_X(3) + p_X(6) + p_X(9) = 4/10, \\ &p_Y(1) = p_X(1) + p_X(4) + p_X(7) = 3/10, \\ &p_Y(2) = p_X(2) + p_X(5) + p_X(8) = 3/10, \\ &p_Y(y) = 0 \quad \text{if } y \notin \{0,1,2\}. \end{aligned}$$

(b) Similarly, using the formula $p_Z(z) = \sum_{\{x \,|\, 5 \ \mathrm{mod(x+1)=z}\}} p_X(x)$, we obtain

$$p_Z(z) = \begin{cases} 2/10 & \text{if } z = 0, \\ 2/10 & \text{if } z = 1, \\ 1/10 & \text{if } z = 2, \\ 5/10 & \text{if } z = 5, \\ 0 & \text{otherwise.} \end{cases}$$

SECTION 2.4. Expectation, Mean, and Variance

Problem 10. Consider the random variable X with PMF

$$p_X(x) = \begin{cases} \frac{x^2}{a} & \text{if } x = -3, -2, -1, 0, 1, 2, 3, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find a and $\mathbf{E}[X]$.

(b) What is the PMF of the random variable $z = (x - \mathbf{E}[X])^2$?

(c) Using part (b) compute the variance of X.

(d) Compute the variance of X using the expected value rule formula $\text{var}(X) = \sum_x (x - \mathbf{E}[X])^2 p_X(x)$.

Solution.(a) $a = 28$ and $\mathbf{E}[X] = 0$.

(b) If $z \in \{1, 4, 9\}$, then

$$p_Z(z) = \sum_{\{x \mid x^2 = z\}} p_X(x) = \frac{z}{28} + \frac{z}{28} = \frac{z}{14}.$$

Otherwise $p_Z(z) = 0$.

(c) $\text{var}(X) = \mathbf{E}[Z] = \sum_{z \in \{1,4,9\}} \frac{z^2}{14} = 7$.

(d) We have

$$\begin{aligned} \text{var}(X) &= \sum_x (x - \mathbf{E}[X])^2 p_X(x) \\ &= 1^2 \cdot \big(p_X(-1) + p_X(1)\big) + 2^2 \cdot \big(p_X(-2) + p_X(2)\big) + 3^2 \cdot \big(p_X(-3) + p_X(3)\big) \\ &= 2 \cdot \frac{1}{28} + 8 \cdot \frac{4}{28} + 18 \cdot \frac{9}{28} \\ &= 7. \end{aligned}$$

Problem 11. A city's temperature is modeled as a random variable with mean and standard deviation both equal to 10 degrees Celcius. A day is described as "normal" if the temperature within that day ranges with one standard deviation of the mean. What would be the temperature range for a normal day if temperature were expressed in Fahreneit degrees?

*Solution.*If X is the temperature in Celcius, the temperature in Fahreneit is $Y = 32 + 9X/5$. Therefore, $\mathbf{E}[Y] = 32 + 9\mathbf{E}[X]/5 = 32 + 18 = 40$. Also $\text{var}(Y) = (9/5)^2\text{var}(X)$, so the standard deviation of Y is $(9/5) \cdot 10 = 18$. Hence a normal day in Fahreneit is one for which the temperature is in the range $[22, 58]$.

Problem 12. Let a and b be positive integers with $a \leq b$, and let X be the random variable that takes as values, with equal probability, the powers of 2 in the interval $[2^a, 2^b]$. Find the expected value and the variance of X.

Solution. We have

$$p_X(x) = \begin{cases} \frac{1}{b-a+1} & \text{if } x \text{ is a power of 2 in the interval } [2^a, 2^b], \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathbf{E}[X] = \sum_{k=a}^{b} \frac{1}{b-a+1} 2^k = \frac{2^a}{b-a+1}(1 + 2 + \cdots + 2^{b-a}) = \frac{2^{b+1} - 2^a}{b-a+1}.$$

Similarly,

$$\mathbf{E}[X^2] = \sum_{k=a}^{b} \frac{1}{b-a+1}(2^k)^2 = \frac{4^{b+1} - 4^a}{3(b-a+1)},$$

and finally

$$\text{var}(X) = \frac{4^{b+1} - 4^a}{3(b-a+1)} - \left(\frac{2^{b+1} - 2^a}{b-a+1}\right)^2.$$

Problem 13. Your friend proposes to play the following game with you. He will hide a $10 bill randomly in one of ten numbered boxes. You are to find the bill, and if you find it, you can keep it. You must search for the bill by asking yes-no questions, and for each question you ask, you must pay $1. The maximum amount of money that you would you pay to play this game in order not to lose money on the average depends on the playing strategy. What would be this amount if you were to play according to the following strategies?

(a) Your questions must be: is it in box 1? is it in box 2? etc...

(b) On each guess you eliminate as close to half of the remaining numbers as possible.

*Solution.*We will find the expected gain according to each strategy, by computing the expected number of questions until we find the bill.

(a) For this strategy, the expected number of guesses is

$$\frac{1}{10}\sum_{i=1}^{10} i = \frac{1}{10} \cdot 55 = 5.5,$$

so we should be willing to pay no more than $4.50 to play with this strategy.

(b) For this strategy, the expected number of guesses is

$$\frac{4}{10} \cdot 4 + \frac{6}{10} \cdot 3 = 3.4,$$

so we should be willing to pay $6.60 to play with this strategy. The above value was computed by enumeration, i.e., counting the number of numbers out of 10 that take 4 guesses to find, and then counting the ones that take 3 guesses to find.

Problem 14. As an advertising campaign, a chocolate factory places golden tickets in some of its candy bars, with the promise that a golden ticket is worth a trip through the chocolate factory, and all the chocolate you can eat for life. If the probability of finding a gold ticket is p, find the mean and the variance of the number of candy bars you need to eat to find a ticket.

*Solution.*The number C of candy bars you need to eat is a geometric random variable. Thus the mean is $\mathbf{E}[C] = 1/p$, and the variance is $\text{var}(C) = (1-p)/p^2$.

Problem 15. St. Petersburg paradox. Consider the following game: you flip a coin, until the first tail appears. If the tail appears on the nth flip of the coin, you receive 2^n dollars. What is the expected gain for playing the game? How much would you be willing to pay to play this game?

*Solution.*The expected value of the gain for a single game is infinite since if X is your gain, then

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} 2^k \cdot 2^{-k} = \sum_{k=1}^{\infty} 1 = \infty.$$

Thus if you are faced with the choice of playing for given fee f or not playing at all, and your objective is to make the choice that maximizes your expected net gain, you would be willing to pay any value of f. However, this is in strong disagreement with the behavior of individuals. In fact experiments have shown that most people are willing to pay only about \$20 to \$30 to play the game. The flaw in formulating the problem as one of choosing the option that maximizes your expected net gain does not take into account your attitude towards risk taking.

Problem 16. * Mean and variance of the discrete uniform. Consider the random variable X that takes the integer values $a, a+1, \ldots, b$ with equal probability $1/(b-a+1)$. Compute its mean and variance.

*Solution.*By symmetry, the mean is $\mathbf{E}[X] = (a+b)/2$. Consider a translation of X, and in particular the special case where $a = 1$ and $b = n$. Then the mean is $\mathbf{E}[X] = (1+n)/2$. Let us show by induction that the second moment is

$$\mathbf{E}[X^2] = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{1}{6}(n+1)(2n+1).$$

For $n = 1$ we have $\mathbf{E}[X^2] = 1$ and the above formula holds. Assume that the formula holds for X uniformly distributed in the range $1, \ldots, n$, i.e.,

$$\frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{1}{6}(n+1)(2n+1).$$

We must show the formula for X uniformly distributed in the range $1, \ldots, n+1$, i.e.,

$$\frac{1}{n+1}\sum_{k=1}^{n+1} k^2 = \frac{1}{6}(n+2)(2n+3).$$

Indeed, we have using the induction hypothesis

$$\begin{aligned}\frac{1}{n+1}\sum_{k=1}^{n+1} k^2 &= \frac{(n+1)^2}{n+1} + \frac{n}{n+1}\frac{1}{n}\sum_{k=1}^{n} k^2 \\ &= n+1+\frac{n}{n+1}\frac{1}{6}(n+1)(2n+1) \\ &= n+1+\frac{n(2n+1)}{6} \\ &= \frac{1}{6}\big(6(n+1)+n(2n+1)\big) \\ &= \frac{1}{6}(n+2)(2n+3).\end{aligned}$$

The variance is

$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{1}{6}(n+1)(2n+1) - \frac{(1+n)^2}{4} = \frac{n^2-1}{12},$$

as shown in Example 2.5. For the case of general values a and b, we note that the variance of a random variable is unaffected by a translation, so as in Example 2.5, $\text{var}(X)$ is obtained by replacing n by $b-a+1$ in the above formula.

Problem 17. * Mean and variance of the Poisson. Consider the Poisson random variable with parameter λ. Compute the second moment and the variance.

*Solution.*As shown in the text, the mean is given by

$$\mathbf{E}[X] = \lambda.$$

Also

$$\begin{aligned}\mathbf{E}[X^2] &= \sum_{k=1}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \lambda\sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda\sum_{m=0}^{\infty}(m+1)\frac{e^{-\lambda}\lambda^m}{m!} \\ &= \lambda\big(\mathbf{E}[X]+1\big) = \lambda(\lambda+1),\end{aligned}$$

from which

$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

SECTION 2.5. Joint PMFs of Multiple Random Variables

Problem 18. PMF of the minimum of several random variables. On any given day your golf score is any integer from 101 to 110, each with probability 0.1. Determined to improve your score, you decide to play with three balls and declare as your score the minimum X of the scores X_1, X_2, and X_3 obtained with balls 1, 2, and 3, respectively.

(a) Calculate the PMF of X.

(b) By how much has your expected score improved as a result of using three balls?

Solution.(a) Consider the possible values of x_1, x_2 and x_3 as coordinates in a cubical lattice of side 10. We can count all the points corresponding to $\min(x_1, x_2, x_3) = 101$. These points are trivial to count geometrically! They are the points in a cubical lattice of side 10, minus the points in a cubical lattice of side 9. Similarly, we can count the number of points corresponding to $\min(x_1, x_2, x_3) = (111 - k)$ as the number of points in a cubical lattice of side k, minus the number of points in a cubical lattice of side $(k - 1)$. Since this problem follows a discrete uniform law, this counting solution is sufficient to find the PMF:

$$p_X(k) = \begin{cases} \frac{(111-k)^3-(110-k)^3}{10^3} & 100 < k \leq 110, \\ 0 & \text{otherwise.} \end{cases}$$

(An alternative solution can be based on the notion of a CDF, which will be introduced in Chapter 3.)

(b) By symmetry, the expected value of the score on any particular ball is 105.5. This can also be seen using the formula for expected value:

$$\mathbf{E}[X_i] = \sum_{k=101}^{110} k \cdot p_{X_i}(k) = \sum_{k=101}^{110} k\frac{1}{10} = 105.5.$$

We now need to calculate the expected value of X and compare it to the above:

$$\mathbf{E}[X] = \sum_{k=-\infty}^{\infty} k \cdot p_X(k) = \sum_{k=101}^{110} k \cdot p_x(k) = \sum_{k=101}^{110} k\frac{(111-k)^3-(110-k)^3}{10^3} = 103.025.$$

The expected improvement is therefore 105.5 - 103.025 = 2.475.

Problem 19. * The quiz problem. Consider a quiz contest where a person is given a list of N questions and can answer these questions in any order he or she chooses. Question i will be answered correctly with probability p_i, and the person will then receive a reward V_i. At the first incorrect answer, the quiz terminates and the person is allowed to keep his or her previous rewards. The problem is to choose the ordering of questions so as to maximize the expected value of total reward obtained. Show that it is optimal to answer questions in a nonincreasing order of $p_iV_i/(1-p_i)$.

*Solution.*We will use a so-called interchange argument, which is often useful in scheduling-like problems. Let i and j be the kth and $(k+1)$st questions in an optimally ordered list

$$L = (i_1, \ldots, i_{k-1}, i, j, i_{k+2}, \ldots, i_N).$$

Consider the list

$$L' = (i_1, \ldots, i_{k-1}, j, i, i_{k+2}, \ldots, i_N)$$

obtained from L by interchanging the order of questions i and j. We compute the expected values of the rewards of L and L', and argue that since L is optimally ordered, we have

$$\mathbf{E}[\text{reward of } L] \geq \mathbf{E}[\text{reward of } L'].$$

Define the weight of question i to be

$$W(i) = \frac{p_iV_i}{(1-p_i)}.$$

We have to show that any permutation of the question in a non-increasing order of weights maximizes the expected reward.

If $L = (i_1, \ldots, i_N)$ is a permutation of the questions, define $L^{(k)}$ to be the permutation obtained from L by interchanging questions i_k and i_{k+1}. Let us first compute the difference between the expected reward of L and that of $L^{(k)}$. We have

$$\mathbf{E}[\text{reward of } L] = p_{i_1}V_{i_1} + p_{i_1}p_{i_2}V_{i_2} + \cdots + p_{i_1}\cdots p_{i_N}V_{i_N}$$

and

$$\begin{aligned}\mathbf{E}[\text{reward of } L^{(k)}] =& p_{i_1}V_{i_1} + p_{i_1}p_{i_2}V_{i_2} + \cdots + p_{i_1}\cdots p_{i_{k-1}}V_{i_{k-1}}\\ &+ p_{i_1}\cdots p_{i_{k-1}}p_{i_{k+1}}V_{i_{k+1}} + p_{i_1}\cdots p_{i_{k-1}}p_{i_{k+1}}p_{i_k}V_{i_k}\\ &+ p_{i_1}\cdots p_{i_{k+1}}V_{i_{k+1}} + \cdots + p_{i_1}\cdots p_{i_N}V_{i_N}\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{E}[\text{reward of } L^{(k)}] - \mathbf{E}[\text{reward of } L] &= p_{i_1}\cdots p_{i_{k-1}}\big(p_{i_{k+1}}V_{i_{k+1}} + p_{i_{k+1}}p_{i_k}V_{i_k}\\ &\quad - p_{i_k}V_{i_k} - p_{i_k}p_{i_{k+1}}V_{i_{k+1}}\big)\\ &= p_{i_1}\cdots p_{i_{k-1}}(1-p_{i_k})(1-p_{i_{k+1}})\big(W(i_{k+1}) - W(i_k)\big).\end{aligned}$$

Now, let us go back to our problem. Consider any permutation L of the questions. If $W(i_k) < W(i_{k+1})$ for some k, it follows from the above equation that the permutation $L^{(k)}$ has an expected reward larger than that of L. So, an optimal permutation of the questions must be in a nonincreasing order of weights.

We still have to argue that any two such permutations have equal expected rewards. Assume that L is such a permutation and say that $W(i_k) = W(i_{k+1})$ for some k. We know that interchanging i_k and i_{k+1} preserve the expected reward. So, the expected reward of any permutation L' in a non-increasing order of weights is equal to that of L because L' can be obtained from L by repeatedly interchanging adjacent questions having equal weights.

Problem 20. * Verify the expected value rule

$$\mathbf{E}\big[g(X,Y)\big] = \sum_{x,y} g(x,y)p_{X,Y}(x,y),$$

and its linear special case:

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y],$$

where a and b are given scalars.

*Solution.*Using the total expectation theorem to reduce the problem to the case of a single random variable, we obtain

$$\begin{aligned}\mathbf{E}\big[g(X,Y)\big] &= \sum_y p_Y(y)\mathbf{E}\big[g(X,Y)\,|\,y\big]\\ &= \sum_y p_Y(y)\sum_x g(x,y)p_{X|Y}(x|y)\\ &= \sum_{x,y} g(x,y)p_{X,Y}(x,y)\end{aligned}$$

as desired. For the linear special case, we have

$$\begin{aligned}\mathbf{E}[aX+bY] &= \sum_{x,y}(ax+by)p_{X,Y}(x,y)\\ &= a\sum_x x\sum_y p_{X,Y}(x,y) + b\sum_y y\sum_x p_{X,Y}(x,y)\\ &= a\sum_x xp_X(x) + b\sum_y yp_Y(y)\\ &= a\mathbf{E}[X] + b\mathbf{E}[Y].\end{aligned}$$

SECTION 2.6. Conditioning

Problem 21. We have m married couples ($2m$ individuals). After a number of years, each person has died, independently, with probability p. Let N be the number of surviving individuals. Let C be the number of couples in which both individuals are alive. Find $E[C \,|\, N=n]$.

*Solution.*Let $X_i = 1$ if the male of the ith couple survives, $Y_i = 1$ if the female of the ith couple survives. (Let X_i and Y_i be zero otherwise.) Then, $C = \sum_{i=1}^m X_iY_i$. So, it suffices to find $E[X_iY_i \,|\, N=n]$. Given $N=n$, X_i has probability $n/2m$ of being alive. Given that $N=n$ and X_i is alive, Y_i has probability $(n-1)/(2m-1)$ of being alive. So, $E[X_iY_i \,|\, N=n] = n(n-1)/(2m)(2m-1)$, and $E[C \,|\, N=n] = n(n-1)/2(2m-1)$.

Problem 22. A scalper is considering buying tickets for a particular game. The price of the tickets is \$75, and the scalper will sell them at \$150. However, if she can't sell them at \$150, she won't sell them at all. Given that the demand for tickets is a binomial random variable with parameters $n = 10$ and $p = 1/2$, how many tickets should she buy in order to maximize her expected profit?

*Solution.*Let Q and b be the numbers of tickets demanded and bought, respectively. If S is the number of tickets sold, then $S = \min(Q, b)$. The scalper's expected profit is

$$r(b) = \mathbf{E}[150S - 75b] = 150\mathbf{E}[S] - 75b.$$

We first find $\mathbf{E}[S]$. We assume that $b \le 10$, since clearly buying more than the maximum number of demanded tickets, which is 10, cannot be optimal. We have

$$\begin{aligned}\mathbf{E}[S] &= \mathbf{E}[Q \,|\, Q \le b)]\mathbf{P}(Q \le b) + \mathbf{E}[Q \,|\, Q > b)]\mathbf{P}(Q > b)\\ &= \sum_{i=0}^{b} i\binom{10}{i}\left(\frac{1}{2}\right)^{10} + b\sum_{i=b+1}^{10}\binom{10}{i}\left(\frac{1}{2}\right)^{10}\\ &= \left(\frac{1}{2}\right)^{10}\left(\sum_{i=0}^{b} i\binom{10}{i} + b\sum_{i=0}^{b}\binom{10}{i}\right).\end{aligned}$$

Thus

$$r(b) = 150\left(\frac{1}{2}\right)^{10}\left(\sum_{i=0}^{b} i\binom{10}{i} + b\sum_{i=0}^{b}\binom{10}{i}\right) - 75b.$$

Intuitively, $r(b)$ first increases with n and then decreases. Thus, to find when $r(b)$ is maximized, we find the first value of b for which $r(b+1) - r(b) \leq 0$.

Problem 23. * D. Bernoulli's problem of joint lives. Consider $2m$ persons forming m couples who live together at a given time. Suppose that at some later time, the probability of each person being alive is p, independently of other persons. At that later time, let A be the number of persons that are alive and let S be the number of couples in which both partners are alive. For any survivor number a, find $\mathbf{E}[S \,|\, A = a]$.

*Solution.*Let X_i be the random variable taking the value 1 or 0 depending on whether the first partner of the ith couple has survived or not. Let Y_i be the corresponding random variable for the second partner of the ith couple. Then we have $S = \sum_{i=1}^m X_i Y_i$ and by using the total expectation theorem,

$$\begin{aligned}
\mathbf{E}[S \,|\, A = a] &= \sum_{i=1}^{m} \mathbf{E}[X_i Y_i \,|\, A = a] \\
&= m\mathbf{E}[X_1 Y_1 \,|\, A = a] \\
&= m\mathbf{E}[Y_1 = 1 \,|\, X_1 = 1,\, A = a]\mathbf{P}(X_1 = 1 \,|\, A = a) \\
&= m\mathbf{P}[Y_1 = 1 \,|\, X_1 = 1,\, A = a]\mathbf{P}(X_1 = 1 \,|\, A = a).
\end{aligned}$$

We have

$$\mathbf{P}(Y_1 = 1 \,|\, X_1 = 1,\, A = a) = \frac{a-1}{2m-1}, \qquad \mathbf{P}(X_1 = 1 \,|\, A = a) = \frac{a}{2m}.$$

Thus

$$\mathbf{E}[S \,|\, A = a] = m\,\frac{a-1}{2m-1}\,\frac{a}{2m}.$$

Note that $\mathbf{E}[S \,|\, A = a]$ does not depend on p.

Problem 24. * A biased coin is tossed successively until a head comes twice in a row or a tail comes twice in a row. The probability of a head is p and the probability of a tail is $q = 1 - p$. Find the expected value of the number of tosses until the game is over.

*Solution.*One possibility here is to calculate the PMF of X, the number of tosses until the game is over, and use it to compute $\mathbf{E}[X]$. However, this turns out to be cumbersome, so we argue by using the total expectation theorem and a suitable partition of the sample space. Let H_k (or T_k) be the event that a head (or a tail, respectively) comes at the kth toss. Since H_1 and T_1 form a partition of the sample space, and $\mathbf{P}(H_1) = p$ and $\mathbf{P}(T_1) = q$, we have

$$\mathbf{E}[X] = p\mathbf{E}[X \,|\, H_1] + q\mathbf{E}[X \,|\, T_1].$$

Using again the total expectation theorem, we have

$$\mathbf{E}[X \,|\, H_1] = p\mathbf{E}[X \,|\, H_1 \cap H_2] + q\mathbf{E}[X \,|\, H_1 \cap T_2] = 2p + q\big(1 + \mathbf{E}[X \,|\, T_1]\big),$$

where we have used the fact $\mathbf{E}[X \,|\, H_1 \cap H_2] = 2$ (since the game ends after two successive heads), and $\mathbf{E}[X \,|\, H_1 \cap T_2) = 1 + \mathbf{E}[X \,|\, T_1]$ (since if the game is not over, only the last toss matters in determining the number of additional tosses up to termination). Similarly, we obtain

$$\mathbf{E}[X \mid T_1] = 2q + p\big(1 + \mathbf{E}[X \mid H_1]\big).$$

Combining the above two relations, collecting terms, and using the fact $p + q = 1$, we obtain after some calculation

$$\mathbf{E}[X \mid T_1] = \frac{2 + p^2}{1 - pq},$$

and similarly

$$\mathbf{E}[X \mid H_1] = \frac{2 + q^2}{1 - pq}.$$

Thus,

$$\mathbf{E}[X] = p \cdot \frac{2 + q^2}{1 - pq} + q \cdot \frac{2 + p^2}{1 - pq},$$

and finally, using the fact $p + q = 1$,

$$\mathbf{E}[X] = \frac{2 + pq}{1 - pq}.$$

In the case of a fair coin ($p = q = 1/2$), we obtain $\mathbf{E}[X] = 3$. It can also be verified that $2 \leq \mathbf{E}[X] \leq 3$ for all values of p.

Problem 25. * A spider and a fly move along a straight line. At each second, the fly moves a unit step to the right or to the left with equal probability p, and stays where it is with probability $1 - 2p$. The spider always takes a unit step in the direction of the fly. The spider and the fly start D units apart, where D is a random variable taking positive integer values with a given PMF. If the spider lands on top of the fly, it's the end. What is the expected value of T, the time it takes for this to happen?

*Solution.*Denote:

A_d the event that initially the spider and the fly are d units apart,

B_d the event that after one second the spider and the fly are d units apart.

Our approach will be to first apply the (conditional version of the) total expectation theorem to compute $\mathbf{E}[T \mid A_1]$, then use the result to compute $\mathbf{E}[T \mid A_2]$, and similarly compute sequentially $\mathbf{E}[T \mid A_d]$ for all relevant values of d. We will then apply the (unconditional version of the) total expectation theorem to compute $\mathbf{E}[T]$.

We have

$$A_d = (A_d \cap B_d) \cup (A_d \cap B_{d-1}) \cup (A_d \cap B_{d-2}), \qquad \text{if } d > 1.$$

This is because if the spider and the fly are at a distance $d > 1$ apart, then one second later their distance will be d (if the fly moved away from the spider) or $d - 1$ (if the fly did not move) or $d - 2$ (if the fly moved towards the spider). We also have, for the case where the spider and the fly start one unit apart,

$$A_1 = (A_1 \cap B_1) \cup (A_1 \cap B_0).$$

Using the total expectation theorem, we obtain

$$\begin{aligned}\mathbf{E}[T \mid A_d] = {} & \mathbf{P}(A_d \cap B_d)\mathbf{E}[T \mid A_d \cap B_d] \\ & + \mathbf{P}(A_d \cap B_{d-1})\mathbf{E}[T \mid A_d \cap B_{d-1}] \qquad \text{if } d > 1, \\ & + \mathbf{P}(A_d \cap B_{d-2})\mathbf{E}[T \mid A_d \cap B_{d-2}]\end{aligned}$$

while for the case $d = 1$,

$$\mathbf{E}[T \mid A_1] = \mathbf{P}(A_1 \cap B_1)\mathbf{E}[T \mid A_1 \cap B_1] + \mathbf{P}(A_1 \cap B_0)\mathbf{E}[T \mid A_1 \cap B_0].$$

It can be seen based on the problem data that

$$\mathbf{P}(A_1 \cap B_1) = 2p, \qquad \mathbf{P}(A_1 \cap B_0) = 1 - 2p,$$

$$\mathbf{E}[T \mid A_1 \cap B_1] = 1 + \mathbf{E}[T \mid A_1], \qquad \mathbf{E}[T \mid A_1 \cap B_0] = 1,$$

so by applying the theorem with $d = 1$, we obtain

$$\mathbf{E}[T \mid A_1] = 2p\big(1 + \mathbf{E}[T \mid A_1]\big) + (1 - 2p),$$

or

$$\mathbf{E}[T \mid A_1] = \frac{1}{1 - 2p}.$$

By applying the theorem with $d = 2$, we obtain

$$\mathbf{E}[T \mid A_2] = p\mathbf{E}[T \mid A_2 \cap B_2] + (1 - 2p)\mathbf{E}[T \mid A_2 \cap B_1] + p\mathbf{E}[T \mid A_2 \cap B_0].$$

We have

$$\begin{aligned}\mathbf{E}[T \mid A_2 \cap B_0] &= 1, \\ \mathbf{E}[T \mid A_2 \cap B_1] &= 1 + \mathbf{E}[T \mid A_1], \\ \mathbf{E}[T \mid A_2 \cap B_2] &= 1 + \mathbf{E}[T \mid A_2],\end{aligned}$$

so by substituting these relations in the expression for $\mathbf{E}[T \mid A_2]$, we obtain

$$\begin{aligned}\mathbf{E}[T \mid A_2] &= p\big(1 + \mathbf{E}[T \mid A_2]\big) + (1 - 2p)\big(1 + \mathbf{E}[T \mid A_1]\big) + p \\ &= p\big(1 + \mathbf{E}[T \mid A_2]\big) + (1 - 2p)\left(1 + \frac{1}{1 - 2p}\right) + p.\end{aligned}$$

This equation yields after some calculation

$$\mathbf{E}[T \mid A_2] = \frac{2}{1 - p}.$$

Generalizing, we obtain

$$\mathbf{E}[T \mid A_d] = p\big(1 + \mathbf{E}[T \mid A_d]\big) + (1 - 2p)\big(1 + \mathbf{E}[T \mid A_{d-1}]\big) + p\big(1 + \mathbf{E}[T \mid A_{d-2}]\big)$$

so $\mathbf{E}[T \mid A_d]$ can be generated recursively for any initial distance d, using as initial conditions the values of $\mathbf{E}[T \mid A_1]$ and $\mathbf{E}[T \mid A_2]$ obtained earlier.

Finally, the expected value of T can be obtained using the given PMF for the initial distance D and the total expectation theorem:

$$\mathbf{E}[T] = \sum_d p_D(d)\mathbf{E}[T \mid A_d].$$

Problem 26. * The multiplication rule for conditional PMFs. Let X, Y, and Z be random variables.

(a) Show that

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y \mid X}(y \mid x)p_{Z \mid X,Y}(z \mid x,y).$$

(b) How can we interpret this formula as a special case of the multiplication rule given in Section 1.3?

(c) Generalize to the case of more than three random variables.

Solution. (a) We have

$$\begin{aligned} p_{X,Y,Z}(x,y,z) &= \mathbf{P}\big(\{X=x\} \cap \{Y=y\} \cap \{Z=z\}\big) \\ &= \mathbf{P}(X=x)\mathbf{P}\big(\{Y=y\} \cap \{Z=z\} \mid X=x\big) \\ &= \mathbf{P}(X=x)\mathbf{P}\big(Y=y \mid X=x\big)\mathbf{P}\big(Z=z \mid \{X=x\} \cap \{Y=y\}\big) \\ &= p_X(x)p_{Y \mid X}(y \mid x)p_{Z \mid X,Y}(z \mid x,y). \end{aligned}$$

(b) The formula can be written as

$$\begin{aligned} &\mathbf{P}\big(\{X=x\} \cap \{Y=y\} \cap \{Z=z\}\big) \\ &\qquad = \mathbf{P}(X=x)\mathbf{P}\big(Y=y \mid X=x\big)\mathbf{P}\big(Z=z \mid \{X=x\} \cap \{Y=y\}\big), \end{aligned}$$

which is a special case of the multiplication rule.

(c) The generalization is

$$\begin{aligned} &p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \\ &\qquad = p_{X_1}(x_1)p_{X_2|X_1}(x_2|x_1)\cdots p_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1,\ldots,x_{n-1}). \end{aligned}$$

Problem 27. * Splitting a Poisson random variable. A transmitter sends out either a 1 with probability p, or a 0 with probability $(1-p)$, independently of earlier transmissions. If the number of transmissions within a given time interval is Poisson with parameter λ, show that the number of 1's transmitted in that same time interval is also Poisson, and has parameter $p\lambda$.

*Solution.*Let X and Y be the numbers of 1's and 0's transmitted, respectively. Let $Z = X + Y$ be the total number of symbols transmitted. We have

$$\begin{aligned} \mathbf{P}(X=n, Y=m) &= \mathbf{P}(X=n, Y=m \mid Z=n+m)\mathbf{P}(Z=n+m) \\ &= \binom{n+m}{n} p^n (1-p)^m \frac{e^{-\lambda}\lambda^{n+m}}{(n+m)!} \\ &= \frac{e^{-\lambda p}(\lambda p)^n}{n!} \, \frac{e^{-\lambda(1-p)}\big(\lambda(1-p)\big)^m}{m!}. \end{aligned}$$

Thus

$$\begin{aligned}\mathbf{P}(X=n) &= \sum_{m=0}^{\infty}\mathbf{P}(X=n, Y=m)\\ &= \frac{e^{-\lambda p}(\lambda p)^n}{n!}e^{-\lambda(1-p)}\sum_{m=0}^{\infty}\frac{\big(\lambda(1-p)\big)^m}{m!}\\ &= \frac{e^{-\lambda p}(\lambda p)^n}{n!}e^{-\lambda(1-p)}e^{\lambda(1-p)}\\ &= \frac{e^{-\lambda p}(\lambda p)^n}{n!},\end{aligned}$$

so that X is Poisson with parameter λp.

SECTION 2.7. Independence

Problem 28. Alice passes through four traffic lights on her way to work, and each light is equally likely to be green or red, independently of the others.

(a) What is the PMF, the mean, and the variance of the number of red lights that Alice encounters?

(b) Suppose that each red light delays Alice by exactly two minutes. What is the variance of Alice's commuting time?

Solution.(a) Let X be the number of red lights that Alice encounters. The PMF of X is binomial with $n = 4$ and $p = 1/2$. The mean and the variance of X are $\mathbf{E}[X] = np = 2$ and $\text{var}(X) = np(1-p) = 4 \cdot (1/2) \cdot (1/2) = 1$.

(b) The variance of Alice's commuting time is the same as the variance of the time by which Alice is delayed by the red lights. This is equal to the variance of $2X$, which is $4\text{var}(X) = 4$.

Problem 29. Your computer has been acting very strangely lately, and you suspect that it might have a virus on it. Unfortunately, all 12 of the different virus detection programs you own are somewhat outdated. You know that if your computer really does have a virus, each of the programs, independently of the others, has a 0.8 chance of correctly identifying your computer to be infected, and a 0.2 chance of thinking your computer is fine. On the other hand, if your computer does not have a virus, each program has a 0.9 chance of believing your computer to be free from viruses, and a 0.1 chance of wrongly thinking your computer is infected. Given that your computer has a 0.65 chance of being infected with some virus, and given that you will believe your virus protection programs only if 9 or more of them agree, find the probability that your detection programs will lead you to the right answer.

*Solution.*Let A denote the event that your detection programs lead you to the correct conclusion about your computer. Let V be the event that your computer has a virus, and let V^c be the event that your computer does not have a virus. We have

$$\mathbf{P}(A) = \mathbf{P}(V)\mathbf{P}(A \,|\, V) + \mathbf{P}(V^c)\mathbf{P}(A \,|\, V^c),$$

and $\mathbf{P}(A\,|\,V)$ and $\mathbf{P}(A\,|\,V^c)$ can be found using the binomial PMF. Thus we have

$$\begin{aligned}\mathbf{P}(A\,|\,V) = &\binom{12}{9}\cdot(0.8)^9\cdot(0.2)^3+\binom{12}{10}\cdot(0.8)^{10}\cdot(0.2)^2\\ &+\binom{12}{11}\cdot(0.8)^{11}\cdot(0.2)^1+\binom{12}{12}\cdot(0.8)^{12}\cdot(0.2)^0\\ &=0.7899.\end{aligned}$$

Similarly we find that $\mathbf{P}(A\,|\,V^c)=0.9742$, so that

$$\mathbf{P}(A)=0.65\cdot 0.7899+0.35\cdot 0.9742=0.8544.$$

Problem 30. A particular professor is known for his arbitrary grading policies. In fact, on any given paper, he gives any grade from the set $\{A, A-, B+, B, B-, C+\}$ with equal probability. How many papers would you expect to hand in before having received at least one of every grade in the range?

*Solution.*Let X_6 be the number of papers until each grade is received at least once. Let X_5 be the number of papers until a certain 5 out of the six grades is received at least once, and so on with X_4, X_3, etc. Then we have

$$\begin{aligned}\mathbf{E}[X_6]&=1+\mathbf{E}[X_5],\\ \mathbf{E}[X_5]&=\frac{1}{6}\big(1+\mathbf{E}[X_5]\big)+\frac{5}{6}\big(1+\mathbf{E}[X_4]\big),\\ \mathbf{E}[X_4]&=\frac{1}{3}\big(1+\mathbf{E}[X_4]\big)+\frac{2}{3}\big(1+\mathbf{E}[X_3]\big),\\ \mathbf{E}[X_3]&=\frac{1}{2}\big(1+\mathbf{E}[X_3]\big)+\frac{1}{2}\big(1+\mathbf{E}[X_2]\big),\\ \mathbf{E}[X_2]&=\frac{2}{3}\big(1+\mathbf{E}[X_2]\big)+\frac{1}{3}\big(1+\mathbf{E}[X_1]\big),\\ \mathbf{E}[X_1]&=\frac{5}{6}\big(1+\mathbf{E}[X_1]\big)+\frac{1}{6}(1),\end{aligned}$$

and solving for $\mathbf{E}[X_6]$, gives $\mathbf{E}[X_6]=14.7$.

Alternative solution: Associate a success with a paper if its grade has not been received before. Let X_i be the number of papers between the ith success and the $(i+1)$st success. Then we have $X=1+\sum_{i=1}^{5}X_i$ and hence

$$\mathbf{E}[X]=1+\sum_{i=1}^{5}\mathbf{E}[X_i].$$

The random variable X_i is geometric with parameter $p_i=\frac{6-i}{6}$, so $\mathbf{E}[X_i]=\frac{6}{6-i}$. It follows that

$$\mathbf{E}[X]=1+\sum_{i=1}^{5}\frac{6}{6-i}=1+6\sum_{i=1}^{5}\frac{1}{i}=14.7.$$

Problem 31. Each morning, Hungry Harry eats some eggs. On any given morning, the number of eggs he eats is equally likely to be 1, 2, 3, 4, 5, or 6, independently of what he has done in the past. Let X be the number of eggs Harry eats in 10 days. Find the mean and variance of X.

Solution. Let X_i be the number of eggs Harry eats on day i. Then $X = \sum_{i=1}^{10} X_i$, and therefore

$$\mathbf{E}[X] = \mathbf{E}\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} \mathbf{E}[X_i] = 35.$$

Similarly, we have

$$\text{var}(X) = \text{var}\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} \text{var}(X_i)$$

since the X_i are independent. We can verify using the formula of Example 2.5, that $\text{var}(X_i) \approx 2.9167$, so we have $\text{var}(X) \approx 29.167$.

Problem 32. Computational problem. Here is a probabilistic method for computing the area of a given subset S of the unit square. The method uses a sequence of independent random selections of points in the unit square $[0,1] \times [0,1]$, according to a uniform probability law. If the ith point belongs to the subset S the value of a random variable X_i is set to 1, and otherwise it is set to 0. Let $X_1, X_2, \ldots$ be the sequence of random variables thus defined, and for any n, let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

(a) Show that $\mathbf{E}[S_n]$ is equal to the area of the subset S, and that $\text{var}(S_n)$ diminishes to 0 as n increases.

(b) Show that if S_{n-1} and X_n are known, this is enough to calculate S_n, so the past X_k, $k = 1, \ldots, n-1$, do not need to be remembered. Give a formula.

(c) Write a computer program to generate S_n for $n = 1, 2, \ldots, 10000$ using the computer's random number generator, for the case where the subset S is the circle inscribed within the unit square. How can you use your program to measure experimentally the value of π?

(d) Use a similar computer program to calculate approximately the area of the set of all (x, y) that lie within the unit square and satisfy $0 \le \cos \pi x + \sin \pi y \le 1$.

Solution.(a) Noting that

$$\mathbf{P}(X_i = 1) = \frac{\text{Area}(S)}{\text{Area}\big([0,1] \times [0,1]\big)} = \text{Area}(S),$$

we obtain

$$\mathbf{E}[S_n] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[X_i] = \mathbf{E}[X_i] = \text{Area}(S),$$

and

$$\mathrm{var}(S_n) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n}\mathrm{var}(X_i) = \frac{1}{n}\big(1 - \mathrm{Area}(S)\big)\mathrm{Area}(S),$$

which tends to zero as n tends to infinity.

(b)

$$S_n = \frac{n-1}{n}S_{n-1} + \frac{1}{n}X_n.$$

Problem 33. * Two biased coins are simultaneously tossed until one of them comes up a head and the other a tail. The first coin comes up a head with probability p and the second with probability q.

(a) Find the PMF, the expected value, and the variance of the number of tosses until this happens.

(b) Assuming that $p = q$, what is the probability that the last toss of the first coin is a head?

(c) (Von Neumann coin) Suggest a method for generating a fair coin flip from a coin whose probability of coming up heads is unknown. *Hint:* Use part (b).

Solution.(a) Let X be the number of tosses until the game is over. Noting that X is geometric with probability of success

$$\mathbf{P}(\{HT, TH\}) = p(1-q) + q(1-p),$$

we obtain

$$p_X(k) = (pq + (1-p)(1-q))^{k-1}(p(1-q) + q(1-p)), \qquad k = 1, 2, \ldots$$

Therefore

$$\mathbf{E}[X] = \frac{1}{p(1-q) + q(1-p)}$$

and

$$\mathrm{var}(X) = \frac{pq + (1-p)(1-q)}{\big(p(1-q) + q(1-p)\big)^2}.$$

(b) The probability that the last toss of the first coin is a head is

$$\mathbf{P}(HT \mid \{HT, TH\}) = \frac{p(1-p)}{p(1-p) + (1-p)p} = \frac{1}{2}.$$

(c) Keep on flipping the coin twice until two different outcomes appears. If the first flip in the last pair of flips is a head declare a head, else declare a tail. The resulting flip is fair, because we know from part (b) that the probability of declaring a head is 1/2.

Problem 34. *

(a) A fair coin is tossed successively until two consecutive heads or two consecutive tails appear. Find the PMF, the expected value, and the variance of the number of tosses until this happens.

(b) Assume now that the coin is tossed successively until a head followed by a tail appear. Find the PMF and the expected value of the number of tosses until until this happens.

*Solution.*Let X be the total number of tosses.

(a) If $x \geq 2$, there are only two possible sequences of outcomes that lead to the event $\{X = x\}$, the sequences HH and TT. Therefore,

$$\mathbf{P}(X = x) = 2(1/2)^x = (1/2)^{x-1}.$$

It follows that

$$p_X(x) = \begin{cases} (1/2)^{x-1} & \text{if } x \geq 2, \\ 0 & \text{otherwise.} \end{cases}$$

and

$$\mathbf{E}[X] = \sum_{x=2}^{\infty} x(1/2)^{x-1} = 2\left(-1/2 + \sum_{x=1}^{\infty} x(1/2)^x\right) = 2\big(-1/2 + 1/(1/2)\big) = 3.$$

To find the variance of X, we first compute $\mathbf{E}[X^2]$. We have

$$\mathbf{E}[X^2] = \sum_{x=2}^{\infty} x^2(1/2)^{x-1} = 2\left(-1/2 + \sum_{x=1}^{\infty} x^2(1/2)^x\right) = -1 + 2 \cdot 6 = 11.$$

Thus

$$\text{var}(X) = 11 - 3^2 = 2.$$

(b) If $x > 2$, there are $x - 1$ sequences of outcomes that lead to the event $\{X = x\}$, the first contains only heads in the first $x - 2$ trials and the others are determined by the index of last occurrence of tail in the first $x - 2$ trials. For the case when $x = 2$, there is only one (hence $x - 1$) possible sequences of outcomes that lead to the event $\{X = x\}$. Therefore, for any $x \geq 2$,

$$\mathbf{P}(X = x) = (x - 1)(1/2)^x.$$

It follows that

$$p_X(x) = \begin{cases} (x-1)(1/2)^x & \text{if } x \geq 2, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathbf{E}[X] = \sum_{x=2}^{\infty} x(x-1)(1/2)^x = \sum_{x=1}^{\infty} x^2(1/2)^x - \sum_{x=1}^{\infty} x(1/2)^x = 4.$$

Problem 35. * Suppose that X and Y are independent, identically distributed, geometric random variables with parameter p. Show that

$$\mathbf{P}(X = i \,|\, X + Y = n) = \frac{1}{n-1}, \qquad i = 1, \ldots, n-1.$$

*Solution.*We can interpret $\mathbf{P}(X = i \,|\, X + Y = n)$ as the probability that a coin will come up a head for the first time on the ith toss given that it came up a head for the second time on the nth toss. We can then argue, intuitively, that given that the second head occurred on the nth toss, the first head is equally likely to have come up at any toss between 1 and $n - 1$. To establish this precisely, note that we have

$$\mathbf{P}(X = i \,|\, X + Y = n) = \frac{\mathbf{P}(X = i \text{ and } X + Y = n)}{\mathbf{P}(X + Y = n)} = \frac{\mathbf{P}(X = i)\mathbf{P}(Y = n - i)}{\mathbf{P}(X + Y = n)}.$$

Also $\mathbf{P}(X = i) = p(1-p)^{i-1}$ for $i \geq 1$, and $\mathbf{P}(Y = n-i) = p(1-p)^{n-i-1}$ for $n - i \geq 1$. It follows that $\mathbf{P}(X = i)\mathbf{P}(Y = n - i) = p^2(1 - p)^{n-2}$ for $i = 1, \ldots, n - 1$ and 0 otherwise. Therefore $\mathbf{P}(X = i \,|\, X + Y = n) = \mathbf{P}(X = j \,|\, X + Y = n)$ for any i and j between 1 and $n - 1$. Hence

$$\mathbf{P}(X = i \,|\, X + Y = n) = \frac{1}{n-1}, \qquad i = 1, \ldots, n-1.$$

Problem 36. * Let X and Y be two random variables with given joint PMF, and let g and h be two functions of X and Y, respectively. Show that if X and Y are independent, then the same is true for the random variables $g(X)$ and $h(Y)$.

*Solution.*Let $G = g(X)$ and $H = h(Y)$. The independence of G and H follows from the calculation

$$\begin{aligned} p_{G,H}(g,h) &= \sum_{\{x,y \,|\, g(x)=g,\, h(y)=h\}} p_{X,Y}(x,y) \\ &= \sum_{\{x,y \,|\, g(x)=g,\, h(y)=h\}} p_X(x)p_Y(y) \\ &= \sum_{\{x \,|\, g(x)=g\}} p_X(x) \sum_{\{x \,|\, h(y)=h\}} p_Y(y) \\ &= p_G(g)p_H(h). \end{aligned}$$

Problem 37. * Variability extremes. Let $X_1, \ldots, X_n$ be independent random variables and let $X = X_1 + \cdots + X_n$ be their sum.

(a) Suppose that each X_i is Bernoulli with parameter p_i, and that $p_1, \ldots, p_n$ are chosen so that the mean of X is a given $\mu > 0$. Show that the variance of X is maximized if the p_i are chosen to be all equal to μ/n.

(b) Suppose that each X_i is geometric with parameter p_i, and that $p_1, \ldots, p_n$ are chosen so that the mean of X is a given $\mu > 0$. Show that the variance of X is minimized if the p_i are chosen to be all equal to n/μ. [Note the strikingly different character of the results of parts (a) and (b).]

Solution.(a) We have

$$\text{var}(X) = \sum_{i=1}^{n} \text{var}(X_i) = \sum_{i=1}^{n} p_i(1-p_i) = \mu - \sum_{i=1}^{n} p_i^2.$$

Thus maximizing the variance is equivalent to minimizing $\sum_{i=1}^{n} p_i^2$. It can be seen (using the constraint that $\sum_{i=1}^{n} p_i = \mu$) that

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} (\mu/n)^2 + \sum_{i=1}^{n} (p_i - \mu/n)^2,$$

so $\sum_{i=1}^{n} p_i^2$ is minimized when $p_i = \mu/n$ for all i.

(b) We have

$$\mu = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} \frac{1}{p_i},$$

and

$$\text{var}(X) = \sum_{i=1}^{n} \text{var}(X_i) = \sum_{i=1}^{n} \frac{1-p_i}{p_i^2}.$$

Introducing the change of variables $y_i = 1/p_i = \mathbf{E}[X_i]$, we see that the constraint becomes

$$\sum_{i=1}^{n} y_i = \mu,$$

and that we must minimize

$$\sum_{i=1}^{n} y_i(y_i - 1) = \sum_{i=1}^{n} y_i^2 - \mu$$

subject to that constraint. This is the same problem as the one of part (a), so the method of proof given there applies.

Problem 38. * **Entropy and uncertainty.** Consider a random variable X that can take n values, $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$. The **entropy** of X is defined to be

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i,$$

and is a measure of the uncertainty about the experimental value of X. To get a sense of this, note that $H(X) \geq 0$ and that $H(X)$ is very close to 0 when X is "nearly deterministic," i.e., takes one of its possible values with probability very close to 1 (since we have $p \log p \approx 0$ if either $p \approx 0$ or $p \approx 1$). The notion of entropy is fundamental in information theory, which originated with C. Shannon's famous work and is described in many specialized textbooks. For example, it can be shown that $H(X)$ is a lower bound to the average number of yes-no questions (such as "is $x = x_1$?" or "is $x < x_5$?") that must be asked in order to determine the experimental value of X that has occurred.

Furthermore, with a suitable strategy for asking questions and assuming a very long string of experimental values of X, the average number of questions required per value can be made as close to $H(X)$ as desired.

(a) Show that if $q_1, \ldots, q_n$ are nonnegative numbers such that $\sum_{i=1}^n q_i = 1$, then

$$H(X) \le -\sum_{i=1}^n p_i \log q_i,$$

with equality if and only if $p_i = q_i$ for all i. As a special case, show that $H(X) \le \log n$, with equality if and only if $p_i = 1/n$ for all i. *Hint*: Use the inequality $\ln \alpha \le \alpha - 1$, for $\alpha > 0$, which holds with equality if and only if $\alpha = 1$. (To see this, write $\ln \alpha = \int_1^\alpha \beta^{-1} \delta\beta < \int_1^\alpha \delta\beta = \alpha - 1$ for $\alpha > 1$, and write $\ln \alpha = -\int_\alpha^1 \beta^{-1}\delta\beta < -\int_\alpha^1 \delta\beta = \alpha - 1$ for $0 < \alpha < 1$.)

(b) Let X and Y be random variables taking a finite number of values, and having joint PMF $p_{X,Y}(x, y)$. Define

$$I(X, Y) = \sum_x \sum_y p_{X,Y}(x, y) \log\left(\frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)}\right).$$

Show that $I(X, Y) \ge 0$, and that $I(X, Y) = 0$ if and only if X and Y are independent.

(c) Show that

$$I(X, Y) = H(X) + H(Y) - H(X, Y),$$

where

$$H(X, Y) = -\sum_x \sum_y p_{X,Y}(x, y) \log p_{X,Y}(x, y),$$

$$H(X) = -\sum_x p_X(x) \log p_X(x), \qquad H(Y) = -\sum_y p_Y(y) \log p_Y(y).$$

(d) Show that

$$I(X, Y) = H(X) - H(X \mid Y),$$

where

$$H(X \mid Y) = -\sum_x \sum_y p_Y(y) p_{X \mid Y}(x \mid y) \log p_{X \mid Y}(x \mid y).$$

[Note that $H(X \mid Y)$ may be viewed as the conditional entropy of X given Y. If we interpret $I(X, Y)$ as the information about X conveyed by Y, we can use the formula $H(X \mid Y) = H(X) - I(X, Y)$ to view the conditional entropy $H(X \mid Y)$ as the uncertainty about X reduced by the information about X conveyed by Y.]

Solution.(a) Using the inequality $\ln \alpha \le \alpha - 1$, we have

$$-\sum_{i=1}^n p_i \ln p_i + \sum_{i=1}^n p_i \ln q_i = \sum_{i=1}^n p_i \ln\left(\frac{q_i}{p_i}\right) \le \sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1\right) = 0,$$

with equality if and only if $p_i = q_i$ for all i. Since $\ln p = \log p \ln 2$, we obtain the desired relation $H(X) \leq -\sum_{i=1}^{n} p_i \log q_i$. The inequality $H(X) \leq \log n$ is obtained by setting $q_i = 1/n$ for all i.

(b) The numbers $p_X(x)p_Y(y)$ satisfy $\sum_x \sum_y p_X(x)p_Y(y) = 1$, so by part (a), we have

$$\sum_x \sum_y p_{X,Y}(x,y) \log\big(p_{X,Y}(x,y)\big) \geq \sum_x \sum_y p_{X,Y}(x,y) \log\big(p_X(x)p_Y(y)\big),$$

with equality if and only if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \qquad \text{for all } x \text{ and } y,$$

which is equivalent to X and Y being independent.

(c) We have

$$I(X,Y) = \sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) - \sum_x \sum_y p_{X,Y}(x,y) \log\big(p_X(x)p_Y(y)\big),$$

and

$$\sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) = -H(X,Y),$$

$$\begin{aligned} -\sum_x \sum_y p_{X,Y}(x,y) \log\big(p_X(x)p_Y(y)\big) &= -\sum_x \sum_y p_{X,Y}(x,y) \log p_X(x) \\ &\quad - \sum_x \sum_y p_{X,Y}(x,y) \log p_Y(y) \\ &= -\sum_x p_X(x) \log p_X(x) - \sum_y p_Y(y) \log p_Y(y) \\ &= H(X) + H(Y). \end{aligned}$$

Combining the above three relations, we obtain $I(X,Y) = H(X) + H(Y) - H(X,Y)$.

(d) From the calculation in part (c), we have

$$\begin{aligned} I(X,Y) &= \sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) - \sum_x p_X(x) \log p_X(x) \\ &\quad - \sum_x \sum_y p_{X,Y}(x,y) \log p_Y(y) \\ &= H(X) + \sum_x \sum_y p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_Y(y)}\right) \\ &= H(X) + \sum_x \sum_y p_Y(y) p_{X\,|\,Y}(x\,|\,y) \log p_{X\,|\,Y}(x\,|\,y) \\ &= H(X) - H(X\,|\,Y). \end{aligned}$$

3

General Random Variables

Contents

Random variables with a continuous range of possible experimental values are quite common – the velocity of a vehicle traveling along the highway could be one example. If such a velocity is measured by a digital speedometer, the speedometer's reading is a discrete random variable. But if we also wish to model the exact velocity, a continuous random variable is called for. Models involving continuous random variables can be useful for several reasons. Besides being finer-grained and possibly more accurate, they allow the use of powerful tools from calculus and often admit an insightful analysis that would not be possible under a discrete model.

All of the concepts and methods introduced in Chapter 2, such as expectation, PMFs, and conditioning, have continuous counterparts. Developing and interpreting these counterparts is the subject of this chapter.

3.1 CONTINUOUS RANDOM VARIABLES AND PDFS

A random variable X is called **continuous** if its probability law can be described in terms of a nonnegative function f_X, called the **probability density function of** X, or PDF for short, which satisfies

$$\mathbf{P}(X \in B) = \int_B f_X(x)\,dx,$$

for every subset B of the real line.† In particular, the probability that the value of X falls within an interval is

$$\mathbf{P}(a \le X \le b) = \int_a^b f_X(x)\,dx,$$

and can be interpreted as the area under the graph of the PDF (see Fig. 3.1). For any single value a, we have $\mathbf{P}(X = a) = \int_a^a f_X(x)\,dx = 0$. For this reason, including or excluding the endpoints of an interval has no effect on its probability:

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

Note that to qualify as a PDF, a function f_X must be nonnegative, i.e., $f_X(x) \ge 0$ for every x, and must also satisfy the normalization equation

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \mathbf{P}(-\infty < X < \infty) = 1.$$

† The integral $\int_B f_X(x)\,dx$ is to be interpreted in the usual calculus/Riemann sense and we implicitly assume that it is well-defined. For highly unusual functions and sets, this integral can be harder – or even impossible – to define, but such issues belong to a more advanced treatment of the subject. In any case, it is comforting to know that mathematical subtleties of this type do not arise if f_X is a piecewise continuous function with a finite number of points of discontinuity, and B is the union of a finite or countable number of intervals.

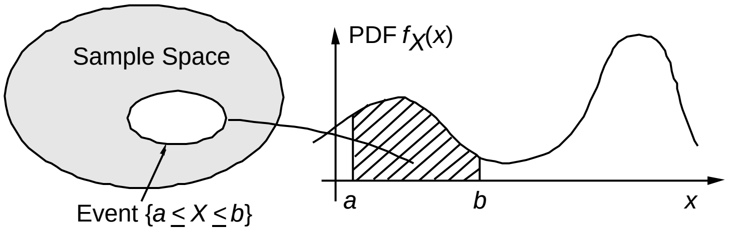


Figure 3.1: Illustration of a PDF. The probability that X takes value in an interval $[a,b]$ is $\int_a^b f_X(x)\,dx$, which is the shaded area in the figure.

Graphically, this means that the entire area under the graph of the PDF must be equal to 1.

To interpret the PDF, note that for an interval $[x, x+\delta]$ with very small length δ, we have

$$\mathbf{P}\big([x, x+\delta]\big) = \int_x^{x+\delta} f_X(t)\,dt \approx f_X(x)\cdot\delta,$$

so we can view $f_X(x)$ as the "probability mass per unit length" near x (cf. Fig. 3.2). It is important to realize that even though a PDF is used to calculate event probabilities, $f_X(x)$ is not the probability of any particular event. In particular, it is not restricted to be less than or equal to one.

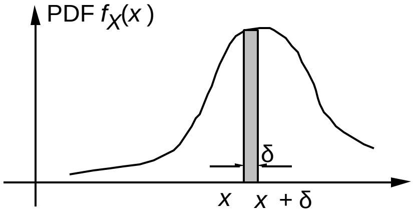


Figure 3.2: Interpretation of the PDF $f_X(x)$ as "probability mass per unit length" around x. If δ is very small, the probability that X takes value in the interval $[x, x+\delta]$ is the shaded area in the figure, which is approximately equal to $f_X(x)\cdot\delta$.

Example 3.1. Continuous Uniform Random Variable. A gambler spins a wheel of fortune, continuously calibrated between 0 and 1, and observes the resulting number. Assuming that all subintervals of [0,1] of the same length are equally likely, this experiment can be modeled in terms a random variable X with PDF

$$f_X(x) = \begin{cases} c & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

for some constant c. This constant can be determined by using the normalization property

$$1 = \int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^1 c\,dx = c\int_0^1 dx = c$$

so that $c = 1$.

More generally, we can consider a random variable X that takes values in an interval $[a, b]$, and again assume that all subintervals of the same length are equally likely. We refer to this type of random variable as **uniform** or **uniformly distributed**. Its PDF has the form

$$f_X(x) = \begin{cases} c & \text{if } a \le x \le b, \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant. This is the continuous analog of the discrete uniform random variable discussed in Chapter 2. For f_X to satisfy the normalization property, we must have (cf. Fig. 3.3)

$$1 = \int_a^b c\,dx = c\int_a^b dx = c(b - a),$$

so that

$$c = \frac{1}{b - a}.$$

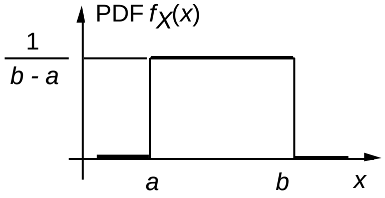


Figure 3.3: The PDF of a uniform random variable.

Note that the probability $\mathbf{P}(X \in I)$ that X takes value in a set I is

$$\mathbf{P}(X \in I) = \int_{[a,b]\cap I} \frac{1}{b - a}\,dx = \frac{1}{b - a}\int_{[a,b]\cap I} dx = \frac{\text{length of } [a, b] \cap I}{\text{length of } [a, b]}.$$

The uniform random variable bears a relation to the discrete uniform law, which involves a sample space with a finite number of equally likely outcomes. The difference is that to obtain the probability of various events, we must now calculate the "length" of various subsets of the real line instead of counting the number of outcomes contained in various events.

Example 3.2. Piecewise Constant PDF. Alvin's driving time to work is between 15 and 20 minutes if the day is sunny, and between 20 and 25 minutes if the day is rainy, with all times being equally likely in each case. Assume that a day is sunny with probability 2/3 and rainy with probability 1/3. What is the PDF of the driving time, viewed as a random variable X?

We interpret the statement that "all times are equally likely" in the sunny and the rainy cases, to mean that the PDF of X is constant in each of the intervals $[15, 20]$ and $[20, 25]$. Furthermore, since these two intervals contain all possible driving times, the PDF should be zero everywhere else:

$$f_X(x) = \begin{cases} c_1 & \text{if } 15 \le x < 20, \\ c_2 & \text{if } 20 \le x \le 25, \\ 0 & \text{otherwise,} \end{cases}$$

where c_1 and c_2 are some constants. We can determine these constants by using the given probabilities of a sunny and of a rainy day:

$$\frac{2}{3} = \mathbf{P}(\text{sunny day}) = \int_{15}^{20} f_X(x)\,dx = \int_{15}^{20} c_1\,dx = 5c_1,$$

$$\frac{1}{3} = \mathbf{P}(\text{rainy day}) = \int_{20}^{25} f_X(x)\,dx = \int_{20}^{25} c_2\,dx = 5c_2,$$

so that

$$c_1 = \frac{2}{15}, \qquad c_2 = \frac{1}{15}.$$

Generalizing this example, consider a random variable X whose PDF has the piecewise constant form

$$f_X(x) = \begin{cases} c_i & \text{if } a_i \le x < a_{i+1}, \quad i = 1, 2, \ldots, n-1, \\ 0 & \text{otherwise,} \end{cases}$$

where $a_1, a_2, \ldots, a_n$ are some scalars with $a_i < a_{i+1}$ for all i, and $c_1, c_2, \ldots, c_n$ are some nonnegative constants (cf. Fig. 3.4). The constants c_i may be determined by additional problem data, as in the case of the preceding driving context. Generally, the c_i must be such that the normalization property holds:

$$1 = \int_{a_1}^{a_n} f_X(x)\,dx = \sum_{i=1}^{n-1} \int_{a_i}^{a_{i+1}} c_i\,dx = \sum_{i=1}^{n-1} c_i(a_{i+1} - a_i).$$

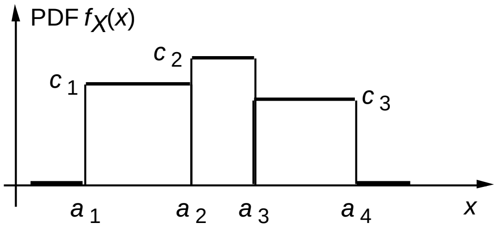


Figure 3.4: A piecewise constant PDF involving three intervals.

Example 3.3. A PDF can be arbitrarily large. Consider a random variable X with PDF

$$f_X(x) = \begin{cases} \dfrac{1}{2\sqrt{x}} & \text{if } 0 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Even though $f_X(x)$ becomes infinitely large as x approaches zero, this is still a valid PDF, because

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^1 \frac{1}{2\sqrt{x}}\,dx = \sqrt{x}\Big|_0^1 = 1.$$

Summary of PDF Properties

Let X be a continuous random variable with PDF f_X.

- $f_X(x) \geq 0$ for all x.
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.
- If δ is very small, then $\mathbf{P}\big([x, x+\delta]\big) \approx f_X(x) \cdot \delta$.
- For any subset B of the real line,

$$\mathbf{P}(X \in B) = \int_B f_X(x)\,dx.$$

Expectation

The **expected value** or **mean** of a continuous random variable X is defined by[†]

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

[†] One has to deal with the possibility that the integral $\int_{-\infty}^{\infty} x f_X(x)\,dx$ is infinite or undefined. More concretely, we will say that the expectation is well-defined if $\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty$. In that case, it is known that the integral $\int_{-\infty}^{\infty} x f_X(x)\,dx$ takes a finite and unambiguous value.

For an example where the expectation is not well-defined, consider a random variable X with PDF $f_X(x) = c/(1+x^2)$, where c is a constant chosen to enforce the normalization condition. The expression $|x| f_X(x)$ is approximately the same as $1/|x|$ when $|x|$ is large. Using the fact $\int_1^{\infty}(1/x)\,dx = \infty$, one can show that $\int_{-\infty}^{\infty} |x| f_X(x)\,dx = \infty$. Thus, $\mathbf{E}[X]$ is left undefined, despite the symmetry of the PDF around zero.

Throughout this book, in lack of an indication to the contrary, we implicitly assume that the expected value of the random variables of interest is well-defined.

This is similar to the discrete case except that the PMF is replaced by the PDF, and summation is replaced by integration. As in Chapter 2, $\mathbf{E}[X]$ can be interpreted as the "center of gravity" of the probability law and, also, as the anticipated average value of X in a large number of independent repetitions of the experiment. Its mathematical properties are similar to the discrete case – after all, an integral is just a limiting form of a sum.

If X is a continuous random variable with given PDF, any real-valued function $Y = g(X)$ of X is also a random variable. Note that Y can be a continuous random variable: for example, consider the trivial case where $Y = g(X) = X$. But Y can also turn out to be discrete. For example, suppose that $g(x) = 1$ for $x > 0$, and $g(x) = 0$, otherwise. Then $Y = g(X)$ is a discrete random variable. In either case, the mean of $g(X)$ satisfies the **expected value rule**

$$\mathbf{E}\big[g(X)\big] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx,$$

in complete analogy with the discrete case.

The n**th moment** of a continuous random variable X is defined as $\mathbf{E}[X^n]$, the expected value of the random variable X^n. The **variance**, denoted by $\text{var}(X)$, is defined as the expected value of the random variable $\big(X - \mathbf{E}[X]\big)^2$.

We now summarize this discussion and list a number of additional facts that are practically identical to their discrete counterparts.

Expectation of a Continuous Random Variable and its Properties

Let X be a continuous random variable with PDF f_X.

- The expectation of X is defined by

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

- The expected value rule for a function $g(X)$ has the form

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

- The variance of X is defined by

$$\text{var}(X) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] = \int_{-\infty}^{\infty} \big(x - \mathbf{E}[X]\big)^2 f_X(x)\,dx.$$

- We have
$$0 \le \mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.$$

- If $Y = aX + b$, where a and b are given scalars, then
$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \qquad \mathrm{var}(Y) = a^2\mathrm{var}(X).$$

Example 3.4. Mean and Variance of the Uniform Random Variable. Consider the case of a uniform PDF over an interval $[a, b]$, as in Example 3.1. We have

$$\begin{aligned}
\mathbf{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx \\
&= \int_a^b x \cdot \frac{1}{b-a}\,dx \\
&= \frac{1}{b-a} \cdot \frac{1}{2}x^2\Big|_a^b \\
&= \frac{1}{b-a} \cdot \frac{b^2-a^2}{2} \\
&= \frac{a+b}{2},
\end{aligned}$$

as one expects based on the symmetry of the PDF around $(a+b)/2$.

To obtain the variance, we first calculate the second moment. We have

$$\begin{aligned}
\mathbf{E}[X^2] &= \int_a^b \frac{x^2}{b-a}\,dx \\
&= \frac{1}{b-a}\int_a^b x^2\,dx \\
&= \frac{1}{b-a} \cdot \frac{1}{3}x^3\Big|_a^b \\
&= \frac{b^3-a^3}{3(b-a)} \\
&= \frac{a^2+ab+b^2}{3}.
\end{aligned}$$

Thus, the variance is obtained as

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12},$$

after some calculation.

Suppose now that $[a, b] = [0, 1]$, and consider the function $g(x) = 1$ if $x \leq 1/3$, and $g(x) = 2$ if $x > 1/3$. The random variable $Y = g(X)$ is a discrete one with PMF $p_Y(1) = \mathbf{P}(X \leq 1/3) = 1/3$, $p_Y(2) = 1 - p_Y(1) = 2/3$. Thus,

$$\mathbf{E}[Y] = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 2 = \frac{5}{3}.$$

The same result could be obtained using the expected value rule:

$$\mathbf{E}[Y] = \int_0^1 g(x) f_X(x)\,dx = \int_0^{1/3} dx + \int_{1/3}^1 2\,dx = \frac{5}{3}.$$

Exponential Random Variable

An **exponential** random variable has a PDF of the form

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where λ is a positive parameter characterizing the PDF (see Fig. 3.5). This is a legitimate PDF because

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^{\infty} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_0^{\infty} = 1.$$

Note that the probability that X exceeds a certain value falls exponentially. Indeed, for any $a \geq 0$, we have

$$\mathbf{P}(X \geq a) = \int_a^{\infty} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_a^{\infty} = e^{-\lambda a}.$$

An exponential random variable can be a very good model for the amount of time until a piece of equipment breaks down, until a light bulb burns out, or until an accident occurs. It will play a major role in our study of random processes in Chapter 5, but for the time being we will simply view it as an example of a random variable that is fairly tractable analytically.

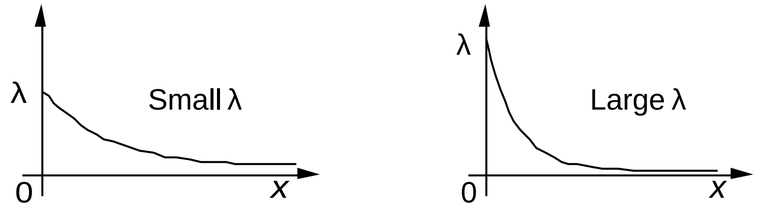


Figure 3.5: The PDF $\lambda e^{-\lambda x}$ of an exponential random variable.

The mean and the variance can be calculated to be

$$\mathbf{E}[X] = \frac{1}{\lambda}, \qquad \mathrm{var}(X) = \frac{1}{\lambda^2}.$$

These formulas can be verified by straightforward calculation, as we now show. We have, using integration by parts,

$$\begin{aligned}\mathbf{E}[X] &= \int_0^\infty x\lambda e^{-\lambda x}\,dx \\ &= (-xe^{-\lambda x})\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \\ &= 0 - \frac{e^{-\lambda x}}{\lambda}\Bigg|_0^\infty \\ &= \frac{1}{\lambda}.\end{aligned}$$

Using again integration by parts, the second moment is

$$\begin{aligned}\mathbf{E}[X^2] &= \int_0^\infty x^2\lambda e^{-\lambda x}\,dx \\ &= (-x^2e^{-\lambda x})\Big|_0^\infty + \int_0^\infty 2xe^{-\lambda x}\,dx \\ &= 0 + \frac{2}{\lambda}\mathbf{E}[X] \\ &= \frac{2}{\lambda^2}.\end{aligned}$$

Finally, using the formula $\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2$, we obtain

$$\mathrm{var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Example 3.5. The time until a small meteorite first lands anywhere in the Sahara desert is modeled as an exponential random variable with a mean of 10 days. The time is currently midnight. What is the probability that a meteorite first lands some time between 6am and 6pm of the first day?

Let X be the time elapsed until the event of interest, measured in days. Then, X is exponential, with mean $1/\lambda = 10$, which yields $\lambda = 1/10$. The desired probability is

$$\mathbf{P}(1/4 \leq X \leq 3/4) = \mathbf{P}(X \geq 1/4) - \mathbf{P}(X > 3/4) = e^{-1/40} - e^{-3/40} = 0.0476,$$

where we have used the formula $\mathbf{P}(X \geq a) = \mathbf{P}(X > a) = e^{-\lambda a}$.

Let us also derive an expression for the probability that the time when a meteorite first lands will be between 6am and 6pm of some day. For the kth day, this set of times corresponds to the event $k-(3/4) \leq X \leq k-(1/4)$. Since these events are disjoint, the probability of interest is

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbf{P}\Big(k-\frac{3}{4} \leq X \leq k-\frac{1}{4}\Big) &= \sum_{k=1}^{\infty}\Big(\mathbf{P}\Big(X \geq k-\frac{3}{4}\Big)-\mathbf{P}\Big(X > k-\frac{1}{4}\Big)\Big) \\ &= \sum_{k=1}^{\infty}\big(e^{-(4k-3)/40}-e^{-(4k-1)/40}\big).\end{aligned}$$

We omit the remainder of the calculation, which involves using the geometric series formula.

3.2 CUMULATIVE DISTRIBUTION FUNCTIONS

We have been dealing with discrete and continuous random variables in a somewhat different manner, using PMFs and PDFs, respectively. It would be desirable to describe all kinds of random variables with a single mathematical concept. This is accomplished by the **cumulative distribution function**, or CDF for short. The CDF of a random variable X is denoted by F_X and provides the probability $\mathbf{P}(X \leq x)$. In particular, for every x we have

$$F_X(x) = \mathbf{P}(X \leq x) = \begin{cases} \displaystyle\sum_{k \leq x} p_X(k) & X\text{: discrete,} \\ \displaystyle\int_{-\infty}^{x} f_X(t)\,dt & X\text{: continuous.}\end{cases}$$

Loosely speaking, the CDF $F_X(x)$ "accumulates" probability "up to" the value x.

Any random variable associated with a given probability model has a CDF, regardless of whether it is discrete, continuous, or other. This is because $\{X \leq x\}$ is always an event and therefore has a well-defined probability. Figures 3.6 and 3.7 illustrate the CDFs of various discrete and continuous random variables. From these figures, as well as from the definition, some general properties of the CDF can be observed.

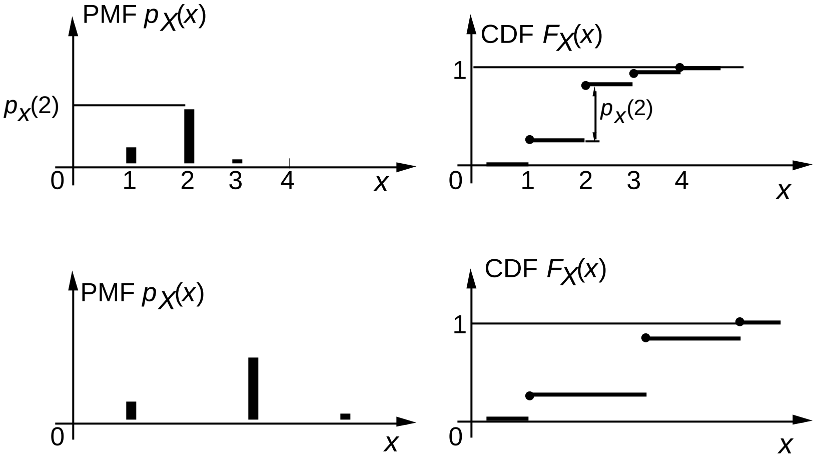


Figure 3.6: CDFs of some discrete random variables. The CDF is related to the PMF through the formula

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{k \leq x} p_X(k),$$

and has a staircase form, with jumps occurring at the values of positive probability mass. Note that at the points where a jump occurs, the value of F_X is the larger of the two correponding values (i.e., F_X is continuous from the right).

Properties of a CDF

The CDF F_X of a random variable X is defined by

$$F_X(x) = \mathbf{P}(X \leq x), \qquad \text{for all } x,$$

and has the following properties.

- F_X is monotonically nondecreasing:

$$\text{if } x \leq y, \text{ then } F_X(x) \leq F_X(y).$$

- $F_X(x)$ tends to 0 as $x \to -\infty$, and to 1 as $x \to \infty$.
- If X is discrete, then F_X has a piecewise constant and staircase-like form.
- If X is continuous, then F_X has a continuously varying form.

- If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^{k} p_X(i),$$

$$p_X(k) = \mathbf{P}(X \leq k) - \mathbf{P}(X \leq k-1) = F_X(k) - F_X(k-1),$$

  for all integers k.

- If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt,$$
$$f_X(x) = \frac{dF_X}{dx}(x).$$

  (The latter relation is valid for those x for which the CDF has a derivative.)

Because the CDF is defined for any type of random variable, it provides a convenient means for exploring the relations between continuous and discrete random variables. This is illustrated in the following example, which shows that there is a close relation between the geometric and the exponential random variables.

Example 3.6. The Geometric and Exponential CDFs. Let X be a geometric random variable with parameter p; that is, X is the number of trials to obtain the first success in a sequence of independent Bernoulli trials, where the probability of success is p. Thus, for $k = 1, 2, \ldots$, we have $\mathbf{P}(X = k) = p(1-p)^{k-1}$ and the CDF is given by

$$F^{\text{geo}}(n) = \sum_{k=1}^{n} p(1-p)^{k-1} = p\frac{1-(1-p)^n}{1-(1-p)} = 1-(1-p)^n, \qquad \text{for } n = 1, 2, \ldots$$

Suppose now that X is an exponential random variable with parameter $\lambda > 0$. Its CDF is given by

$$F^{exp}(x) = \mathbf{P}(X \leq x) = 0, \qquad \text{for } x \leq 0,$$

and

$$F^{\text{exp}}(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x}, \qquad \text{for } x > 0.$$

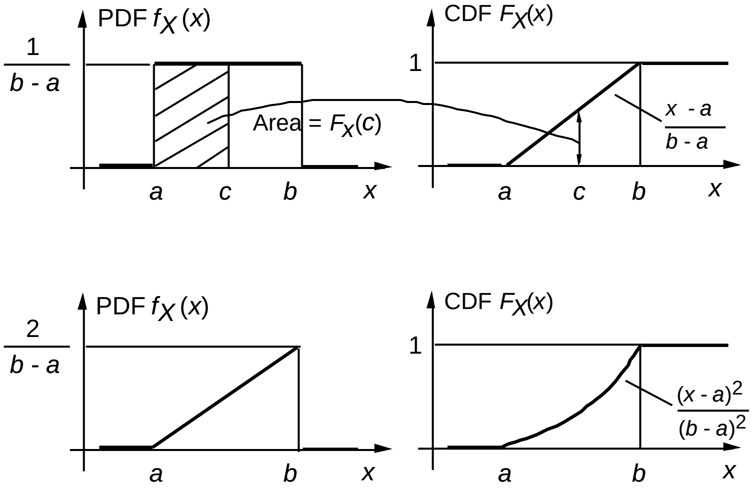


Figure 3.7: CDFs of some continuous random variables. The CDF is related to the PDF through the formula

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

Thus, the PDF f_X can be obtained from the CDF by differentiation:

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

For a continuous random variable, the CDF has no jumps, i.e., it is continuous.

To compare the two CDFs above, let $\delta = -\ln(1-p)/\lambda$, so that

$$e^{-\lambda\delta} = 1 - p.$$

Then we see that the values of the exponential and the geometric CDFs are equal for all $x = n\delta$, where $n = 1, 2, \ldots$, i.e.,

$$F^{\text{exp}}(n\delta) = F^{\text{geo}}(n), \qquad n = 1, 2, \ldots,$$

as illustrated in Fig. 3.8.

If δ is very small, there is close proximity of the exponential and the geometric CDFs, provided that we scale the values taken by the geometric random variable by δ. This relation is best interpreted by viewing X as time, either continuous, in the case of the exponential, or δ-discretized, in the case of the geometric. In particular, suppose that δ is a small number, and that every δ seconds, we flip a coin with the probability of heads being a small number p. Then, the time of the first occurrence of heads is well approximated by an exponential random variable. The parameter

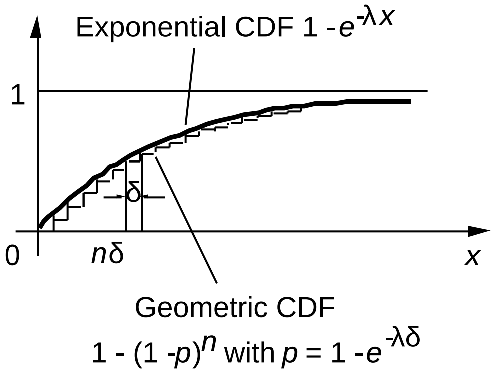


Figure 3.8: Relation of the geometric and the exponential CDFs. We have

$$F^{\rm exp}(n\delta) = F^{\rm geo}(n), \qquad n = 1, 2, \ldots,$$

if the interval δ is such that $e^{-\lambda\delta} = 1 - p$. As δ approaches 0, the exponential random variable can be interpreted as the "limit" of the geometric.

λ of this exponential is such that $e^{-\lambda\delta} = 1 - p$ or $\lambda = -\ln(1 - p)/\delta$. This relation between the geometric and the exponential random variables will play an important role in the theory of the Bernoulli and Poisson stochastic processes in Chapter 5.

Sometimes, in order to calculate the PMF or PDF of a discrete or continuous random variable, respectively, it is more convenient to first calculate the CDF and then use the preceding relations. The systematic use of this approach for the case of a continuous random variable will be discussed in Section 3.6. The following is a discrete example.

Example 3.7. The Maximum of Several Random Variables. You are allowed to take a certain test three times, and your final score will be the maximum of the test scores. Thus,

$$X = \max\{X_1, X_2, X_3\},$$

where X_1, X_2, X_3 are the three test scores and X is the final score. Assume that your score in each test takes one of the values from 1 to 10 with equal probability 1/10, independently of the scores in other tests. What is the PMF p_X of the final score?

We calculate the PMF indirectly. We first compute the CDF $F_X(k)$ and then obtain the PMF as

$$p_X(k) = F_X(k) - F_X(k-1), \qquad k = 1, \ldots, 10.$$

We have

$$\begin{aligned} F_X(k) &= \mathbf{P}(X \le k) \\ &= \mathbf{P}(X_1 \le k,\, X_2 \le k,\, X_3 \le k) \\ &= \mathbf{P}(X_1 \le k)\mathbf{P}(X_2 \le k)\mathbf{P}(X_3 \le k) \\ &= \left(\frac{k}{10}\right)^3, \end{aligned}$$

where the third equality follows from the independence of the events $\{X_1 \le k\}$, $\{X_2 \le k\}$, $\{X_3 \le k\}$. Thus the PMF is given by

$$p_X(k) = \left(\frac{k}{10}\right)^3 - \left(\frac{k-1}{10}\right)^3, \qquad k = 1, \ldots, 10.$$

3.3 NORMAL RANDOM VARIABLES

A continuous random variable X is said to be **normal** or **Gaussian** if it has a PDF of the form (see Fig. 3.9)

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

where μ and σ are two scalar parameters characterizing the PDF, with σ assumed nonnegative. It can be verified that the normalization property

$$\frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2}\, dx = 1$$

holds (see the end-of-chapter problems).

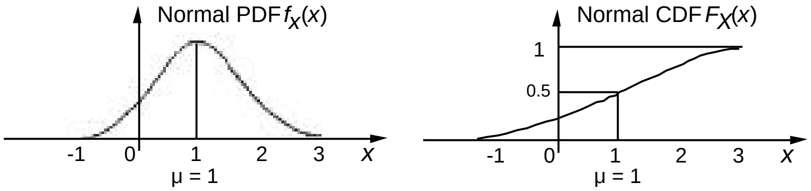


Figure 3.9: A normal PDF and CDF, with $\mu = 1$ and $\sigma^2 = 1$. We observe that the PDF is symmetric around its mean μ, and has a characteristic bell-shape. As x gets further from μ, the term $e^{-(x-\mu)^2/2\sigma^2}$ decreases very rapidly. In this figure, the PDF is very close to zero outside the interval $[-1, 3]$.

The mean and the variance can be calculated to be

$$\mathbf{E}[X] = \mu, \qquad \mathrm{var}(X) = \sigma^2.$$

To see this, note that the PDF is symmetric around μ, so its mean must be μ. Furthermore, the variance is given by

$$\mathrm{var}(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}\,dx.$$

Using the change of variables $y = (x-\mu)/\sigma$ and integration by parts, we have

$$\begin{aligned}
\mathrm{var}(X) &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2}\,dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(-ye^{-y^2/2}\right)\Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\
&= \sigma^2.
\end{aligned}$$

The last equality above is obtained by using the fact

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,dy = 1,$$

which is just the normalization property of the normal PDF for the case where $\mu = 0$ and $\sigma = 1$.

The normal random variable has several special properties. The following one is particularly important and will be justified in Section 3.6.

Normality is Preserved by Linear Transformations

If X is a normal random variable with mean μ and variance σ^2, and if a, b are scalars, then the random variable

$$Y = aX + b$$

is also normal, with mean and variance

$$\mathbf{E}[Y] = a\mu + b, \qquad \mathrm{var}(Y) = a^2\sigma^2.$$

The Standard Normal Random Variable

A normal random variable Y with zero mean and unit variance is said to be a **standard normal**. Its CDF is denoted by Φ,

$$\Phi(y) = \mathbf{P}(Y \le y) = \mathbf{P}(Y < y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2}\, dt.$$

It is recorded in a table (given in the next page), and is a very useful tool for calculating various probabilities involving normal random variables; see also Fig. 3.10.

Note that the table only provides the values of $\Phi(y)$ for $y \ge 0$, because the omitted values can be found using the symmetry of the PDF. For example, if Y is a standard normal random variable, we have

$$\begin{aligned}\Phi(-0.5) &= \mathbf{P}(Y \le -0.5) = \mathbf{P}(Y \ge 0.5) = 1 - \mathbf{P}(Y < 0.5)\\ &= 1 - \Phi(0.5) = 1 - .6915 = 0.3085.\end{aligned}$$

Let X be a normal random variable with mean μ and variance σ^2. We "standardize" X by defining a new random variable Y given by

$$Y = \frac{X - \mu}{\sigma}.$$

Since Y is a linear transformation of X, it is normal. Furthermore,

$$\mathbf{E}[Y] = \frac{\mathbf{E}[X] - \mu}{\sigma} = 0, \qquad \text{var}(Y) = \frac{\text{var}(X)}{\sigma^2} = 1.$$

Thus, Y is a standard normal random variable. This fact allows us to calculate the probability of any event defined in terms of X: we redefine the event in terms of Y, and then use the standard normal table.

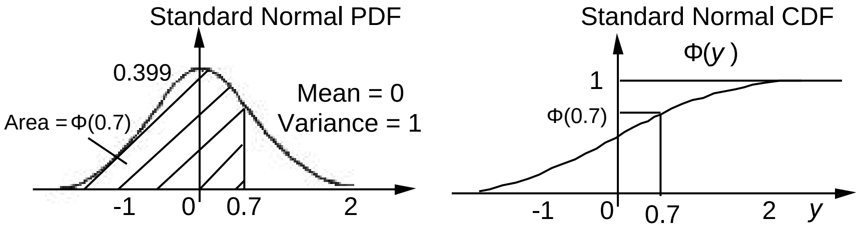


Figure 3.10: The PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

of the standard normal random variable. Its corresponding CDF, which is denoted by $\Phi(y)$, is recorded in a table.

Example 3.8. Using the Normal Table. The annual snowfall at a particular geographic location is modeled as a normal random variable with a mean of $\mu = 60$ inches, and a standard deviation of $\sigma = 20$. What is the probability that this year's snowfall will be at least 80 inches?

Let X be the snow accumulation, viewed as a normal random variable, and let

$$Y = \frac{X - \mu}{\sigma} = \frac{X - 60}{20},$$

be the corresponding standard normal random variable. We want to find

$$\mathbf{P}(X \geq 80) = \mathbf{P}\left(\frac{X - 60}{20} \geq \frac{80 - 60}{20}\right) = \mathbf{P}\left(Y \geq \frac{80 - 60}{20}\right) = \mathbf{P}(Y \geq 1) = 1 - \Phi(1),$$

where Φ is the CDF of the standard normal. We read the value $\Phi(1)$ from the table:

$$\Phi(1) = 0.8413,$$

so that

$$\mathbf{P}(X \geq 80) = 1 - \Phi(1) = 0.1587.$$

Generalizing the approach in the preceding example, we have the following procedure.

CDF Calculation of the Normal Random Variable

The CDF of a normal random variable X with mean μ and variance σ^2 is obtained using the standard normal table as

$$\mathbf{P}(X \leq x) = \mathbf{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \mathbf{P}\left(Y \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

where Y is a standard normal random variable.

The normal random variable is often used in signal processing and communications engineering to model noise and unpredictable distortions of signals. The following is a typical example.

Example 3.9. Signal Detection. A binary message is transmitted as a signal that is either -1 or $+1$. The communication channel corrupts the transmission with additive normal noise with mean $\mu = 0$ and variance σ^2. The receiver concludes that the signal -1 (or $+1$) was transmitted if the value received is < 0 (or ≥ 0, respectively); see Fig. 3.11. What is the probability of error?

An error occurs whenever -1 is transmitted and the noise N is at least 1 so that $N + S = N - 1 \geq 0$, or whenever $+1$ is transmitted and the noise N is smaller

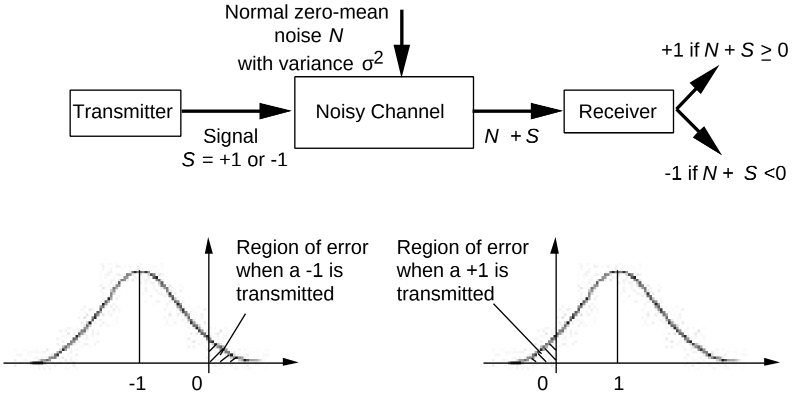

Normal zero-mean noise N with variance σ^2
Transmitter
Noisy Channel
Receiver
Signal S = +1 or -1
$N + S$
+1 if $N + S \geq 0$
-1 if $N + S < 0$
Region of error when a -1 is transmitted
Region of error when a +1 is transmitted
-1
0
0
1

Figure 3.11: The signal detection scheme of Example 3.9. The area of the shaded region gives the probability of error in the two cases where -1 and $+1$ is transmitted.

than -1 so that $N + S = N + 1 < 0$. In the former case, the probability of error is

$$\begin{aligned}\mathbf{P}(N \geq 1) = 1 - \mathbf{P}(N < 1) &= 1 - \mathbf{P}\left(\frac{N-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right)\\ &= 1 - \Phi\left(\frac{1-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1}{\sigma}\right).\end{aligned}$$

In the latter case, the probability of error is the same, by symmetry. The value of $\Phi(1/\sigma)$ can be obtained from the normal table. For $\sigma = 1$, we have $\Phi(1/\sigma) = \Phi(1) = 0.8413$, and the probability of the error is 0.1587.

The normal random variable plays an important role in a broad range of probabilistic models. The main reason is that, generally speaking, it models well the additive effect of many independent factors, in a variety of engineering, physical, and statistical contexts. Mathematically, the key fact is that *the sum of a large number of independent and identically distributed (not necessarily normal) random variables has an approximately normal CDF, regardless of the CDF of the individual random variables.* This property is captured in the celebrated *central limit theorem*, which will be discussed in Chapter 7.

3.4 CONDITIONING ON AN EVENT

The **conditional PDF** of a continuous random variable X, conditioned on a particular event A with $\mathbf{P}(A) > 0$, is a function $f_{X|A}$ that satisfies

$$\mathbf{P}(X \in B \mid A) = \int_B f_{X|A}(x)\,dx,$$

for any subset B of the real line. It is the same as an ordinary PDF, except that it now refers to a new universe in which the event A is known to have occurred.

An important special case arises when we condition on X belonging to a subset A of the real line, with $\mathbf{P}(X \in A) > 0$. We then have

$$\mathbf{P}(X \in B \,|\, X \in A) = \frac{\mathbf{P}(X \in B \text{ and } X \in A)}{\mathbf{P}(X \in A)} = \frac{\displaystyle\int_{A \cap B} f_X(x)\,dx}{\mathbf{P}(X \in A)}.$$

This formula must agree with the earlier one, and therefore,[†]

$$f_{X|A}(x \,|\, A) = \begin{cases} \dfrac{f_X(x)}{\mathbf{P}(X \in A)} & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

As in the discrete case, the conditional PDF is zero outside the conditioning set. Within the conditioning set, the conditional PDF has exactly the same shape as the unconditional one, except that it is scaled by the constant factor $1/\mathbf{P}(X \in A)$. This normalization ensures that $f_{X|A}$ integrates to 1, which makes it a legitimate PDF; see Fig. 3.13.

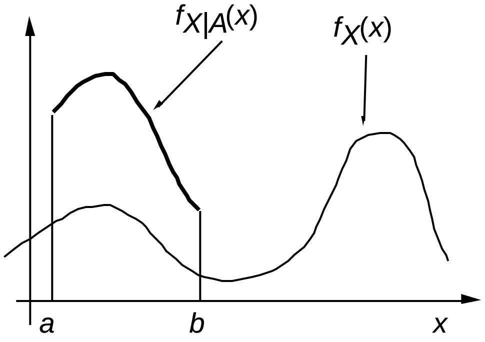


Figure 3.13: The unconditional PDF f_X and the conditional PDF $f_{X|A}$, where A is the interval $[a, b]$. Note that within the conditioning event A, $f_{X|A}$ retains the same shape as f_X, except that it is scaled along the vertical axis.

Example 3.10. The exponential random variable is memoryless. Alvin goes to a bus stop where the time T between two successive buses has an exponential PDF with parameter λ. Suppose that Alvin arrives t secs after the preceding bus arrival and let us express this fact with the event $A = \{T > t\}$. Let X be the time that Alvin has to wait for the next bus to arrive. What is the conditional CDF $F_{X|A}(x \,|\, A)$?

[†] We are using here the simpler notation $f_{X|A}(x)$ in place of $f_{X|X \in A}$, which is more accurate.

We have

$$\begin{aligned}
\mathbf{P}(X > x \mid A) &= \mathbf{P}(T > t + x \mid T > t) \\
&= \frac{\mathbf{P}(T > t + x \text{ and } T > t)}{\mathbf{P}(T > t)} \\
&= \frac{\mathbf{P}(T > t + x)}{\mathbf{P}(T > t)} \\
&= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} \\
&= e^{-\lambda x},
\end{aligned}$$

where we have used the expression for the CDF of an exponential random variable derived in Example 3.6.

Thus, the conditional CDF of X is exponential with parameter λ, regardless the time t that elapsed between the preceding bus arrival and Alvin's arrival. This is known as the *memorylessness property* of the exponential. Generally, if we model the time to complete a certain operation by an exponential random variable X, this property implies that as long as the operation has not been completed, the remaining time up to completion has the same exponential CDF, no matter when the operation started.

For a continuous random variable, the conditional expectation is defined similar to the unconditional case, except that we now need to use the conditional PDF. We summarize the discussion so far, together with some additional properties in the table that follows.

Conditional PDF and Expectation Given an Event

- The conditional PDF $f_{X|A}$ of a continuous random variable X given an event A with $\mathbf{P}(A) > 0$, satisfies

$$\mathbf{P}(X \in B \mid A) = \int_B f_{X|A}(x)\,dx.$$

- If A be a subset of the real line with $\mathbf{P}(X \in A) > 0$, then

$$f_{X|A}(x) = \begin{cases} \dfrac{f_X(x)}{\mathbf{P}(X \in A)} & \text{if } x \in A, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathbf{P}(X \in B \mid X \in A) = \int_B f_{X|A}(x)\,dx,$$

for any set B.

- The corresponding conditional expectation is defined by

$$\mathbf{E}[X \mid A] = \int_{-\infty}^{\infty} x f_{X|A}(x)\,dx.$$

- The expected value rule remains valid:

$$\mathbf{E}\big[g(X) \mid A\big] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x)\,dx.$$

- If $A_1, A_2, \ldots, A_n$ are disjoint events with $\mathbf{P}(A_i) > 0$ for each i, that form a partition of the sample space, then

$$f_X(x) = \sum_{i=1}^{n} \mathbf{P}(A_i) f_{X|A_i}(x)$$

  (a version of the total probability theorem), and

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{E}[X \mid A_i]$$

  (the total expectation theorem). Similarly,

$$\mathbf{E}\big[g(X)\big] = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{E}\big[g(X) \mid A_i\big].$$

To justify the above version of the total probability theorem, we use the total probability theorem from Chapter 1, to obtain

$$\mathbf{P}(X \le x) = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{P}(X \le x \mid A_i).$$

This formula can be rewritten as

$$\int_{-\infty}^{x} f_X(t)\,dt = \sum_{i=1}^{n} \mathbf{P}(A_i) \int_{-\infty}^{x} f_{X|A_i}(t)\,dt.$$

We take the derivative of both sides, with respect to x, and obtain the desired relation

$$f_X(x) = \sum_{i=1}^{n} \mathbf{P}(A_i) f_{X|A_i}(x).$$

If we now multiply both sides by x and then integrate from $-\infty$ to ∞, we obtain the total expectation theorem for continuous random variables.

The total expectation theorem can often facilitate the calculation of the mean, variance, and other moments of a random variable, using a divide-and-conquer approach.

Example 3.11. Mean and Variance of a Piecewise Constant PDF. Suppose that the random variable X has the piecewise constant PDF

$$f_X(x) = \begin{cases} 1/3 & \text{if } 0 \le x \le 1, \\ 2/3 & \text{if } 1 < x \le 2, \\ 0 & \text{otherwise,} \end{cases}$$

(see Fig. 3.14). Consider the events

$$\begin{aligned} A_1 &= \big\{X \text{ lies in the first interval } [0,1]\big\}, \\ A_2 &= \big\{X \text{ lies in the second interval } (1,2]\big\}. \end{aligned}$$

We have from the given PDF,

$$\mathbf{P}(A_1) = \int_0^1 f_X(x)\,dx = \frac{1}{3}, \qquad \mathbf{P}(A_2) = \int_1^2 f_X(x)\,dx = \frac{2}{3}.$$

Furthermore, the conditional mean and second moment of X, conditioned on A_1 and A_2, are easily calculated since the corresponding conditional PDFs $f_{X|A_1}$ and $f_{X|A_2}$ are uniform. We recall from Example 3.4 that the mean of a uniform random variable on an interval $[a,b]$ is $(a+b)/2$ and its second moment is $(a^2+ab+b^2)/3$. Thus,

$$\mathbf{E}[X\,|\,A_1] = \frac{1}{2}, \qquad \mathbf{E}[X\,|\,A_2] = \frac{3}{2},$$

$$\mathbf{E}\big[X^2\,|\,A_1\big] = \frac{1}{3}, \qquad \mathbf{E}\big[X^2\,|\,A_2\big] = \frac{7}{3}.$$

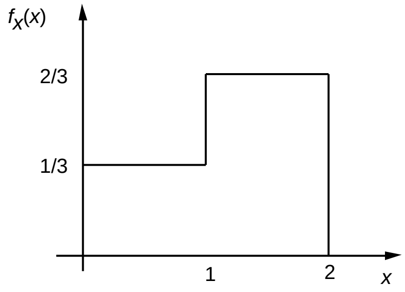


Figure 3.14: Piecewise constant PDF for Example 3.11.

We now use the total expectation theorem to obtain

$$\mathbf{E}[X] = \mathbf{P}(A_1)\mathbf{E}[X \mid A_1] + \mathbf{P}(A_2)\mathbf{E}[X \mid A_2] = \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{3}{2} = \frac{7}{6},$$

$$\mathbf{E}[X^2] = \mathbf{P}(A_1)\mathbf{E}[X^2 \mid A_1] + \mathbf{P}(A_2)\mathbf{E}[X^2 \mid A_2] = \frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{7}{3} = \frac{15}{9}.$$

The variance is given by

$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{15}{9} - \frac{49}{36} = \frac{11}{36}.$$

Note that this approach to the mean and variance calculation is easily generalized to piecewise constant PDFs with more than two pieces.

The next example illustrates a divide-and-conquer approach that uses the total probability theorem to calculate a PDF.

Example 3.12. The metro train arrives at the station near your home every quarter hour starting at 6:00 AM. You walk into the station every morning between 7:10 and 7:30 AM, with the time in this interval being a uniform random variable. What is the PDF of the time you have to wait for the first train to arrive?

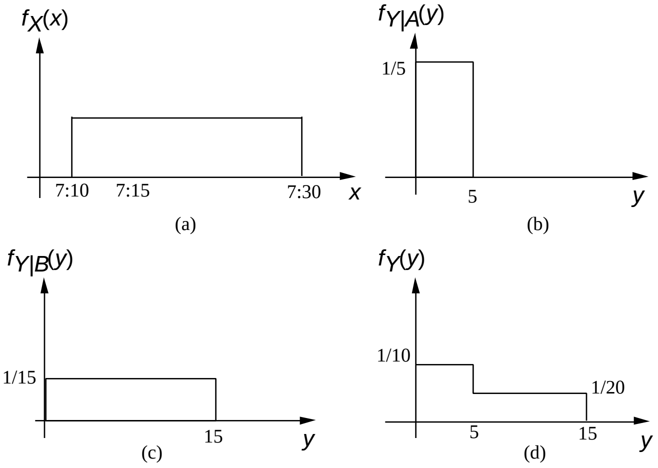


Figure 3.15: The PDFs f_X, $f_{Y|A}$, $f_{Y|B}$, and f_Y in Example 3.12.

The time of your arrival, denoted by X, is a uniform random variable on the interval from 7:10 to 7:30; see Fig. 3.15(a). Let Y be the waiting time. We calculate the PDF f_Y using a divide-and-conquer strategy. Let A and B be the events

$$A = \{7{:}10 \le X \le 7{:}15\} = \{\text{you board the 7:15 train}\},$$

$$B = \{7{:}15 < X \le 7{:}30\} = \{\text{you board the 7:30 train}\}.$$

Conditioned on the event A, your arrival time is uniform on the interval from 7:10 to 7:15. In that case, the waiting time Y is also uniform and takes values between 0 and 5 minutes; see Fig. 3.15(b). Similarly, conditioned on B, Y is uniform and takes values between 0 and 15 minutes; see Fig. 3.15(c). The PDF of Y is obtained using the total probability theorem,

$$f_Y(y) = \mathbf{P}(A) f_{Y|A}(y) + \mathbf{P}(B) f_{Y|B}(y),$$

and is shown in Fig. 3.15(d). In particular,

$$f_Y(y) = \frac{1}{4}\cdot\frac{1}{5} + \frac{3}{4}\cdot\frac{1}{15} = \frac{1}{10}, \quad \text{for } 0 \le y \le 5,$$

and

$$f_Y(y) = \frac{1}{4}\cdot 0 + \frac{3}{4}\cdot\frac{1}{15} = \frac{1}{20}, \quad \text{for } 5 < y \le 15.$$

3.5 MULTIPLE CONTINUOUS RANDOM VARIABLES

We will now extend the notion of a PDF to the case of multiple random variables. In complete analogy with discrete random variables, we introduce joint, marginal, and conditional PDFs. Their intuitive interpretation as well as their main properties parallel the discrete case.

We say that two continuous random variables associated with a common experiment are **jointly continuous** and can be described in terms of a **joint PDF** $f_{X,Y}$, if $f_{X,Y}$ is a nonnegative function that satisfies

$$\mathbf{P}\big((X,Y) \in B\big) = \int\!\!\int_{(x,y)\in B} f_{X,Y}(x,y)\,dx\,dy,$$

for every subset B of the two-dimensional plane. The notation above means that the integration is carried over the set B. In the particular case where B is a rectangle of the form $B = [a,b] \times [c,d]$, we have

$$\mathbf{P}(a \le X \le b,\, c \le Y \le d) = \int_c^d \int_a^b f_{X,Y}(x,y)\,dx\,dy.$$

Furthermore, by letting B be the entire two-dimensional plane, we obtain the normalization property

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1.$$

To interpret the PDF, we let δ be very small and consider the probability of a small rectangle. We have

$$\mathbf{P}(a \leq X \leq a+\delta,\, c \leq Y \leq c+\delta) = \int_c^{c+\delta} \int_a^{a+\delta} f_{X,Y}(x,y)\,dx\,dy \approx f_{X,Y}(a,c)\cdot\delta^2,$$

so we can view $f_{X,Y}(a,c)$ as the "probability per unit area" in the vicinity of (a,c).

The joint PDF contains all conceivable probabilistic information on the random variables X and Y, as well as their dependencies. It allows us to calculate the probability of any event that can be defined in terms of these two random variables. As a special case, it can be used to calculate the probability of an event involving only one of them. For example, let A be a subset of the real line and consider the event $\{X \in A\}$. We have

$$\mathbf{P}(X \in A) = \mathbf{P}\big(X \in A \text{ and } Y \in (-\infty,\infty)\big) = \int_A \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx.$$

Comparing with the formula

$$\mathbf{P}(X \in A) = \int_A f_X(x)\,dx,$$

we see that the **marginal** PDF f_X of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

Example 3.13. Two-Dimensional Uniform PDF. Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour (recall the example given in Section 1.2). Let X and Y denote the delays of Romeo and Juliet, respectively. Assuming that no pairs (x,y) in the square $[0,1] \times [0,1]$ are more likely than others, a natural model involves a joint PDF of the form

$$f_{X,Y}(x,y) = \begin{cases} c & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant. For this PDF to satisfy the normalization property

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = \int_0^1 \int_0^1 c\,dx\,dy = 1,$$

we must have

$$c = 1.$$

This is an example of a uniform PDF on the unit square. More generally, let us fix some subset S of the two-dimensional plane. The corresponding uniform joint PDF on S is defined to be

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\text{area of } S} & \text{if } (x,y) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

For any set $A \subset S$, the probability that the experimental value of (X,Y) lies in A is

$$\mathbf{P}\big((X,Y) \in A\big) = \int\!\!\int_{(x,y)\in A} f_{X,Y}(x,y)\,dx\,dy = \frac{1}{\text{area of } S} \int\!\!\int_{(x,y)\in A\cap S} dx\,dy = \frac{\text{area of } A\cap S}{\text{area of } S}.$$

Example 3.14. We are told that the joint PDF of the random variables X and Y is a constant c on the set S shown in Fig. 3.16 and is zero outside. Find the value of c and the marginal PDFs of X and Y.

The area of the set S is equal to 4 and, therefore, $f_{X,Y}(x,y) = c = 1/4$, for $(x,y) \in S$. To find the marginal PDF $f_X(x)$ for some particular x, we integrate (with respect to y) the joint PDF over the vertical line corresponding to that x. The resulting PDF is shown in the figure. We can compute f_Y similarly.

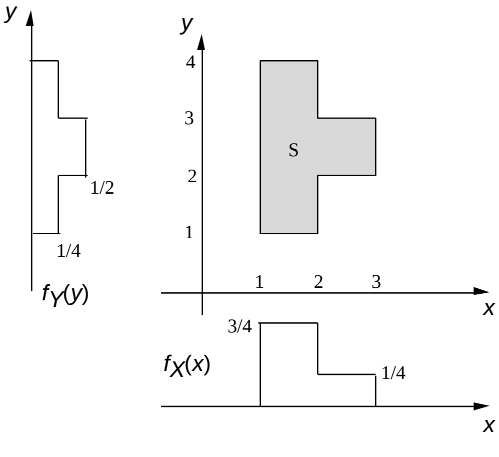


Figure 3.16: The joint PDF in Example 3.14 and the resulting marginal PDFs.

Example 3.15. Buffon's Needle. This is a famous example, which marks the origin of the subject of geometrical probability, that is, the analysis of the geometrical configuration of randomly placed objects.

A surface is ruled with parallel lines, which are at distance d from each other (see Fig. 3.17). Suppose that we throw a needle of length l on the surface at random. What is the probability that the needle will intersect one of the lines?

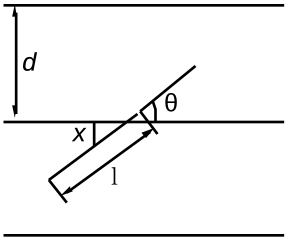


Figure 3.17: Buffon's needle. The length of the line segment between the midpoint of the needle and the point of intersection of the axis of the needle with the closest parallel line is $x/\sin\theta$. The needle will intersect the closest parallel line if and only if this length is less than $l/2$.

We assume here that $l < d$ so that the needle cannot intersect two lines simultaneously. Let X be the distance from the midpoint of the needle to the nearest of the parallel lines, and let Θ be the acute angle formed by the axis of the needle and the parallel lines (see Fig. 3.17). We model the pair of random variables (X, Θ) with a uniform joint PDF over the rectangle $[0, d/2] \times [0, \pi/2]$, so that

$$f_{X,\Theta}(x,\theta) = \begin{cases} 4/(\pi d) & \text{if } x \in [0, d/2] \text{ and } \theta \in [0, \pi/2], \\ 0 & \text{otherwise.} \end{cases}$$

As can be seen from Fig. 3.17, the needle will intersect one of the lines if and only if

$$X \le \frac{l}{2}\sin\Theta,$$

so the probability of intersection is

$$\begin{aligned}
\mathbf{P}\big(X \le (l/2)\sin\Theta\big) &= \int\!\!\int_{x \le (l/2)\sin\theta} f_{X,\Theta}(x,\theta)\, dx\, d\theta \\
&= \frac{4}{\pi d}\int_0^{\pi/2}\int_0^{(l/2)\sin\theta} dx\, d\theta \\
&= \frac{4}{\pi d}\int_0^{\pi/2} \frac{l}{2}\sin\theta\, d\theta \\
&= \frac{2l}{\pi d}(-\cos\theta)\Big|_0^{\pi/2} \\
&= \frac{2l}{\pi d}.
\end{aligned}$$

The probability of intersection can be empirically estimated, by repeating the experiment a large number of times. Since it is equal to $2l/\pi d$, this provides us with a method for the experimental evaluation of π.

Expectation

If X and Y are jointly continuous random variables, and g is some function, then $Z = g(X,Y)$ is also a random variable. We will see in Section 3.6 methods for computing the PDF of Z, if it has one. For now, let us note that the expected value rule is still applicable and

$$\mathbf{E}\big[g(X,Y)\big] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy.$$

As an important special case, for any scalars a, b, we have

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

Conditioning One Random Variable on Another

Let X and Y be continuous random variables with joint PDF $f_{X,Y}$. For any fixed y with $f_Y(y) > 0$, the conditional PDF of X given that $Y = y$, is defined by

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

This definition is analogous to the formula $p_{X|Y} = p_{X,Y}/p_Y$ for the discrete case.

When thinking about the conditional PDF, it is best to view y as a fixed number and consider $f_{X|Y}(x\,|\,y)$ as a function of the single variable x. As a function of x, the conditional PDF $f_{X|Y}(x\,|\,y)$ has the same shape as the joint PDF $f_{X,Y}(x,y)$, because the normalizing factor $f_Y(y)$ does not depend on x; see Fig. 3.18. Note that the normalization ensures that

$$\int_{-\infty}^{\infty} f_{X|Y}(x\,|\,y)\,dx = 1,$$

so *for any fixed* y, $f_{X|Y}(x\,|\,y)$ is a legitimate PDF.

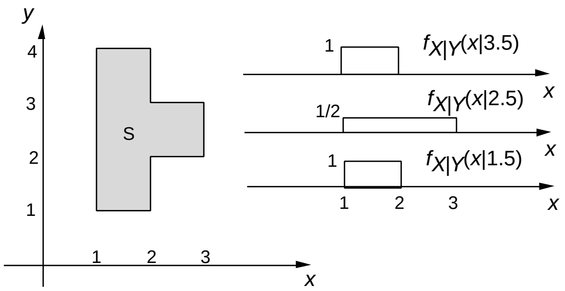


Figure 3.18: Visualization of the conditional PDF $f_{X|Y}(x\,|\,y)$. Let X, Y have a joint PDF which is uniform on the set S. For each fixed y, we consider the joint PDF along the slice $Y = y$ and normalize it so that it integrates to 1.

Example 3.16. Circular Uniform PDF. John throws a dart at a circular target of radius r (see Fig. 3.19). We assume that he always hits the target, and that all points of impact (x, y) are equally likely, so that the joint PDF of the random variables X and Y is uniform. Following Example 3.13, and since the area of the circle is πr^2, we have

$$
\begin{aligned}
f_{X,Y}(x,y) &= \begin{cases} \dfrac{1}{\text{area of the circle}} & \text{if } (x,y) \text{ is in the circle,} \\ 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} \dfrac{1}{\pi r^2} & \text{if } x^2 + y^2 \le r^2, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

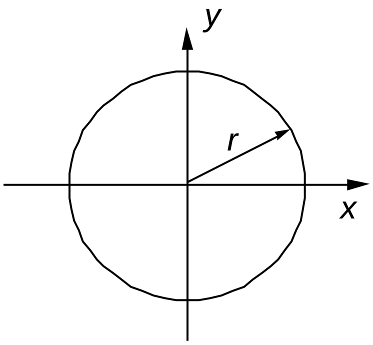


Figure 3.19: Circular target for Example 3.16.

To calculate the conditional PDF $f_{X|Y}(x \mid y)$, let us first calculate the marginal PDF $f_Y(y)$. For $|y| > r$, it is zero. For $|y| \le r$, it can be calculated as follows:

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \\
&= \frac{1}{\pi r^2} \int_{x^2+y^2 \le r^2} dx \\
&= \frac{1}{\pi r^2} \int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} dx \\
&= \frac{2}{\pi r^2} \sqrt{r^2 - y^2}.
\end{aligned}
$$

Note that the marginal $f_Y(y)$ is not a uniform PDF.

The conditional PDF is

$$\begin{aligned} f_{X|Y}(x\,|\,y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{\dfrac{1}{\pi r^2}}{\dfrac{2}{\pi r^2}\sqrt{r^2-y^2}} \\ &= \frac{1}{2\sqrt{r^2-y^2}}. \end{aligned}$$

Thus, for a fixed value of y, the conditional PDF $f_{X|Y}$ is uniform.

To interpret the conditional PDF, let us fix some small positive numbers δ_1 and δ_2, and condition on the event $B = \{y \le Y \le y + \delta_2\}$. We have

$$\begin{aligned} \mathbf{P}(x \le X \le x+\delta_1 \,|\, y \le Y \le y+\delta_2) &= \frac{\mathbf{P}(x \le X \le x+\delta_1 \text{ and } y \le Y \le y+\delta_2)}{\mathbf{P}(y \le Y \le y+\delta_2)} \\ &\approx \frac{f_{X,Y}(x,y)\delta_1\delta_2}{f_Y(y)\delta_2} = f_{X|Y}(x\,|\,y)\delta_1. \end{aligned}$$

In words, $f_{X|Y}(x\,|\,y)\delta_1$ provides us with the probability that X belongs in a small interval $[x, x+\delta_1]$, given that Y belongs in a small interval $[y, y+\delta_2]$. Since $f_{X|Y}(x\,|\,y)\delta_1$ does not depend on δ_2, we can think of the limiting case where δ_2 decreases to zero and write

$$\mathbf{P}(x \le X \le x+\delta_1 \,|\, Y = y) \approx f_{X|Y}(x\,|\,y)\delta_1, \qquad (\delta_1 \text{ small}),$$

and, more generally,

$$\mathbf{P}(X \in A \,|\, Y = y) = \int_A f_{X|Y}(x\,|\,y)\,dx.$$

Conditional probabilities, given the zero probability event $\{Y = y\}$, were left undefined in Chapter 1. But the above formula provides a natural way of defining such conditional probabilities in the present context. In addition, it allows us to view the conditional PDF $f_{X|Y}(x\,|\,y)$ (as a function of x) as a description of the probability law of X, given that the event $\{Y = y\}$ has occurred.

As in the discrete case, the conditional PDF $f_{X|Y}$, together with the marginal PDF f_Y are sometimes used to calculate the joint PDF. Furthermore, this approach can be also used for modeling: instead of directly specifying $f_{X,Y}$, it is often natural to provide a probability law for Y, in terms of a PDF f_Y, and then provide a conditional probability law $f_{X|Y}(x,y)$ for X, given any possible value y of Y.

Example 3.17. Let X be exponentially distributed with mean 1. Once we observe the experimental value x of X, we generate a normal random variable Y with zero mean and variance $x+1$. What is the joint PDF of X and Y?

We have $f_X(x) = e^{-x}$, for $x \geq 0$, and

$$f_{Y|X}(y\,|\,x) = \frac{1}{\sqrt{2\pi(x+1)}} e^{-y^2/2(x+1)}.$$

Thus,

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y\,|\,x) = e^{-x} \frac{1}{\sqrt{2\pi(x+1)}} e^{-y^2/2(x+1)},$$

for all $x \geq 0$ and all y.

Having defined a conditional probability law, we can also define a corresponding conditional expectation by letting

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

The properties of (unconditional) expectation carry though, with the obvious modifications, to conditional expectation. For example the conditional version of the expected value rule

$$\mathbf{E}[g(X)\,|\,Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x\,|\,y)\,dx$$

remains valid.

Summary of Facts About Multiple Continuous Random Variables

Let X and Y be jointly continuous random variables with joint PDF $f_{X,Y}$.

- The joint, marginal, and conditional PDFs are related to each other by the formulas

$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x\,|\,y),$$
$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X\,|\,Y}(x\,|\,y)\,dy.$$

  The conditional PDF $f_{X|Y}(x\,|\,y)$ is defined only for those y for which $f_Y(y) > 0$.

- They can be used to calculate probabilities:

$$\begin{aligned}\mathbf{P}\big((X,Y)\in B\big) &= \int\!\!\int_{(x,y)\in B} f_{X,Y}(x,y)\,dx\,dy,\\ \mathbf{P}(X\in A) &= \int_A f_X(x)\,dx,\\ \mathbf{P}(X\in A\,|\,Y=y) &= \int_A f_{X|Y}(x\,|\,y)\,dx.\end{aligned}$$

- They can also be used to calculate expectations:

$$\begin{aligned}\mathbf{E}[g(X)] &= \int g(x)f_X(x)\,dx,\\ \mathbf{E}\big[g(X,Y)\big] &= \int\int g(x,y)f_{X,Y}(x,y)\,dx\,dy,\\ \mathbf{E}\big[g(X)\,|\,Y=y\big] &= \int g(x)f_{X|Y}(x\,|\,y)\,dx,\\ \mathbf{E}\big[g(X,Y)\,|\,Y=y\big] &= \int g(x,y)f_{X|Y}(x\,|\,y)\,dx.\end{aligned}$$

- For any event A, we have the following version of the total probability theorem:

$$\mathbf{P}(A) = \int \mathbf{P}(A\,|\,Y=y)f_Y(y)\,dy.$$

- We have the following versions of the total expectation theorem:

$$\begin{aligned}\mathbf{E}[X] &= \int \mathbf{E}[X\,|\,Y=y]f_Y(y)\,dy,\\ \mathbf{E}\big[g(X)\big] &= \int \mathbf{E}\big[g(X)\,|\,Y=y]f_Y(y)\,dy,\\ \mathbf{E}\big[g(X,Y)\big] &= \int \mathbf{E}\big[g(X,Y)\,|\,Y=y]f_Y(y)\,dy.\end{aligned}$$

To justify the first version of the total expectation theorem, we observe that

$$\int \mathbf{E}[X\,|\,Y=y]f_Y(y)\,dy = \int\left[\int xf_{X|Y}(x\,|\,y)\,dx\right]f_Y(y)\,dy$$

$$
\begin{aligned}
&= \int \int x f_{X|Y}(x\,|\,y) f_Y(y)\,dx\,dy \\
&= \int \int x f_{X,Y}(x,y)\,dx\,dy \\
&= \int x \left[\int f_{X,Y}(x,y)\,dy \right] dx \\
&= \int x f_X(x)\,dx \\
&= \mathbf{E}[X].
\end{aligned}
$$

The other two versions are justified similarly. The total probability equation $\mathbf{P}(A) = \int \mathbf{P}(A\,|\,Y = y) f_Y(y)\,dy$ follows from the total expectation theorem by letting X be the random variable that takes the value 1 if $X \in A$ and the value 0 otherwise, since then $\mathbf{E}[X] = \mathbf{P}(A)$ and $\mathbf{E}[X\,|\,Y = y] = \mathbf{P}(A\,|\,Y = y)$.

Inference and the Continuous Bayes' Rule

In many situations, we have a model of an underlying but unobserved phenomenon, represented by a random variable X with PDF f_X, and we make noisy measurements Y. The measurements are supposed to provide information about X and are modeled in terms of a conditional PDF $f_{Y|X}$. For example, if Y is the same as X, but corrupted by zero-mean normally distributed noise, one would let the conditional PDF $f_{Y|X}(y\,|\,x)$ of Y, given that $X = x$, be normal with mean equal to x. Once the experimental value of Y is measured, what information does this provide on the unknown value of X?

This setting is similar to that encountered in Section 1.4, when we introduced the Bayes rule and used it to solve inference problems. The only difference is that we are now dealing with continuous random variables.

Note that the information provided by the event $\{Y = y\}$ is described by the conditional PDF $f_{X|Y}(x\,|\,y)$. It thus suffices to evaluate the latter PDF. A calculation analogous to the original derivation of the Bayes' rule, based on the formulas $f_X f_{Y|X} = f_{X,Y} = f_Y f_{X|Y}$, yields

$$
f_{X|Y}(x\,|\,y) = \frac{f_X(x) f_{Y|X}(y\,|\,x)}{f_Y(y)} = \frac{f_X(x) f_{Y|X}(y\,|\,x)}{\displaystyle\int f_X(t) f_{Y|X}(y\,|\,t) dt},
$$

which is the desired formula.

Example 3.18. A lightbulb produced by the General Illumination Company is known to have an exponentially distributed lifetime Y. However, the company has been experiencing quality control problems. On any given day, the parameter λ of the PDF of Y is actually a random variable, uniformly distributed in the interval $[0, 1/2]$. We test a lightbulb and record the experimental value y of its lifetime. What can we say about the underlying parameter λ?

We model the parameter λ as a random variable X, with a uniform distribution. All available information about X is contained in the conditional PDF $f_{X|X}(x\,|\,y)$. We view y as a constant (equal to the observed value of Y) and concentrate on the dependence of the PDF on x. Note that $f_X(x) = 2$, for $0 \le x \le 1/2$. By the continuous Bayes rule, we have

$$f_{X|Y}(x\,|\,y) = \frac{2xe^{-xy}}{\int_0^{1/2} 2te^{-ty}dt}, \qquad \text{for } 0 \le x \le \frac{1}{2}.$$

In some cases, the unobserved phenomenon is inherently discrete. For example, if a binary signal is observed in the presence of noise with a normal distribution. Or if a medical diagnosis is to be made on the basis of continuous measurements like temperature and blood counts. In such cases, a somewhat different version of Bayes' rule applies.

Let X be a discrete random variable that takes values in a finite set $\{1, \ldots, n\}$ and which represents the different discrete possibilities for the unobserved phenomenon of interest. The PMF p_X of X is assumed to be known. Let Y be a continuous random variable which, for any given value x, is described by a conditional PDF $f_{Y\,|\,X}(y\,|\,x)$. We are interested in the conditional PMF of X given the experimental value y of Y.

Instead of working with conditioning event $\{Y = y\}$ which has zero probability, let us instead condition on the event $\{y \le Y \le y + \delta\}$, where δ is a small positive number, and then take the limit as δ tends to zero. We have, using the Bayes rule

$$\begin{aligned}
\mathbf{P}(X = x\,|\,Y = y) &\approx \mathbf{P}(X = x\,|\,y \le Y \le y + \delta) \\
&= \frac{p_X(x)\mathbf{P}(y \le Y \le y + \delta\,|\,X = x)}{\mathbf{P}(y \le Y \le y + \delta)} \\
&\approx \frac{p_X(x) f_{Y|X}(y\,|\,x)\delta}{f_Y(y)\delta} \\
&= \frac{p_X(x) f_{Y|X}(y\,|\,x)}{f_Y(y)}.
\end{aligned}$$

The denominator can be evaluated using a version of the total probability theorem introduced in Section 3.4. We have

$$f_Y(y) = \sum_{i=1}^{n} p_X(i) f_{Y|X}(y\,|\,i).$$

Example 3.19. Let us revisit the signal detection problem considered in 3.9. A signal S is transmitted and we are given that $\mathbf{P}(S = 1) = p$ and $\mathbf{P}(S = -1) = 1 - p$.

The received signal is $Y = N+S$, where N is zero mean normal noise, with variance σ^2, independent of S. What is the probability that $S = 1$, as a function of the observed value y of Y?

Conditioned on $S = s$, the random variable Y has a normal distribution with mean s and variance σ^2. Applying the formula developed above, we obtain

$$\mathbf{P}(S = 1 \,|\, Y = y) = \frac{p_S(1)f_{Y|S}(y\,|\,1)}{f_Y(y)} = \frac{\frac{p}{\sqrt{2\pi}\,\sigma}e^{-(y-1)^2/2\sigma^2}}{\frac{p}{\sqrt{2\pi}\,\sigma}e^{-(y-1)^2/2\sigma^2} + \frac{1-p}{\sqrt{2\pi}\,\sigma}e^{-(y+1)^2/2\sigma^2}}.$$

Independence

In full analogy with the discrete case, we say that two continuous random variables X and Y are **independent** if their joint PDF is the product of the marginal PDFs:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad \text{for all } x, y.$$

Comparing with the formula $f_{X,Y}(x,y) = f_{X|Y}(x\,|\,y)f_Y(y)$, we see that independence is the same as the condition

$$f_{X|Y}(x\,|\,y) = f_X(x), \qquad \text{for all } x \text{ and all } y \text{ with } f_Y(y) > 0,$$

or, symmetrically,

$$f_{Y|X}(y\,|\,x) = f_Y(y), \qquad \text{for all } y \text{ and all } x \text{ with } f_X(x) > 0.$$

If X and Y are independent, then any two events of the form $\{X \in A\}$ and $\{Y \in B\}$ are independent. Indeed,

$$\begin{aligned}\mathbf{P}(X \in A \text{ and } Y \in B) &= \int_{x\in A}\int_{y\in B} f_{X,Y}(x,y)\,dy\,dx\\ &= \int_{x\in A}\int_{y\in B} f_X(x)f_Y(y)\,dy\,dx\\ &= \int_{x\in A} f_X(x)\,dx \int_{y\in B} f_Y(y)\,dy\\ &= \mathbf{P}(X \in A)\mathbf{P}(Y \in B).\end{aligned}$$

A converse statement is also true; see the end-of-chapter problems.

A calculation similar to the discrete case shows that if X and Y are independent, then

$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)]\mathbf{E}[h(Y)],$$

for any two functions g and h. Finally, the variance of the sum of *independent* random variables is again equal to the sum of the variances.

Independence of Continuous Random Variables

Suppose that X and Y are independent, that is,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad \text{for all } x, y.$$

We then have the following properties.

- The random variables $g(X)$ and $h(Y)$ are independent, for any functions g and h.
- We have
$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y],$$
and, more generally,
$$\mathbf{E}\big[g(X)h(Y)\big] = \mathbf{E}\big[g(X)\big]\mathbf{E}\big[h(Y)\big],$$
- We have
$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

Joint CDFs

If X and Y are two random variables associated with the same experiment, we define their joint CDF by

$$F_{X,Y}(x,y) = \mathbf{P}(X \le x,\, Y \le y).$$

As in the case of one random variable, the advantage of working with the CDF is that it applies equally well to discrete and continuous random variables. In particular, if X and Y are described by a joint PDF $f_{X,Y}$, then

$$F_{X,Y}(x,y) = \mathbf{P}(X \le x,\, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\,ds\,dt.$$

Conversely, the PDF can be recovered from the PDF by differentiating:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y).$$

Example 3.20. Let X and Y be described by a uniform PDF on the unit square. The joint CDF is given by

$$F_{X,Y}(x,y) = \mathbf{P}(X \le x,\, Y \le y) = xy, \qquad \text{for } 0 \le x, y \le 1.$$

We then verify that

$$\frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y) = \frac{\partial^2 (xy)}{\partial x \partial y}(x,y) = 1 = f_{X,Y}(x,y),$$

for all (x, y) in the unit square.

More than Two Random Variables

The joint PDF of three random variables X, Y, and Z is defined in analogy with the above. For example, we have

$$\mathbf{P}\big((X,Y,Z) \in B\big) = \int\int\int_{(x,y,z)\in B} f_{X,Y,Z}(x,y,z)\,dx\,dy\,dz,$$

for any set B. We also have relations such as

$$f_{X,Y}(x,y) = \int f_{X,Y,Z}(x,y,z)\,dz,$$

and

$$f_X(x) = \int\int f_{X,Y,Z}(x,y,z)\,dy\,dz.$$

One can also define conditional PDFs by formulas such as

$$f_{X,Y|Z}(x,y\,|\,z) = \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)}, \qquad \text{for } f_Z(z) > 0,$$

$$f_{X|Y,Z}(x\,|\,y,z) = \frac{f_{X,Y,Z}(x,y,z)}{f_{Y,Z}(y,z)}, \qquad \text{for } f_{Y,Z}(y,z) > 0.$$

There is an analog of the multiplication rule:

$$f_{X,Y,Z}(x,y,z) = f_{X|Y,Z}(x\,|\,y,z) f_{Y|Z}(y\,|\,z) f_Z(z).$$

Finally, we say that the three random variables X, Y, and Z are independent if

$$f_{X,Y,Z}(x,y,z) = f_X(x) f_Y(y) f_Z(z), \qquad \text{for all } x,y,z.$$

The expected value rule for functions takes the form

$$\mathbf{E}\big[g(X,Y,Z)\big] = \int\int\int g(x,y,z) f_{X,Y,Z}(x,y,z)\,dx\,dy\,dz,$$

and if g is linear and of the form $aX + bY + cZ$, then

$$\mathbf{E}[aX + bY + cZ] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c\mathbf{E}[Z].$$

Furthermore, there are obvious generalizations of the above to the case of more than three random variables. For example, for any random variables $X_1, X_2, \ldots, X_n$ and any scalars $a_1, a_2, \ldots, a_n$, we have

$$\mathbf{E}[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1\mathbf{E}[X_1] + a_2\mathbf{E}[X_2] + \cdots + a_n\mathbf{E}[X_n].$$

3.6 DERIVED DISTRIBUTIONS

We have seen that the mean of a function $Y = g(X)$ of a continuous random variable X, can be calculated using the expected value rule

$$\mathbf{E}[Y] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx,$$

without first finding the PDF f_Y of Y. Still, in some cases, we may be interested in an explicit formula for f_Y. Then, the following two-step approach can be used.

Calculation of the PDF of a Function $Y = g(X)$ of a Continuous Random Variable X

1. Calculate the CDF F_Y of Y using the formula

$$F_Y(y) = \mathbf{P}\big(g(X) \le y\big) = \int_{\{x \,|\, g(x) \le y\}} f_X(x)\, dx.$$

2. Differentiate to obtain the PDF of Y:

$$f_Y(y) = \frac{dF_Y}{dy}(y).$$

Example 3.21. Let X be uniform on $[0, 1]$. Find the PDF of $Y = \sqrt{X}$. Note that Y takes values between 0 and 1. For every $y \in [0, 1]$, we have

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}(\sqrt{X} \le y) = \mathbf{P}(X \le y^2) = y^2, \quad 0 \le y \le 1.$$

We then differentiate and obtain

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{d(y^2)}{dy} = 2y, \qquad 0 \le y \le 1.$$

Outside the range $[0, 1]$, the CDF $F_Y(y)$ is constant, with $F_Y(y) = 0$ for $y \le 0$, and $F_Y(y) = 1$ for $y \ge 1$. By differentiating, we see that $f_Y(y) = 0$ for y outside $[0, 1]$.

Example 3.22. John Slow is driving from Boston to the New York area, a distance of 180 miles at a constant speed, whose value is uniformly distributed between 30 and 60 miles per hour. What is the PDF of the duration of the trip?

Let X be the speed and let $Y = g(X)$ be the trip duration:

$$g(X) = \frac{180}{X}.$$

To find the CDF of Y, we must calculate

$$\mathbf{P}(Y \le y) = \mathbf{P}\left(\frac{180}{X} \le y\right) = \mathbf{P}\left(\frac{180}{y} \le X\right).$$

We use the given uniform PDF of X, which is

$$f_X(x) = \begin{cases} 1/30 & \text{if } 30 \le x \le 60, \\ 0 & \text{otherwise,} \end{cases}$$

and the corresponding CDF, which is

$$F_X(x) = \begin{cases} 0 & \text{if } x \le 30, \\ (x-30)/30 & \text{if } 30 \le x \le 60, \\ 1 & \text{if } 60 \le x. \end{cases}$$

Thus,

$$\begin{aligned} F_Y(y) &= \mathbf{P}\left(\frac{180}{y} \le X\right) \\ &= 1 - F_X\left(\frac{180}{y}\right) \\ &= \begin{cases} 0 & \text{if } y \le 180/60, \\ 1 - \dfrac{\dfrac{180}{y} - 30}{30} & \text{if } 180/60 \le y \le 180/30, \\ 1 & \text{if } 180/30 \le y, \end{cases} \\ &= \begin{cases} 0 & \text{if } y \le 3, \\ 2 - (6/y) & \text{if } 3 \le y \le 6, \\ 1 & \text{if } 6 \le y, \end{cases} \end{aligned}$$

(see Fig. 3.20). Differentiating this expression, we obtain the PDF of Y:

$$f_Y(y) = \begin{cases} 0 & \text{if } y \le 3, \\ 6/y^2 & \text{if } 3 \le y \le 6, \\ 0 & \text{if } 6 \le y. \end{cases}$$

Example 3.23. Let $Y = g(X) = X^2$, where X is a random variable with known PDF. For any $y \ge 0$, we have

$$\begin{aligned} F_Y(y) &= \mathbf{P}(Y \le y) \\ &= \mathbf{P}(X^2 \le y) \\ &= \mathbf{P}(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}), \end{aligned}$$

and therefore, by differentiating and using the chain rule,

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}), \qquad y \ge 0.$$

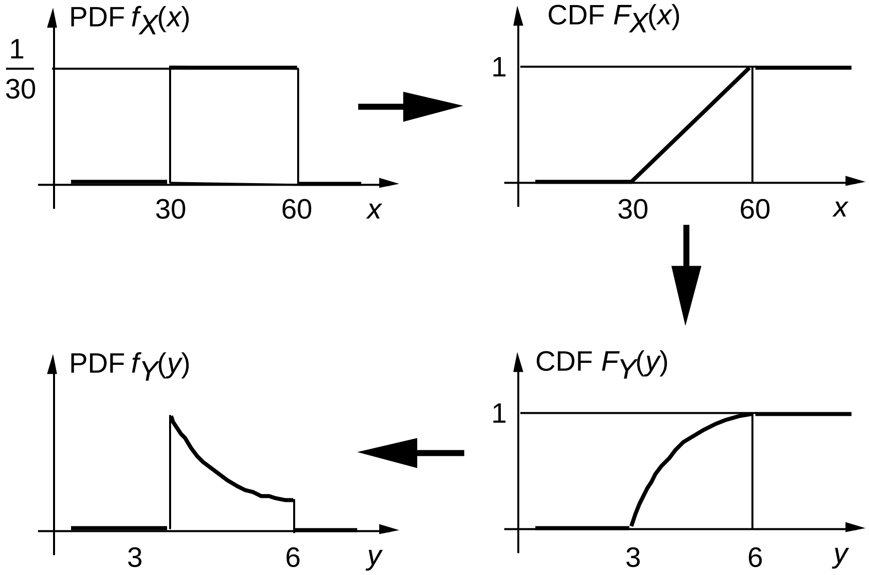


Figure 3.20: The calculation of the PDF of $Y = 180/X$ in Example 3.22. The arrows indicate the flow of the calculation.

The Linear Case

An important case arises when Y is a linear function of X. See Fig. 3.21 for a graphical interpretation.

The PDF of a Linear Function of a Random Variable

Let X be a continuous random variable with PDF f_X, and let

$$Y = aX + b,$$

for some scalars $a \neq 0$ and b. Then,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

To verify this formula, we use the two-step procedure. We only show the

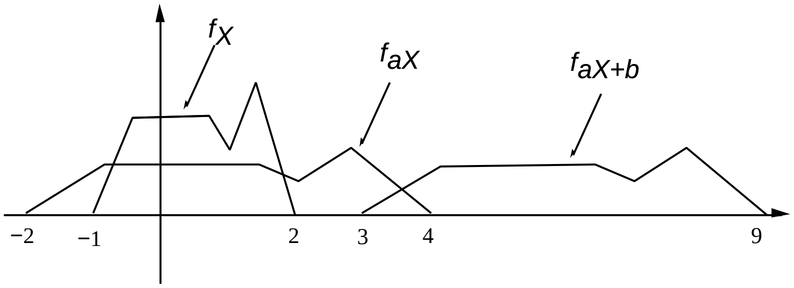


Figure 3.21: The PDF of $aX + b$ in terms of the PDF of X. In this figure, $a = 2$ and $b = 5$. As a first step, we obtain the PDF of aX. The range of Y is wider than the range of X, by a factor of a. Thus, the PDF f_X must be stretched (scaled horizontally) by this factor. But in order to keep the total area under the PDF equal to 1, we need to scale the PDF (vertically) by the same factor a. The random variable $aX + b$ is the same as aX except that its values are shifted by b. Accordingly, we take the PDF of aX and shift it (horizontally) by b. The end result of these operations is the PDF of $Y = aX + b$ and is given mathematically by

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

If a were negative, the procedure would be the same except that the PDF of X would first need to be reflected around the vertical axis ("flipped") yielding f_{-X}. Then a horizontal and vertical scaling (by a factor of $|a|$ and $1/|a|$, respectively) yields the PDF of $-|a|X = aX$. Finally, a horizontal shift of b would again yield the PDF of $aX + b$.

steps for the case where $a > 0$; the case $a < 0$ is similar. We have

$$\begin{aligned} F_Y(y) &= \mathbf{P}(Y \le y) \\ &= \mathbf{P}(aX + b \le y) \\ &= \mathbf{P}\left(X \le \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

We now differentiate this equality and use the chain rule, to obtain

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{1}{a} \cdot \frac{dF_X}{dx}\left(\frac{y-b}{a}\right) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right).$$

Example 3.24. A linear function of an exponential random variable. Suppose that X is an exponential random variable with PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

where λ is a positive parameter. Let $Y = aX + b$. Then,

$$f_Y(y) = \begin{cases} \dfrac{\lambda}{|a|} e^{-\lambda(y-b)/a} & \text{if } (y-b)/a \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that if $b = 0$ and $a > 0$, then Y is an exponential random variable with parameter λ/a. In general, however, Y need not be exponential. For example, if $a < 0$ and $b = 0$, then the range of Y is the negative real axis.

Example 3.25. A linear function of a normal random variable is normal. Suppose that X is a normal random variable with mean μ and variance σ^2, and let $Y = aX + b$, where a and b are some scalars. We have

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

Therefore,

$$\begin{aligned} f_Y(y) &= \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{|a|} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-((y-b)/a)-\mu)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}\,|a|\sigma} e^{-(y-b-a\mu)^2/2a^2\sigma^2}. \end{aligned}$$

We recognize this as a normal PDF with mean $a\mu + b$ and variance $a^2\sigma^2$. In particular, Y is a normal random variable.

The Monotonic Case

The calculation and the formula for the linear case can be generalized to the case where g is a monotonic function. Let X be a continuous random variable and suppose that its range is contained in a certain interval I, in the sense that $f_X(x) = 0$ for $x \notin I$. We consider the random variable $Y = g(X)$, and assume that g is **strictly monotonic** over the interval I. That is, either

(a) $g(x) < g(x')$ for all $x, x' \in I$ satisfying $x < x'$ (monotonically increasing case), or

(b) $g(x) > g(x')$ for all $x, x' \in I$ satisfying $x < x'$ (monotonically decreasing case).

Furthermore, we assume that the function g is differentiable. Its derivative will necessarily be nonnegative in the increasing case and nonpositive in the decreasing case.

An important fact is that a monotonic function can be "inverted" in the sense that there is some function h, called the inverse of g, such that for all $x \in I$, we have $y = g(x)$ if and only if $x = h(y)$. For example, the inverse of the function $g(x) = 180/x$ considered in Example 3.22 is $h(y) = 180/y$, because we have $y = 180/x$ if and only if $x = 180/y$. Other such examples of pairs of inverse functions include

$$g(x) = ax + b, \qquad h(y) = \frac{y - b}{a},$$

where a and b are scalars with $a \neq 0$ (see Fig. 3.22), and

$$g(x) = e^{ax}, \qquad h(y) = \frac{\ln y}{a},$$

where a is a nonzero scalar.

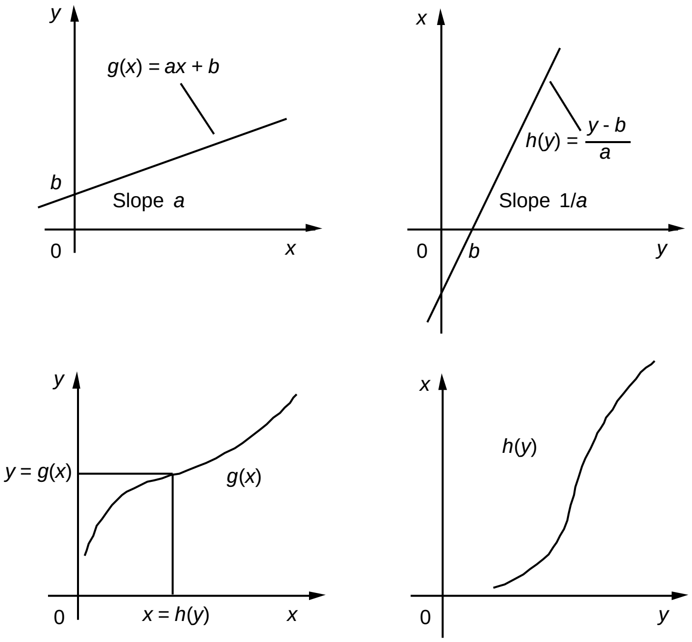


Figure 3.22: A monotonically increasing function g (on the left) and its inverse (on the right). Note that the graph of h has the same shape as the graph of g, except that it is rotated by 90 degrees and then reflected (this is the same as interchanging the x and y axes).

For monotonic functions g, the following is a convenient analytical formula for the PDF of the function $Y = g(X)$.

PDF Formula for a Monotonic Function of a Continuous Random Variable

Suppose that g is monotonic and that for some function h and all x in the range I of X we have

$$y = g(x) \qquad \text{if and only if} \qquad x = h(y).$$

Assume that h has first derivative $(dh/dy)(y)$. Then the PDF of Y in the region where $f_Y(y) > 0$ is given by

$$f_Y(y) = f_X\big(h(y)\big)\left|\frac{dh}{dy}(y)\right|.$$

For a verification of the above formula, assume first that g is monotonically increasing. Then, we have

$$F_Y(y) = \mathbf{P}\big(g(X) \le y\big) = \mathbf{P}\big(X \le h(y)\big) = F_X\big(h(y)\big),$$

where the second equality can be justified using the monotonically increasing property of g (see Fig. 3.23). By differentiating this relation, using also the chain rule, we obtain

$$f_Y(y) = \frac{dF_Y}{dy}(y) = f_X\big(h(y)\big)\frac{dh}{dy}(y).$$

Because g is monotonically increasing, h is also monotonically increasing, so its derivative is positive:

$$\frac{dh}{dy}(y) = \left|\frac{dh}{dy}(y)\right|.$$

This justifies the PDF formula for a monotonically increasing function g. The justification for the case of monotonically decreasing function is similar: we differentiate instead the relation

$$F_Y(y) = \mathbf{P}\big(g(X) \le y\big) = \mathbf{P}\big(X \ge h(y)\big) = 1 - F_X\big(h(y)\big),$$

and use the chain rule.

There is a similar formula involving the derivative of g, rather than the derivative of h. To see this, differentiate the equality $g\big(h(y)\big) = y$, and use the chain rule to obtain

$$\frac{dg}{dh}\big(h(y)\big) \cdot \frac{dh}{dy}(y) = 1.$$

Let us fix some x and y that are related by $g(x) = y$, which is the same as $h(y) = x$. Then,

$$\frac{dg}{dx}(x) \cdot \frac{dh}{dy}(y) = 1,$$

which leads to

$$f_Y(y) = f_X(x) \Big/ \left| \frac{dg}{dx}(x) \right|.$$

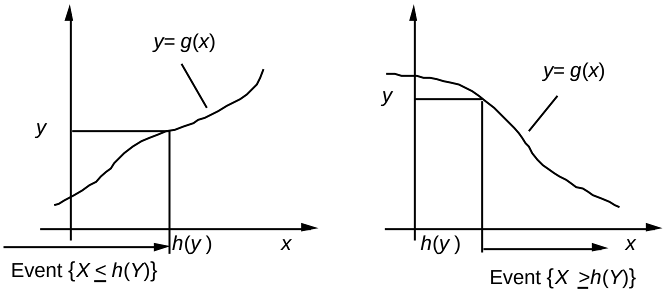


Figure 3.23: Calculating the probability $\mathbf{P}\big(g(X) \le y\big)$. When g is monotonically increasing (left figure), the event $\{g(X) \le y\}$ is the same as the event $\{X \le h(y)\}$. When g is monotonically decreasing (right figure), the event $\{g(X) \le y\}$ is the same as the event $\{X \ge h(y)\}$.

Example 3.22. (Continued) To check the PDF formula, let us apply it to the problem of Example 3.22. In the region of interest, $x \in [30, 60]$, we have $h(y) = 180/y$, and

$$\frac{dF_X}{dh}\big(h(y)\big) = \frac{1}{30}, \qquad \left| \frac{dh}{dy}(y) \right| = \frac{180}{y^2}.$$

Thus, in the region of interest $y \in [3, 6]$, the PDF formula yields

$$f_Y(y) = f_X\big(h(y)\big) \left| \frac{dh}{dy}(y) \right| = \frac{1}{30} \cdot \frac{180}{y^2} = \frac{6}{y^2},$$

consistently with the expression obtained earlier.

Example 3.26. Let $Y = g(X) = X^2$, where X is a continuous uniform random variable in the interval $(0, 1]$. Within this interval, g is monotonic, and its inverse is $h(y) = \sqrt{y}$. Thus, for any $y \in (0, 1]$, we have

$$\left| \frac{dh}{dy}(y) \right| = \frac{1}{2\sqrt{y}}, \qquad f_X(\sqrt{y}) = 1,$$

and

$$f_Y(y) = \begin{cases} \dfrac{1}{2\sqrt{y}} & \text{if } y \in (0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

We finally note that if we interpret PDFs in terms of probabilities of small intervals, the content of our formulas becomes pretty intuitive; see Fig. 3.24.

Functions of Two Random Variables

The two-step procedure that first calculates the CDF and then differentiates to obtain the PDF also applies to functions of more than one random variable.

Example 3.27. Two archers shoot at a target. The distance of each shot from the center of the target is uniformly distributed from 0 to 1, independently of the other shot. What is the PDF of the distance of the losing shot from the center?

Let X and Y be the distances from the center of the first and second shots, respectively. Let also Z be the distance of the losing shot:

$$Z = \max\{X, Y\}.$$

We know that X and Y are uniformly distributed over $[0, 1]$, so that for all $z \in [0, 1]$, we have

$$\mathbf{P}(X \le z) = \mathbf{P}(Y \le z) = z.$$

Thus, using the independence of X and Y, we have for all $z \in [0, 1]$,

$$\begin{aligned} F_Z(z) &= \mathbf{P}\big(\max\{X, Y\} \le z\big) \\ &= \mathbf{P}(X \le z,\, Y \le z) \\ &= \mathbf{P}(X \le z)\mathbf{P}(Y \le z) \\ &= z^2. \end{aligned}$$

Differentiating, we obtain

$$f_Z(z) = \begin{cases} 2z & \text{if } 0 \le z \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Example 3.28. Let X and Y be independent random variables that are uniformly distributed on the interval $[0, 1]$. What is the PDF of the random variable $Z = Y/X$?

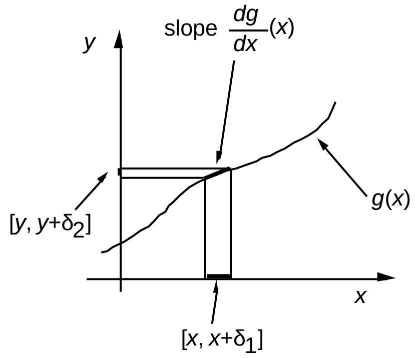


Figure 3.24: Illustration of the PDF formula for a monotonically increasing function g. Consider an interval $[x, x+\delta_1]$, where δ_1 is a small number. Under the mapping g, the image of this interval is another interval $[y, y+\delta_2]$. Since $(dg/dx)(x)$ is the slope of g, we have

$$\frac{\delta_2}{\delta_1} \approx \frac{dg}{dx}(x),$$

or in terms of the inverse function,

$$\frac{\delta_1}{\delta_2} \approx \frac{dh}{dy}(y),$$

We now note that the event $\{x \le X \le x+\delta_1\}$ is the same as the event $\{y \le Y \le y+\delta_2\}$. Thus,

$$\begin{aligned} f_Y(y)\delta_2 &\approx \mathbf{P}(y \le Y \le y+\delta_2) \\ &= \mathbf{P}(x \le X \le x+\delta_1) \\ &\approx f_X(x)\delta_1. \end{aligned}$$

We move δ_1 to the left-hand side and use our earlier formula for the ratio δ_2/δ_1, to obtain

$$f_Y(y)\frac{dg}{dx}(x) = f_X(x).$$

Alternatively, if we move δ_2 to the right-hand side and use the formula for δ_1/δ_2, we obtain

$$f_Y(y) = f_X\big(h(y)\big) \cdot \frac{dh}{dy}(y).$$

We will find the PDF of Z by first finding its CDF and then differentiating. We consider separately the cases $0 \le z \le 1$ and $z > 1$. As shown in Fig. 3.25, we have

$$F_Z(z) = \mathbf{P}\left(\frac{Y}{X} \le z\right) = \begin{cases} z/2 & \text{if } 0 \le z \le 1, \\ 1 - 1/(2z) & \text{if } z > 1, \\ 0 & \text{otherwise.} \end{cases}$$

By differentiating, we obtain

$$f_Z(z) = \begin{cases} 1/2 & \text{if } 0 \le z \le 1, \\ 1/(2z^2) & \text{if } z > 1, \\ 0 & \text{otherwise.} \end{cases}$$

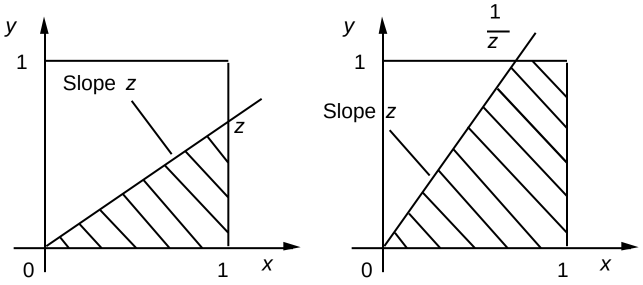


Figure 3.25: The calculation of the CDF of $Z = Y/X$ in Example 3.28. The value $\mathbf{P}(Y/X \le z)$ is equal to the shaded subarea of the unit square. The figure on the left deals with the case where $0 \le z \le 1$ and the figure on the right refers to the case where $z > 1$.

Example 3.29. Romeo and Juliet have a date at a given time, and each, independently, will be late by an amount of time that is exponentially distributed with parameter λ. What is the PDF of the difference between their times of arrival?

Let us denote by X and Y the amounts by which Romeo and Juliet are late, respectively. We want to find the PDF of $Z = X - Y$, assuming that X and Y are independent and exponentially distributed with parameter λ. We will first calculate the CDF $F_Z(z)$ by considering separately the cases $z \ge 0$ and $z < 0$ (see Fig. 3.26).

For $z \ge 0$, we have (see the left side of Fig. 3.26)

$$\begin{aligned} F_Z(z) &= \mathbf{P}(X - Y \le z) \\ &= 1 - \mathbf{P}(X - Y > z) \\ &= 1 - \int_0^\infty \left(\int_{z+y}^\infty f_{X,Y}(x,y)\,dx \right) dy \\ &= 1 - \int_0^\infty \lambda e^{-\lambda y} \left(\int_{z+y}^\infty \lambda e^{-\lambda x}\,dx \right) dy \\ &= 1 - \int_0^\infty \lambda e^{-\lambda y} e^{-\lambda(z+y)}\,dy \\ &= 1 - e^{-\lambda z} \int_0^\infty \lambda e^{-2\lambda y}\,dy \\ &= 1 - \frac{1}{2} e^{-\lambda z}. \end{aligned}$$

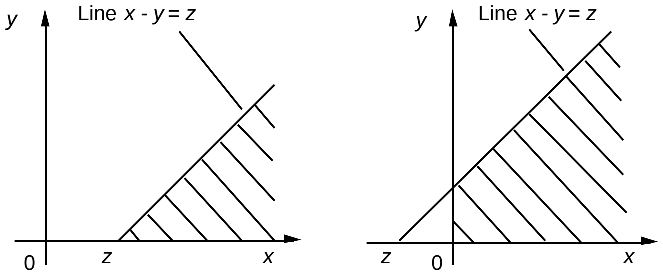


Figure 3.26: The calculation of the CDF of $Z = X - Y$ in Example 3.29. To obtain the value $\mathbf{P}(X - Y > z)$ we must integrate the joint PDF $f_{X,Y}(x, y)$ over the shaded area in the above figures, which correspond to $z \geq 0$ (left side) and $z < 0$ (right side).

For the case $z < 0$, we can use a similar calculation, but we can also argue using symmetry. Indeed, the symmetry of the situation implies that the random variables $Z = X - Y$ and $-Z = Y - X$ have the same distribution. We have

$$F_Z(z) = \mathbf{P}(Z \leq z) = \mathbf{P}(-Z \geq -z) = \mathbf{P}(Z \geq -z) = 1 - F_Z(-z).$$

With $z < 0$, we have $-z \geq 0$ and using the formula derived earlier,

$$F_Z(z) = 1 - F_Z(-z) = 1 - \left(1 - \frac{1}{2}e^{-\lambda(-z)}\right) = \frac{1}{2}e^{\lambda z}.$$

Combining the two cases $z \geq 0$ and $z < 0$, we obtain

$$F_Z(z) = \begin{cases} 1 - \dfrac{1}{2}e^{-\lambda z} & \text{if } z \geq 0, \\ \dfrac{1}{2}e^{\lambda z} & \text{if } z < 0, \end{cases}$$

We now calculate the PDF of Z by differentiating its CDF. We obtain

$$f_Z(z) = \begin{cases} \dfrac{\lambda}{2}e^{-\lambda z} & \text{if } z \geq 0, \\ \\ \dfrac{\lambda}{2}e^{\lambda z} & \text{if } z < 0, \end{cases}$$

or

$$f_Z(z) = \frac{\lambda}{2}e^{-\lambda|z|}.$$

This is known as a **two-sided exponential PDF**, also known as the **Laplace PDF**.

3.7 SUMMARY AND DISCUSSION

Continuous random variables are characterized by PDFs and arise in many applications. PDFs are used to calculate event probabilities. This is similar to the use of PMFs for the discrete case, except that now we need to integrate instead of adding. Joint PDFs are similar to joint PMFs and are used to determine the probability of events that are defined in terms of multiple random variables. Finally, conditional PDFs are similar to conditional PMFs and are used to calculate conditional probabilities, given the value of the conditioning random variable.

We have also introduced a few important continuous probability laws and derived their mean and variance. A summary is provided in the table that follows.

Summary of Results for Special Random Variables

Continuous Uniform Over $[a, b]$:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \text{var}(X) = \frac{(b-a)^2}{12}.$$

Exponential with Parameter λ:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{1}{\lambda}, \qquad \text{var}(X) = \frac{1}{\lambda^2}.$$

Normal with Parameters μ and σ^2:

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

$$\mathbf{E}[X] = \mu, \qquad \text{var}(X) = \sigma^2.$$

SOLVED PROBLEMS

SECTION 3.1. Continuous Random Variables and PDFs

Problem 1. * Show that the expected value of a continuous random variable X satisfies

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}(X > x)\,dx - \int_0^\infty \mathbf{P}(X < -x)\,dx.$$

Solution. We have

$$\begin{aligned}
\int_0^\infty \mathbf{P}(X > x)\,dx &= \int_0^\infty \left(\int_x^\infty f_X(y)\,dy\right)\,dx \\
&= \int_0^\infty \left(\int_0^y f_X(y)\,dx\right)\,dy \\
&= \int_0^\infty f_X(y)\left(\int_0^y dx\right)\,dy \\
&= \int_0^\infty y f_X(y)\,dy,
\end{aligned}$$

and similarly we show that

$$\int_0^\infty \mathbf{P}(X < -x)\,dx = -\int_{-\infty}^0 y f_X(y)\,dy.$$

Combining the two relations above, we obtain the desired result.

Problem 2. * **Laplace random variable.** Let X have the PDF

$$f_X(x) = \frac{\lambda}{2}e^{-\lambda|x|},$$

where λ is a positive scalar. Verify that f_X satisfies the normalization condition. Find $\mathbf{E}[X]$ and $\text{var}(X)$.

Solution. We have

$$\int_{-\infty}^\infty f_X(x)dx = \int_{-\infty}^\infty \frac{\lambda}{2}e^{-\lambda|x|}dx = 2\cdot\frac{1}{2}\int_0^\infty \lambda e^{-\lambda x}dx = 2\cdot\frac{1}{2} = 1,$$

where we have used the normalization property of the exponential PDF, $\int_0^\infty \lambda e^{-\lambda x}dx = 1$.

By symmetry of the PDF, we have $\mathbf{E}[X] = 0$. We also have

$$\mathbf{E}[X^2] = \int_{-\infty}^{\infty} x^2 \frac{\lambda}{2} e^{-\lambda|x|} dx = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2},$$

where we have used the fact that the second moment of the exponential PDF is $2/\lambda^2$. Thus $\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = 2/\lambda^2$.

SECTION 3.2. Cumulative Distribution Functions

Problem 3. Find the PDF, the mean, and the variance of the random variable X with CDF

$$F_X(x) = \begin{cases} 1 - \frac{a^3}{x^3} & \text{if } x \geq a, \\ 0 & \text{if } x < a, \end{cases}$$

where a is a positive constant.

Solution. We have

$$f_X(x) = \frac{dF_X}{dx}(x) = \begin{cases} 3a^3 x^{-4} & \text{if } x \geq a, \\ 0 & \text{if } x < a. \end{cases}$$

Also

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^{\infty} x \cdot 3a^3 x^{-4} dx = 3a^3 \int_a^{\infty} x^{-3} dx = 3a^3 \left(-\frac{1}{2}x^{-2}\right)\Bigg|_a^{\infty} = \frac{3a}{2}.$$

Finally, we have

$$\mathbf{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_a^{\infty} x^2 \cdot 3a^3 x^{-4} dx = 3a^3 \int_a^{\infty} x^{-2} dx = 3a^3 \left(-x^{-1}\right)\Bigg|_a^{\infty} = 3a^2,$$

so the variance is

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = 3a^2 - \left(\frac{3a}{2}\right)^2 = \frac{3a^2}{4}.$$

Problem 4. Consider a triangle with height h. Let X be the distance from the base of a point randomly chosen within the triangle. What is the CDF and the PDF of X?

Solution. Let $A = bh/2$ be the area of the given triangle, where b is the length of the base. From the randomly chosen point, draw a line parallel to the base, and let A_x be the area of the triangle thus formed. The height of this triangle is $h - x$ and its base has length $b(h-x)/h$. Thus $A_x = b(h-x)^2/(2h)$. For $x \in [0, h]$, we have

$$F_X(x) = 1 - \mathbf{P}(X > x) = 1 - \frac{A_x}{A} = 1 - \frac{b(h-x)^2/(2h)}{bh/2} = 1 - \left(\frac{h-x}{h}\right)^2,$$

while for $x \notin [0, h]$, we have $F_X(x) = 0$.

The PDF is obtained by differentiating the CDF. We have

$$f_X(x) = \frac{dF_X}{dx}(x) = \begin{cases} \frac{2(h-x)}{h^2} & \text{if } 0 \le x \le h, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 5. * For a nonnegative, integer-valued random variable X, show

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} \mathbf{P}(X \ge i) = \sum_{i=0}^{\infty} \bigl(1 - F_X(i)\bigr).$$

Solution. We have

$$\begin{aligned} \sum_{i=1}^{\infty} \mathbf{P}(X \ge i) &= \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} \mathbf{P}(X = k) \\ &= \sum_{k=1}^{\infty} \sum_{i=1}^{k} \mathbf{P}(X = k) \\ &= \sum_{k=1}^{\infty} k\mathbf{P}(X = k). \end{aligned}$$

Problem 6. * The *median* of a random variable X is the number μ that satisfies $F_X(\mu) = \frac{1}{2}$. Find the median of the exponential random variable with parameter λ.

Solution. We are given that $F_X(\mu) = \frac{1}{2}$ or

$$\frac{1}{2} = \int_0^{\mu} \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}\Big|_0^{\mu} = 1 - e^{-\lambda \mu},$$

or

$$-\lambda \mu = \ln \frac{1}{2},$$

from which

$$\mu = \frac{\ln 2}{\lambda}.$$

Problem 7. * **Simulating a continuous random variable.** Computers have subroutines that can generate experimental values of a random variable X that is uniformly distributed in the interval $[0, 1]$. Such a subroutine can be used to generate experimental values of a continuous random variable with given CDF $F(y)$ as follows: Each time X takes a value $x \in (0, 1)$, we generate the unique value y for which $F(y) = x$. (We can neglect the zero probability event that X takes the value 0 or 1.)

(a) Show that the CDF $F_Y(y)$ of the random variable Y thus generated is indeed equal to the given $F(y)$.

(b) Describe how the procedure can be used to simulate an exponential random variable with parameter λ.

(c) How can the procedure be generalized to simulate a discrete integer-valued random variable.

Solution. (a) For the random variable Y thus generated, we have

$$F_Y(y) = \mathbf{P}(Y \leq y) \leq \mathbf{P}\big(X \leq F(y)\big).$$

But since X is uniformly distributed in $[0, 1]$, we have that $\mathbf{P}\big(X \leq F(y)\big)$ is equal to $F(y)$.

(b) The exponential PDF has the form $F(y) = 1 - e^{-\lambda y}$ for $y \geq 0$. Thus to generate values of Y, we should generate values $x \in (0, 1)$ of a uniformly distributed random variable X, and set y to the value for which $1 - e^{-\lambda y} = x$, or $y = -\ln(1 - x)/\lambda$.

(c) For a generated value x, let y be the integer such that $F(y) < x \leq F(y + 1)$.

SECTION 3.3. Normal Random Variables

Problem 8. Let X and Y be normal random variables with means 0 and 1, respectively, and variances 1 and 4, respectively.

(a) Find $\mathbf{P}(X \leq 1.5)$ and $\mathbf{P}(X \leq -1)$.

(b) What is the PDF of $Z = (Y - 1)/2$.

(c) Find $\mathbf{P}(-1 \leq Y \leq 1)$.

Solution. (a) X is a standard normal, so we have $\mathbf{P}(X \leq 1.5) \leq \Phi(1.5) = 0.9332$. Also $\mathbf{P}(X \leq -1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$.

(b) By subtracting from Y its mean and dividing by the standard deviation, we obtain the standard normal.

(c) We have $\mathbf{P}(-1 \leq Y \leq 1) = \mathbf{P}(-1 \leq (Y - 1)/2 \leq 0) = \mathbf{P}(-1 \leq Z \leq 0) = \mathbf{P}(0 \leq Z \leq 1) = \Phi(1) - \Phi(0) = 0.8413 - 0.5 = 0.3413$.

Problem 9. Let X be a normal random variable with zero mean and standard deviation σ. Use the normal tables to compute the probabilities of the events $\{X \geq k\sigma\}$ and $\{|X| \leq k\sigma\}$ for $k = 1, 2, 3$.

Solution. Let Z be the standard normal random variable. We have $X = Z/\sigma$, so

$$\mathbf{P}(X \geq k\sigma) = \mathbf{P}(Z \geq k) = 1 - \Phi(k).$$

From the normal tables we have

$$\Phi(1) = 0.8413, \qquad \Phi(2) = 0.9772, \qquad \Phi(3) = 0.9986.$$

Thus $\mathbf{P}(X \geq \sigma) = 0.1587$, $\mathbf{P}(X \geq 2\sigma) = 0.0228$, $\mathbf{P}(X \geq 3\sigma) = 0.0014$.

We also have

$$\mathbf{P}(|X| \leq k\sigma) = \mathbf{P}(|Z| \leq k) = \Phi(k) - \mathbf{P}(Z \leq -k) = \Phi(k) - \big(1 - \Phi(k)\big) = 2\Phi(k) - 1.$$

Using the normal table values above, we obtain $\mathbf{P}(|X| \leq \sigma) = 0.6826$, $\mathbf{P}(|X| \leq 2\sigma) = 0.9544$, $\mathbf{P}(|X| \leq 3\sigma) = 0.9972$.

Problem 10. A city's temperature is modeled as a normal random variable with mean and standard deviation both equal to 10 degrees Celcius. What is the probability that the temperature at a randomly chosen time will be less or equal to 60 degrees Fahreneit?

Solution. If X is the temperature in Celcius, the temperature in Fahreneit is $Y = 32 + 9X/5$. Therefore, 60 degrees Fahreneit corresponds to 15 degrees Celcius. So if Z is the standard normal, we have using $\mathbf{E}[X] = \sigma_X = 10$,

$$\mathbf{P}(Y \geq 60) = \mathbf{P}(X \geq 15) = \mathbf{P}\left(Z \geq \frac{15 - \mathbf{E}[X]}{\sigma_X}\right) = \mathbf{P}(Z \geq 0.5) = 1 - \Phi(0.5).$$

From the normal tables we have $\Phi(0.5) = 0.6915$, so $\mathbf{P}(Y \geq 60) = 0.3085$.

Problem 11. * Show that the normal PDF satisfies the normalization property.

Solution. We note that

$$\begin{aligned}\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx\right)^2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy \\ &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r dr d\theta \\ &= \int_0^{\infty} e^{-\frac{r^2}{2}} r dr \\ &= -e^{-u}\Big|_0^{\infty} \\ &= 1.\end{aligned}$$

Thus we have

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx = 1,$$

because the integral is positive. It follows that

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\,du = 1.$$

SECTION 3.4. Conditioning on an Event

Problem 12. Let X be a random variable with PDF

$$f_X(x) = \begin{cases} \frac{x}{4} & \text{if } 1 < x \leq 3, \\ 0 & \text{otherwise,} \end{cases}$$

and let A be the event $\{X \geq 2\}$.

(a) Find $\mathbf{E}[X]$, $\mathbf{P}(A)$, and $\mathbf{E}[X \,|\, A]$.

(b) Let $Y = X^2$. Find $\mathbf{E}[Y]$ and $\mathrm{var}(Y)$.

Solution. (a) We have

$$\mathbf{E}[X] = \int_1^3 \frac{x^2}{4} dx = \frac{13}{6},$$

$$\mathbf{P}(A) = \int_2^3 \frac{x}{4} dx = \frac{5}{8}.$$

To obtain $\mathbf{E}[X \,|\, A]$, we compute $f_{X \,|\, A}(x \,|\, A)$. We have

$$f_{X \,|\, A}(x \,|\, A) = \begin{cases} \frac{f_X(x)}{\mathbf{P}(A)} & \text{if } x \in A, \\ 0 & \text{otherwise,} \end{cases} = \begin{cases} \frac{2x}{5} & \text{if } 2 \leq x \leq 3, \\ 0 & \text{otherwise,} \end{cases}$$

from which

$$\mathbf{E}[X \,|\, A] = \int_2^3 x \cdot \frac{2x}{5} dx = \frac{38}{15}.$$

(b) We have

$$\mathbf{E}[Y] = \mathbf{E}[X^2] = \int_1^3 \frac{x^3}{4} dx = 5.$$

We also have

$$\mathbf{E}[Y^2] = \mathbf{E}[X^4] = \int_1^3 \frac{x^5}{4} dx = \frac{91}{3}.$$

Thus

$$\mathrm{var}(Y) = \mathbf{E}[Y^2] - \big(\mathbf{E}[Y]\big)^2 = \frac{91}{3} - 5^2 = \frac{16}{3}.$$

Problem 13. The random variable X has the PDF

$$f_X(x) = \begin{cases} cx^{-2} & \text{if } 1 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Determine the numerical value of c.

(b) Let A be the event $\{X > 1.5\}$. Calculate $\mathbf{P}(A)$ and sketch the conditional PDF for X given that A has occurred.

(c) Let $Y = X^2$. Calculate the conditional expectation and the conditional variance for Y given A.

Solution. (a)

$$c = \frac{1}{\int_1^2 x^{-2} dx} = 2.$$

(b)

$$\mathbf{P}(A) = \int_{1.5}^2 2x^{-2} dx = \frac{1}{3}.$$

$$f_{X\,|\,A}(x\,|\,A) = \begin{cases} 6x^{-2} & \text{if } 1.5 < x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

(c) We have

$$\mathbf{E}[Y\,|\,A] = \mathbf{E}[X^2\,|\,A] = \int_{1.5}^{2} 6x^{-2}x^2 dx = 3,$$

$$\mathbf{E}[Y^2\,|\,A] = \mathbf{E}[X^4\,|\,A] = \int_{1.5}^{2} 6x^{-2}x^4 dx = \frac{37}{4},$$

and

$$\text{var}(Y\,|\,A) = \frac{37}{4} - 3^2 = \frac{1}{4}.$$

Problem 14. Calamity Jane goes to the bank to make a deposit, and is equally likely to find 0 or 1 customer ahead of her. The times of service of these customers are independent and exponentially distributed with parameter λ. What is the CDF of Jane's waiting time?

Solution. Let X be the waiting time and Y be the number of customers found. For $x < 0$, we have $F_X(x) = 0$. For $x \ge 0$,

$$F_X(x) = \mathbf{P}(X \le x) = \frac{1}{2}\big(\mathbf{P}(X \le x\,|\,Y = 0) + \mathbf{P}(X \le x\,|\,Y = 1)\big).$$

We have

$$\mathbf{P}(X \le x\,|\,Y = 0) = 1,$$

$$\mathbf{P}(X \le x\,|\,Y = 1) = 1 - e^{-\lambda x}.$$

Thus

$$F_X(x) = \frac{1}{2}(2 - e^{-\lambda x}), \qquad x \ge 0.$$

Problem 15. An absent-minded professor schedules two student appointments for the same time. The appointment durations are independent and exponentially distributed with mean thirty minutes. The first student arrives on time, but the second student arrives five minutes late. What is the expected value of the time between the arrival of the first student and the departure of the second student?

Solution. The expected value in question is

$$\begin{aligned}\mathbf{E}[\text{Time}] = {} & (5 + \mathbf{E}[\text{stay of 2nd student}]) \cdot \mathbf{P}(\text{1st stays less or equal to 5 minutes}) \\ & + \big(\mathbf{E}[\text{stay of 1st}\,|\,\text{stay of 1st } \ge 5] + \mathbf{E}[\text{stay of 2nd}]\big) \\ & \cdot \mathbf{P}(\text{1st stays more than 5 minutes}).\end{aligned}$$

We have $\mathbf{E}[\text{stay of 2nd student}] = 30$, and, using the memoryless property of the exponential distribution,

$$\mathbf{E}[\text{stay of 1st}\,|\,\text{stay of 1st } \ge 5] = 5 + \mathbf{E}[\text{stay of 1st}] = 35.$$

Also

$$\mathbf{P}(\text{1st student stays less or equal to 5 minutes}) = 1 - e^{-5/30},$$

$$\mathbf{P}(\text{1st student stays more than 5 minutes}) = e^{-5/30}.$$

By substitution we obtain

$$\mathbf{E}[\text{Time}] = (5+30)\cdot(1-e^{-5/30}) + (35+30)\cdot e^{-5/30} = 35 + 30\cdot e^{-5/30} = 60.394.$$

Problem 16. Alvin throws darts at a circular target of radius R and is equally likely to hit any point in the target. Let X be the distance of Alvin's hit from the center.

(a) Find the PDF, the mean, and the variance of X.

(b) The target has an inner circle of radius T. If $X \le T$, Alvin gets a score of $S = 1/X$. Otherwise his score is $S = 0$. Find the CDF of S. Is S a continuous random variable?

Solution. (a) For $x \in [0, R]$, we have

$$\mathbf{P}(X \le x) = \frac{\pi x^2}{\pi R^2} = \left(\frac{x}{R}\right)^2.$$

By differentiating, we obtain the PDF

$$f_X(x) = \begin{cases} \frac{2x}{R^2} & \text{if } 0 \le x \le R, \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$\mathbf{E}[X] = \int_0^R \frac{2x^2}{R^2}dx = \frac{2R}{3}.$$

Also

$$\mathbf{E}[X^2] = \int_0^R \frac{2x^3}{R^2}dx = \frac{R^2}{2},$$

so

$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{R^2}{2} - \frac{4R^2}{9} = \frac{R^2}{18}.$$

(b) If $s < 0$, we have $\mathbf{P}(S \le s) = 0$, and if $0 \le s < 1/T$, we have

$$\mathbf{P}(S \le s) = \mathbf{P}(\text{Alvin's hit is outside the inner circle}) = 1 - \mathbf{P}(X \le T) = 1 - \frac{T^2}{R^2}.$$

Finally if $1/T \le s$, we have

$$\mathbf{P}(S \le s) = \mathbf{P}(X \le T)\mathbf{P}(S \le s \,|\, X \le T) + \mathbf{P}(X > T)\mathbf{P}(S \le s \,|\, X > T).$$

We have

$$\mathbf{P}(X \le T) = \frac{T^2}{R^2}, \qquad \mathbf{P}(X > T) = 1 - \frac{T^2}{R^2},$$

and since $S = 0$ when $X > T$, we have

$$\mathbf{P}(S \le s \,|\, X > T) = 1.$$

Finally, we have

$$\mathbf{P}(S \le s \,|\, X \le T) = \mathbf{P}(1/X \le s \,|\, X \le T) = \frac{\mathbf{P}(1/s \le X \le T)}{\mathbf{P}(X \le T)} = \frac{\frac{\pi T^2 - \pi(1/s)^2}{\pi R^2}}{\frac{\pi T^2}{\pi R^2}} = 1 - \frac{1}{s^2 T^2}.$$

Combining the above relations, we obtain

$$\mathbf{P}(S \le s) = \frac{T^2}{R^2}\left(1 - \frac{1}{s^2T^2}\right) + 1 - \frac{T^2}{R^2} = 1 - \frac{1}{s^2R^2}.$$

Collecting the results of the preceding calculations, the CDF of S is

$$F_S(s) = \begin{cases} 0 & \text{if } s < 0, \\ 1 - \frac{T^2}{R^2} & \text{if } 0 \le s < 1/T, \\ 1 - \frac{1}{s^2R^2} & \text{if } 1/T \le s. \end{cases}$$

Because F_S has a discontinuity at $S = 0$, the random variable S is not continuous.

Problem 17. * Consider the following two-sided exponential PDF

$$f_X(x) = \begin{cases} p\lambda e^{-\lambda x} & \text{if } x \ge 0, \\ (1-p)\lambda e^{\lambda x} & \text{if } x < 0, \end{cases}$$

where λ and p are scalars with $\lambda > 0$ and $p \in [0, 1]$. Find the mean and the variance of X in two ways:

(a) By straightforward calculation of the associated expected values.

(b) By using a divide-and-conquer strategy and the mean and variance of the (one-sided) exponential random variable.

Solution. (a)

$$\begin{aligned} \mathbf{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx \\ &= \int_{-\infty}^{0} x(1-p)\lambda e^{\lambda x}\,dx + \int_0^{\infty} xp\lambda e^{-\lambda x}\,dx \\ &= -\frac{1-p}{\lambda} + \frac{p}{\lambda} \\ &= \frac{2p-1}{\lambda}, \end{aligned}$$

$$\begin{aligned} \mathbf{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \\ &= \int_{-\infty}^{0} x^2(1-p)\lambda e^{\lambda x}\,dx + \int_0^{\infty} x^2 p\lambda e^{-\lambda x}\,dx \\ &= \frac{2(1-p)}{\lambda^2} + \frac{2p}{\lambda^2} = \frac{2}{\lambda^2}, \end{aligned}$$

and

$$\mathrm{var}(X) = \frac{2}{\lambda^2} - \left(\frac{2p-1}{\lambda}\right)^2.$$

(b) Let Y be an exponential random variable with parameter λ and let Z be a 0-1 random variable with a probability p of being 1. Noting that

$$f_X(x) = pf_Y(x) + (1-p)f_{-Y}(x),$$

we can express X as

$$X = \begin{cases} Y & \text{if } Z = 1, \\ -Y & \text{if } Z = 0. \end{cases}$$

It follows that

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{P}(Z=1)\mathbf{E}[X \mid Z=1] + \mathbf{P}(Z=0)\mathbf{E}[X \mid Z=0] \\ &= p\mathbf{E}[Y] + (1-p)\mathbf{E}[-Y] \\ &= \frac{p}{\lambda} - \frac{1-p}{\lambda} \\ &= \frac{2p-1}{\lambda}, \end{aligned}$$

$$\begin{aligned} \mathbf{E}[X^2] &= \mathbf{P}(Z=1)\mathbf{E}[X^2 \mid Z=1] + \mathbf{P}(Z=0)\mathbf{E}[X^2 \mid Z=0] \\ &= p\mathbf{E}[Y^2] + (1-p)\mathbf{E}[(-Y)^2] \\ &= \frac{2p}{\lambda^2} + \frac{2(1-p)}{\lambda^2} \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

and

$$\mathrm{var}(X) = \frac{2}{\lambda^2} - \left(\frac{2p-1}{\lambda}\right)^2.$$

Problem 18. * **Mixed random variables.** Probability models sometimes involve random variables which can be viewed as a mixture of a discrete random variable Y and a continuous random variable Z. By this we mean that the experimental value of X is obtained according to the probability law of Y with a given probability p, and according to the probability law of Z with the complementary probability $1-p$. Then, X is called a *mixed random variable* and its CDF is given, using the total probability theorem, by

$$\begin{aligned} F_X(x) &= \mathbf{P}(X \le x) \\ &= p\mathbf{P}(Y \le x) + (1-p)\mathbf{P}(Z \le x)\,. \\ &= p \cdot F_Y(x) + (1-p) \cdot F_Z(x) \end{aligned}$$

Its expected value is defined in a way that conforms to the total expectation theorem:

$$\mathbf{E}[X] = p\mathbf{E}[Y] + (1-p)\mathbf{E}[Z].$$

The taxi stand and the bus stop near Al's home are in the same location. Al goes there at a given time and if a taxi is waiting (this happens with probability 2/3) he boards it. Otherwise he waits for a taxi or a bus to come, whichever comes first. The next taxi will arrive in a time that is uniformly distributed between 0 and 10 minutes, while the next bus will arrive in exactly 5 minutes. The probability of boarding the next bus, given that Al has to wait, is

$$\mathbf{P}(\text{a taxi will take more than 5 minutes to arrive}) = \frac{1}{2}.$$

Find the CDF and the expected value of Al's waiting time.

Solution. Let A be the event that Al will find a taxi waiting or will be picked up by the bus after 5 minutes. Al's waiting time, call it X, is a mixed random variable. With probability

$$\mathbf{P}(A) = \frac{2}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{6},$$

it is equal to its discrete component Y (corresponding to either finding a taxi waiting, or boarding the bus), which has PMF

$$p_Y(y) = \begin{cases} \frac{2}{3\mathbf{P}(A)} & \text{if } y = 0, \\ \frac{1}{6\mathbf{P}(A)} & \text{if } y = 5, \end{cases} = \begin{cases} \frac{12}{15} & \text{if } y = 0, \\ \frac{3}{15} & \text{if } y = 5. \end{cases}$$

[This equation follows from the following calculation:

$$p_Y(0) = \mathbf{P}(Y = 0 \,|\, A) = \frac{\mathbf{P}(Y = 0, A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(Y = 0)}{\mathbf{P}(A)} = \frac{2}{3\mathbf{P}(A)},$$

and the calculation for $p_Y(1)$ is similar.] With the complementary probability $1-\mathbf{P}(A)$, the waiting time is equal to its continuous component Z (corresponding to boarding a taxi after having to wait for some time less than 5 minutes), which has PDF

$$f_Z(z) = \begin{cases} 1/5 & \text{if } 0 \le z \le 5, \\ 0 & \text{otherwise.} \end{cases}$$

The CDF is given by $f_X(x) = \mathbf{P}(A) f_Y(y) + \big(1 - \mathbf{P}(A)\big) f_Z(z)$ or

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{5}{6}\frac{12}{15} + \frac{1}{6}\frac{z}{5} & \text{if } 0 \le x < 5, \\ 1 & \text{if } 5 \le x. \end{cases}$$

The expected value of the waiting time is

$$\mathbf{E}[X] = \mathbf{P}(A)\mathbf{E}[Y] + \big(1 - \mathbf{P}(A)\big)\mathbf{E}[Z] = \frac{5}{6} \cdot \frac{3}{15} \cdot 5 + \frac{1}{6} \cdot \frac{5}{2} = \frac{15}{12}.$$

SECTION 3.5. Multiple Continuous Random Variables

Problem 19. A history professor is so disorganized, that on any given day, he is equally likely to be lecturing about any event in the past 1000 years, regardless of what he lectured on at any point in the past. Find the CDF of the time separating the subject matter of any two lectures.

Solution. The date of the events upon which the professor lectures on any given day, is uniformly distributed from 0 to 1000, independently of other lectures. Let X_1 be the date of the material covered in lecture 1, and X_2 be the date of the material of lecture 2. Let A be the event $\{X_1 < X_2\}$ and let $T = |X_1 - X_2|$. Then the CDF of T is

$$F_T(t) = \mathbf{P}(|X_1 - X_2| \le t) = \frac{1}{2}\mathbf{P}(X_1 - X_2 \le t \,|\, A^c) + \frac{1}{2}\mathbf{P}(X_2 - X_1 \le t \,|\, A).$$

We have

$$\mathbf{P}(X_1 - X_2 \le t \,|\, A^c) = \mathbf{P}(X_2 - X_1 \le t \,|\, A) = \begin{cases} 0 & \text{if } t < 0, \\ 1 - \frac{(1000-t)^2}{(1000)^2} & \text{if } 0 \le t \le 1000, \\ 1 & \text{if } t > 1000, \end{cases}$$

so $F_T(t)$ is also equal to the above expression.

Problem 20. The random variables X and Y describe the cartesian coordinates of a random point within a unit circle centered at the origin. Let $W = \max\{X, Y\}$. Find the CDF of W. You need not evaluate integrals.

Solution. To find the CDF of W, we need to find

$$\mathbf{P}(W \le w) = \mathbf{P}(\max\{X, Y\} \le w) = \mathbf{P}(X \le w, Y \le w).$$

To calculate this we need to integrate the joint density of X, Y over the region where both are smaller than w. This region is the infinite quarter of the plane, with upper right hand corner at (w, w). Clearly, there are several cases to consider, depending upon the value of w, since the joint PDF of X and Y, is only nonnegative within the unit circle centered at the origin. Thus, we have

$$F_W(w) = \begin{cases} 0 & \text{if } w \le -\frac{\sqrt{2}}{2}, \\ \frac{1}{\pi}\int_{-\sqrt{1-w^2}}^{w}(w + \sqrt{1-x^2})\,dx & \text{if } -\frac{\sqrt{2}}{2} < w \le 0, \\ \frac{1}{4} + \frac{w^2 + 2w\sqrt{1-w^2}}{\pi} + \frac{2}{\pi}\int_{-1}^{-\sqrt{1-w^2}}\sqrt{1-x^2}\,dx & \text{if } 0 < w \le \frac{\sqrt{2}}{2}, \\ 1 - \frac{4}{\pi}\int_{w}^{1}\sqrt{1-x^2}\,dx & \text{if } \frac{\sqrt{2}}{2} < w \le 1, \\ 1 & \text{if } w > 1. \end{cases}$$

Problem 21. Consider the random variables X and Y with joint PDF of the form

$$f_{X,Y}(x, y) = \gamma x^2 y, \qquad \text{for } x \ge 0, y \ge 0, x + 3y \le 10.$$

(a) Find the value of the constant γ.

(b) Find $\mathbf{E}\big[\max\{X, Y\}\big]$.

Solution. (a) The normalization property requires that

$$\gamma = \left(\int_0^{10}\int_0^{\frac{10-x}{3}} x^2 y\,dy\,dx\right)^{-1} = \frac{54}{10{,}000}$$

(b) The expectation of the maximum of X and Y will be the sum of two integrals, since we have

$$\max\{x,y\} = \begin{cases} x & \text{if } x \geq y, \\ y & \text{otherwise.} \end{cases}$$

Let A be the region where $x \geq y$ and where the joint PDF is nonzero, and let B be the region where $x < y$. Then we have

$$\mathbf{E}\big[\max\{X,Y\}\big] = \int\int_{(x,y)\in A} x f_{X,Y}(x,y)\, d(Area) + \int\int_{(x,y)\in B} y f_{X,Y}(x,y)\, d(Area),$$

or

$$\begin{aligned} \mathbf{E}\big[\max\{X,Y\}\big] &= \int_A x f_{X,Y}(x,y)\, d(Area) + \int_B y f_{X,Y}(x,y)\, d(Area) \\ &= \int_0^{10/4} \int_y^{10-3y} x\gamma x^2 y\, dx\, dy + \int_0^{10/4} \int_x^{(10-x)/3} y\gamma x^2 y\, dy\, dx. \end{aligned}$$

Problem 22. A point is chosen at random from a semicircle of radius R. The semicircle is centered at the origin, and is in the upper half plane.

(a) Find the joint PDF of its coordinates X and Y.

(b) Find the marginal PDF of Y and use it to find $\mathbf{E}[Y]$.

(c) Now check your answer in (b) by computing $\mathbf{E}[Y]$ without using the marginal PDF of Y.

Solution. (a) The point is randomly chosen from the half circle. We take this to mean that it is uniformly distributed there. Thus the joint PDF of X, Y must be $f_{X,Y}(x,y) = 2/\pi R^2$.

(b) To find the marginal PDF of Y, we integrate the joint PDF over the range of X:

$$f_Y(y) = \int_{-A}^{A} \frac{2}{\pi R^2}\, dx = \begin{cases} \frac{4A}{\pi R^2} & \text{if } 0 \leq y \leq R, \\ 0 & \text{otherwise.} \end{cases}$$

where $A = \sqrt{R^2 - y^2}$. We have

$$\mathbf{E}[Y] = \frac{4}{\pi R^2} \int_0^R y\sqrt{R^2 - y^2}\, dy = \frac{4R}{3\pi},$$

where the integration is performed using the substitution $z = R^2 - y^2$.

(c) There is no need to find the marginal PDF f_Y in order to find the expectation. Let D denote the semicircle. We have

$$\mathbf{E}[Y] = \int_D y f_{X,Y}(x,y)\, dxdy = \int_0^{\pi} \int_0^R \frac{2}{\pi R^2} r \sin\theta\, r dr d\theta = \frac{4R}{3\pi}.$$

Problem 23. Let X and Y be independent continuous random variables with PDFs f_X and f_Y, respectively.

(a) Show that $f_{X+Y\,|\,X}(z\,|\,x) = f_Y(z-x)$.

(b) Assume now that X and Y are independent exponentially distributed random variables with mean equal to 1. Find the conditional PDF of X, given that $X+Y=z$.

Solution. (a) We have $\mathbf{P}(X+Y\leq z\,|\,X=x) = \mathbf{P}(x+Y\leq z) = \mathbf{P}(Y\leq z-x)$. By differentiating both sides with respect to z, the result follows.

(b) Let $Z = X+Y$. We have, for $0\leq x\leq z$,

$$f_{X\,|\,Z}(x,z) = \frac{f_{Z\,|\,X}(z\,|\,x)f_X(x)}{f_Z(z)} = \frac{f_Y(z-x)f_X(x)}{f_Z(z)} = \frac{e^{-(z-x)}e^{-x}}{f_Z(z)} = \frac{e^{-z}}{f_Z(z)}.$$

Since this is the same for all x, it follows that the conditional distribution of X is uniform on the interval $[0,z]$, with PDF $f_{X\,|\,Z}(x\,|\,z) = 1/z$.

Problem 24. * Let X, Y, and Z be three random variables with joint PDF $f_{X,Y,Z}(x,y,z)$. Show the multiplication rule:

$$f_{X,Y,Z}(x,y,z) = f_{X|Y,Z}(x\,|\,y,z)f_{Y|Z}(y\,|\,z)f_Z(z).$$

Solution. We have using the definition of conditional density

$$f_{X|Y,Z}(x\,|\,y,z) = \frac{f_{X,Y,Z}(x,y,z)}{f_{Y,Z}(y,z)},$$

and

$$f_{Y,Z}(y,z) = f_{Y|Z}(y\,|\,z)f_Z(z).$$

Combining these two relations, we obtain the multiplication rule.

Problem 25. * Consider two continuous random variables with joint PDF $f_{X,Y}$. Show that the two different expressions

$$\int xf_X(x)\,dx, \qquad \text{and} \qquad \int\int xf_{X,Y}(x,y)\,dx\,dy$$

for $\mathbf{E}[X]$ are equal.

Solution. We have

$$\begin{aligned}\int\int xf_{X,Y}(x,y)\,dx\,dy &= \int\int xf_{X,Y}(x,y)\,dy\,dx\\ &= \int x\int f_{X,Y}(x,y)\,dy\,dx\\ &= \int xf_X(x)\,dx,\end{aligned}$$

where the last step makes use of the formula $f_X(x) = \int f_{X,Y}(x,y)\,dy$.

Problem 26. * Estimating an expected value by simulation. Let $f_X(x)$ be a PDF such that for some nonnegative scalars a, b, and M we have $f_X(x) = 0$ for all

$x \notin [a,b]$ and $xf_X(x) \leq M$ for all x. Let Y_i, $i = 1,\ldots,N$, be independent random variables with values generated as follows: a point (v,w) is chosen at random within the rectangle whose corners are $(a,0)$, $(b,0)$, (a,M), and (b,M), and if $w \leq vf_X(v)$, the value of Y_i is set to 1, and otherwise it is set to 0. Consider the random variable

$$Z = \frac{Y_1 + \cdots + Y_N}{N}.$$

Show that

$$\mathbf{E}[Z] = \frac{\mathbf{E}[X]}{M(b-a)}$$

and

$$\mathrm{var}(Z) \leq \frac{1}{4N}.$$

In particular, we have $\mathrm{var}(Z) \to 0$ as $N \to \infty$.

Solution. We have

$$\mathbf{P}(Y_i = 1) = \mathbf{P}\bigl(w \leq vf_X(v)\bigr) = \frac{\int_a^b xf_X(x)dx}{M(b-a)} = \frac{\mathbf{E}[X]}{M(b-a)}.$$

According to the example in Section 2.7, the random variable Z has mean $\mathbf{P}(Y_i = 1)$ and variance $\mathbf{P}(Y_i = 1)\bigl(1 - \mathbf{P}(Y_i = 1)\bigr)/N$. Since $0 \leq (1-2p)^2 = 1 - 4p(1-p)$, we have $p(1-p) \leq 1/4$ for any p, so it follows that $\mathrm{var}(Z) \leq 1/(4N)$.

Problem 27. * Consider the following variant of the Buffon needle problem (Example 3.15). A needle of length l is dropped on a plane surface that is partitioned in rectangles with sides whose lengths are a and b. Suppose that the needle's length l satisfies $l < a$ and $l < b$. What is the expected number of recatangle sides crossed by the needle? What is the probability that the needle will cross at least one side of some rectangle.

Solution. Let A be the event that the needle will cross a horizontal line, and let B be the probability that it will cross a vertical line. From the analysis of Example 3.15, we have that

$$\mathbf{P}(A) = \frac{2l}{\pi a}, \qquad \mathbf{P}(B) = \frac{2l}{\pi b}.$$

Since at most one horizontal (or vertical) line can be crossed, the expected number of horizontal lines crossed is $\mathbf{P}(A)$ [or $\mathbf{P}(B)$, respectively]. Thus the expected number of crossed lines is

$$\mathbf{P}(A) + \mathbf{P}(B) = \frac{2l}{\pi a} + \frac{2l}{\pi b} = \frac{2l(a+b)}{\pi ab}.$$

The probability that at least one line will be crossed is

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

Let X (or Y) be the distance from the needle's center to the nearest horizontal (or vertical) line. Let Θ be the angle formed by the needle's axis and the horizontal lines as in Example 3.15. We have

$$\mathbf{P}(A \cap B) = \mathbf{P}\left(X \leq \frac{l\sin\Theta}{2}, Y \leq \frac{l\cos\Theta}{2}\right),$$

and the triple (X, Y, Θ) is uniformly distributed over the set $[0, a/2] \times [0, b/2] \times [0, \pi/2]$. Hence, within in this set, we have

$$f_{X,Y,\Theta}(x, y, \theta) = \frac{8}{\pi ab}.$$

The probability $\mathbf{P}(A \cap B)$ is

$$\begin{aligned}
\mathbf{P}\big(X \le (l/2)\sin\Theta,\, Y \le (l/2)\cos\Theta\big) &= \iint\limits_{\substack{x \le (l/2)\sin\theta \\ y \le (l/2)\cos\theta}} f_{X,Y,\Theta}(x, y, \theta)\, dx\, dy\, d\theta \\
&= \frac{8}{\pi ab} \int_0^{\pi/2} \int_0^{(l/2)\cos\theta} \int_0^{(l/2)\sin\theta} dx\, dy\, d\theta \\
&= \frac{2l^2}{\pi ab} \int_0^{\pi/2} \cos\theta\, \sin\theta\, d\theta \\
&= \frac{l^2}{\pi ab}.
\end{aligned}$$

Thus we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) = \frac{2l}{\pi a} + \frac{2l}{\pi b} - \frac{l^2}{\pi ab} = \frac{l}{\pi ab}\big(2(a + b) - l\big).$$

Problem 28. * Consider two continuous random variables with joint PDF $f_{X,Y}$. Suppose that for any subsets A and B of the real line, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. Show that the random variables X and Y are independent.

Solution. For any two real numbers x and y, we use the independence of the events $\{X \le x\}$ and $\{Y \le y\}$ to obtain,

$$F_{X,Y}(x, y) = \mathbf{P}(X \le x,\, Y \le y) = \mathbf{P}(X \le x)\mathbf{P}(Y \le y) = F_X(x)F_Y(y).$$

We take derivatives of both sides, to obtain

$$f_{X,Y}(x, y) = \frac{\partial F_{X,Y}}{\partial x \partial y} = \frac{\partial F_X}{\partial x}(x)\frac{\partial F_Y}{\partial y}(y) = f_X(x)f_Y(y),$$

which establishes that the random variables X and Y are independent.

Problem 29. * **Random sums of random variables.** You visit a random number N of stores and in the ith store, you spend a random amount of money X_i. What is the mean and variance of the total amount of money

$$T = X_1 + X_2 + \cdots + X_N$$

that you spend? We assume that N is a discrete random variable with a given PMF and that the X_i are random variables with the same mean $\mathbf{E}[X]$ and variance $\mathrm{var}(X)$. Furthermore, we assume that N and all the X_i are independent.

Solution. Let A_i be the event that you visit i stores, i.e.,

$$A_i = \{N = i\}.$$

Then, we have for all i,

$$\mathbf{E}[T \mid A_i] = i\mathbf{E}[X],$$

since conditional on A_i, you will visit exactly i stores, and you will spend an expected amount of money $\mathbf{E}[X]$ in each.

We now apply the total expectation theorem. We have

$$\begin{aligned}\mathbf{E}[T] &= \sum_{i=1}^{\infty} \mathbf{P}(A_i)\mathbf{E}[T \mid A_i] \\ &= \sum_{i=1}^{\infty} \mathbf{P}(N = i) i\mathbf{E}[X] \\ &= \mathbf{E}[X] \sum_{i=1}^{\infty} i\mathbf{P}(N = i) \\ &= \mathbf{E}[X] \cdot \mathbf{E}[N].\end{aligned}$$

Similarly, using also the independence of the X_i, which implies that $\mathbf{E}[X_iX_j] = \big(\mathbf{E}[X]\big)^2$ if $i \neq j$, the second moment of T is calculated as

$$\begin{aligned}\mathbf{E}[T^2] &= \sum_{i=1}^{\infty} \mathbf{P}(A_i)\mathbf{E}[T^2 \mid A_i] \\ &= \sum_{i=1}^{\infty} \mathbf{P}(A_i)\mathbf{E}\big[(X_1 + \cdots + X_N)^2 \mid A_i\big] \\ &= \sum_{i=1}^{\infty} \mathbf{P}(N = i)\Big(i\mathbf{E}[X^2] + i(i-1)\big(\mathbf{E}[X]\big)^2\Big) \\ &= \mathbf{E}[X^2] \sum_{i=1}^{\infty} i\mathbf{P}(N = i) + \big(\mathbf{E}[X]\big)^2 \sum_{i=1}^{\infty} i(i-1)\mathbf{P}(N = i) \\ &= \mathbf{E}[X^2]\mathbf{E}[N] + \big(\mathbf{E}[X]\big)^2\big(\mathbf{E}[N^2] - \mathbf{E}[N]\big) \\ &= \mathrm{var}(X)\mathbf{E}[N] + \big(\mathbf{E}[X]\big)^2\mathbf{E}[N^2].\end{aligned}$$

The variance is then obtained by

$$\begin{aligned}\mathrm{var}(T) &= \mathbf{E}[T^2] - \big(\mathbf{E}[T]\big)^2 \\ &= \mathrm{var}(X)\mathbf{E}[N] + \big(\mathbf{E}[X]\big)^2\mathbf{E}[N^2] - \big(\mathbf{E}[X]\big)^2\big(\mathbf{E}[N]\big)^2 \\ &= \mathrm{var}(X)\mathbf{E}[N] + \big(\mathbf{E}[X]\big)^2\Big(\mathbf{E}[N^2] - \big(\mathbf{E}[N]\big)^2\Big),\end{aligned}$$

so finally

$$\mathrm{var}(T) = \mathrm{var}(X)\mathbf{E}[N] + \big(\mathbf{E}[X]\big)^2\mathrm{var}(N).$$

Note: The formulas for $\mathbf{E}[T]$ and $\text{var}(T)$ will also be obtained with alternative methods in Chapter 4.

SECTION 3.6. Derived Distributions

Problem 30. The metro train arrives at the station near your home every quarter hour starting at 6:00 AM. You walk into your station every morning between 7:10 and 7:30 AM, with the time in this interval being a random variable with given PDF (cf. Example 3.12). Let X be the difference in minutes between 7:10 and the time of your arrival. Calculate the CDF of Y in terms of the CDF of X and differentiate to obtain the PDF of Y.

Solution. We have

$$F_Y(y) = \begin{cases} 0 & \text{if } y \le 0, \\ \mathbf{P}(5 - y \le X \le 5) + \mathbf{P}(20 - y \le X \le 20) & \text{if } 0 \le y \le 5, \\ \mathbf{P}(20 - y \le X \le 20) & \text{if } 5 < y \le 15, \\ 1 & \text{if } y > 15. \end{cases}$$

Using the CDF of X, we have

$$\mathbf{P}(5 - y \le X \le 5) = F_X(5) - F_X(5 - y),$$

$$\mathbf{P}(20 - y \le X \le 20) = F_X(20) - F_X(20 - y).$$

Thus, we have

$$F_Y(y) = \begin{cases} 0 & \text{if } y \le 0, \\ F_X(5) - F_X(5 - y) + F_X(20) - F_X(20 - y) & \text{if } 0 \le y \le 5, \\ F_X(20) - F_X(20 - y) & \text{if } 5 < y \le 15, \\ 1 & \text{if } y > 15. \end{cases}$$

Differentiating, we obtain

$$f_Y(y) = \begin{cases} f_X(5 - y) + f_X(20 - y) & \text{if } 0 \le y \le 5, \\ f_X(20 - y) & \text{if } 5 < y \le 15, \\ 0 & \text{otherwise,} \end{cases}$$

consistently with the result of Example 3.12.

Problem 31. Let X be a uniformly distributed random variable between 0 and 1. Show that the PDF of $Y = \cos \pi X$ is

$$f_Y(y) = \frac{1}{\pi\sqrt{1 - y^2}}, \qquad \text{for } -1 < y < 1.$$

Solution. We have

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}\left(\frac{1}{\pi}\cos^{-1} y \le X \le 1\right) = 1 - \frac{1}{\pi}\cos^{-1} y$$

and therefore by differentiation

$$f_Y(y) = \frac{1}{\pi\sqrt{1-y^2}}, \qquad \text{for } -1 < y < 1.$$

Problem 32. If the random variable Θ is uniformly distributed between $-\pi$ and π, find the density function for random variable $X = \cos\Theta$.

Solution. We first find the CDF, and then take the derivative to find the PDF.

$$\mathbf{P}(X \le x) = \mathbf{P}(\cos\Theta \le x) = \frac{1}{\pi}(\pi - \cos^{-1} x)$$

and therefore

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-x^2}}.$$

Problem 33. If X is normally distributed, with mean μ and variance σ^2 find the PDF of the random variable $Y = e^X$.

Solution. We first find the CDF, and then take the derivative to find the PDF. We have

$$\mathbf{P}(Y \le y) = \mathbf{P}(e^X \le y) = \begin{cases} \mathbf{P}(X \le \ln y) & \text{if } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$f_Y(y) = \begin{cases} \frac{d}{dx} F_X(\ln y) & \text{if } y > 0, \\ 0 & \text{otherwise,} \end{cases} = \begin{cases} \frac{1}{y} f_X(\ln y) & \text{if } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 34. Suppose X is uniformly distributed between 0 and 1. Find the PDF of $Y = e^X$.

Solution. We have

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}(e^X \le y) = \mathbf{P}(X \le \ln y) = \ln y$$

and thus

$$f_Y(y) = \frac{1}{y}, \qquad \text{for } 0 < y \le e.$$

Problem 35. If X is a random variable that is uniformly distributed between -1 and 1, find the PDF of $Y = \sqrt{|X|}$ and the PDF of $Y = -\ln|X|$.

Solution. Let $Y = \sqrt{|X|}$. We have

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}(\sqrt{|X|} \le y) = \mathbf{P}(-y^2 \le X \le y^2) = y^2$$

and therefore by differentiation

$$f_Y(y) = 2y, \qquad \text{for } 0 \le y \le 1.$$

Let $Y = -\ln|X|$. We have

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(\ln|X| \geq -y) = \mathbf{P}(X \geq e^{-y}) + \mathbf{P}(X \leq -e^{-y}) = 1 - e^{-y},$$

and therefore by differentiation

$$f_Y(y) = e^{-y}, \qquad \text{for } 0 \leq y < \infty.$$

Problem 36. Suppose that X is a standard normal random variable. Find the PDF of $Y = |X|^{\frac{1}{3}}$ and $Y = |X|^{\frac{1}{4}}$.

Solution. Let $Y = |X|^{\frac{1}{3}}$. We have

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(|X|^{\frac{1}{3}} \leq y) = \mathbf{P}(-y^3 \leq X \leq y^3) = F_X(y^3) - F_X(-y^3)$$

and therefore

$$f_Y(y) = 3y^2 f_X(y^3) + 3y^2 f_X(-y^3) = 6y^2 f_X(y^3), \qquad \text{for } y > 0.$$

Let $Y = |X|^{\frac{1}{4}}$. We have

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(|X|^{\frac{1}{4}} \leq y) = \mathbf{P}(-y^4 \leq X \leq y^4) = F_X(y^4) - F_X(-y^4)$$

and therefore

$$f_Y(y) = 4y^3 f_X(y^4) + 4y^3 f_X(-y^4) = 8y^3 f_X(y^4), \qquad \text{for } y > 0.$$

Problem 37. Let X and Y be independent random variables, uniformly distributed on the interval $[0, 1]$. Find the PDF of $X - Y$.

Solution. We have $P(|X - Y| \leq a) = 1 - (1 - a)^2$. (To see this, draw the event of interest as a subset of the unit square and calculate its area.) Taking derivatives, the desired PDF is $f_Z(z) = 2(1 - z)$, for $0 \leq z \leq 1$, and zero otherwise.

Problem 38. Let X and Y be the cartesian coordinates of a random point on the triangle with vertices at $(0, 1)$, $(0, -1)$, and $(1, 0)$. Find the CDF and the PDF of $Z = |X - Y|$.

Solution. To find the CDF, we integrate the joint density of X, and Y over the region where $|X - Y| \leq z$ for a given z. In the case where $z \leq 0$ or $z \geq 1$, the CDF is 0 and 1, respectively. In the case where $0 < z < 1$, we have

$$\begin{aligned} F_Z(z) &= \mathbf{P}(|X - Y| \leq z \,|\, X \geq Y)\mathbf{P}(X \geq Y) + \mathbf{P}(Y - X \leq z \,|\, X < Y)\mathbf{P}(X < Y) \\ &= \left(\frac{z}{2} + \frac{z^2}{4}\right) + \left(\frac{1}{4} - \frac{(1 - z)^2}{4}\right) \\ &= z. \end{aligned}$$

We have

$$F_Z(z) = \begin{cases} 0 & \text{if } z \leq 0, \\ z & \text{if } 0 < z < 1, \\ 1 & \text{if } z \geq 1. \end{cases}$$

By taking the derivative with respect to Z, we obtain

$$f_Z(z) = \begin{cases} 1 & \text{if } 0 \le z \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 39. Consider the Romeo and Juliet Example 3.29, but assume that the times X and Y are independent and exponentially distributed with different parameters λ_X and λ_Y. Find the PDF of $Z = X - Y$.

Solution. We will first calculate the CDF $F_Z(z)$ by considering separately the cases $z \ge 0$ and $z < 0$. For $z \ge 0$, we have (see the left side of Fig. 3.26)

$$\begin{aligned} F_Z(z) &= \mathbf{P}(X - Y \le z) \\ &= 1 - \mathbf{P}(X - Y > z) \\ &= 1 - \int_0^\infty \left(\int_{z+y}^\infty f_{X,Y}(x,y)\,dx \right) dy \\ &= 1 - \int_0^\infty \lambda_Y e^{-\lambda_Y y} \left(\int_{z+y}^\infty \lambda_X e^{-\lambda_X x}\,dx \right) dy \\ &= 1 - \int_0^\infty \lambda_Y e^{-\lambda_Y y} e^{-\lambda_X (z+y)}\,dy \\ &= 1 - e^{-\lambda_X z} \int_0^\infty \lambda_Y e^{-(\lambda_X + \lambda_Y) y}\,dy \\ &= 1 - \frac{\lambda_Y}{\lambda_X + \lambda_Y} e^{-\lambda_X z}. \end{aligned}$$

For the case $z < 0$, we have using the preceding calculation

$$F_Z(z) = 1 - F_Z(-z) = 1 - \left(1 - \frac{\lambda_X}{\lambda_X + \lambda_Y} e^{-\lambda_Y(-z)} \right) = \frac{\lambda_X}{\lambda_X + \lambda_Y} e^{\lambda_Y z}.$$

Combining the two cases $z \ge 0$ and $z < 0$, we obtain

$$F_Z(z) = \begin{cases} 1 - \frac{\lambda_Y}{\lambda_X + \lambda_Y} e^{-\lambda_X z} & \text{if } z \ge 0, \\ \frac{\lambda_X}{\lambda_X + \lambda_Y} e^{\lambda_Y z} & \text{if } z < 0, \end{cases}$$

The PDF of Z is obtained by differentiating its CDF. We have

$$f_Z(z) = \begin{cases} \frac{\lambda_X \lambda_Y}{\lambda_X + \lambda_Y} e^{-\lambda_X z} & \text{if } z \ge 0, \\ \frac{\lambda_X \lambda_Y}{\lambda_X + \lambda_Y} e^{\lambda_Y z} & \text{if } z < 0. \end{cases}$$

Problem 40. * Two points are chosen randomly and independently from the interval $[0, 1]$. Show that the expected value of the distance from each point to the nearest endpoint of the interval is $1/3$, and that the expected value of the distance between the two points is also $1/3$.

Solution. Let X and Y be the two points, and let $Z = \max\{X, Y\}$. For any $t \in [0, 1]$, we have

$$\mathbf{P}(Z \leq t) = \mathbf{P}(X \leq t)\mathbf{P}(Y \leq t) = t^2,$$

and by differentiating, the corresponding PDF is

$$f_Z(z) = \begin{cases} 0 & \text{if } z \leq 0, \\ 2z & \text{if } 0 \leq z \leq 1, \\ 0 & \text{if } z \geq 1. \end{cases}$$

Thus, we have

$$\mathbf{E}[Z] = \int_{-\infty}^{\infty} z f_Z(z) dz = \int_0^1 2z^2 dz = \frac{2}{3}.$$

The distance of the largest of the two points to the nearest endpoint is $1 - Z$, and its expected value is $1 - \mathbf{E}[Z] = 1/3$. A symmetric argument shows that the distance of the smallest of the two points to the nearest endpoint is also $1/3$. Since the expected distances to the nearest endpoints are $1/3$, the expected distance between the two points must also be $1/3$.

Problem 41. * Competing exponentials. Two runners run a long distance course in times X and Y, which are independent and exponentially distributed with parameters λ_X and λ_Y, respectively. The winner's time is

$$Z = \min\{X, Y\}.$$

Show that the CDF of Z is exponential with parameter $\lambda_X + \lambda_Y$.

Solution. For all $z \geq 0$, we have, using the independence of X and Y, and the form of the exponential CDF,

$$\begin{aligned} F_Z(z) &= \mathbf{P}\big(\min\{X, Y\} \leq Z\big) \\ &= 1 - \mathbf{P}\big(\min\{X, Y\} > Z\big) \\ &= 1 - \mathbf{P}(X > Z,\, Y > Z) \\ &= 1 - \mathbf{P}(X > Z)\mathbf{P}(Y > Z) \\ &= 1 - e^{-\lambda_X z} e^{-\lambda_Y z} \\ &= 1 - e^{-(\lambda_X + \lambda_Y)z}. \end{aligned}$$

This is recognized as the exponential CDF with parameter $\lambda_X + \lambda_Y$. Thus, the minimum of two exponentials with parameters λ_X and λ_Y is an exponential with parameter $\lambda_X + \lambda_Y$.

Problem 42. * Cauchy random variable.

(a) Let X be a uniformly distributed random variable between 0 and 1. Show that the PDF of $Y = \tan \pi X$ is

$$f_Y(y) = \frac{1}{\pi} \frac{1}{1 + y^2}, \qquad -\infty < y < \infty.$$

(Y is called a *Cauchy random variable*.)

(b) Let Y be a Cauchy random variable. Find the PDF of the random variable

$$X = \tan^{-1} Y.$$

Solution. (a) For $y < 0$, we have

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}\left(\frac{1}{2} \leq X \leq \frac{1}{\pi}\tan^{-1} y + 1\right) = \frac{1}{\pi}\tan^{-1} y + \frac{1}{2}.$$

Also for $y > 0$, we have

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}\left(0 \leq X \leq \frac{1}{\pi}\tan^{-1} y\right) + \mathbf{P}\left(\frac{1}{2} \leq X \leq 1\right) = \frac{1}{\pi}\tan^{-1} y + \frac{1}{2}.$$

Therefore, by differentiation,

$$f_Y(y) = \frac{1}{\pi}\frac{1}{1+y^2}, \qquad \text{for } -\infty < y < \infty.$$

(b) We first compute the CDF and then differentiate to obtain the PDF of X. We have

$$\mathbf{P}(X \leq x) = \mathbf{P}(\tan^{-1} Y \leq x) = \mathbf{P}(Y \leq \tan x) = \frac{1}{\pi}\int_{-\infty}^{\tan x} \frac{1}{1+y^2}\,dy = \frac{1}{\pi}\cdot\left(x + \frac{\pi}{2}\right).$$

Taking the derivative, we find that X is uniformly distributed on an interval of length π, and by the definition of $\tan^{-1}$ the interval must be $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

4

Further Topics on Random Variables and Expectations

Contents

In this chapter, we develop a number of more advanced topics. We introduce methods that are useful in:

(a) dealing with the sum of independent random variables, including the case where the number of random variables is itself random;

(b) addressing problems of estimation or prediction of an unknown random variable on the basis of observed values of other random variables.

With these goals in mind, we introduce a number of tools, including transforms and convolutions, and refine our understanding of the concept of conditional expectation.

4.1 TRANSFORMS

In this section, we introduce the transform associated with a random variable. The transform provides us with an alternative representation of its probability law (PMF or PDF). It is not particularly intuitive, but it is often convenient for certain types of mathematical manipulations.

The **transform** of the distribution of a random variable X (also referred to as the **moment generating function** of X) is a function $M_X(s)$ of a free parameter s, defined by

$$M_X(s) = \mathbf{E}[e^{sX}].$$

The simpler notation $M(s)$ can also be used whenever the underlying random variable X is clear from the context. In more detail, when X is a discrete random variable, the corresponding transform is given by

$$M(s) = \sum_x e^{sx} p_X(x),$$

while in the continuous case, we have†

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx.$$

Example 4.1. Let

$$p_X(x) = \begin{cases} 1/2, & \text{if } x = 2, \\ 1/6, & \text{if } x = 3, \\ 1/3, & \text{if } x = 5. \end{cases}$$

† The reader who is familiar with Laplace transforms may recognize that the transform associated with a continuous random variable is essentially the same as the Laplace transform of its PDF, the only difference being that Laplace transforms usually involve e^{-sx} rather than e^{sx}. For the discrete case, a variable z is sometimes used in place of e^s and the resulting transform $M(z) = \sum_x z^x p_X(x)$ is known as the *z-transform.* However, we will not be using z-transforms in this book.

Then, the corresponding transform is

$$M(s) = \frac{1}{2}e^{2s} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}$$

(see Fig. 4.1).

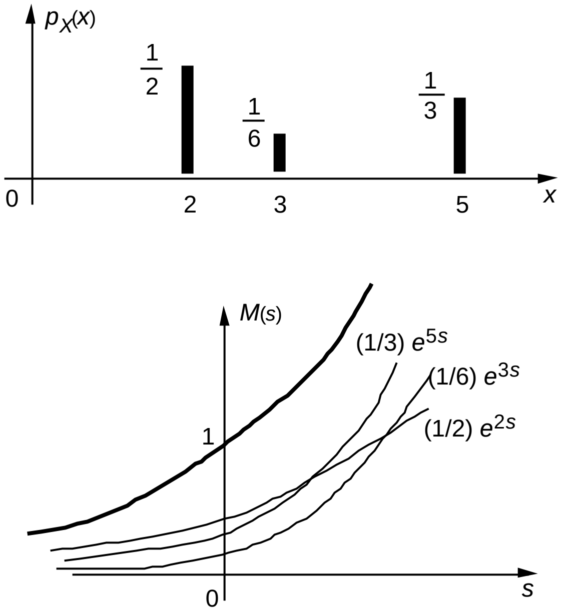


Figure 4.1: The PMF and the corresponding transform for Example 4.1. The transform $M(s)$ consists of the weighted sum of the three exponentials shown. Note that at $s = 0$, the transform takes the value 1. This is generically true since

$$M(0) = \sum_x e^{0\cdot x} p_X(x) = \sum_x p_X(x) = 1.$$

Example 4.2. The Transform of a Poisson Random Variable. Consider a Poisson random variable X with parameter λ:

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots$$

The corresponding transform is given by

$$M(s) = \sum_{x=0}^{\infty} e^{sx} \frac{\lambda^x e^{-\lambda}}{x!}.$$

We let $a = e^s\lambda$ and obtain

$$M(s) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{a^x}{x!} = e^{-\lambda}e^{a} = e^{a-\lambda} = e^{\lambda(e^s-1)}.$$

Example 4.3. The Transform of an Exponential Random Variable. Let X be an exponential random variable with parameter λ:

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

Then,

$$\begin{aligned} M(s) =& \lambda\int_0^{\infty} e^{sx}e^{-\lambda x}\,dx \\ =& \lambda\int_0^{\infty} e^{(s-\lambda)x}\,dx \\ =& \lambda\frac{e^{(s-\lambda)x}}{s-\lambda}\bigg|_0^{\infty} \qquad \text{(if } s<\lambda) \\ =& \frac{\lambda}{\lambda-s}. \end{aligned}$$

The above calculation and the formula for $M(s)$ is correct only if the integrand $e^{(s-\lambda)x}$ decays as x increases, which is the case if and only if $s < \lambda$; otherwise, the integral is infinite.

It is important to realize that the transform is not a number but rather a *function* of a free variable or parameter s. Thus, we are dealing with a transformation that starts with a function, e.g., a PDF $f_X(x)$ (which is a function of a free variable x) and results in a new function, this time of a real parameter s. Strictly speaking, $M(s)$ is only defined for those values of s for which $\mathbf{E}[e^{sX}]$ is finite, as noted in the preceding example.

Example 4.4. The Transform of a Linear Function of a Random Variable. Let $M_X(s)$ be the transform associated with a random variable X. Consider a new random variable $Y = aX + b$. We then have

$$M_Y(s) = \mathbf{E}[e^{s(aX+b)}] = e^{sb}\mathbf{E}[e^{saX}] = e^{sb}M_X(sa).$$

For example, if X is exponential with parameter $\lambda = 1$, so that $M_X(s) = 1/(1-s)$, and if $Y = 2X + 3$, then

$$M_Y(s) = e^{3s}\frac{1}{1-2s}.$$

Example 4.5. The Transform of a Normal Random Variable. Let X be a normal random variable with mean μ and variance σ^2. To calculate the corresponding transform, we first consider the special case of the standard normal random variable Y, where $\mu = 0$ and $\sigma^2 = 1$, and then use the formula of the preceding example. The PDF of the standard normal is

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

and its transform is

$$\begin{aligned} M_Y(s) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, e^{sy}\, dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)+sy} dy \\ &= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)+sy-(s^2/2)} dy \\ &= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y-s)^2/2} dy \\ &= e^{s^2/2}, \end{aligned}$$

where the last equality follows by using the normalization property of a normal PDF with mean s and unit variance.

A general normal random variable with mean μ and variance σ^2 is obtained from the standard normal via the linear transformation

$$X = \sigma Y + \mu.$$

The transform of the standard normal is $M_Y(s) = e^{s^2/2}$, as verified above. By applying the formula of Example 4.4, we obtain

$$M_X(s) = e^{s\mu} M_Y(s\sigma) = e^{\frac{\sigma^2 s^2}{2} + \mu s}.$$

From Transforms to Moments

The reason behind the alternative name "moment generating function" is that the moments of a random variable are easily computed once a formula for the associated transform is available. To see this, let us take the derivative of both sides of the definition

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx,$$

with respect to s. We obtain

$$\begin{aligned}\frac{d}{ds}M(s) &= \frac{d}{ds}\int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx\\ &= \int_{-\infty}^{\infty} \frac{d}{ds} e^{sx} f_X(x)\,dx\\ &= \int_{-\infty}^{\infty} x e^{sx} f_X(x)\,dx.\end{aligned}$$

This equality holds for all values of s. By considering the special case where $s = 0$, we obtain†

$$\frac{d}{ds}M(s)\bigg|_{s=0} = \int_{-\infty}^{\infty} x f_X(x)\,dx = \mathbf{E}[X].$$

More generally, if we differentiate n times the function $M(s)$ with respect to s, a similar calculation yields

$$\frac{d^n}{ds^n}M(s)\bigg|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(x)\,dx = \mathbf{E}[X^n].$$

Example 4.6. We saw earlier (Example 4.1) that the PMF

$$p_X(x) = \begin{cases} 1/2, & \text{if } x = 2,\\ 1/6, & \text{if } x = 3,\\ 1/3, & \text{if } x = 5,\end{cases}$$

has the transform

$$M(s) = \frac{1}{2}e^{2s} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}.$$

Thus,

$$\begin{aligned}\mathbf{E}[X] &= \frac{d}{ds}M(s)\bigg|_{s=0}\\ &= \frac{1}{2}2e^{2s} + \frac{1}{6}3e^{3s} + \frac{1}{3}5e^{5s}\bigg|_{s=0}\\ &= \frac{1}{2}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{3}\cdot 5\\ &= \frac{19}{6}.\end{aligned}$$

† This derivation involves an interchange of differentiation and integration. The interchange turns out to be justified for all of the applications to be considered in this book. Furthermore, the derivation remains valid for general random variables, including discrete ones. In fact, it could be carried out more abstractly, in the form

$$\frac{d}{ds}M(s) = \frac{d}{ds}\mathbf{E}[e^{sX}] = \mathbf{E}\left[\frac{d}{ds}e^{sX}\right] = \mathbf{E}[Xe^{sX}],$$

leading to the same conclusion.

Also,

$$\begin{aligned}\mathbf{E}[X^2] &= \frac{d^2}{ds^2}M(s)\bigg|_{s=0}\\ &= \frac{1}{2}4e^{2s}+\frac{1}{6}9e^{3s}+\frac{1}{3}25e^{5s}\bigg|_{s=0}\\ &= \frac{1}{2}\cdot 4+\frac{1}{6}\cdot 9+\frac{1}{3}\cdot 25\\ &= \frac{71}{6}.\end{aligned}$$

For an exponential random variable with PDF

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geq 0,$$

we found earlier that

$$M(s) = \frac{\lambda}{\lambda - s}.$$

Thus,

$$\frac{d}{ds}M(s) = \frac{\lambda}{(\lambda - s)^2}, \qquad \frac{d^2}{ds^2}M(s) = \frac{2\lambda}{(\lambda - s)^3}.$$

By setting $s = 0$, we obtain

$$\mathbf{E}[X] = \frac{1}{\lambda}, \qquad \mathbf{E}[X^2] = \frac{2}{\lambda^2},$$

which agrees with the formulas derived in Chapter 3.

Inversion of Transforms

A very important property of transforms is the following.

Inversion Property

The transform $M_X(s)$ completely determines the probability law of the random variable X. In particular, if $M_X(s) = M_Y(s)$ for all s, then the random variables X and Y have the same probability law.

This property is a rather deep mathematical fact that we will use frequently.† There exist explicit formulas that allow us to recover the PMF or PDF of a random variable starting from the associated transform, but they are quite difficult to use. In practice, transforms are usually inverted by "pattern matching," based on tables of known distribution-transform pairs. We will see a number of such examples shortly.

† In fact, the probability law of a random variable is completely determined even if we only know the transform $M(s)$ for values of s in some interval of positive length.

Example 4.7. We are told that the transform associated with a random variable X is

$$M(s) = \frac{1}{4}e^{-s} + \frac{1}{2} + \frac{1}{8}e^{4s} + \frac{1}{8}e^{5s}.$$

Since $M(s)$ is a sum of terms of the form e^{sx}, we can compare with the general formula

$$M(s) = \sum_x e^{sx} p_X(x),$$

and infer that X is a discrete random variable. The different values that X can take can be read from the corresponding exponents and are -1, 0, 4, and 5. The probability of each value x is given by the coefficient multiplying the corresponding e^{sx} term. In our case, $\mathbf{P}(X = -1) = 1/4$, $\mathbf{P}(X = 0) = 1/2$, $\mathbf{P}(X = 4) = 1/8$, $\mathbf{P}(X = 5) = 1/8$.

Generalizing from the last example, the distribution of a finite-valued discrete random variable can be always found by inspection of the corresponding transform. The same procedure also works for discrete random variables with an infinite range, as in the example that follows.

Example 4.8. The Transform of a Geometric Random Variable. We are told that the transform associated with random variable X is of the form

$$M(s) = \frac{pe^s}{1-(1-p)e^s},$$

where p is a constant in the range $0 < p < 1$. We wish to find the distribution of X. We recall the formula for the geometric series:

$$\frac{1}{1-\alpha} = 1 + \alpha + \alpha^2 + \cdots,$$

which is valid whenever $|\alpha| < 1$. We use this formula with $\alpha = (1-p)e^s$, and for s sufficiently close to zero so that $(1-p)e^s < 1$. We obtain

$$M(s) = pe^s\Big(1 + (1-p)e^s + (1-p)^2e^{2s} + (1-p)^3e^{3s} + \cdots\Big).$$

As in the previous example, we infer that this is a discrete random variable that takes positive integer values. The probability $\mathbf{P}(X = k)$ is found by reading the coefficient of the term e^{ks}. In particular, $\mathbf{P}(X = 1) = p$, $\mathbf{P}(X = 2) = p(1-p)$, etc., and

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

We recognize this as the geometric distribution with parameter p.

Note that

$$\frac{d}{ds}M(s) = \frac{pe^s}{1-(1-p)e^s} + \frac{(1-p)pe^s}{(1-(1-p)e^s)^2}.$$

If we set $s = 0$, the above expression evaluates to $1/p$, which agrees with the formula for $\mathbf{E}[X]$ derived in Chapter 2.

Example 4.9. The Transform of a Mixture of Two Distributions. The neighborhood bank has three tellers, two of them fast, one slow. The time to assist a customer is exponentially distributed with parameter $\lambda = 6$ at the fast tellers, and $\lambda = 4$ at the slow teller. Jane enters the bank and chooses a teller at random, each one with probability 1/3. Find the PDF of the time it takes to assist Jane and its transform.

We have

$$f_X(x) = \frac{2}{3} \cdot 6e^{-6x} + \frac{1}{3} \cdot 4e^{-4x}, \qquad x \geq 0.$$

Then,

$$\begin{aligned} M(s) &= \int_0^\infty e^{sx} \left(\frac{2}{3} 6e^{-6x} + \frac{1}{3} 4e^{-4x} \right) dx \\ &= \frac{2}{3} \int_0^\infty e^{sx} 6e^{-6x}\, dx + \frac{1}{3} \int_0^\infty e^{sx} 4e^{-4x}\, dx \\ &= \frac{2}{3} \cdot \frac{6}{6-s} + \frac{1}{3} \cdot \frac{4}{4-s} \qquad \text{(for } s < 4\text{).} \end{aligned}$$

More generally, let $X_1, \ldots, X_n$ be continuous random variables with PDFs $f_{X_1}, \ldots f_{X_n}$, and let Y be a random variable, which is equal to X_i with probability p_i. Then,

$$f_Y(y) = p_1 f_{X_1}(y) + \cdots + p_n f_{X_n}(y),$$

and

$$M_Y(s) = p_1 M_{X_1}(s) + \cdots + p_n M_{X_n}(s).$$

The steps in this problem can be reversed. For example, we may be told that the transform associated with a random variable Y is of the form

$$\frac{1}{2} \cdot \frac{1}{2-s} + \frac{3}{4} \cdot \frac{1}{1-s}.$$

We can then rewrite it as

$$\frac{1}{4} \cdot \frac{2}{2-s} + \frac{3}{4} \cdot \frac{1}{1-s},$$

and recognize that Y is the mixture of two exponential random variables with parameters 2 and 1, which are selected with probabilities 1/4 and 3/4, respectively.

Sums of Independent Random Variables

Transform methods are particularly convenient when dealing with a sum of random variables. This is because it turns out that *addition of independent random variables corresponds to multiplication of transforms*, as we now show.

Let X and Y be independent random variables, and let $W = X + Y$. The transform associated with W is, by definition,

$$M_W(s) = \mathbf{E}[e^{sW}] = \mathbf{E}[e^{s(X+Y)}] = \mathbf{E}[e^{sX}e^{sY}].$$

Consider a fixed value of the parameter s. Since X and Y are independent, e^{sX} and e^{sY} are independent random variables. Hence, the expectation of their product is the product of the expectations, and

$$M_W(s) = \mathbf{E}[e^{sX}]\mathbf{E}[e^{sY}] = M_X(s)M_Y(s).$$

By the same argument, if $X_1, \ldots, X_n$ is a collection of independent random variables, and

$$W = X_1 + \cdots + X_n,$$

then

$$M_W(s) = M_{X_1}(s) \cdots M_{X_n}(s).$$

Example 4.10. The Transform of the Binomial. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with a common parameter p. Then,

$$M_{X_i}(s) = (1-p)e^{0s} + pe^{1s} = 1 - p + pe^s, \qquad \text{for all } i.$$

The random variable $Y = X_1 + \cdots + X_n$ is binomial with parameters n and p. Its transform is given by

$$M_Y(s) = \big(1 - p + pe^s\big)^n.$$

Example 4.11. The Sum of Independent Poisson Random Variables is Poisson. Let X and Y be independent Poisson random variables with means λ and μ, respectively, and let $W = X + Y$. Then,

$$M_X(s) = e^{\lambda(e^s-1)}, \qquad M_Y(s) = e^{\mu(e^s-1)},$$

and

$$M_W(s) = M_X(s)M_Y(s) = e^{\lambda(e^s-1)}e^{\mu(e^s-1)} = e^{(\lambda+\mu)(e^s-1)}.$$

Thus, W has the same transform as a Poisson random variable with mean $\lambda + \mu$. By the uniqueness property of transforms, W is Poisson with mean $\lambda + \mu$.

Example 4.12. The Sum of Independent Normal Random Variables is Normal. Let X and Y be independent normal random variables with means μ_x, μ_y, and variances σ_x^2, σ_y^2, respectively. Let $W = X + Y$. Then,

$$M_X(s) = e^{\frac{\sigma_x^2 s^2}{2} + \mu_x s}, \qquad M_Y(s) = e^{\frac{\sigma_y^2 s^2}{2} + \mu_y s},$$

and

$$M_W(s) = e^{\frac{(\sigma_x^2 + \sigma_y^2) s^2}{2} + (\mu_x + \mu_y) s}.$$

Thus, W has the same transform as a normal random variable with mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$. By the uniqueness property of transforms, W is normal with these parameters.

Summary of Transforms and their Properties

- The transform associated with the distribution of a random variable X is given by

$$M_X(s) = \mathbf{E}[e^{sX}] = \begin{cases} \displaystyle\sum_x e^{sx} p_X(x), & x \text{ discrete,} \\ \displaystyle\int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx, & x \text{ continuous.} \end{cases}$$

- The distribution of a random variable is completely determined by the corresponding transform.
- Moment generating properties:

$$M_X(0) = 1, \qquad \frac{d}{ds} M_X(s)\bigg|_{s=0} = \mathbf{E}[X], \qquad \frac{d^n}{ds^n} M_X(s)\bigg|_{s=0} = \mathbf{E}[X^n].$$

- If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(as)$.
- If X and Y are independent, then $M_{X+Y}(s) = M_X(s) M_Y(s)$.

We have derived formulas for the transforms of a few common random variables. Such formulas can be derived with a moderate amount of algebra for many other distributions. Some of the most useful ones are summarized in the tables that follow.

Transforms of Joint Distributions

If two random variables X and Y are described by some joint distribution (e.g., a joint PDF), then each one is associated with a transform $M_X(s)$ or $M_Y(s)$. These

Transforms for Common Discrete Random Variables

Bernoulli(p)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1-p, & \text{if } k = 0. \end{cases} \qquad M_X(s) = 1 - p + pe^s.$$

Binomial(n, p)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

$$M_X(s) = (1 - p + pe^s)^n.$$

Geometric(p)

$$p_X(k) = p(1-p)^{k-1}, \quad k = 1, 2, \ldots \qquad M_X(s) = \frac{pe^s}{1-(1-p)e^s}.$$

Poisson(λ)

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, \ldots \qquad M_X(s) = e^{\lambda(e^s - 1)}.$$

Uniform(a, b)

$$p_X(k) = \frac{1}{b-a+1}, \quad k = a, a+1, \ldots, b.$$

$$M_X(s) = \frac{e^{as}}{b-a+1} \frac{e^{(b-a+1)s} - 1}{e^s - 1}.$$

are the transforms of the marginal distributions and do not convey information on the dependence between the two random variables. Such information is contained in a multivariate transform, which we now define.

Consider n random variables $X_1, \ldots, X_n$ related to the same experiment. Let $s_1, \ldots, s_n$ be scalar free parameters. The associated multivariate transform is a function of these n parameters and is defined by

$$M_{X_1,\ldots,X_n}(s_1, \ldots, s_n) = \mathbf{E}\big[e^{sX_1 + \cdots + s_nX_n}\big].$$

The inversion property of transforms discussed earlier extends to the multivariate case. That is, if $Y_1, \ldots, Y_n$ is another set of random variables and $M_{X_1,\ldots,X_n}(s_1, \ldots, s_n)$, $M_{Y_1,\ldots,Y_n}(s_1, \ldots, s_n)$ are the same functions of $s_1, \ldots, s_n$,

Transforms for Common Continuous Random Variables

Uniform(a, b)

$$f_X(x) = \frac{1}{b-a}, \quad a \le x \le b. \qquad M_X(s) = \frac{1}{b-a}\frac{e^{sb}-e^{sa}}{s}.$$

Exponential(λ)

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \ge 0. \qquad M_X(s) = \frac{\lambda}{\lambda - s}, \qquad (s > \lambda).$$

Normal(μ, σ^2)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty. \qquad M_X(s) = e^{\frac{\sigma^2 s^2}{2}+\mu s}.$$

then the joint distribution of $X_1, \ldots, X_n$ is the same as the joint distribution of $Y_1, \ldots, Y_n$.

4.2 SUMS OF INDEPENDENT RANDOM VARIABLES — CONVOLUTIONS

If X and Y are independent random variables, the distribution of their sum $W = X + Y$ can be obtained by computing and then inverting the transform $M_W(s) = M_X(s)M_Y(s)$. But it can also be obtained directly, using the method developed in this section.

The Discrete Case

Let $W = X+Y$, where X and Y are independent integer-valued random variables with PMFs $p_X(x)$ and $p_Y(y)$. Then, for any integer w,

$$\begin{aligned} p_W(w) =& \mathbf{P}(X+Y=w) \\ =& \sum_{(x,y):\ x+y=w} \mathbf{P}(X=x \text{ and } Y=y) \\ =& \sum_x \mathbf{P}(X=x \text{ and } Y=w-x) \\ =& \sum_x p_X(x)p_Y(w-x). \end{aligned}$$

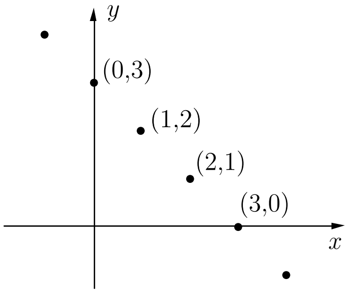


Figure 4.2: The probability $p_W(3)$ that $X+Y=3$ is the sum of the probabilities of all pairs (x, y) such that $x+y=3$, which are the points indicated in the figure. The probability of a generic such point is of the form $p_{X,Y}(x, 3-x) = p_X(x)p_Y(3-x)$.

The resulting PMF $p_W(w)$ is called the **convolution** of the PMFs of X and Y. See Fig. 4.2 for an illustration.

Example 4.13. Let X and Y be independent and have PMFs given by

$$p_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 1, 2, 3, \\ 0 & \text{otherwise,} \end{cases} \qquad p_Y(y) = \begin{cases} \frac{1}{2} & \text{if } x = 0, \\ \frac{1}{3} & \text{if } x = 1, \\ \frac{1}{6} & \text{if } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

To calculate the PMF of $W = X + Y$ by convolution, we first note that the range of possible values of w are the integers from the range $[1, 5]$. Thus we have

$$p_W(w) = 0 \quad \text{if } w \neq 1, 2, 3, 4, 5.$$

We calculate $p_W(w)$ for each of the values $w = 1, 2, 3, 4, 5$ using the convolution formula. We have

$$p_W(1) = \sum_x p_X(x)p_Y(1-x) = p_X(1) \cdot p_Y(0) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6},$$

where the second equality above is based on the fact that for $x \neq 1$ either $p_X(x)$ or $p_Y(1-x)$ (or both) is zero. Similarly, we obtain

$$p_W(2) = p_X(1) \cdot p_Y(1) + p_X(2) \cdot p_Y(0) = \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{18},$$

$$p_W(3) = p_X(1) \cdot p_Y(2) + p_X(2) \cdot p_Y(1) + p_X(3) \cdot p_Y(0) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{3},$$

$$p_W(4) = p_X(2) \cdot p_Y(2) + p_X(3) \cdot p_Y(1) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{6},$$

$$p_W(5) = p_X(3) \cdot p_Y(2) = \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}.$$

The Continuous Case

Let X and Y be independent continuous random variables with PDFs $f_X(x)$ and $f_Y(y)$. We wish to find the PDF of $W = X + Y$. Since W is a function of two random variables X and Y, we can follow the method of Chapter 3, and start by deriving the CDF $F_W(w)$ of W. We have

$$\begin{aligned} F_W(w) &= \mathbf{P}(W \le w) \\ &= \mathbf{P}(X + Y \le w) \\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{w-x} f_X(x) f_Y(y)\, dy\, dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) \left[\int_{y=-\infty}^{w-x} f_Y(y)\, dy \right] dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) F_Y(w-x)\, dx. \end{aligned}$$

The PDF of W is then obtained by differentiating the CDF:

$$\begin{aligned} f_W(w) &= \frac{dF_W}{dw}(w) \\ &= \frac{d}{dw} \int_{x=-\infty}^{\infty} f_X(x) F_Y(w-x)\, dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) \frac{dF_Y}{dw}(w-x)\, dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx. \end{aligned}$$

This formula is entirely analogous to the formula for the discrete case, except that the summation is replaced by an integral and the PMFs are replaced by PDFs. For an intuitive understanding of this formula, see Fig. 4.3.

Example 4.14. The random variables X and Y are independent and uniformly distributed in the interval $[0, 1]$. The PDF of $W = X + Y$ is

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx.$$

The integrand $f_X(x) f_Y(w-x)$ is nonzero (and equal to 1) for $0 \le x \le 1$ and $0 \le w - x \le 1$. Combining these two inequalities, the integrand is nonzero for $\max\{0, w-1\} \le x \le \min\{1, w\}$. Thus,

$$f_W(w) = \begin{cases} \min\{1, w\} - \max\{0, w-1\}, & 0 \le w \le 2, \\ 0, & \text{otherwise,} \end{cases}$$

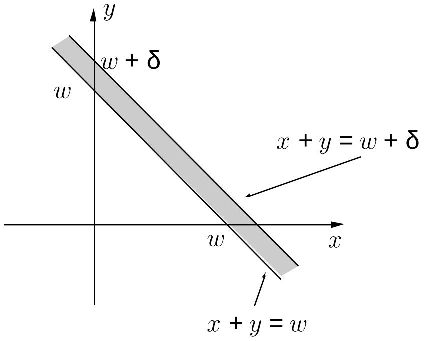


Figure 4.3: Illustration of the convolution formula for the case of continuous random variables (compare with Fig. 4.2). For small δ, the probability of the strip indicated in the figure is $\mathbf{P}(w \le X + Y \le w + \delta) \approx f_W(w) \cdot \delta$. Thus,

$$\begin{aligned} f_W(w) \cdot \delta =& \mathbf{P}(w \le X + Y \le w + \delta) \\ &= \int_{x=-\infty}^{\infty} \int_{y=w-x}^{w-x+\delta} f_X(x) f_Y(y)\, dy\, dx \\ &\approx \int_{x=-\infty}^{\infty} f_X(x) f_Y(w-x) \delta\, dx. \end{aligned}$$

The desired formula follows by canceling δ from both sides.

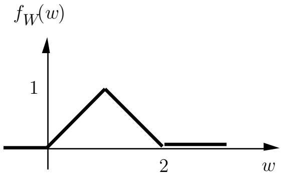


Figure 4.4: The PDF of the sum of two independent uniform random variables in $[0, 1]$.

which has the triangular shape shown in Fig. 4.4.

The calculation in the last example was based on a literal application of the convolution formula. The most delicate step was to determine the correct limits for the integration. This is often tedious and error prone, but can be bypassed using a graphical method described next.

Graphical Calculation of Convolutions

We will use a dummy variable t as the argument of the different functions involved in this discussion; see also Fig. 4.5. Consider a PDF $f_X(t)$ which is zero outside the range $a \leq t \leq b$ and a PDF $f_Y(t)$ which is zero outside the range $c \leq t \leq d$. Let us fix a value w, and plot $f_Y(w-t)$ as a function of t. This plot has the same shape as the plot of $f_Y(t)$ except that it is first "flipped" and then shifted by an amount w. (If $w > 0$, this is a shift to the right, if $w < 0$, this is a shift to the left.) We then place the plots of $f_X(t)$ and $f_Y(w-t)$ on top of each other. The value of $f_W(w)$ is equal to the integral of the product of these two plots. By varying the amount w by which we are shifting, we obtain $f_W(w)$ for any w.

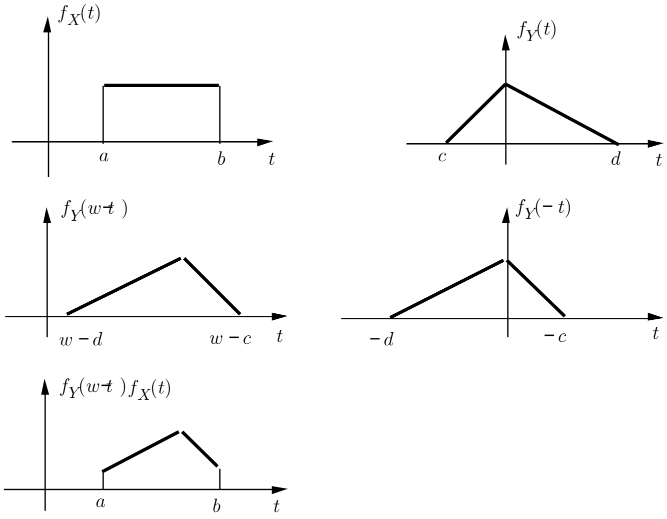


Figure 4.5: Illustration of the convolution calculation. For the value of w under consideration, $f_W(w)$ is equal to the integral of the function shown in the last plot.

4.3 CONDITIONAL EXPECTATION AS A RANDOM VARIABLE

The value of the conditional expectation $\mathbf{E}[X \mid Y = y]$ of a random variable X given another random variable Y depends on the realized experimental value y of Y. This makes $\mathbf{E}[X \mid Y]$ a function of Y, and therefore a random variable. In this section, we study the expectation and variance of $\mathbf{E}[X \mid Y]$. In the process, we obtain some useful formulas (the **law of iterated expectations** and the **law of conditional variances**) that are often convenient for the calculation of expected values and variances.

Recall that the conditional expectation $\mathbf{E}[X \mid Y = y]$ is defined by

$$\mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y), \qquad \text{(discrete case)},$$

and

$$\mathbf{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx, \qquad \text{(continuous case)}.$$

Once a value of y is given, the above summation or integration yields a numerical value for $\mathbf{E}[X \mid Y = y]$.

Example 4.15. Let the random variables X and Y have a joint PDF which is equal to 2 for (x, y) belonging to the triangle indicated in Fig. 4.6(a), and zero everywhere else. In order to compute $\mathbf{E}[X \mid Y = y]$, we first need to obtain the conditional density of X given $Y = y$.

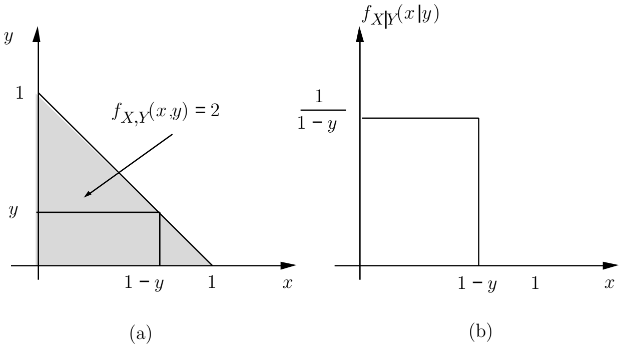


Figure 4.6: (a) The joint PDF in Example 4.15. (b) The conditional density of X.

We have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx = \int_0^{1-y} 2\, dx = 2(1 - y), \qquad 0 \le y \le 1,$$

and

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{1}{1 - y}, \qquad 0 \le x \le 1 - y.$$

The conditional density is shown in Fig. 4.6(b).

Intuitively, since the joint PDF is constant, the conditional PDF (which is a "slice" of the joint, at some fixed y) is also a constant. Therefore, the conditional PDF must be a uniform distribution. Given that $Y = y$, X ranges from 0 to $1 - y$. Therefore, for the PDF to integrate to 1, its height must be equal to $1/(1 - y)$, in agreement with Fig. 4.6(b).

For $y > 1$ or $y < 0$, the conditional PDF is undefined, since these values of y are impossible. For $y = 1$, X must be equal to 0, with certainty, and $\mathbf{E}[X \,|\, Y = 1] = 0$.

For $0 \le y < 1$, the conditional mean $\mathbf{E}[X \,|\, Y = y]$ is the expectation of the uniform PDF in Fig. 4.6(b), and we have

$$\mathbf{E}[X \,|\, Y = y] = \frac{1-y}{2}, \qquad 0 \le y < 1.$$

Since $\mathbf{E}[X \,|\, Y = 1] = 0$, the above formula is also valid when $y = 1$. The conditional expectation is undefined when y is outside $[0, 1]$.

For any number y, $\mathbf{E}[X \,|\, Y = y]$ is also a number. As y varies, so does $\mathbf{E}[X \,|\, Y = y]$, and we can therefore view $\mathbf{E}[X \,|\, Y = y]$ as a function of y. Since y is the experimental value of the random variable Y, we are dealing with a function of a random variable, hence a new random variable. More precisely, we **define** $\mathbf{E}[X \,|\, Y]$ to be the random variable whose value is $\mathbf{E}[X \,|\, Y = y]$ when the outcome of Y is y.

Example 4.15. (continued) We saw that $\mathbf{E}[X \,|\, Y = y] = (1 - y)/2$. Hence, $\mathbf{E}[X \,|\, Y]$ is the random variable $(1 - Y)/2$:

$$\mathbf{E}[X \,|\, Y] = \frac{1-Y}{2}.$$

Since $\mathbf{E}[X \,|\, Y]$ is a random variable, it has an expectation $\mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big]$ of its own. Applying the expected value rule, this is given by

$$\mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big] = \begin{cases} \displaystyle\sum_y \mathbf{E}[X \,|\, Y = y] p_Y(y), & Y \text{ discrete}, \\ \displaystyle\int_{-\infty}^{\infty} \mathbf{E}[X \,|\, Y = y] f_Y(y)\, dy, & Y \text{ continuous}. \end{cases}$$

Both expressions in the right-hand side should be familiar from Chapters 2 and 3, respectively. By the corresponding versions of the total expectation theorem, they are equal to $\mathbf{E}[X]$. This brings us to the following conclusion, which is actually valid for every type of random variable Y (discrete, continuous, mixed, etc.), as long as X has a well-defined and finite expectation $\mathbf{E}[X]$.

Law of iterated expectations: $\mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big] = \mathbf{E}[X]$.

Example 4.15 (continued) In Example 4.15, we found $\mathbf{E}[X \mid Y] = (1-Y)/2$ [see Fig. 4.6(b)]. Taking expectations of both sides, and using the law of iterated expectations to evaluate the left-hand side, we obtain $\mathbf{E}[X] = \big(1-\mathbf{E}[Y]\big)/2$. Because of symmetry, we must have $\mathbf{E}[X] = \mathbf{E}[Y]$. Therefore, $\mathbf{E}[X] = \big(1-\mathbf{E}[X]\big)/2$, which yields $\mathbf{E}[X] = 1/3$. In a slightly different version of this example, where there is no symmetry between X and Y, we would use a similar argument to express $\mathbf{E}[Y]$.

Example 4.16. We start with a stick of length ℓ. We break it at a point which is chosen randomly and uniformly over its length, and keep the piece that contains the left end of the stick. We then repeat the same process on the stick that we were left with. What is the expected length of the stick that we are left with, after breaking twice?

Let Y be the length of the stick after we break for the first time. Let X be the length after the second time. We have $\mathbf{E}[X \mid Y] = Y/2$, since the breakpoint is chosen uniformly over the length Y of the remaining stick. For a similar reason, we also have $\mathbf{E}[Y] = \ell/2$. Thus,

$$\mathbf{E}[X] = \mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}\left[\frac{Y}{2}\right] = \frac{\mathbf{E}[Y]}{2} = \frac{\ell}{4}.$$

Example 4.17. Averaging Quiz Scores by Section. A class has n students and the quiz score of student i is x_i. The average quiz score is

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The class consists of S sections, with n_s students in section s. The average score in section s is

$$m_s = \frac{1}{n_s} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i.$$

The average score over the whole class can be computed by taking the average score m_s of each section, and then forming a *weighted average*; the weight given to section s is proportional to the number of students in that section, and is n_s/n. We verify that this gives the correct result:

$$\begin{aligned}\sum_{s=1}^{S} \frac{n_s}{n} m_s &= \sum_{s=1}^{S} \frac{n_s}{n} \cdot \frac{1}{n_s} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i \\ &= \frac{1}{n}\sum_{s=1}^{S} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i. \\ &= m.\end{aligned}$$

How is this related to conditional expectations? Consider an experiment in which a student is selected at random, each student having probability $1/n$ of being selected. Consider the following two random variables:

$$X = \text{quiz score of a student,}$$
$$Y = \text{section of a student,}\ \ (Y \in \{1, \ldots, S\}).$$

We then have

$$\mathbf{E}[X] = m.$$

Conditioning on $Y = s$ is the same as assuming that the selected student is in section s. Conditional on that event, every student in that section has the same probability $1/n_s$ of being chosen. Therefore,

$$\mathbf{E}[X \,|\, Y = s] = \frac{1}{n_s} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i = m_s.$$

A randomly selected student belongs to section s with probability n_s/n, i.e., $\mathbf{P}(Y = s) = n_s/n$. Hence,

$$\mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big] = \sum_{s=1}^{S} \mathbf{E}[X \,|\, Y = s]\mathbf{P}(Y = s) = \sum_{s=1}^{S} \frac{n_s}{n} m_s.$$

As shown earlier, this is the same as m. Thus, averaging by section can be viewed as a special case of the law of iterated expectations.

Example 4.18. Forecast Revisions. Let Y be the sales of a company in the first semester of the coming year, and let X be the sales over the entire year. The company has constructed a statistical model of sales, and so the joint distribution of X and Y is assumed to be known. In the beginning of the year, the expected value $\mathbf{E}[X]$ serves as a forecast of the actual sales X. In the middle of the year, the first semester sales have been realized and the experimental value of the random value Y is now known. This places us in a new "universe," where everything is conditioned on the realized value of Y. We then consider the mid-year revised forecast of yearly sales, which is $\mathbf{E}[X \,|\, Y]$.

We view $\mathbf{E}[X \,|\, Y] - \mathbf{E}[X]$ as the forecast revision, in light of the mid-year information. The law of iterated expectations implies that

$$\mathbf{E}\big[\mathbf{E}[X \,|\, Y] - \mathbf{E}[X]\big] = 0.$$

This means that, in the beginning of the year, we do not expect our forecast to be revised in any specific direction. Of course, the actual revision will usually be positive or negative, but the probabilities are such that it is zero on the average. This is quite intuitive. For example, if a positive revision was expected, the original forecast should have been higher in the first place.

The Conditional Variance

The conditional distribution of X given $Y=y$ has a mean, which is $\mathbf{E}[X\,|\,Y=y]$, and by the same token, it also has a variance. This is defined by the same formula as the unconditional variance, except that everything is conditioned on $Y=y$:

$$\mathrm{var}(X\,|\,Y=y) = \mathbf{E}\Big[\big(X-\mathbf{E}[X\,|\,Y=y]\big)^2 \;\big|\; Y=y\Big].$$

Note that the conditional variance is a function of the experimental value y of the random variable Y. Hence, it is a function of a random variable, and is itself a random variable that will be denoted by $\mathrm{var}(X\,|\,Y)$.

Arguing by analogy to the law of iterated expectations, we may conjecture that the expectation of the conditional variance $\mathrm{var}(X\,|\,Y)$ is related to the unconditional variance $\mathrm{var}(X)$. This is indeed the case, but the relation is more complex.

Law of Conditional Variances:

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X\,|\,Y)\big] + \mathrm{var}\big(\mathbf{E}[X\,|\,Y]\big)$$

To verify the law of conditional variances, we start with the identity

$$X-\mathbf{E}[X] = \big(X-\mathbf{E}[X\,|\,Y]\big) + \big(\mathbf{E}[X\,|\,Y]-\mathbf{E}[X]\big).$$

We square both sides and then take expectations to obtain

$$\begin{aligned}\mathrm{var}(X) &= \mathbf{E}\Big[\big(X-\mathbf{E}[X]\big)^2\Big]\\ &= \mathbf{E}\Big[\big(X-\mathbf{E}[X\,|\,Y]\big)^2\Big] + \mathbf{E}\Big[\big(\mathbf{E}[X\,|\,Y]-\mathbf{E}[X]\big)^2\Big]\\ &\quad + 2\mathbf{E}\Big[\big(X-\mathbf{E}[X\,|\,Y]\big)\big(\mathbf{E}[X\,|\,Y]-\mathbf{E}[X]\big)\Big].\end{aligned}$$

Using the law of iterated expectations, the first term in the right-hand side of the above equation can be written as

$$\mathbf{E}\Big[\mathbf{E}\Big[\big(X-\mathbf{E}[X\,|\,Y]\big)^2 \;\big|\; Y\Big]\Big],$$

which is the same as $\mathbf{E}\big[\mathrm{var}(X\,|\,Y)\big]$. The second term is equal to $\mathrm{var}\big(\mathbf{E}[X\,|\,Y]\big)$, since $\mathbf{E}[X]$ is the mean of $\mathbf{E}[X\,|\,Y]$. Finally, the third term is zero, as we now show. Indeed, if we define $h(Y) = 2\big(\mathbf{E}[X\,|\,Y]-\mathbf{E}[X]\big)$, the third term is

$$\begin{aligned}\mathbf{E}\Big[\big(X-\mathbf{E}[X\,|\,Y]\big)h(Y)\Big] &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[\mathbf{E}[X\,|\,Y]h(Y)\big]\\ &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\Big[\mathbf{E}\big[Xh(Y)\,|\,Y\big]\Big]\\ &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[Xh(Y)\big]\\ &= 0.\end{aligned}$$

Example 4.16. (continued) Consider again the problem where we break twice a stick of length ℓ, at randomly chosen points, with Y being the length of the stick after the first break and X being the length after the second break. We calculated the mean of X as $\ell/4$, and now let us use the law of conditional variances to calculate $\mathrm{var}(X)$. We have $\mathbf{E}[X \mid Y] = Y/2$, so since Y is uniformly distributed between 0 and ℓ,

$$\mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = \mathrm{var}(Y/2) = \frac{1}{4}\mathrm{var}(Y) = \frac{1}{4} \cdot \frac{\ell^2}{12} = \frac{\ell^2}{48}.$$

Also, since X is uniformly distributed between 0 and Y, we have

$$\mathrm{var}(X \mid Y) = \frac{Y^2}{12}.$$

Thus, since Y is uniformly distributed between 0 and ℓ,

$$\mathbf{E}\big[\mathrm{var}(X \mid Y)\big] = \frac{1}{12}\int_0^\ell \frac{1}{\ell}y^2 dy = \frac{1}{12}\frac{1}{3\ell}y^3\Big|_0^\ell = \frac{\ell^2}{36}.$$

Using now the law of conditional variances, we obtain

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = \frac{\ell^2}{48} + \frac{\ell^2}{36} = \frac{7\ell^2}{144}.$$

Example 4.19. Averaging Quiz Scores by Section – Variance. The setting is the same as in Example 4.17 and we consider the random variables

$$X = \text{quiz score of a student},$$
$$Y = \text{section of a student}, \quad (Y \in \{1, \ldots, S\}).$$

Let n_s be the number of students in section s, and let n be the total number of students. We interpret the different quantities in the formula

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big).$$

In this context, $\mathrm{var}(X \mid Y = s)$ is the variance of the quiz scores within section s. Then, $\mathbf{E}\big[\mathrm{var}(X \mid Y)\big]$ is the average of the section variances. This latter expectation is an average over the probability distribution of Y, i.e.,

$$\mathbf{E}\big[\mathrm{var}(X \mid Y)\big] = \sum_{s=1}^{S} \frac{n_s}{n}\mathrm{var}(X \mid Y = s).$$

Recall that $\mathbf{E}[X \mid Y = s]$ is the average score in section s. Then, $\mathrm{var}\big(\mathbf{E}[X \mid Y]\big)$ is a measure of the variability of the averages of the different sections. The law of conditional variances states that the total quiz score variance can be broken into two parts:

(a) The average score variability $\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big]$ *within* individual sections.

(b) The variability $\mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big)$ *between* sections.

We have seen earlier that the law of iterated expectations (in the form of the total expectation theorem) can be used to break down complicated expectation calculations, by considering different cases. A similar method applies to variance calculations.

Example 4.20. Computing Variances by Conditioning. Consider a continuous random variable X with the PDF given in Fig. 4.7. We define an auxiliary random variable Y as follows:

$$Y = \begin{cases} 1, & \text{if } x < 1, \\ 2, & \text{of } x \geq 1. \end{cases}$$

Here, $\mathbf{E}[X \,|\, Y]$ takes the values 1/2 and 3/2, with probabilities 1/3 and 2/3, respectively. Thus, the mean of $\mathbf{E}[X \,|\, Y]$ is 7/6. Therefore,

$$\mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big) = \frac{1}{3}\left(\frac{1}{2} - \frac{7}{6}\right)^2 + \frac{2}{3}\left(\frac{3}{2} - \frac{7}{6}\right)^2 = \frac{2}{9}.$$

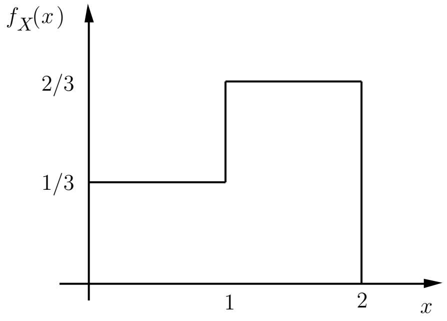


Figure 4.7: The PDF in Example 4.20.

Conditioned on either value of Y, X is uniformly distributed on a unit length interval. Therefore, $\mathrm{var}(X \,|\, Y = y) = 1/12$ for each of the two possible values of y, and $\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] = 1/12$. Putting everything together, we obtain

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] + \mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big) = \frac{1}{12} + \frac{2}{9} = \frac{11}{36}.$$

We summarize the main points in this section.

The Mean and Variance of a Conditional Expectation

- $\mathbf{E}[X \mid Y = y]$ is a number, whose value depends on y.
- $\mathbf{E}[X \mid Y]$ is a function of the random variable Y, hence a random variable. Its experimental value is $\mathbf{E}[X \mid Y = y]$ whenever the experimental value of Y is y.
- $\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}[X]$ (law of iterated expectations).
- $\text{var}(X \mid Y)$ is a random variable whose experimental value is $\text{var}(X \mid Y = y)$, whenever the experimental value of Y is y.
- $\text{var}(X) = \mathbf{E}\big[\text{var}(X \mid Y)\big] + \text{var}\big(\mathbf{E}[X \mid Y]\big)$.

4.4 SUM OF A RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES

In our discussion so far of sums of random variables, we have always assumed that the number of variables in the sum is known and fixed, i.e., it is nonrandom. In this section we will consider the case where the number of random variables being added is itself random. In particular, we consider the sum

$$Y = X_1 + \cdots + X_N,$$

where N is a random variable that takes nonnegative integer values, and $X_1, X_2, \ldots$ are identically distributed random variables. We assume that $N, X_1, X_2, \ldots$ are independent, meaning that any finite subcollection of these random variables are independent.

We first note that the randomness of N can affect significantly the character of the random sum $Y = X_1 + \cdots + X_N$. In particular, the PMF/PDF of $Y = \sum_{i=1}^{N} Y_i$ is much different from the PMF/PDF of the sum $\overline{Y} = \sum_{i=1}^{\mathbf{E}[N]} Y_i$ where N has been replaced by its expected value (assuming that $\mathbf{E}[N]$ is integer). For example, let X_i be uniformly distributed in the interval $[0, 1]$, and let N be equal to 1 or 3 with probability 1/2 each. Then the PDF of the random sum Y takes values in the interval $[0, 3]$, whereas if we replace N by its expected value $\mathbf{E}[N] = 2$, the sum $\overline{Y} = X_1 + X_2$ takes values in the interval $[0, 2]$. Furthermore, using the total probability theorem, we see that the PDF of Y is a mixture of the uniform PDF and the PDF of $X_1 + X_2 + X_3$, and has considerably different character than the triangular PDF of $\overline{Y} = X_1 + X_2$ which is given in Fig. 4.4.

Let us denote by μ and σ^2 the common mean and the variance of the X_i. We wish to derive formulas for the mean, variance, and the transform of Y . The method that we follow is to first condition on the event $N = n$, under which we have the sum of a *fixed* number of random of random variables, a case that we already know how to handle.

Fix some number n. The random variable $X_1 + \cdots + X_n$ is independent of N and, therefore, independent of the event $\{N = n\}$. Hence,

$$\begin{aligned}\mathbf{E}[Y \mid N = n] &= \mathbf{E}[X_1 + \cdots + X_N \mid N = n] \\ &= \mathbf{E}[X_1 + \cdots + X_n \mid N = n] \\ &= \mathbf{E}[X_1 + \cdots + X_n] \\ &= n\mu.\end{aligned}$$

This is true for every nonnegative integer n and, therefore,

$$\mathbf{E}[Y \mid N] = N\mu.$$

Using the law of iterated expectations, we obtain

$$\mathbf{E}[Y] = \mathbf{E}\big[\mathbf{E}[Y \mid N]\big] = \mathbf{E}[\mu N] = \mu\mathbf{E}[N].$$

Similarly,

$$\begin{aligned}\text{var}(Y \mid N = n) &= \text{var}(X_1 + \cdots + X_N \mid N = n) \\ &= \text{var}(X_1 + \cdots + X_n) \\ &= n\sigma^2.\end{aligned}$$

Since this is true for every nonnegative integer n, the random variable $\text{var}(Y \mid N)$ is equal to $N\sigma^2$. We now use the law of conditional variances to obtain

$$\begin{aligned}\text{var}(Y) &= \mathbf{E}\big[\text{var}(Y \mid N)\big] + \text{var}\big(\mathbf{E}[Y \mid N]\big) \\ &= \mathbf{E}[N]\sigma^2 + \text{var}(N\mu) \\ &= \mathbf{E}[N]\sigma^2 + \mu^2\text{var}(N).\end{aligned}$$

The calculation of the transform proceeds along similar lines. The transform associated with Y, conditional on $N = n$, is $\mathbf{E}[e^{sY} \mid N = n]$. However, conditioned on $N = n$, Y is the sum of the independent random variables $X_1, \ldots, X_n$, and

$$\begin{aligned}\mathbf{E}[e^{sY} \mid N = n] &= \mathbf{E}\big[e^{sX_1} \cdots e^{sX_N} \mid N = n\big] = \mathbf{E}\big[e^{sX_1} \cdots e^{sX_n}\big] \\ &= \mathbf{E}[e^{sX_1}] \cdots \mathbf{E}[e^{sX_n}] = \big(M_X(s)\big)^n.\end{aligned}$$

Using the law of iterated expectations, the (unconditional) transform associated with Y is

$$\mathbf{E}[e^{sY}] = \mathbf{E}\big[\mathbf{E}[e^{sY} \mid N]\big] = \mathbf{E}\big[\big(M_X(s)\big)^N\big] = \sum_{n=0}^{\infty}(M_X(s))^n p_N(n).$$

This is similar to the transform $M_N(s) = \mathbf{E}[e^{sN}]$ associated with N, except that e^s is replaced by $M_X(s)$.

Example 4.21. A remote village has three gas stations, and each one of them is open on any given day with probability 1/2, independently of the others. The amount of gas available in each gas station is unknown and is uniformly distributed between 0 and 1000 gallons. We wish to characterize the distribution of the total amount of gas available at the gas stations that are open.

The number N of open gas stations is a binomial random variable with $p = 1/2$ and the corresponding transform is

$$M_N(s) = (1 - p + pe^s)^3 = \frac{1}{8}(1 + e^s)^3.$$

The transform $M_X(s)$ associated with the amount of gas available in an open gas station is

$$M_X(s) = \frac{e^{1000s} - 1}{1000s}.$$

The transform associated with the total amount Y available is the same as $M_N(s)$, except that each occurrence of e^s is replaced with $M_X(s)$, i.e.,

$$M_Y(s) = \frac{1}{8}\left(1 + \left(\frac{e^{1000s} - 1}{1000s}\right)\right)^3.$$

Example 4.22. Sum of a Geometric Number of Independent Exponential Random Variables. Jane visits a number of bookstores, looking for *Great Expectations*. Any given bookstore carries the book with probability p, independently of the others. In a typical bookstore visited, Jane spends a random amount of time, exponentially distributed with parameter λ, until she either finds the book or she decides that the bookstore does not carry it. Assuming that Jane will keep visiting bookstores until she buys the book and that the time spent in each is independent of everything else, we wish to determine the mean, variance, and PDF of the total time spent in bookstores.

The total number N of bookstores visited is geometrically distributed with parameter p. Hence, the total time Y spent in bookstores is the sum of a geometrically distributed number N of independent exponential random variables $X_1, X_2, \ldots$. We have

$$\mathbf{E}[Y] = \mathbf{E}[N]\mathbf{E}[X] = \frac{1}{p} \cdot \frac{1}{\lambda}.$$

Using the formulas for the variance of geometric and exponential random variables, we also obtain

$$\text{var}(Y) = \mathbf{E}[N]\text{var}(X) + \big(\mathbf{E}[X]\big)^2\text{var}(N) = \frac{1}{p} \cdot \frac{1}{\lambda^2} + \frac{1}{\lambda^2} \cdot \frac{1-p}{p^2} = \frac{1}{\lambda^2 p^2}.$$

In order to find the transform $M_Y(s)$, let us recall that

$$M_X(s) = \frac{\lambda}{\lambda - s}, \qquad M_N(s) = \frac{pe^s}{1-(1-p)e^s}.$$

Then, $M_Y(s)$ is found by starting with $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$. This yields

$$M_Y(s) = \frac{pM_X(s)}{1-(1-p)M_X(s)} = \frac{\dfrac{p\lambda}{\lambda - s}}{1-(1-p)\dfrac{\lambda}{\lambda - s}},$$

which simplifies to

$$M_Y(s) = \frac{p\lambda}{p\lambda - s}.$$

We recognize this as the transform of an exponentially distributed random variable with parameter $p\lambda$, and therefore,

$$f_Y(y) = p\lambda e^{-p\lambda y}, \qquad y \geq 0.$$

This result can be surprising because the sum of a *fixed* number n of independent exponential random variables is not exponentially distributed. For example, if $n = 2$, the transform associated with the sum is $\big(\lambda/(\lambda - s)\big)^2$, which does not correspond to the exponential distribution.

Example 4.23. Sum of a Geometric Number of Independent Geometric Random Variables. This example is a discrete counterpart of the preceding one. We let N be geometrically distributed with parameter p. We also let each random variable X_i be geometrically distributed with parameter q. We assume that all of these random variables are independent. Let $Y = X_1 + \cdots + X_N$. We have

$$M_N(s) = \frac{pe^s}{1-(1-p)e^s}, \qquad M_X(s) = \frac{qe^s}{1-(1-q)e^s}.$$

To determine $M_Y(s)$, we start with the formula for $M_N(s)$ and replace each occurrence of e^s with $M_X(s)$. This yields

$$M_Y(s) = \frac{pM_X(s)}{1-(1-p)M_X(s)},$$

and, after some algebra,

$$M_Y(s) = \frac{pqe^s}{1-(1-pq)e^s}.$$

We conclude that Y is geometrically distributed, with parameter pq.

Properties of Sums of a Random Number of Independent Random Variables

Let $X_1, X_2, \ldots$ be random variables with common mean μ and common variance σ^2. Let N be a random variable that takes nonnegative integer values. We assume that all of these random variables are independent, and consider

$$Y = X_1 + \cdots + X_N.$$

Then,

- $\mathbf{E}[Y] = \mu \mathbf{E}[N]$.
- $\text{var}(Y) = \sigma^2 \mathbf{E}[N] + \mu^2 \text{var}(N)$.
- The transform $M_Y(s)$ is found by starting with the transform $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$.

4.5 COVARIANCE AND CORRELATION

The **covariance** of two random variables X and Y is denoted by $\text{cov}(X,Y)$, and is defined by

$$\text{cov}(X,Y) = \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big].$$

When $\text{cov}(X,Y) = 0$, we say that X and Y are **uncorrelated**.

Roughly speaking, a positive or negative covariance indicates that the values of $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ obtained in a single experiment "tend" to have the same or the opposite sign, respectively (see Fig. 4.8). Thus the sign of the covariance provides an important qualitative indicator of the relation between X and Y.

If X and Y are independent, then

$$\text{cov}(X,Y) = \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] = \mathbf{E}\big[X - \mathbf{E}[X]\big]\mathbf{E}\big[Y - \mathbf{E}[Y]\big] = 0.$$

Thus if X and Y are independent, they are also uncorrelated. However, the reverse is not true, as illustrated by the following example.

Example 4.24. The pair of random variables (X,Y) takes the values $(1,0)$, $(0,1)$, $(-1,0)$, and $(0,-1)$, each with probability $1/4$ (see Fig. 4.9). Thus, the marginal PMFs of X and Y are symmetric around 0, and $\mathbf{E}[X] = \mathbf{E}[Y] = 0$. Furthermore, for all possible value pairs (x,y), either x or y is equal to 0, which implies that $XY = 0$ and $\mathbf{E}[XY] = 0$. Therefore,

$$\text{cov}(X,Y) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\Big] = \mathbf{E}[XY] = 0,$$

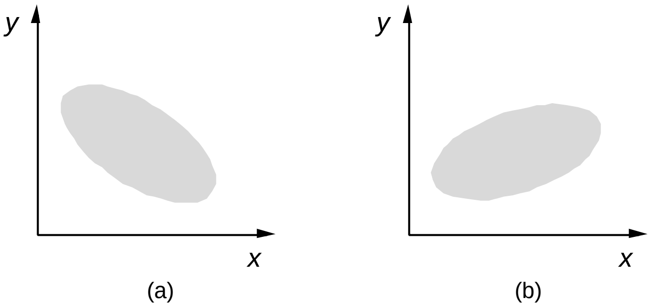


Figure 4.8: Examples of positively and negatively correlated random variables. Here X and Y are uniformly distributed over the ellipses shown. In case (a) the covariance $\mathrm{cov}(X,Y)$ is negative, while in case (b) it is positive.

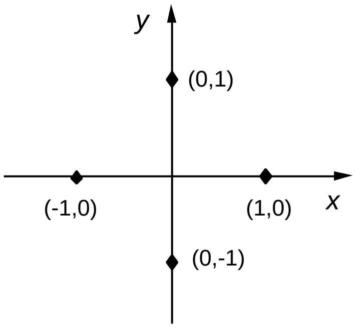


Figure 4.9: Joint PMF of X and Y for Example 4.21. Each of the four points shown has probability 1/4. Here X and Y are uncorrelated but not independent.

and X and Y are uncorrelated. However, X and Y are not independent since, for example, a nonzero value of X fixes the value of Y to zero.

The **correlation coefficient** ρ of two random variables X and Y that have nonzero variances is defined as

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}.$$

It may be viewed as a normalized version of the covariance $\mathrm{cov}(X,Y)$, and in fact it can be shown that ρ ranges from -1 to 1 (see the end-of-chapter problems).

If $\rho > 0$ (or $\rho < 0$), then the values of $x - \mathbf{E}[X]$ and $y - \mathbf{E}[Y]$ "tend" to have the same (or opposite, respectively) sign, and the size of $|\rho|$ provides a normalized measure of the extent to which this is true. In fact, always assuming that X and Y have positive variances, it can be shown that $\rho = 1$ (or $\rho = -1$) if and only if there exists a positive (or negative, respectively) constant c such that

$$y - \mathbf{E}[Y] = c\big(x - \mathbf{E}[X]\big), \qquad \text{for all possible numerical values } (x,y)$$

(see the end-of-chapter problems). The following example illustrates in part this property.

Example 4.25. Consider n independent tosses of a biased coin with probability of a head equal to p. Let X and Y be the numbers of heads and of tails, respectively, and let us look at the correlation of X and Y. Here, for all possible pairs of values (x, y), we have $x + y = n$, and we also have $\mathbf{E}[X] + \mathbf{E}[Y] = n$. Thus,

$$x - \mathbf{E}[X] = -\big(y - \mathbf{E}[Y]\big), \qquad \text{for all possible } (x, y).$$

We will calculate the correlation coefficient of X and Y, and verify that it is indeed equal to -1.

We have

$$\begin{aligned} \mathrm{cov}(X, Y) &= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] \\ &= -\mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\ &= -\mathrm{var}(X). \end{aligned}$$

Hence, the correlation coefficient is

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{-\mathrm{var}(X)}{\sqrt{\mathrm{var}(X)\mathrm{var}(X)}} = -1.$$

The covariance can be used to obtain a formula for the variance of the sum of several (not necessarily independent) random variables. In particular, if $X_1, X_2, \ldots, X_n$ are random variables with finite variance, we have

$$\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathrm{cov}(X_i, X_j).$$

This can be seen from the following calculation, where for brevity, we denote $\tilde{X}_i = X_i - \mathbf{E}[X_i]$:

$$\begin{aligned} \mathrm{var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbf{E}\left[\left(\sum_{i=1}^{n} \tilde{X}_i\right)^2\right] \\ &= \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{X}_i \tilde{X}_j\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1} \mathbf{E}[\tilde{X}_i^{\,2}] + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1}^{n} \mathrm{var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathrm{cov}(X_i, X_j). \end{aligned}$$

The following example illustrates the use of this formula.

Example 4.26. Consider the hat problem discussed in Section 2.5, where n people throw their hats in a box and then pick a hat at random. Let us find the variance of X, the number of people that pick their own hat. We have

$$X = X_1 + \cdots + X_n,$$

where X_i is the random variable that takes the value 1 if the ith person selects his/her own hat, and takes the value 0 otherwise. Noting that X_i is Bernoulli with parameter $p = \mathbf{P}(X_i = 1) = 1/n$, we obtain

$$\text{var}(X_i) = \frac{1}{n}\left(1 - \frac{1}{n}\right).$$

For $i \neq j$, we have

$$\begin{aligned}
\text{cov}(X_i, X_j) &= \mathbf{E}\Big[\big(X_i - \mathbf{E}[X_i]\big)\big(X_j - \mathbf{E}[X_j]\big)\Big] \\
&= \mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\mathbf{E}[X_j] \\
&= \mathbf{P}(X_i = 1 \text{ and } X_j = 1) - \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1) \\
&= \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1 \,|\, X_i = 1) - \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1) \\
&= \frac{1}{n}\frac{1}{n-1} - \frac{1}{n^2} \\
&= \frac{1}{n^2(n-1)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\text{var}(X) &= \text{var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{n} \text{var}(X_i) + 2\sum_{\substack{i,j=1 \\ i<j}}^{n} \text{cov}(X_i, X_j) \\
&= n\frac{1}{n}\left(1 - \frac{1}{n}\right) + 2\frac{n(n-1)}{2}\frac{1}{n^2(n-1)} \\
&= 1.
\end{aligned}$$

4.6 LEAST SQUARES ESTIMATION

In many practical contexts, we want to form an estimate of the value of a random variable X given the value of a related random variable Y, which may be viewed as some form of "measurement" of X. For example, X may be the range of an aircraft and Y may be a noise-corrupted measurement of that range. In this section we discuss a popular formulation of the estimation problem, which is based on finding the estimate c that minimizes the expected value of the squared error $(X-c)^2$ (hence the name "least squares").

If the value of Y is not available, we may consider finding an estimate (or prediction) c of X. The estimation error $X-c$ is random (because X is random), but the mean squared error $\mathbf{E}\big[(X-c)^2\big]$ is a number that depends on c and can be minimized over c. With respect to this criterion, it turns out that the best possible estimate is $c=\mathbf{E}[X]$, as we proceed to verify.

Let $m=\mathbf{E}[X]$. For any estimate c, we have

$$\begin{aligned}\mathbf{E}\big[(X-c)^2\big] &= \mathbf{E}\big[(X-m+m-c)^2\big]\\ &= \mathbf{E}\big[(X-m)^2\big]+2\mathbf{E}\big[(X-m)(m-c)\big]+\mathbf{E}\big[(m-c)^2\big]\\ &= \mathbf{E}\big[(X-m)^2\big]+2\mathbf{E}[X-m](m-c)+(m-c)^2\\ &= \mathbf{E}\big[(X-m)^2\big]+(m-c)^2,\end{aligned}$$

where we used the fact $\mathbf{E}[X-m]=0$. The first term in the right-hand side is the variance of X and is unaffected by our choice of c. Therefore, we should choose c in a way that minimizes the second term, which leads to $c=m=\mathbf{E}[X]$ (see Fig. 4.10).

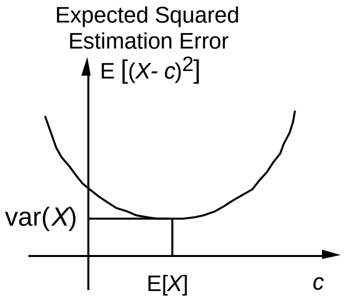


Figure 4.10: The mean squared error $\mathbf{E}\big[(X-c)^2\big]$, as a function of the estimate c, is a quadratic in c and is minimized when $c=\mathbf{E}[X]$. The minimum value of the mean squared error is $\text{var}(X)$.

Suppose now that we observe the experimental value y of some related random variable Y, before forming an estimate of X. How can we exploit this additional information? Once we are told that Y takes a particular value y, the situation is identical to the one considered earlier, except that we are now in a new "universe," where everything is conditioned on $Y=y$. We can therefore adapt our earlier conclusion and assert that $c=\mathbf{E}[X\,|\,Y=y]$ minimizes the *conditional* mean squared error $\mathbf{E}\big[(c-X)^2 \,|\, Y=y\big]$. Note that the resulting estimate c depends on the experimental value y of Y (as it should). Thus, we call $\mathbf{E}[X \,|\, Y=y]$ the *least-squares estimate* of X given the experimental value y.

Example 4.27. Let X be uniformly distributed in the interval $[4,10]$ and suppose that we observe X with some random error W, that is, we observe the experimental value of the random variable

$$Y = X + W.$$

We assume that W is uniformly distributed in the interval $[-1,1]$, and independent of X. What is the least squares estimate of X given the experimental value of Y?

We have $f_X(x) = 1/6$ for $4 \le x \le 10$, and $f_X(x) = 0$, elsewhere. Conditioned on X being equal to some x, Y is the same as $x + W$, and is uniform over the interval $[x-1, x+1]$. Thus, the joint PDF is given by

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y \,|\, x) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12},$$

if $4 \le x \le 10$ and $x - 1 \le y \le x + 1$, and is zero for all other values of (x,y). The slanted rectangle in the right-hand side of Fig. 4.11 is the set of pairs (x,y) for which $f_{X,Y}(x,y)$ is nonzero.

Given an experimental value y of Y, the conditional PDF $f_{X|Y}$ of X is uniform on the corresponding vertical section of the slanted rectangle. The optimal estimate $\mathbf{E}[X \,|\, Y=y]$ is the midpoint of that section. In the special case of the present example, it happens to be a piecewise linear function of y.

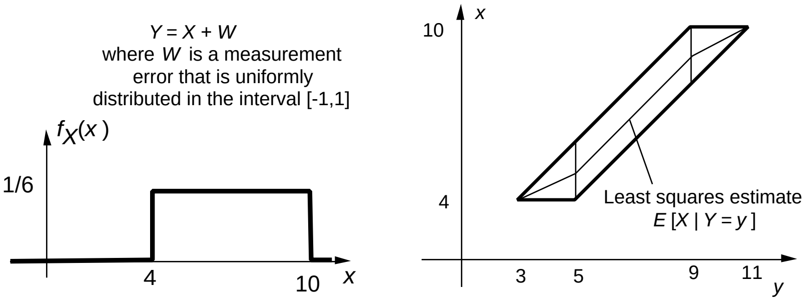


Figure 4.11: The PDFs in Example 4.27. The least squares estimate of X given the experimental value y of the random variable $Y = X + W$ depends on y and is represented by the piecewise linear function shown in the figure on the right.

As Example 4.27 illustrates, the estimate $\mathbf{E}[X \,|\, Y=y]$ depends on the observed value y and should be viewed as a function of y; see Fig. 4.12. To amplify this point, we refer to any function of the available information as an **estimator**. Given an experimental outcome y of Y, an estimator $g(\cdot)$ (which is a function) produces an estimate $g(y)$ (which is a number). However, if y is left unspecified, then the estimator results in a random variable $g(Y)$. The expected value of the squared estimation error associated with an estimator $g(Y)$ is

$$\mathbf{E}\Big[\big(X - g(Y)\big)^2\Big].$$

Out of all estimators, it turns out that the mean squared estimation error is minimized when $g(Y) = \mathbf{E}[X \,|\, Y]$. To see this, note that if c is any number, we have

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y = y]\big)^2 \;\big|\; Y = y\Big] \leq \mathbf{E}\big[(X - c)^2 \,|\, Y = y\big].$$

Consider now an estimator $g(Y)$. For a given value y of Y, $g(y)$ is a number and, therefore,

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y = y]\big)^2 \,|\, Y = y\Big] \leq \mathbf{E}\Big[\big(X - g(y)\big)^2 \;\big|\; Y = y\Big].$$

This inequality is true for *every* possible experimental value y of Y. Thus,

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y]\big)^2 \;\big|\; Y\Big] \leq \mathbf{E}\Big[\big(X - g(Y)\big)^2 \;\big|\; Y\Big],$$

which is now an inequality between random variables (functions of Y). We take expectations of both sides, and use the law of iterated expectations, to conclude that

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y]\big)^2\Big] \leq \mathbf{E}\Big[\big(X - g(Y)\big)^2\Big]$$

for all functions $g(Y)$.

y
LEAST SQUARES
ESTIMATOR
E[X |Y = y]

Figure 4.12: The least squares estimator.

Key Facts about Least Mean Squares Estimation

- $\mathbf{E}\big[(X-c)^2\big]$ is minimized when $c = \mathbf{E}[X]$:

$$\mathbf{E}\Big[\big(X-\mathbf{E}[X]\big)^2\Big] \le \mathbf{E}\big[(X-c)^2\big], \qquad \text{for all } c.$$

- $\mathbf{E}\big[(X-c)^2 \,|\, Y=y\big]$ is minimized when $c = \mathbf{E}[X \,|\, Y=y]$:

$$\mathbf{E}\Big[\big(X-\mathbf{E}[X \,|\, Y=y]\big)^2 \;\big|\; Y=y\Big] \le \mathbf{E}\big[(X-c)^2 \,|\, Y=y\big], \quad \text{for all } c.$$

- Out of all estimators $g(Y)$ of X based on Y, the mean squared estimation error $\mathbf{E}\Big[\big(X-g(Y)\big)^2\Big]$ is minimized when $g(Y) = \mathbf{E}[X \,|\, Y]$:

$$\mathbf{E}\Big[\big(X-\mathbf{E}[X \,|\, Y]\big)^2\Big] \le \mathbf{E}\Big[\big(X-g(Y)\big)^2\Big], \qquad \text{for all functions } g(Y).$$

Some Properties of the Estimation Error

Let us introduce the notation

$$\hat{X} = \mathbf{E}[X \,|\, Y], \qquad\qquad \tilde{X} = X - \hat{X},$$

for the (optimal) estimator and the associated estimation error, respectively. Note that both $\hat{X}$ and $\tilde{X}$ are random variables, and by the law of iterated expectations,

$$\mathbf{E}[\tilde{X}] = \mathbf{E}\big[X - \mathbf{E}[X \,|\, Y]\big] = \mathbf{E}[X] - \mathbf{E}[X] = 0.$$

The equation $\mathbf{E}[\tilde{X}] = 0$ remains valid even if we condition on Y, because

$$\mathbf{E}[\tilde{X} \,|\, Y] = \mathbf{E}[X - \hat{X} \,|\, Y] = \mathbf{E}[X \,|\, Y] - \mathbf{E}[\hat{X} \,|\, Y] = \hat{X} - \hat{X} = 0.$$

We have used here the fact that $\hat{X}$ is completely determined by Y and therefore $\mathbf{E}[\hat{X} \,|\, Y] = \hat{X}$. For similar reasons,

$$\mathbf{E}\big[\big(\hat{X} - \mathbf{E}[X]\big)\tilde{X} \,|\, Y\big] = \big(\hat{X} - \mathbf{E}[X]\big)\mathbf{E}[\tilde{X} \,|\, Y] = 0.$$

Taking expectations and using the law of iterated expectations, we obtain

$$\mathbf{E}\big[\big(\hat{X} - \mathbf{E}[X]\big)\tilde{X}\big] = 0.$$

Note that $X = \hat{X} + \tilde{X}$, which yields $X - \mathbf{E}[X] = \hat{X} - \mathbf{E}[X] + \tilde{X}$. We square both sides of the latter equality and take expectations to obtain

$$\begin{aligned}\mathrm{var}(X) &= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)^2\big] \\ &= \mathbf{E}\left[\big(\hat{X} - \mathbf{E}[X] + \tilde{X}\big)^2\right] \\ &= \mathbf{E}\left[\big(\hat{X} - \mathbf{E}[X]\big)^2\right] + \mathbf{E}[\tilde{X}^2] + 2\mathbf{E}\big[\big(\hat{X} - \mathbf{E}[X]\big)\tilde{X}\big] \\ &= \mathbf{E}\left[\big(\hat{X} - \mathbf{E}[X]\big)^2\right] + \mathbf{E}[\tilde{X}^2] \\ &= \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X}).\end{aligned}$$

(The last equality holds because $\mathbf{E}[\hat{X}] = \mathbf{E}[X]$ and $\mathbf{E}[\tilde{X}] = 0$.) In summary, we have established the following important formula, which is just another version of the law of conditional variances introduced in Section 4.3.

$$\mathrm{var}(X) = \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X}).$$

Example 4.28. Let us say that the observed random variable Y is *uninformative* if the mean squared estimation error $\mathbf{E}[\tilde{X}^2] = \mathrm{var}(\tilde{X})$ is the same as the unconditional variance $\mathrm{var}(X)$ of X. When is this the case?

Using the formula

$$\mathrm{var}(X) = \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X}),$$

we see that Y is uninformative if and only if $\mathrm{var}\big(\hat{X}\big) = 0$. The variance of a random variable is zero if and only if that random variable is a constant, equal to its mean. We conclude that Y is uninformative if and only if $\hat{X} = \mathbf{E}[X \mid Y] = \mathbf{E}[X]$, for every realization of Y.

If X and Y are independent, we have $\mathbf{E}[X \mid Y] = \mathbf{E}[X]$ and Y is indeed uninformative, which is quite intuitive. The converse, however, is not true. That is, it is possible for $\mathbf{E}[X \mid Y]$ to be always equal to the constant $\mathbf{E}[X]$, without X and Y being independent. (Can you construct an example?)

Estimation Based on Several Measurements

So far, we have discussed the case where we estimate one random variable X on the basis of another random variable Y. In practice, one often has access to the experimental values of several random variables $Y_1, \ldots, Y_n$, that can be used to estimate X. Generalizing our earlier discussion, and using essentially the same argument, the mean squared estimation error is minimized if we use $\mathbf{E}[X \mid Y_1, \ldots, Y_n]$ as our estimator. That is,

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y_1, \ldots, Y_n]\big)^2\Big] \le \mathbf{E}\Big[\big(X - g(Y_1, \ldots, Y_n)\big)^2\Big],$$

$$\text{for all functions } g(Y_1, \ldots, Y_n).$$

This provides a complete solution to the general problem of least squares estimation, but is sometimes difficult to implement, because:

(a) In order to compute the conditional expectation $\mathbf{E}[X \mid Y_1, \ldots, Y_n]$, we need a complete probabilistic model, that is, the joint PDF $f_{X,Y_1,\ldots,Y_n}(\cdot)$ of $n+1$ random variables.

(b) Even if this joint PDF is available, $\mathbf{E}[X \mid Y_1, \ldots, Y_n]$ can be a very complicated function of $Y_1, \ldots, Y_n$.

As a consequence, practitioners often resort to approximations of the conditional expectation or focus on estimators that are not optimal but are simple and easy to implement. The most common approach involves *linear estimators*, of the form

$$a_1 Y_1 + \cdots + a_n Y_n + b.$$

Given a particular choice of $a_1, \ldots, a_n, b$, the corresponding mean squared error is

$$\mathbf{E}\big[(X - a_1 Y_1 - \cdots - a_n Y_n - b)^2\big],$$

and it is meaningful to choose the coefficients $a_1, \ldots, a_n, b$ in a way that minimizes the above expression. This problem is relatively easy to solve and only requires knowledge of the means, variances, and covariances of the different random variables. We develop the solution for the case where $n = 1$.

Linear Least Mean Squares Estimation Based on a Single Measurement

We are interested in finding a and b that minimize the mean squared estimation error $\mathbf{E}\big[(X - aY - b)^2\big]$, associated with a linear estimator $aY + b$ of X. Suppose that a has already been chosen. How should we choose b? This is the same as having to choose a constant b to estimate the random variable $aX - Y$ and, by our earlier results, the best choice is to let $b = \mathbf{E}[X - aY] = \mathbf{E}[X] - a\mathbf{E}[Y]$.

It now remains to minimize, with respect to a, the expression

$$\mathbf{E}\Big[\big(X - aY - \mathbf{E}[X] + a\mathbf{E}[Y]\big)^2\Big],$$

which is the same as

$$\begin{aligned}
&\mathbf{E}\Big[\big((X - \mathbf{E}[X]) - a(Y - \mathbf{E}[Y])\big)^2\Big] \\
&\quad = \mathbf{E}\big[(X - \mathbf{E}[X])^2\big] + a^2\mathbf{E}\big[(Y - \mathbf{E}[Y])^2\big] - 2a\mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] \\
&\quad = \sigma_X^2 + a^2\sigma_Y^2 - 2a \cdot \mathrm{cov}(X, Y),
\end{aligned}$$

where $\text{cov}(X, Y)$ is the covariance of X and Y:

$$\text{cov}(X,Y) = \mathbf{E}\big[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\big].$$

This is a quadratic function of a, which is minimized at the point where its derivative is zero, that is, if

$$a = \frac{\text{cov}(X,Y)}{\sigma_Y^2} = \frac{\rho\sigma_X\sigma_Y}{\sigma_Y^2} = \rho\frac{\sigma_X}{\sigma_Y},$$

where

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}$$

is the correlation coefficient. With this choice of a, the mean squared estimation error is given by

$$\begin{aligned}\sigma_X^2 + a^2\sigma_Y^2 - 2a\cdot\text{cov}(X,Y) =& \sigma_X^2 + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}\sigma_Y^2 - 2\rho\frac{\sigma_X}{\sigma_y}\rho\sigma_X\sigma_Y \\ =& (1-\rho^2)\sigma_X^2.\end{aligned}$$

Linear Least Mean Squares Estimation Formulas

The least mean squares linear estimator of X based on Y is

$$\mathbf{E}[X] + \frac{\text{cov}(X,Y)}{\sigma_Y^2}\big(Y - \mathbf{E}[Y]\big).$$

The resulting mean squared estimation error is equal to

$$(1-\rho^2)\text{var}(X).$$

4.7 THE BIVARIATE NORMAL DISTRIBUTION

We say that two random variables X and Y have a *bivariate normal* distribution if there are two independent normal random variables U and V and some scalars a, b, c, d, such that

$$X = aU + bV, \qquad Y = cU + dV.$$

To keep the discussion simple, we restrict ourselves to the case where U, V (and therefore, X and Y as well) have zero mean.

A most important property of the bivariate normal distribution is the following:

If two random variables X and Y have a bivariate normal distribution and are uncorrelated, then they are independent.

This property can be verified using multivariate transforms. We assume that X and Y have a bivariate normal distribution and are uncorrelated. Recall that if z is a zero-mean normal random variable with variance σ_Z^2, then $\mathbf{E}[e^Z] = M_Z(1) = \sigma_Z^2/2$. Fix some scalars s_1, s_2 and let $Z = s_1X + s_2Y$. Then, Z is the sum of the independent normal random variables $(as_1 + cs_2)U$ and $(bs_1 + ds_2)V$, and is therefore normal. Since X and Y are uncorrelated, the variance of Z is $s_1^2\sigma_X^2 + s_2^2\sigma_Y^2$. Then,

$$\begin{aligned} M_{X,Y}(s_1, s_2) &= \mathbf{E}\left[e^{s_1X+s_2Y}\right] \\ &= \mathbf{E}[e^Z] \\ &= e^{(s_1^2\sigma_X^2+s_2^2\sigma_Y^2)/2.} \end{aligned}$$

Let $\overline{X}$ and $\overline{Y}$ be *independent* zero-mean normal random variables with the same variances σ_X^2 and σ_Y^2 as X and Y. Since they are independent, they are uncorrelated, and the same argument as above yields

$$M_{\overline{X},\overline{Y}}(s_1, s_2) = e^{(s_1^2\sigma_X^2+s_2^2\sigma_Y^2)/2.}.$$

Thus, the two pairs of random variables (X, Y) and $(\overline{X}, \overline{Y})$ are associated with the same multivariate transform. Since the multivariate transform completely determines the joint PDF, it follows that the pair (X, Y) has the same joint PDF as the pair $(\overline{X}, \overline{Y})$. Since $\overline{X}$ and $\overline{Y}$ are independent, X and Y must also be independent.

Let us define

$$\hat{X} = \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}Y, \qquad \tilde{X} = X - \hat{X}.$$

Thus, $\hat{X}$ is the best *linear* estimator of X given Y, and $\tilde{X}$ is the estimation error. Since X and Y are linear combinations of independent normal random variables U and V, it follows that Y and $\tilde{X}$ are also linear combinations of U and V. In particular, Y and $\tilde{X}$ have a bivariate normal distribution. Furthermore,

$$\mathrm{cov}(Y, \tilde{X}) = \mathbf{E}[Y\tilde{X}] = \mathbf{E}[YX] - \mathbf{E}[Y\hat{X}] = \mathbf{E}[YX] - \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}\mathbf{E}[Y^2] = 0.$$

Thus, Y and $\tilde{X}$ are uncorrelated and, therefore, independent. Since $\hat{X}$ is a scalar multiple of Y, we also see that $\hat{X}$ and $\tilde{X}$ are independent.

We now start from the identity

$$X = \hat{X} + \tilde{X},$$

which implies that

$$\mathbf{E}[X \,|\, Y] = \mathbf{E}[\hat{X} \,|\, Y] + \mathbf{E}[\tilde{X} \,|\, Y].$$

But $\mathbf{E}[\hat{X} \,|\, Y] = \hat{X}$ because $\hat{X}$ is completely determined by Y. Also, $\tilde{X}$ is independent of Y and

$$\mathbf{E}[\tilde{X} \,|\, Y] = \mathbf{E}[\tilde{X}] = \mathbf{E}[X - \hat{X}] = 0.$$

(The last equality was obtained because X and Y are assumed to have zero mean and $\hat{X}$ is a constant multiple of Y.) Putting everything together, we come to the important conclusion that the best linear estimator $\hat{X}$ is of the form

$$\hat{X} = \mathbf{E}[X \,|\, Y].$$

Differently said, the optimal estimator $\mathbf{E}[X \,|\, Y]$ turns out to be linear.

Let us now determine the conditional density of X, conditioned on Y. We have $X = \hat{X} + \tilde{X}$. After conditioning on Y, the value of the random variable $\hat{X}$ is completely determined. On the other hand, $\tilde{X}$ is independent of Y and its distribution is not affected by conditioning. Therefore, the conditional distribution of X given Y is the same as the distribution of $\tilde{X}$, shifted by $\hat{X}$. Since $\tilde{X}$ is normal with mean zero and some variance $\sigma^2_{\tilde{X}}$, we conclude that the conditional distribution of X is also normal with mean $\hat{X}$ and variance $\sigma^2_{\tilde{X}}$.

We summarize our conclusions below. Although our discussion used the zero-mean assumption, these conclusions also hold for the non-zero mean case and we state them with this added generality.

Properties of the Bivariate Normal Distribution

Let X and Y have a bivariate normal distribution. Then:

- X and Y are independent if and only if they are uncorrelated.
- The conditional expectation is given by

$$\mathbf{E}[X \,|\, Y] = \mathbf{E}[X] + \frac{\mathrm{cov}(X, Y)}{\sigma^2_Y}(Y - \mathbf{E}[Y]).$$

  It is a linear function of Y and has a normal distribution.

- The conditional distribution of X given Y is normal with mean $\mathbf{E}[X \,|\, Y]$ and variance
$$\sigma_{\tilde{X}}^2 = (1 - \rho^2)\sigma_X^2.$$

Finally, let us note that while if X and Y have a bivariate normal distribution, then X and Y are (individually) normal random variables, the reverse is not true even if X and Y are uncorrelated. This is illustrated in the following example.

Example 4.29. Let X have a normal distribution with zero mean and unit variance. Let z be independent of X, with $\mathbf{P}(Z = 1) = \mathbf{P}(Z = -1) = 1/2$. Let $Y = ZX$, which is also normal with zero mean (why?). Furthermore,

$$\mathbf{E}[XY] = \mathbf{E}[ZX^2] = \mathbf{E}[Z]\mathbf{E}[X^2] = 0 \times 1 = 0,$$

so X and Y are uncorrelated. On the other hand X and Y are clearly dependent. (For example, if $X = 1$, then Y must be either -1 or 1.) This may seem to contradict our earlier conclusion that zero correlation implies independence? However, in this example, the joint PDF of X and Y is *not* multivariable normal, even though both marginal distributions are normal.

SOLVED PROBLEMS

SECTION 4.1. Transforms

Problem 1. Let X be a random variable that takes the values 1, 2, and 3 with the following probabilities:

$$\mathbf{P}(X=1)=\frac{1}{2},\qquad \mathbf{P}(X=2)=\frac{1}{4},\qquad \mathbf{P}(X=3)=\frac{1}{4}.$$

Find the transform of X and use it to obtain the first three moments, $\mathbf{E}[X]$, $\mathbf{E}[X^2]$, $\mathbf{E}[X^3]$.

Solution. The transform is given by

$$M(s)=\mathbf{E}[e^{sX}]=\frac{1}{2}e^s+\frac{1}{4}e^{2s}+\frac{1}{4}e^{3s}.$$

We have

$$\mathbf{E}[X]=\frac{d}{ds}M(s)\Big|_{s=0}=\frac{1}{2}+\frac{2}{4}+\frac{3}{4}=\frac{7}{4},$$

$$\mathbf{E}[X^2]=\frac{d^2}{ds^2}M(s)\Big|_{s=0}=\frac{1}{2}+\frac{4}{4}+\frac{27}{4}=\frac{15}{4},$$

$$\mathbf{E}[X^3]=\frac{d^3}{ds^3}M(s)\Big|_{s=0}=\frac{1}{2}+\frac{8}{4}+\frac{27}{4}=\frac{37}{4}.$$

Problem 2. A nonnegative integer-valued random variable X has one of the following two expressions as its transform:

1. $M(s)=e^{2(e^{e^s-1}-1)}$.
2. $M(s)=e^{2(e^{e^s}-1)}$.

(a) Explain why one of the two cannot possibly be the transform, and indicate which one is the true transform.

(b) Find $\mathbf{P}(X=0)$.

Solution. (a) By the definition of the transform,

$$M(s)=\mathbf{P}(X=0)+e^s\mathbf{P}(X=1)+e^{2s}\mathbf{P}(X=2)+\cdots$$

so when evaluated at $s=0$, the transform should equal 1. Only the first option satisfies this requirement.

(b) By the above expansion, we see that if we take the limit of the transform as $s\to-\infty$, we obtain $\mathbf{P}(X=0)$. Thus

$$\mathbf{P}(X=0)=\lim_{s\to-\infty}M(s)=e^{2(e^{-1}-1)}\approx 0.2825.$$

Problem 3. Find the PDF of the continuous random variable X that has the transform

$$M(s) = \frac{1}{3} \cdot \frac{2}{2-s} + \frac{2}{3} \cdot \frac{3}{3-s}.$$

Solution. We recognize this transform as corresponding to the following mixture of exponential PDFs:

$$f_X(x) = \begin{cases} \frac{1}{3} \cdot 2e^{-2x} + \frac{2}{3} \cdot 3e^{-3x} & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

By the inversion theorem, this must be the desired PDF.

Problem 4. Let X be a random variable that takes nonnegative integer values, and has a transform of the form

$$M_X(s) = K \frac{14 + 5e^s - 3e^{2s}}{8(2 - e^s)},$$

where K is some scalar. Find $\mathbf{E}[X]$, $p_X(1)$, and $\mathbf{E}[X \mid X \neq 0]$.

Solution. We find K by using the equation $M_X(0) = 1$:

$$K = \left(\frac{8(2-e^s)}{14 + 5e^s - 3e^{2s}} \right) \bigg|_{s=0} = \frac{8}{16} = \frac{1}{2}.$$

We have

$$\begin{aligned}
\mathbf{E}[X] &= \frac{dM_X(s)}{ds}\bigg|_{s=0} \\
&= \frac{K}{8}\Big((5e^s - 6e^{2s})(2-e^s)^{-1} + (-1)(-1)e^s(2-e^s)^{-2}(14 + 5e^s - 3e^{2s})\Big)\bigg|_{s=0} \\
&= \frac{K}{8}\big((-1)(+1) + (1)(14+5-3)\big) \\
&= \frac{15K}{8} \\
&= \frac{15(1/2)}{8} \\
&= \frac{15}{16}.
\end{aligned}$$

Since X takes nonnegative integer values, we have

$$M_X(s) = p_X(0) \cdot e^0 + p_X(1) \cdot e^s + p_X(2) \cdot e^{2s} + p_X(3) \cdot e^{3s} + \cdots$$

$$\frac{dM_X(s)}{ds} = p_X(1) \cdot e^s + 2p_X(2) \cdot e^{2s} + 3p_X(3) \cdot e^{3s} + \cdots$$

$$\frac{dM_X(s)}{ds} \cdot \frac{1}{e^s} = p_X(1) + 2p_X(2) \cdot e^s + 3p_X(3) \cdot e^{2s} + \cdots$$

Setting $e^s = 0$ eliminates all but the first term on the right hand side, so we have

$$\begin{aligned}
p_X(1) &= \left(\frac{dM_X(s)}{ds} \cdot \frac{1}{e^s} \right) \bigg|_{e^s=0} \\
&= \frac{1}{16}\big((5-6e^s)(2-e^s)^{-1} + (-1)(-1)(2-e^s)^{-2}(14+5e^s-3e^{2s}) \big) \Big|_{e^s=0} \\
&= \frac{3}{8}.
\end{aligned}$$

Let $A = \{X \neq 0\}$. We have

$$p_{X\,|\,A}(x\,|\,A) = \begin{cases} \frac{p_X(x)}{\mathbf{P}(A)} & \text{if } x \in A, \\ 0 & \text{otherwise,} \end{cases}$$

so that

$$\begin{aligned}
\mathbf{E}[X\,|\,X \neq 0] &= \sum_{x=1}^{\infty} x \cdot p_{X\,|\,A}(x\,|\,A) \\
&= \sum_{x=1}^{\infty} \frac{x \cdot p_X(x)}{\mathbf{P}(A)} \\
&= \frac{\mathbf{E}[X]}{\mathbf{P}(A)} \\
&= \frac{15/16}{\mathbf{P}(A)}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathbf{P}(A) &= 1 - \mathbf{P}\{X = 0\} \\
&= 1 - M_X(s)\Big|_{e^s=0} \\
&= 1 - \left(\frac{14}{16}\right)\left(\frac{1}{2}\right) \\
&= 1 - \frac{7}{16} \\
&= \frac{9}{16}.
\end{aligned}$$

Therefore,

$$\mathbf{E}[X\,|\,X \neq 0] = \frac{15/16}{9/16} = \frac{5}{3}.$$

Problem 5. Let X, Y, and Z be independent random variables. X is Bernoulli with parameter $1/3$. Y is exponential with parameter 2. Z is Poisson with parameter 3.

(a) Consider the new random variable $U = XY + (1-X)Z$. Find the transform associated with U.

(b) Find the transform of $2Z + 3$.

(c) Find the transform of $Y + Z$.

Solution. (a) We have $U = X$ if $X = 1$, which happens with probability 1/3, and $U = Z$ if $X = 0$, which happens with probability 2/3. Therefore, U is mixture of random variables and its transform is

$$M_U(s) = \mathbf{P}(X = 1)M_Y(s) + \mathbf{P}(X = 0)M_Z(s) = \frac{1}{3}\frac{2}{2-s} + \frac{2}{3}e^{3(e^s-1)}.$$

(b) Let $V = 2Z + 3$. We have

$$M_V(s) = e^{3s}M_Z(2s) = e^{3s}e^{3(e^{2s}-1)} = e^{3(s-1+e^{2s})}.$$

(c) Let $W = Y + Z$. We have

$$M_W(s) = M_Y(s)M_Z(s) = \frac{2}{2-s}e^{3(e^s-1)}.$$

Problem 6. * Let X be a discrete random variable taking nonnegative integer values. Let $M(s)$ be the transform of X.

(a) Show that
$$\mathbf{P}(X = 0) = \lim_{s\to-\infty} M(s).$$

(b) Use part (a) to verify that if X is a binomial random variable with parameters n and p, we have $\mathbf{P}(X = 0) = (1 - p)^n$. Furthermore, if X is a Poisson random variable with parameter λ, we have $\mathbf{P}(X = 0) = e^{-\lambda}$.

(c) Suppose that X is instead known to take only integer values that are greater or equal to a given integer $\overline{k}$. How can we calculate $P(X = \overline{k})$ using the transform of X?

Solution. (a) We have
$$M(s) = \sum_{k=0}^{\infty} \mathbf{P}(X = k)e^{ks}.$$

As $s \to -\infty$, all the terms e^{ks} with $k > 0$ tend to 0, so we obtain $\lim_{s\to-\infty} M(s) = \mathbf{P}(X = 0)$.

(b) In the case of the binomial, we have from the transform tables

$$M(s) = (1 - p + pe^s)^n,$$

so that $\lim_{s\to-\infty} M(s) = (1 - p)^n$. In the case of the Poisson, we have

$$M(s) = e^{\lambda(e^s-1)},$$

so that $\lim_{s\to-\infty} M(s) = e^{-\lambda}$.

(c) The random variable $Y = X - \overline{k}$ takes only nonnegative integer values and has transform $M_Y(s) = e^{-s\overline{k}}M(s)$ (cf. Example 4.4). Since $\mathbf{P}(Y = 0) = \mathbf{P}(X = \overline{k})$, we have from part (a)
$$\mathbf{P}(X = \overline{k}) = \lim_{s\to-\infty} e^{-s\overline{k}}M(s).$$

Problem 7. * Transform of the discrete uniform. Find the transform of the integer-valued random variable X that is uniformly distributed in the range $[M, M+N]$.

Solution. The PMF of X is

$$p_X(k) = \begin{cases} \frac{1}{N+1} & \text{if } k = M, M+1, \ldots, M+N, \\ 0 & \text{otherwise.} \end{cases}$$

The transform is

$$\begin{aligned} M(s) &= \sum_{k=-\infty}^{\infty} e^{sk}\mathbf{P}(X = k) \\ &= \sum_{k=M}^{M+N} \frac{1}{N+1} e^{sk} \\ &= \frac{e^{sM}}{N+1} \sum_{k=0}^{N} e^{sk} \\ &= \frac{e^{sM}}{N+1} \cdot \frac{1 - e^{s(N+1)}}{1 - e^{s}}. \end{aligned}$$

Problem 8. * Transform of the continuous uniform. Let X be a continuous random variable that is uniformly distributed between a and b.

(a) Find the transform of X.

(b) Use the transform in (a) to find the mean and the variance of X.

Solution. (a) The transform is given by

$$M(s) = \mathbf{E}[e^{sX}] = \int_a^b \frac{e^{sx}}{b-a} dx = \frac{e^{sb} - e^{sa}}{s(b-a)}.$$

(b) We have

$$\mathbf{E}[X^n] = \frac{d^n}{ds^n} M(s)\Big|_{s=0}.$$

Thus for $n = 1$ we have using L' Hopital's rule

$$\begin{aligned} \mathbf{E}[X] &= \frac{d}{ds} M(s)\Big|_{s=0} \\ &= \left\{ -\left(\frac{1}{b-a}\right) \frac{e^{sb} - e^{sa}}{s^2} + \left(\frac{1}{b-a}\right) \frac{be^{sb} - ae^{sa}}{s} \right\}\Bigg|_{s=0} \\ &= -\frac{b^2 - a^2}{2(b-a)} + \frac{b^2 - a^2}{b-a} \\ &= \frac{b+a}{2}. \end{aligned}$$

To find var(X), we calculate $\mathbf{E}[X^2]$ using the second derivative of the transform. We have using L' Hopital's rule

$$\begin{aligned}\mathbf{E}[X^2] &= \frac{d^2}{ds^2}M(s)\bigg|_{s=0} \\ &= \left\{\left(\frac{2}{b-a}\right)\frac{e^{sb}-e^{sa}}{s^3} - \left(\frac{2}{b-a}\right)\frac{be^{sb}-ae^{sa}}{s^2} + \left(\frac{1}{b-a}\right)\frac{b^2e^{sb}-a^2e^{sa}}{s}\right\}\Bigg|_{s=0} \\ &= \frac{1}{3}\frac{b^3-a^3}{b-a} + \frac{a^3-b^3}{b-a} + \frac{b^3-a^3}{b-a} \\ &= \frac{1}{3}(b^2+ab+a^2).\end{aligned}$$

Therefore

$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{1}{3}(b^2+ab+a^2) - \left(\frac{b+a}{2}\right)^2.$$

Problem 9. * Mean and variance of the Poisson. Use the formula for the transform of a Poisson random variable X to calculate $E[X]$ and $E[X^2]$.

Solution. Let λ be the parameter of the Poisson random variable X. Its transform is given by

$$M_X(s) = e^{\lambda(e^s-1)}, \; s < \lambda.$$

Using the equation

$$\frac{d^n}{ds^n}M(s)\bigg|_{s=0} = E[X^n],$$

we obtain

$$E[X] = (\lambda e^s e^{\lambda(e^s-1)})\bigg|_{s=0} = \lambda \cdot 1 \cdot e^{\lambda(1-1)} = \lambda,$$

$$E[X^2] = \left(\lambda(e^s e^{\lambda(e^s-1)} + e^s \cdot \lambda e^s \cdot e^{\lambda(e^s-1)})\right)\bigg|_{s=0} = \lambda + \lambda^2.$$

Problem 10. * Find the third, fourth, and fifth moments of the exponential random variable with parameter λ.

Solution. The transform is

$$M(s) = \frac{\lambda}{\lambda - s}.$$

Thus,

$$\frac{d}{ds}M(s) = \frac{\lambda}{(\lambda-s)^2}, \qquad \frac{d^2}{ds^2}M(s) = \frac{2\lambda}{(\lambda-s)^3}, \qquad \frac{d^3}{ds^3}M(s) = \frac{6\lambda}{(\lambda-s)^4},$$

$$\frac{d^4}{ds^4}M(s) = \frac{24\lambda}{(\lambda-s)^5}, \qquad \frac{d^5}{ds^5}M(s) = \frac{120\lambda}{(\lambda-s)^6}.$$

By setting $s = 0$, we obtain

$$\mathbf{E}[X^3] = \frac{6}{\lambda^3}, \qquad \mathbf{E}[X^4] = \frac{24}{\lambda^4}, \qquad \mathbf{E}[X^5] = \frac{120}{\lambda^5}.$$

Problem 11. * Using partial fraction expansions. Suppose that the transform of the random variable X has the form

$$M(s) = \frac{A(e^s)}{B(e^s)},$$

where $A(t)$ and $B(t)$ are polynomials of the generic variable t. Assume that $A(t)$ and $B(t)$ have no common roots and the degree of $A(t)$ is smaller than the degree of $B(t)$. Assume also that $B(t)$ has distinct, real, and nonzero roots. Then it can be seen that $M(s)$ can be written in the form

$$M(s) = \frac{a_1}{1 - r_1 e^s} + \cdots + \frac{a_m}{1 - r_m e^s},$$

where $1/r_1, \ldots, 1/r_m$ are the roots of $B(t)$ and the a_i are constants that are equal to $\lim_{e^s \to \frac{1}{r_i}} (1 - r_i e^s) M(s)$, $i = 1, \ldots, m$.

(a) Show that the PMF of X has the form

$$\mathbf{P}(X = k) = \begin{cases} \sum_{i=1}^m a_i r_i^k & \text{if } k = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Note: For large k, the PMF of X can be approximated by $a_{\bar{i}} r_{\bar{i}}^k$, where $\bar{i}$ is the index corresponding to the largest $|r_i|$ (assuming $\bar{i}$ is unique).

(b) Extend the result of part (a) to the case where $M(s) = e^{bs} A(t)/B(t)$ and b is an integer.

Solution. (a) We have for all s such that $|r_i| e^s < 1$

$$\frac{1}{1 - r_i e^s} = 1 + r_i e^s + r_i^2 e^{2s} + \cdots$$

Therefore

$$M(s) = \sum_{i=1}^m a_i + \left(\sum_{i=1}^m a_i r_i \right) e^s + \left(\sum_{i=1}^m a_i r_i^2 \right) e^{2s} + \cdots$$

and by inverting this transform, we see that

$$\mathbf{P}(X = k) = \sum_{i=1}^m a_i r_i^k$$

for $k \geq 0$ and $\mathbf{P}(X = k) = 0$ for $k < 0$.

(b) In this case $M(s)$ corresponds to the translation by b of the random variable whose transform is $A(t)/B(t)$ (cf. Example 4.4), so we have

$$\mathbf{P}(X = k) = \begin{cases} \sum_{i=1}^{m} a_i r_i^{(k-b)} & \text{if } k = b, b+1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

SECTION 4.2. Sums of Independent Random Variables - Convolutions

Problem 12. A soccer team has three designated players who take turns striking penalty shots. The probability of success of the ith player is a given p_i, independently of the success of the other players. Let X be the number of successful penalty shots after each player has had one turn. Use convolution to calculate the PMF of X. Confirm your answer by first calculating the transform of X and then obtaining the PMF from the transform.

Solution. Let X_i, $i = 1, 2, 3$, be the Bernoulli random variable that take the value 1 if the ith player is successful. We have $X = X_1 + X_2 + X_3$. Let $q_i = 1 - p_i$. Convolution of X_1 and X_2 yields the PMF of $W = X_1 + X_2$:

$$p_W(w) = \begin{cases} q_1 q_2 & \text{if } w = 0, \\ q_1 p_2 + p_1 q_2 & \text{if } w = 1, \\ p_1 p_2 & \text{if } w = 2, \\ 0 & \text{otherwise.} \end{cases}$$

Convolution of W and X_3 yields the PMF of $X = X_1 + X_2 + X_3$:

$$p_X(x) = \begin{cases} q_1 q_2 q_3 & \text{if } w = 0, \\ p_1 q_2 q_3 + q_1 p_2 q_3 + q_1 q_2 p_3 & \text{if } w = 1, \\ q_1 p_2 p_3 + p_1 q_2 p_3 + p_1 p_2 q_3 & \text{if } w = 2, \\ p_1 p_2 p_3 & \text{if } w = 3, \\ 0 & \text{otherwise.} \end{cases}$$

The transform of X is the product of the transforms of X_i, $i = 1, 2, 3$. We have

$$M_X(s) = (q_1 + p_1 e^s)(q_2 + p_2 e^s)(q_3 + p_3 e^s).$$

By carrying out the multiplications above, and by examining the coefficients of the terms e^{ks}, we obtain the probabilities $\mathbf{P}(X = k)$. These probabilities are seen to coincide with the ones computed by convolution.

Problem 13. Harry and Larry approach each other from very far away. Harry will see Larry at a distance that is exponentially distributed with parameter 1 km, while Larry will see Harry at a distance that is uniformly distributed between 0 and 1 km. Use convolution to find the PDF of X, the distance during which only one of the two persons will be seeing the other.

Solution. Let Y and Z be the distances at which Harry will first see Larry, and Larry will first see Harry, respectively. We have $X = |W|$, where $W = Y - Z$. We find the

PDF of W by convolution of the exponential with parameter 1 with the uniform in the interval $[-1, 0]$. Then we note that

$$f_X(x) = \begin{cases} f_W(x) + f_W(-x) & \text{if } x \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 14. The random variables X, Y, and Z are independent and uniformly distributed between zero and one. Find the PDF of $W = X + Y + Z$.

Solution. Let $V = X + Y$. As in Example 4.14, the PDF of V is

$$f_V(v) = \begin{cases} v & 0 \le v \le 1, \\ 2 - v & 1 \le v \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

We have $W = V + Z$. By convolution of the PDFs f_V and f_Z, we obtain

$$f_W(w) = \begin{cases} w^2/2 & 0 \le w \le 1, \\ 1 - (w-1)^2/2 - (2-w)^2/2 & 1 \le w \le 2, \\ (3-w)^2/2 & 2 \le w \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 15. Consider a PDF that is positive only within an interval $[a, b]$ and is symmetric around the mean $(a + b)/2$. Let X and Y be independent random variables that both have this PDF. Suppose that you have calculated the PDF and the transform of $X + Y$. How can you easily obtain the PDF and the transform of $X - Y$?

Solution. We have $X - Y = X + Z - (a + b)$, where $Z = a + b - Y$ is distributed identically with X and Y.

SECTION 4.3. Conditional Expectation as a Random Variable

Problem 16. Oscar is an engineer who is equally likely to work between zero and one hundred hours each week (i.e., the time he works is uniformly distributed between zero and one hundred). He gets paid one dollar an hour. If Oscar works more than fifty hours during a week, there is a probability of 1/2 that he will actually be paid overtime, which means he will receive two dollars an hour for each hour he works longer than fifty hours. Otherwise, he will just get his normal pay for all of his hours that week. Independently of receiving overtime pay, if Oscar works more than seventy five hours in a week, there is a probability of 1/2 that he will receive a one hundred dollar bonus, in addition to whatever else he earns. Find the expected value and variance of Oscar's weekly salary.

Solution. Let W be the number of hours Oscar works in a week, and T be the total amount of Oscar's earnings in a week (including overtime and bonus). Let also O be the event that Oscar receives overtime pay, and B be the event that Oscar receives a bonus. We will find $\mathbf{E}[T]$ and $\text{var}(T)$ by using iterated expectations.

We have

$$\mathbf{E}[T] = \sum_{i=1}^{7} \mathbf{P}(A_i)\mathbf{E}[T \mid A_i]$$

where $A_1, \ldots, A_7$ are the events

$$\begin{aligned}
A_1 &= \{0 \le w \le 50\},\\
A_2 &= \{50 < w \le 75\} \cap O',\\
A_3 &= \{50 < w \le 75\} \cap O,\\
A_4 &= \{75 < w \le 100\} \cap O' \cap B,\\
A_5 &= \{75 < w \le 100\} \cap O' \cap B',\\
A_6 &= \{75 < w \le 100\} \cap O \cap B,\\
A_7 &= \{75 < w \le 100\} \cap O \cap B'.
\end{aligned}$$

Note that $\{A_1, A_2, ..., A_7\}$ form a partition of the sample space.

Over each A_i, the conditional PDF $f_{T\,|\,A_i}(t\,|\,A_i)$ is constant because any linear function $aX + b$ of a uniformly distributed RV X is also uniformly distributed. Therefore,

$$\begin{aligned}
f_{T\,|\,A_1}(t\,|\,A_1) &= \frac{1}{50} && \text{for } 0 \le t \le 50,\\
f_{T\,|\,A_2}(t\,|\,A_2) &= \frac{1}{25} && \text{for } 50 < t \le 75,\\
f_{T\,|\,A_3}(t\,|\,A_3) &= \frac{1}{50} && \text{for } 50 < t \le 100,\\
f_{T\,|\,A_4}(t\,|\,A_4) &= \frac{1}{25} && \text{for } 175 < t \le 200,\\
f_{T\,|\,A_5}(t\,|\,A_5) &= \frac{1}{25} && \text{for } 75 < t \le 100,\\
f_{T\,|\,A_6}(t\,|\,A_6) &= \frac{1}{50} && \text{for } 200 < t \le 250,\\
f_{T\,|\,A_7}(t\,|\,A_7) &= \frac{1}{50} && \text{for } 100 < t \le 150.
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E}[T\,|\,A_1] &= 25,\\
\mathbf{E}[T\,|\,A_2] &= \frac{125}{2},\\
\mathbf{E}[T\,|\,A_3] &= 75,\\
\mathbf{E}[T\,|\,A_4] &= \frac{375}{2},\\
\mathbf{E}[T\,|\,A_5] &= \frac{175}{2},\\
\mathbf{E}[T\,|\,A_6] &= 225,\\
\mathbf{E}[T\,|\,A_7] &= 125.
\end{aligned}$$

The expected salary per week is then equal to

$$\mathbf{E}[T] = \frac{1}{2}\cdot 25 + \frac{1}{8}\cdot\frac{125}{2} + \frac{1}{8}\cdot 75 + \frac{1}{16}\cdot\frac{375}{2} + \frac{1}{16}\cdot\frac{175}{2} + \frac{1}{16}\cdot 225 + \frac{1}{16}\cdot 125 = 68.75.$$

We use a similar argument to find $\mathbf{E}[T^2]$. We have

$$\mathbf{E}[T^2] = \sum_{i=1}^{7} \mathbf{P}(A_i)\mathbf{E}[T^2\,|\,A_i].$$

Using the fact that $\mathbf{E}[X^2] = (a^2 + ab + b^2)/3 = ((a+b)^2 - ab)/3$ for any uniformly distributed RV X ranging from a to b, we obtain

$$\begin{aligned}
\mathbf{E}[T^2 \,|\, A_1] &= 50^2/3, \\
\mathbf{E}[T^2 \,|\, A_2] &= (125^2 - 50 \cdot 75)/3, \\
\mathbf{E}[T^2 \,|\, A_3] &= (150^2 - 50 \cdot 100)/3, \\
\mathbf{E}[T^2 \,|\, A_4] &= (375^2 - 175 \cdot 200)/3, \\
\mathbf{E}[T^2 \,|\, A_5] &= (175^2 - 75 \cdot 100)/3, \\
\mathbf{E}[T^2 \,|\, A_6] &= (450^2 - 200 \cdot 250)/3, \\
\mathbf{E}[T^2 \,|\, A_7] &= (250^2 - 100 \cdot 150)/3.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{E}[T^2] &= \frac{1}{2} \cdot \frac{2500}{3} + \frac{1}{8} \cdot \frac{11875}{3} + \frac{1}{8} \cdot \frac{17500}{3} + \frac{1}{16} \cdot \frac{105625}{3} \\
&\quad + \frac{1}{16} \cdot \frac{23125}{3} + \frac{1}{16} \cdot \frac{152500}{3} + \frac{1}{16} \cdot \frac{47500}{3} \\
&= \frac{101875}{12}.
\end{aligned}$$

and

$$\mathrm{var}(T) = \mathbf{E}[T^2] - (\mathbf{E}[T])^2 = \frac{180625}{48} \approx 3763.$$

Problem 17. Pat and Nat are dating, and all of their dates are scheduled to start at 9pm. Nat always arrives promptly at 9pm. Pat is highly disorganized and arrives at a time that is uniformly distributed between 8pm and 10pm. Let X be the time in hours between 8pm and the time when Pat arrives. If Pat arrives after 9pm, their date will last exactly 3 hours. If Pat arrives after 9pm, their date will last for a time that is uniformly distributed between 0 and $3 - X$ hours. The date starts at the time they meet. Nat gets irritated when Pat is late and will the relationship after the second date on which Pat is late by more than 45 minutes. All dates are independent of any other dates.

(a) What is the expected number of hours Pat waits for Nat to arrive? (Note: The waiting time must be nonnegative.)

(b) What is the expected duration of any particular date?

(c) What is the expected number of dates they will have before breaking up?

Solution. (a) Let W be the number of hours that Pat waits. We have

$$\mathbf{E}[X] = \mathbf{P}(0 \le X \le 1)\mathbf{E}[W \,|\, 0 \le X \le 1] + \mathbf{P}(X > 1)\mathbf{E}[W \,|\, X > 1].$$

Since $W > 0$ only if $X > 1$, we have

$$E[W] = \mathbf{P}(X > 1)\mathbf{E}[W \,|\, X > 1] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

(b) Let D be the duration of a date. We have $\mathbf{E}[D \,|\, 0 \leq X \leq 1] = 3$ and by using iterated expectations,

$$\mathbf{E}[D \,|\, X > 1] = \mathbf{E}\Big[\mathbf{E}[D \,|\, X, X > 1] \,|\, X > 1\Big] = \mathbf{E}\left[\frac{3-X}{2} \,|\, X > 1\right].$$

Hence

$$\begin{aligned} \mathbf{E}[D] &= \mathbf{P}(0 \leq X \leq 1)\mathbf{E}[D \,|\, 0 \leq X \leq 1] + \mathbf{P}(X > 1)\mathbf{E}[D \,|\, X > 1] \\ &= \frac{1}{2}\cdot 3 + \frac{1}{2}\cdot \mathbf{E}\left[\frac{3-X}{2} \,|\, X > 1\right] \\ &= \frac{3}{2} + \frac{1}{2}\left(\frac{3}{2} - \frac{\mathbf{E}[X \,|\, X > 1]}{2}\right) \\ &= \frac{3}{2} + \frac{1}{2}\left(\frac{3}{4} - \frac{3/2}{2}\right) \\ &= \frac{15}{8}. \end{aligned}$$

(c) The probability that Pat will be late by more than 45 minutes is 1/8. The number of dates before breaking up is the sum of two geometrically distributed random variables with parameter 1/, and its expected value is $2 \cdot 8 = 16$.

Problem 18. A retired professor comes to his office at any time between 9 AM and 1 PM, with all times in that interval being equally likely, and performs a single task. The duration of the task is exponentially distributed with parameter $\lambda(y) = 1/(5-y)$, where y is the length of the time interval between 9 AM and the time of his arrival.

(a) What is the expected time that the professor devotes to his task?

(b) What is the expected time that the professor leaves his office?

(c) The professor has a PhD student who on a given day comes to see him at a time that is uniformly distributed between 9 AM and 5 PM. If the student does not find the professor, he leaves and does not return. If he finds the professor, he spends an amount of time that is uniformly distributed between 0 and 1 hour. The professor will spend the same amount of total time on his task regardless of whether he is interrupted by the student. What is the expected amount of time that the professor will spend with the student and what is the expected time that he will leave his office?

Solution. (a) Consider the following two random variables:

$X =$ amount of time the professor devotes to his task [exponentially distributed with parameter $\lambda = 1/(5-y)$]

$Y =$ length of time between 9 AM and his arrival (uniformly distributed between 0 and 4)

Since the random variable X depends on the value y of Y, we have

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

We have

$$\mathbf{E}[X \,|\, Y = y] = \frac{1}{\lambda} = 5 - y.$$

Also, the PDF for the random variable Y is

$$f_Y(y) = \begin{cases} 1/4 & \text{if } 0 \le y \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\mathbf{E}[X] = \int_0^4 \frac{1}{4}(5-y)dy = 3 \text{ hours.}$$

(b) Let Z be the length of time from 9 AM until the professor leaves his office. Then

$$Z = X + Y.$$

So,

$$\mathbf{E}[Z] = \mathbf{E}[X] + \mathbf{E}[Y].$$

We already know $\mathbf{E}[X]$ from part (a). Since the random variable Y is uniformly distributed between 0 and 4, $\mathbf{E}[Y] = 2$. Therefore,

$$\mathbf{E}[Z] = 3 + 2 = 5.$$

Thus the expected time that the professor leaves his office is 5 hours after 9 AM.

(c) We define the following random variables:

$W =$ length of time between 9 AM and arrival of the PhD student (uniformly distributed between 9 AM and 5 PM).

$R =$ amount of time the student will spend with the professor, if he finds the professor (uniformly distributed between 0 and 1 hour).

$T =$ amount of time the professor will spend with the student.

Let also F be the event that the student finds the professor.

To find $\mathbf{E}[T]$, we write

$$\mathbf{E}[T] = \mathbf{P}(F)\mathbf{E}[T \,|\, F] + \mathbf{P}(F')\mathbf{E}[T \,|\, F']$$

Using the problem data,

$$\mathbf{E}[T \,|\, F] = \mathbf{E}[R] = \frac{1}{2}$$

(this is the expected value of a uniformly distribution ranging from 0 to 1),

$$\mathbf{E}[T \,|\, F'] = 0$$

(since the student leaves if he does not find the professor). We have

$$\mathbf{E}[T] = \mathbf{E}[T \,|\, F]\mathbf{P}(F) = \frac{1}{2}\mathbf{P}(F),$$

so we need to find $\mathbf{P}(F)$.

In order that the student finds the professor, his arrival should be between the arrival and the departure of the professor. Thus

$$\mathbf{P}(F) = \mathbf{P}(Y \leq W \leq X + Y).$$

We have that W can be between 0 (9AM) and 8 (5PM), but $X + Y$ can be any value greater than 0. In particular, it may happen that the sum is greater than the upper bound for W. We write

$$\mathbf{P}(F) = \mathbf{P}(Y \leq W \leq X + Y) = 1 - \big(\mathbf{P}(W < Y) + \mathbf{P}(W > X + Y)\big)$$

We have

$$\mathbf{P}(W < Y) = \int_{y=0}^{y=4} \frac{1}{4} \int_{w=0}^{w=y} \frac{1}{8} dw\ dy = \frac{1}{4}$$

and

$$\begin{aligned}
\mathbf{P}(W > X + Y) &= \int_{y=0}^{y=4} \mathbf{P}(W > X + Y \,|\, Y = y) \cdot f_Y(y) dy \\
&= \int_{y=0}^{y=4} \mathbf{P}(X < W - Y \,|\, Y = y) \cdot f_Y(y) dy \\
&= \int_{y=0}^{y=4} \int_{w=y}^{w=8} F_{X\,|\,Y}(w - y) \cdot f_W(w) \cdot f_Y(y) dw\ dy \\
&= \int_{y=0}^{y=4} \frac{1}{4} \int_{w=y}^{w=8} \frac{1}{8} \int_{x=0}^{x=w-y} \frac{1}{5 - y} e^{-x/5-y} dx\ dw\ dy \\
&= \frac{12}{32} + \frac{1}{32} \int_{y=0}^{y=4} (5 - y) e^{-\frac{8-y}{5-y}} dy.
\end{aligned}$$

Integrating numerically, we have

$$\int_{y=0}^{y=4} (5 - y) e^{-\frac{8-y}{5-y}} dy = 1.7584.$$

Thus,

$$\mathbf{P}(Y \leq W \leq X + Y) = 1 - (\mathbf{P}(W < Y) + \mathbf{P}(W > X + Y)) = 1 - 0.68 = 0.32.$$

The expected amount of time the professor will spend with the student is then

$$\mathbf{E}[T] = \frac{1}{2}\mathbf{P}(F) = \frac{1}{2}\ 0.32 = 0.16 = 9.6 \text{ mins.}$$

Next, we want to find the expected time the professor will leave his office. Let Z be the length of time measured from 9AM until he leaves his office. If the professor doesn't spend any time with the student, then Z will be equal to $X + Y$. On the other hand, if the professor is interrupted by the student, then the length of time will be equal to $X + Y + R$. This is because the professor will spend the same amount of total time on the task regardless of whether he is interrupted by the student. Therefore,

$$\mathbf{E}[Z] = \mathbf{P}(F')\mathbf{E}[Z \,|\, F'] + \mathbf{P}(F)\mathbf{E}[Z \,|\, F] = \mathbf{P}(F')\mathbf{E}[X + Y] + \mathbf{P}(F)\mathbf{E}[X + Y + R].$$

Using the results of the earlier calculations,

$$\mathbf{E}[X+Y]=5,$$

$$\mathbf{E}[X+Y+R]=\mathbf{E}[X+Y]+\mathbf{E}[R]=5+\frac{1}{2}=\frac{11}{2}.$$

Therefore,

$$\mathbf{E}[Z]=0.68\cdot 5+0.32\cdot\frac{11}{2}=5.16.$$

Thus the expected time the professor will leave his office is 5.16 hours after 9 AM.

Problem 19. Consider a gambler who at each gamble either wins or loses his bet with probabilities p and $1-p$. When $p>\frac{1}{2}$, a popular gambling system, known as the Kelley strategy, is to always bet the fraction $2p-1$ of the current fortune. Assuming $p>1/2$, compute the expected fortune after n gambles of a gambler who starts with x units and employs the Kelley strategy.

Solution. The problem is simplified by looking at the fraction of the original stake that the gambler has at any given moment. Because the expected value operation is linear, we can compute the expected fraction of the original stake and multiply by the original stake to get the expected total fortune (the original stake is a constant).

If the gambler has a at the beginning of a round, he bets $a(2p-1)$ on the round. If he wins, he will have $a+a(2p-1)$ units. If he loses, he will have $a-a(2p-1)$ units. Thus at the end of the round, he will have $2pa$ following a win, and $2(1-p)a$ following a loss.

Thus, we see that winning multiplies the gambler's fortune by $2p$ and losing multiplies it by $2(1-p)$. Therefore, if he wins k times and loses m times, he will have $(2p)^k\big(2(1-p)\big)^m$ times his original fortune. We can also compute the probability of this event. Let Y be the number of times the gambler wins in the first n gambles. Then Y has the binomial PMF:

$$p_Y(y)=\binom{n}{y}p^y(1-p)^{n-y},\qquad y=0,1,\ldots,n.$$

We can now calculate the expected fraction of the original stake that he has after n gambles. Let Z be a random variable representing this fraction. We know that Z is related to Y via

$$Z=(2p)^Y\big(2(1-p)\big)^{n-Y}.$$

We will calculate the expected value of Z using the density of Y.

$$\begin{aligned}\mathbf{E}[Z] &= \sum_{y=0}^{n} Z(y)p_Y(y)\\ &= \sum_{y=0}^{n}(2p)^y[2(1-p)]^{n-y}\binom{n}{y}p^y(1-p)^{n-y}\\ &= \sum_{y=0}^{n} 2^y p^y 2^{n-y}(1-p)^{n-y}\binom{n}{y}p^y(1-p)^{n-y}\\ &= 2^n\sum_{y=0}^{n} p^y(1-p)^{n-y}\binom{n}{y}p^y(1-p)^{n-y}\\ &= 2^n\sum_{y=0}^{n}\binom{n}{y}\left(p^2\right)^y\left[(1-p)^2\right]^{n-y}\\ &= 2^n\left(p^2+(1-p)^2\right)^n,\end{aligned}$$

where the last equality follows using the generalized binomial formula

$$\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k} = (a+b)^n.$$

Thus the gambler's expected fortune is

$$2^n\left(p^2+(1-p)^2\right)^n x,$$

where x is the fortune at the beginning of the first round.

An alternative method for solving the problem involves using iterated expectations. Let X_k be the fortune after the kth gamble. Again, we use the fact that the expected fortune after the kth gamble is

$$X_k = 2\left(p^2+(1-p)^2\right)X_{k-1}.$$

Therefore, using iterated expectations, the fortune after n gambles is

$$\begin{aligned}\mathbf{E}[X_n] = E\left[\mathbf{E}[X_n\,|\,X_{n-1}]\right] &= 2\left(p^2+(1-p)^2\right)\mathbf{E}[X_{n-1}]\\ &= 2\left(p^2+(1-p)^2\right)E\left[\mathbf{E}[X_{n-1}\,|\,X_{n-2}]\right] = \left(2\left(p^2+(1-p)^2\right)\right)^2\mathbf{E}[X_{n-2}]\\ &= \left(2\left(p^2+(1-p)^2\right)\right)^2 E\left[\mathbf{E}[X_{n-2}\,|\,X_{n-3}]\right] = \left(2\left(p^2+(1-p)^2\right)\right)^3\mathbf{E}[X_{n-3}]\\ &\cdots\\ &= \left(2^n\left(p^2+(1-p)^2\right)\right)^n\mathbf{E}[X] = 2^n\left(p^2+(1-p)^2\right)^n x.\end{aligned}$$

Problem 20. Let X and Y be two random variables that are uniformly distributed over the triangle formed by the points $(0,0)$, $(1,0)$, and $(0,2)$ (this is an asymmetric version of the PDF of Example 4.15). Calculate $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ using the law of iterated expectations and a method similar to the one of Example 4.15.

Solution. The conditional density of X given that $Y = y$ is uniform over the interval $[0, (2-y)/2]$ and we have

$$\mathbf{E}[X \mid Y = y] = \frac{2-y}{4}, \qquad 0 \le y \le 2.$$

Therefore, using the law of iterated expectations,

$$\mathbf{E}[X] = \mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}\left[\frac{2-Y}{4}\right] = \frac{2-\mathbf{E}[Y]}{4}.$$

Similarly, the conditional density of Y given that $X = x$ is uniform over the interval $[0, 2(1-x)]$ and we have

$$\mathbf{E}[Y \mid X = x] = 1 - x, \qquad 0 \le x \le 1.$$

Therefore

$$\mathbf{E}[Y] = \mathbf{E}\big[\mathbf{E}[Y \mid X]\big] = \mathbf{E}[1 - X] = 1 - \mathbf{E}[X].$$

By solving the two equations above for $\mathbf{E}[X]$ and $\mathbf{E}[Y]$, we obtain

$$\mathbf{E}[X] = \frac{1}{3}, \qquad \mathbf{E}[X] = \frac{2}{3}.$$

SECTION 4.4. Sum of a Random Number of Independent Random Variables

Problem 21. At a certain time, the number of people that enter an elevator is a random variable with Poisson PMF, with parameter λ. The weight of each person to get on the elevator is independent of each other person's weight, and is uniformly distributed between 100 and 200 lbs. Let X_i be the fraction of 100 by which the ith person exceeds 100 lbs, e.g. if the 7^{th} person weighs 175 lbs., then $X_7 = 0.75$. Let Y be the sum of the X_i.

(a) Find the transform of Y.

(b) Use the transform to compute the expected value of Y.

(c) Verify your answer of (b) by using the law of iterated expectations.

Solution. (a) Let N be the number of people that enter the elevator. Its transform is $M_N(s) = e^{\lambda(e^s - 1)}$. Let $M_X(s)$ be the common transform of the random variables X_i. Since X_i is uniformly distributed within $[0, 1]$, we have

$$M_X(s) = \frac{e^s - 1}{s}.$$

The transform $M_Y(s)$ is found by starting with the transform $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$. Thus,

$$M_Y(s) = e^{\lambda(M_X(s)-1)} = e^{\lambda\left(\frac{e^s-1}{s}-1\right)}.$$

(b) We have using the chain rule

$$\mathbf{E}[Y] = \frac{d}{ds}M_Y(s)\bigg|_{s=0} = \frac{d}{ds}M_X(s)\bigg|_{s=0} \cdot \lambda e^{\lambda(M_X(s)-1)}\bigg|_{s=0} = \frac{1}{2}\cdot\lambda = \frac{\lambda}{2},$$

where we have used the fact that $M_X(0) = 1$.

(c) From the law of iterated expectations we obtain

$$\mathbf{E}[Y] = \mathbf{E}\big[\mathbf{E}[Y \mid N]\big] = \mathbf{E}\big[N\mathbf{E}[X]\big] = \mathbf{E}[N]\mathbf{E}[X] = \frac{\lambda}{2}.$$

Problem 22. A motorist is going through 4 lights, each of which is found to be red with probability 1/2. The waiting times at each light are modeled as independent normal random variables with mean 1 minute and standard deviation 1/2 minute. Let X be the total waiting time at the red lights.

(a) Use the total probability theorem to find the transform of X and the probability that X exceeds 4 minutes. Is X normal?

(b) Find the transform of X by viewing X as a random sum of random variables.

Solution. (a) Using the total probability theorem, we have

$$\mathbf{P}(X > 4) = \sum_{k=0}^{4} \mathbf{P}(k \text{ lights are red})\mathbf{P}(X > 4 \,|\, k \text{ lights are red}).$$

We have

$$\mathbf{P}(k \text{ lights are red}) = \binom{4}{k}\left(\frac{1}{2}\right)^4.$$

The conditional PDF of X given that k lights are red, is normal with mean k minutes and standard deviation $(1/2)\sqrt{k}$. X is a mixture of normal random variables and the transform of its (unconditional) PDF is the corresponding mixture of the transforms of the (conditional) normal PDFs. However, X is not normal, because a mixture of normal PDFs need not be normal. The probability $\mathbf{P}(X > 4 \,|\, k \text{ lights are red})$ can be computed from the normal tables for each k, and $\mathbf{P}(X > 4)$ is obtained by substituting the results in the total probability formula above.

(b) Let K be the number of traffic lights that are found to be red. We can view X as the sum of K independent normal random variables. Thus the transform of X can be found by replacing in the binomial transform $M_K(s) = (1/2 + (1/2)e^s)^4$ the occurrence of e^s by the normal transform corresponding to $\mu = 1$ and $\sigma = 1/2$. Thus

$$M_X(s) = \left(\frac{1}{2} + \frac{1}{2}\left(e^{\frac{(1/2)^2 s^2}{2} + s}\right)\right)^4.$$

Note that by using the random sum argument of this part, we cannot easily obtain the probability $\mathbf{P}(X > 4)$.

Problem 23. The number C of customers who visit an internet bookstore in a day is Poisson-distributed with parameter λ. The number B of books purchased by any customer is Poisson-distributed with parameter μ. The random variables C and B are independent. The bookstore wants to increase by 10% the expected value of the number of books sold per day. Can this be done either by increasing λ by 10% while keeping μ unchanged, or by increasing μ by 10% while keeping λ unchanged? Which of these possibilities leads to the smallest variance in the number of books sold per day?

Solution. Let B_i be the number of books purchased by the ith customer, and let T be the number of books sold in a day. Then

$$T = B_1 + B_2 + \ldots + B_C$$

and T is the sum of a random number of independently and identically distributed random variables. Therefore, the mean and variance of T can be found using the mean and variance of C and B_i. In particular, we have

$$\mathbf{E}[T] = \mathbf{E}[C]\mathbf{E}[B], \qquad \text{var}(T) = \mathbf{E}[C]\text{var}(B) + \mathbf{E}[B]^2\text{var}(C).$$

Since C and B have Poisson PMFs,

$$\mathbf{E}[C] = \text{var}(C) = \lambda, \qquad \mathbf{E}[B] = \text{var}(B) = \mu.$$

Therefore,

$$\mathbf{E}[T] = \lambda\mu$$

and it is indeed true that the bookstore can obtain a 10% increase in $\mathbf{E}[T]$ by either increasing λ by 10%, or by increasing μ by 10%.

The variance of T is

$$\text{var}(T) = \lambda\mu + \mu^2\lambda = \lambda\mu(1+\mu).$$

If we increase μ by 10%, the variance of the number of books sold is

$$1.1\lambda\mu(1 + 1.1\mu)$$

If we increase λ by 10%, the variance is

$$1.1\lambda\mu(1+\mu).$$

Hence the smallest variance is obtained when λ is increased by 10%.

Problem 24. Construct an example to show that the sum of a random number of normal random variables is not normal (even though a fixed sum is).

Solution. Take X and Y to be normal with means 1 and 2, respectively, and very small variances. Then the random variable that takes the value of X with some probability p and the value of Y with probability $1-p$ takes values near 1 and 2 with relatively high probability, but takes values near its mean (which is $3-2p$) with relatively low probability. Thus, this random variable need not be normal.

Now let N be a random variable taking only the values 1 and 2 with probabilities p and $1-p$, respectively. The sum of a number N of normal random variables with mean equal to 1 and very small variance is a mixture of the type discussed above, which need not be normal.

Problem 25. * Use transforms to show that the sum of a Poisson-distributed number of independent, identically distributed random variables is Poisson.

Solution. Let N be a Poisson-distributed random variable with parameter λ. Let X_i, $i = 1, \ldots, N$, be independent Bernoulli random variables with success probability p, and let

$$L = X_1 + \cdots + X_N$$

be the corresponding random sum. The transform of L is found by starting with the transform associated with N, which is

$$M_N(s) = e^{\lambda(e^s - 1)},$$

and replacing each occurrence of e^s by the transform associated with X_i, which is

$$M_X(s) = 1 - p + pe^s.$$

We obtain

$$M_L(s) = e^{\lambda(1-p+pe^s-1)} = e^{\lambda p(e^s-1)}.$$

This is the transform of a Poisson random variable with parameter λp.

Problem 26. * A coin that comes up heads with probability p is tossed repeatedly and independently.

(a) Find the transform of the total number of tosses until two heads come up in a row by using the fact that the sum of a geometrically distributed number of independent, identically distributed, geometric random variables is geometrically distributed (cf. Example 4.23).

(b) Use a related argument to find the transform of the total number of tosses until three heads come up in a row.

Solution. (a) The head/tail sequence generated has the form

$$T \cdots THT \cdots THT \cdots \cdots HT \cdots THH$$

The number of tail strings is geometrically distributed with parameter p. The number of tosses corresponding to a tail string followed by a head (i.e., $T \cdots TH$) is also geometric with parameter p. Counting the very last head in the entire sequence, we see that the total number of tosses is $Z = 1 + Y$, where Y is the sum of geometrically distributed number of geometric random variables. By the result of Example 4.23, Y is geometrically distributed with parameter p^2. Thus the PMF of $Z = 1 + Y$ is

$$p_Z(k) = \begin{cases} p^2(1-p^2)^{k-2} & \text{if } k = 2, 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The transform of Z is

$$M_Z(s) = e^s \frac{p^2 e^s}{1 - (1-p^2)e^s} = \frac{p^2 e^{2s}}{1 - (1-p^2)e^s}.$$

(b) The head/tail sequence generated is a geometrically distributed number (with parameter p) of strings of the form of part (a) plus a final head to complete a 3-head sequence. Thus, if W is the corresponding number of tosses, we have

$$W = 1 + Z_1 + \cdots + Z_N,$$

where $Z_1, \ldots, Z_N$ are independent random variables that are distributed like the random variable Z of part (a), and N is geometrically distributed with parameter p. Therefore,

$$M_W(s) = e^s \left(\frac{pM_Z(s)}{1 - (1-p)M_Z(s)} \right),$$

where $M_Z(s)$ is the transform found in part (a).

SECTION 4.5. Covariance and Correlation

Problem 27. Consider four random variables, W, X, Y, Z, with

$$\mathbf{E}[W] = \mathbf{E}[X] = \mathbf{E}[Y] = \mathbf{E}[Z] = 0,$$

$$\mathrm{var}(W) = \mathrm{var}(X) = \mathrm{var}(Y) = \mathrm{var}(Z) = 1,$$

and assume that W, X, Y, Z are pairwise uncorrelated. Find the correlation coefficients of $\rho(A,B)$ and $\rho(A,C)$, where $A = W + X$, $B = X + Y$, and $C = Y + Z$.

Solution. We have

$$\mathrm{cov}(A,B) = \mathbf{E}[AB] - \mathbf{E}[A]\mathbf{E}[B] = \mathbf{E}[WX + WY + X^2 + XY] = \mathbf{E}[X^2] = 1,$$

and

$$\mathrm{var}(A) = \mathrm{var}(B) = 2,$$

so

$$\rho(A,B) = \frac{\mathrm{cov}(A,B)}{\sqrt{\mathrm{var}(A)\mathrm{var}(B)}} = \frac{1}{2}.$$

We also have

$$\mathrm{cov}(A,C) = \mathbf{E}[AC] - \mathbf{E}[A]\mathbf{E}[C] = \mathbf{E}[WY + WZ + XY + XZ] = 0,$$

so that

$$\rho(A,C) = 0.$$

Problem 28. * Schwartz inequality. Show that if X and Y are random variables, we have

$$\big(\mathbf{E}[XY]\big)^2 \le \mathbf{E}[X^2]\mathbf{E}[Y^2].$$

Solution. We have

$$\begin{aligned}0 &\le \mathbf{E}\left[\left(X - \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}Y\right)^2\right] \\ &= \mathbf{E}\left[X^2 - 2\frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}XY + \frac{\mathbf{E}[XY]^2}{\mathbf{E}[Y^2]^2}Y^2\right] \\ &= \mathbf{E}[X^2] - 2\frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}\mathbf{E}[XY] + \frac{\mathbf{E}[XY]^2}{\mathbf{E}[Y^2]^2}\mathbf{E}[Y^2] \\ &= \mathbf{E}[X^2] - \frac{\mathbf{E}[XY]^2}{\mathbf{E}[Y^2]},\end{aligned}$$

i.e., $\big(\mathbf{E}[XY]\big)^2 \le \mathbf{E}[X^2]\mathbf{E}[Y^2]$.

Problem 29. * **Correlation coefficient.** Consider the correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$$

of two random variables X and Y. Show that:

(a) $|\rho(X,Y)| \le 1$. *Hint*: Use the Schwartz inequality, cf. the preceding problem.

(b) If for all possible pairs of values (x,y), $y - \mathbf{E}[Y]$ is a positive (or negative) multiple of $x - \mathbf{E}[X]$, then $\rho(X,Y) = 1$ [or $\rho(X,Y) = -1$, respectively]. *Hint*: Use the analysis of Example 4.20.

(c) If $\rho(X,Y) = 1$ [or $\rho(X,Y) = -1$], then for all possible pairs of values (x,y), $y - \mathbf{E}[Y]$ is a positive (or negative, respectively) multiple of $x - \mathbf{E}[X]$.

Solution. (a) Let $\tilde{X} = X - \mathbf{E}[X]$ and $\tilde{Y} = Y - \mathbf{E}[Y]$. Using the Schwartz inequality, we get

$$\big(\rho(X,Y)\big)^2 = \frac{\mathbf{E}[\tilde{X}\tilde{Y}]^2}{\mathbf{E}[\tilde{X}^2]\mathbf{E}[\tilde{Y}^2]} \le 1,$$

and hence $|\rho(X,Y)| \le 1$.

(b) If $\tilde{Y} = a\tilde{X}$, then

$$\rho(X,Y) = \frac{\mathbf{E}[\tilde{X}a\tilde{X}]}{\sqrt{\mathbf{E}[\tilde{X}^2]\mathbf{E}[(a\tilde{X})^2]}} = \frac{a}{|a|}.$$

(c) If $\big(\rho(X,Y)\big)^2 = 1$, then

$$\begin{aligned}\mathbf{E}\left[\left(\tilde{X} - \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]}\tilde{Y}\right)^2\right] &= \mathbf{E}\left[\tilde{X}^2 - 2\frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]}\tilde{X}\tilde{Y} + \frac{\mathbf{E}[\tilde{X}\tilde{Y}]^2}{\mathbf{E}[\tilde{Y}^2]^2}\tilde{Y}^2\right] \\ &= \mathbf{E}[\tilde{X}^2] - 2\frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]}\mathbf{E}[\tilde{X}\tilde{Y}] + \frac{\mathbf{E}[\tilde{X}\tilde{Y}]^2}{\mathbf{E}[\tilde{Y}^2]^2}\mathbf{E}[\tilde{Y}^2] \\ &= \mathbf{E}[\tilde{X}^2] - \frac{\mathbf{E}[\tilde{X}\tilde{Y}]^2}{\mathbf{E}[\tilde{Y}^2]} \\ &= \mathbf{E}[\tilde{X}^2]\Big(1 - \big(\rho(X,Y)\big)^2\Big) \\ &= 0.\end{aligned}$$

Thus $\tilde{X} - \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]}\tilde{Y}$ is a constant, which must be zero because $\mathbf{E}[\tilde{X}] = \mathbf{E}[\tilde{Y}] = 0$. It follows that

$$\tilde{X} = \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]}\tilde{Y} = \sqrt{\frac{\mathbf{E}[\tilde{X}^2]}{\mathbf{E}[\tilde{Y}^2]}}\rho(X,Y)\tilde{Y},$$

i.e., the sign of the constant ratio of $\tilde{X}$ and $\tilde{Y}$ is determined by the sign of $\rho(X,Y)$.

SECTION 4.6. Least Squares Estimation

Problem 30. A police radar always overestimates the speed of incoming cars by an amount that is uniformly distributed between 0 and 5 miles/hour. Assume that car speeds are uniformly distributed from 55 to 75 miles/hour. What is the least squares estimator of the car speed based on the radar's measurement?

Solution. Let X be the car speed and let Y be the radar's measurement. Similar to Example 4.27, the joint PDF of X and Y is uniform in the range of pairs (x,y) such that $x \in [55,75]$ and $x \le y \le x+5$. We have similar to Example 4.27,

$$\mathbf{E}[X \mid Y] = \begin{cases} \frac{y}{2} + 27.5 & \text{if } 55 \le y \le 60, \\ y - 2.5 & \text{if } 60 \le y \le 75, \\ \frac{y}{2} + 35 & \text{if } 75 \le y \le 80, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 31. A carton contains a number of boxes and each box independently contains a number of gadgets that is Poisson-distributed with parameter λ.

(a) We open a carton and count the number of boxes in the carton. What is the least squares estimate of the number of gadgets in the carton given our count.

(b) We open a carton and use computer vision to count the number of boxes in the carton. The computer may overcount or undercount this number by as much as 2 boxes. In particular, the computer returns this number as $n+W$, where n is the correct number and W is a uniformly distributed integer-valued random variable in the interval $[-2,2]$. What is the least squares estimate of the number of gadgets in the carton given the computer's count.

Solution. (a) Let X_i be the number of gadgets in the ith box. The number of gadgets in the carton is

$$Y = X_1 + \cdots + X_N,$$

where N is the number of boxes in the carton. The expected number of gadgets in each box is $\mathbf{E}[X] = \lambda$. Thus, we have

$$\mathbf{E}[Y \mid N = n] = n\mathbf{E}[X] = n\lambda.$$

(b) Let $C = N + W$ be the computer's count. We have

$$\mathbf{E}[Y \mid C = c] = \sum_{i=-2}^{2} \mathbf{P}(W = i)\mathbf{E}[Y \mid C = c, W = i] = \frac{1}{5}\sum_{n=c-2}^{c+2} n\lambda = c\lambda.$$

Problem 32. * Let X and Y be two random variables with positive variances, and let ρ be their correlation coefficient.

(a) Show that $\rho = -1$ if and only if there exists a constant $\mathbf{b}$ and a *negative* constant $\mathbf{a}$ such that $Y = aX + b$.

(b) Let $\hat{X}_L$ be the linear least mean squares estimator of X based on Y. Show that $\mathbf{E}[(X - \hat{X})Y] = 0$. Use this property to show that the estimation error $X - \hat{X}_L$ is uncorrelated with Y.

(b) Let $\hat{X} = \mathbf{E}[X \mid Y]$ be the least mean squares estimator of X given Y. Show that $\mathbf{E}[(X - \hat{X})h(Y)] = 0$ for any function h.

(d) Is it true that the estimation error $X - \mathbf{E}[X \mid Y]$ is independent of Y?

Solution. (a) Let U and V be the random variables

$$U = \frac{X - \mathbf{E}[X]}{\sigma_X}, \qquad V = \frac{Y - \mathbf{E}[Y]}{\sigma_Y}.$$

Suppose there exists a scalar b and a negative scalar a such that $Y = aX + b$. We have $\mathbf{E}[Y] = a\mathbf{E}[X] + b$ and $var(Y) = a^2 var(X)$. Therefore, $\sigma_Y = -a\sigma_X$ (since $a < 0$). We obtain

$$\begin{aligned}
\rho &= \mathbf{E}[UV] \\
&= \mathbf{E}\left[\frac{X - \mathbf{E}[X]}{\sigma_X}\,\frac{Y - \mathbf{E}[Y]}{\sigma_Y}\right] \\
&= \mathbf{E}\left[\frac{X - \mathbf{E}[X]}{\sigma_X}\,\frac{(aX + b - a\mathbf{E}[X] - b)}{-a\sigma_X}\right] \\
&= \mathbf{E}\left[\frac{X - \mathbf{E}[X]}{\sigma_X}\,\frac{X - \mathbf{E}[X]}{-\sigma_X}\right] \\
&= \mathbf{E}[-U^2] \\
&= -\mathbf{E}[U^2] \\
&= -1.
\end{aligned}$$

Now suppose that $\rho = \mathbf{E}[UV] = -1$. We observe that $\mathbf{E}[U] = \mathbf{E}[V] = 0$ and $\mathbf{E}[U^2] = \mathbf{E}[V^2] = 1$. We have

$$\mathbf{E}[(U - \mathbf{E}[UV]V)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (u - \mathbf{E}[UV]v)^2 f_{U,V}(u,v)\,du dv \geq 0.$$

Since the integrand is nonnegative, it must be identically zero in order for the integral to be zero, i.e., $U = \mathbf{E}[UV]V = -V$ or $Y = aX + b$ where $a = \frac{\sigma_Y}{\sigma_X}$ and $b = \mathbf{E}[Y] - \frac{\sigma_Y}{\sigma_X}\mathbf{E}[X]$. We now claim that $\mathbf{E}[(U - \mathbf{E}[UV]V)^2] = 0$. Indeed, we have

$$\begin{aligned}
\mathbf{E}\big[(U - \mathbf{E}[UV]V)^2\big] &= \mathbf{E}[U^2] - 2(\mathbf{E}[UV])^2 + \big(\mathbf{E}[UV]\big)^2\mathbf{E}[V^2] \\
&= 1 - \big(\mathbf{E}[UV]\big)^2 \\
&= 0.
\end{aligned}$$

This completes the proof.

(b) We have

$$\hat{X}_L = \mathbf{E}[X] + \frac{cov(X,Y)}{\sigma_Y^2}\big(Y - \mathbf{E}[Y]\big),$$

so

$$\begin{aligned}
\mathbf{E}\big[(X-\hat{X}_L)Y\big] &= \mathbf{E}\left[XY - \left(\mathbf{E}[X] + \frac{cov(X,Y)}{\sigma_Y^2}\big(Y-\mathbf{E}[Y]\big)\right)Y\right] \\
&= \mathbf{E}\left[XY - \mathbf{E}[X]Y - \frac{cov(X,Y)}{\sigma_Y^2}\big(Y^2 - Y\mathbf{E}[Y]\big)\right] \\
&= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] - \frac{cov(X,Y)\mathbf{E}[Y^2]}{\sigma_Y^2} + \frac{cov(X,Y)\mathbf{E}[Y]^2}{\sigma_Y^2} \\
&= cov(X,Y)\left[1 - \frac{\mathbf{E}[Y^2]}{\sigma_Y^2} + \frac{\mathbf{E}[Y]^2}{\sigma_Y^2}\right] \\
&= cov(X,Y)\left[1 - \frac{\sigma_Y^2}{\sigma_Y^2}\right] \\
&= 0.
\end{aligned}$$

Now, let us consider the correlation of the estimation error $X - \hat{X}_L$ with Y. Since

$$\rho = \frac{cov(X-\hat{X}_L, Y)}{\sigma_{X-\hat{X}_L}\sigma_Y},$$

we must show that $cov(X-\hat{X}_L, Y) = 0$. We have

$$cov(X-\hat{X}_L, Y) = \mathbf{E}\big[(X-\hat{X}_L)Y\big] - \mathbf{E}[X-\hat{X}_L]\mathbf{E}[Y] = -\mathbf{E}[X-\hat{X}_L]\mathbf{E}[Y]$$

and

$$\begin{aligned}
\mathbf{E}[X-\hat{X}_L] &= \mathbf{E}\left[X - \mathbf{E}[X] - \frac{cov(X,Y)}{\sigma_Y^2}(Y-\mathbf{E}[Y])\right] \\
&= \mathbf{E}[X] - \mathbf{E}[X] - \frac{cov(X,Y)}{\sigma_Y^2}\mathbf{E}\big[Y - \mathbf{E}[Y]\big] \\
&= 0.
\end{aligned}$$

The desired property follows.

(c) We have

$$\mathbf{E}\big[(X-\hat{X})h(Y)\big] = \mathbf{E}\big[\big(X - \mathbf{E}[X\,|\,Y]\big)h(Y)\big],$$

by the definition of $\hat{X}$. Now using the linearity of expectation, we obtain

$$\begin{aligned}
\mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[\mathbf{E}\big[X\,|\,Y\big]h(Y)\big] &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[\mathbf{E}[Xh(Y)\,|\,Y]\big] \\
&= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[Xh(Y)\big] \\
&= 0,
\end{aligned}$$

where the first equality is obtained by noticing that $\mathbf{E}[X\,|\,Y]$ is taken with respect to X, thus allowing $h(Y)$ to be pulled into the expectation. The second equality results from the law of iterated expectations.

(d) The answer is no. For a counterexample, let X and Y be discrete random variables with the joint PMF

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{4} & \text{for } (x,y) = (1,0), (0,1), (-1,0), (0,-1) \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathbf{E}[X \,|\, Y = y] = 0$ for all possible values y, so $\mathbf{E}[X \,|\, Y] = 0$. (More precisely, $\mathbf{E}[X \,|\, Y]$ is a random variable that equals zero with probability one.) Thus, we have $X - \mathbf{E}[X \,|\, Y] = X$ (where the equality refers to equality in distribution). Since X and Y are not independent, $X - \mathbf{E}[X \,|\, Y]$ and Y are also not independent.

SECTION 4.7. The Bivariate Normal Distribution

Problem 33. Let X_1 and X_2 be independent standard normal random variables. Define the random variables Y_1 and Y_2 by

$$Y_1 = 2X_1 + X_2, \qquad Y_2 = X_1 - X_2.$$

Find $\mu_{Y_1}, \mu_{Y_2}, \text{cov}(Y_1, Y_2)$ and then write the joint PDF $f_{Y_1,Y_2}(y_1, y_2)$.

Solution. The means are given by

$$\begin{aligned} \mathbf{E}[Y_1] &= \mathbf{E}[2X_1 + X_2] = \mathbf{E}[2X_1] + \mathbf{E}[X_2] = 0, \\ \mathbf{E}[Y_2] &= \mathbf{E}[X_1 - X_2] = \mathbf{E}[X_1] - \mathbf{E}[X_2] = 0. \end{aligned}$$

The covariance is obtained as follows:

$$\begin{aligned} \text{cov}(Y_1, Y_2) &= \mathbf{E}[Y_1 Y_1] - \mu_{Y_1}\mu_{Y_2} \\ &= \mathbf{E}\big[(2X_1 + X_2) \cdot (X_1 - X_2)\big] \\ &= \mathbf{E}\big[2X_1^2 - X_1X_2 - X_2^2\big] \\ &= 1. \end{aligned}$$

The bivariate normal is determined by the means, the variance, and the correlation coefficient, so we need to calculate the variances. We have

$$\sigma_{Y_1}^2 = \mathbf{E}[Y_1^2] - \mu_{Y_1}^2 = \mathbf{E}[4X_1^2 + 4X_1X_2 + X_2^2] = 5.$$

Similarly,

$$\sigma_{Y_2}^2 = \mathbf{E}[Y_2^2] - \mu_{Y_2}^2 = 5.$$

Thus

$$\rho(Y_1, Y_2) = \frac{\text{cov}(Y_1, Y_2)}{\sigma_{Y_1}\sigma_{Y_2}} = \frac{1}{5}.$$

To write the PDF of (Y_1, Y_2) we substitute the above values into the bivariate density function.

Problem 34. The coordinates (X, Y) of a point on the plane have a bivariate normal joint distribution with $\mu_X = \mu_Y = 0$, $\rho = 0$, and $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Given that the point is at least at distance c from the origin, find the conditional joint density function for X, Y.

Solution. Let C denote the event that $X^2 + Y^2 > c$. Then,

$$\begin{aligned}\mathbf{P}(C) &= \frac{1}{2\pi\sigma^2}\int_0^{2\pi}\int_c^{\infty} re^{-\frac{r^2}{2\sigma^2}}\,dr d\theta\\ &= \frac{1}{\sigma^2}\int_c^{\infty} re^{-\frac{r^2}{2\sigma^2}}\,dr\\ &= e^{-\frac{c^2}{2\sigma^2}},\end{aligned}$$

and for $(x, y) \in C$,

$$f_{XY\,|\,C}(x, y\,|\,C) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}(x^2 + y^2 - c^2)\right\}.$$

Problem 35. Suppose that X is a standard normal random variable, and that the random variable Z takes the values 0 or 1 with equal probability. Consider a random variable Y such that

$$Y = \begin{cases} X & \text{if } z = 1,\\ -X & \text{if } z = 0.\end{cases}$$

(a) Are X and Y independent?

(b) Are Y and Z independent?

(c) Show that Y is a standard normal random variable.

(d) Show that $\mathrm{cov}(X, Y) = 0$.

Solution. (a) X and Y cannot be independent, since given X we know that Y can take one of two values that depend on the value of X.

(b) Y and Z are independent because X is symmetric relative to the origin.

(c) We have

$$\begin{aligned}f_{Y,Z}(y, z) &= f_{Y\,|\,Z}(y\,|\,z)\cdot f_Z(z)\\ &= f_X(x)\cdot f_Z(z),\end{aligned}$$

so that

$$f_Y(y) = \sum_I f_X(x)\cdot f_Z(z) = f_X(x),$$

and therefore Y is standard normal.

(d) We want to show that $\mathrm{cov}(X,Y)=0$. Since $\mathbf{E}[X]=\mathbf{E}[Y]=0$, we have

$$\begin{aligned}\mathrm{cov}(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{Y\,|\,X}(y\,|\,x)\cdot f_X(x)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)\frac{1}{2}(\delta(x)+\delta(-x))dxdy \\ &= \frac{1}{2}\int_{-\infty}^{\infty} y[xf_X(x)-xf_X(-x)]dy \\ &= 0\end{aligned}$$

as required. The last equality follows from the fact that since X is a standard normal random variable, we have $f_X(x)=f_X(-x)$. Note that we have two dependent normal random variables X,Y that have zero correlation. There is a small subtlety here. We know that if two random variables have bivariate joint distribution, and are uncorrelated, then they are independent. However in this case, we have two dependent normal random variables, whose correlation is zero. The difference here is that the joint distribution is not bivariate normal.

Problem 36. Suppose that X is a standard normal random variable.

(a) Find the third and fourth moments of X.

(b) Define a new random variable Y such that

$$Y=a+bX+cX^2.$$

Find the correlation coefficient $\rho(X,Y)$.

Solution. (a) The transform of X is

$$M_X(s)=e^{\frac{1}{2}s^2}.$$

By taking derivatives with respect to s, and find that

$$\mathbf{E}[X]=0,\quad \mathbf{E}[X^2]=1,\quad \mathbf{E}[X^3]=0,\quad \mathbf{E}[X^4]=3.$$

(b) To compute the correlation coefficient

$$\rho(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y},$$

we first compute the covariance:

$$\begin{aligned}\mathrm{cov}(X,Y) &= \mathbf{E}[XY]-\mathbf{E}[X]\mathbf{E}[Y] \\ &= \mathbf{E}[aX+bX^2+cX^3]-\mathbf{E}[X]\mathbf{E}[Y] \\ &= a\mathbf{E}[X]+b\mathbf{E}[X^2]+c\mathbf{E}[X^3] \\ &= b.\end{aligned}$$

Now $\mathrm{var}(X) = 1$ so that $\sigma_X = 1$. We have

$$\begin{aligned}\mathrm{var}(Y) &= \mathrm{var}(a + bX + cX^2)\\ &= \mathbf{E}\big[(a + bX + cX^2)^2\big] - \big(\mathbf{E}[a + bX + cX^2]\big)^2\\ &= (a^2 + 2ac + b^2 + 3c^2) - (a^2 + c^2 + 2ac)\\ &= b^2 + 2c^2\end{aligned}$$

and therefore

$$\rho(X, Y) = \frac{b}{\sqrt{b^2 + 2c^2}}.$$

Problem 37. * Let X and Y have a bivariate normal distribution. Show that $\sigma^2_{X\,|\,Y}$ is independent of Y.

Solution. We have

$$\sigma^2_{X\,|\,Y} = \big(1 - \rho(X, Y)^2\big)\,\sigma^2_X,$$

which does not depend on the experimental value of Y.

5

Stochastic Processes

Contents

A stochastic process is a mathematical model of a probabilistic experiment that evolves in time and generates a sequence of numerical values. For example, a stochastic process can be used to model:

(a) the sequence of daily prices of a stock;

(b) the sequence of scores in a football game;

(c) the sequence of failure times of a machine;

(d) the sequence of hourly traffic loads at a node of a communication network;

(e) the sequence of radar measurements of the position of an airplane.

Each numerical value in the sequence is modeled by a random variable, so a stochastic process is simply a (finite or infinite) sequence of random variables and does not represent a major conceptual departure from our basic framework. We are still dealing with a single basic experiment that involves outcomes governed by a probability law, and random variables that inherit their probabilistic properties from that law.[†] However, stochastic processes involve some change in emphasis over our earlier models. In particular:

(a) We tend to focus on the **dependencies** in the sequence of values generated by the process. For example, how do future prices of a stock depend on past values?

(b) We are often interested in **long-term averages**, involving the entire sequence of generated values. For example, what is the fraction of time that a machine is idle?

(c) We sometimes wish to characterize the likelihood or frequency of certain **boundary events.** For example, what is the probability that within a given hour all circuits of some telephone system become simultaneously busy, or what is the frequency with which some buffer in a computer network overflows with data?

In this book, we will discuss two major categories of stochastic processes.

(a) *Arrival-Type Processes*: Here, we are interested in occurrences that have the character of an "arrival," such as message receptions at a receiver, job completions in a manufacturing cell, customer purchases at a store, etc. We will focus on models in which the interarrival times (the times between successive arrivals) are independent random variables. In Section 5.1, we consider the case where arrivals occur in discrete time and the interarrival times are geometrically distributed – this is the *Bernoulli process*. In Section 5.2, we consider the case where arrivals occur in continuous time and

[†] Let us emphasize that all of the random variables arising in a stochastic process refer to a single and common experiment, and are therefore defined on a common sample space. The corresponding probability law can be specified directly or indirectly (by assuming some of its properties), as long as it unambiguously determines the joint CDF of any subset of the random variables involved.

the interarrival times are exponentially distributed – this is the *Poisson process*.

(b) *Markov Processes*: Here, we are looking at experiments that evolve in time and in which the future evolution exhibits a probabilistic dependence on the past. As an example, the future daily prices of a stock are typically dependent on past prices. However, in a Markov process, we assume a very special type of dependence: the next value depends on past values only through the current value. There is a rich methodology that applies to such processes, and which will be developed in Chapter 6.

5.1 THE BERNOULLI PROCESS

The Bernoulli process can be visualized as a sequence of independent coin tosses, where the probability of heads in each toss is a fixed number p in the range $0 < p < 1$. In general, the Bernoulli process consists of a sequence of Bernoulli trials, where each trial produces a 1 (a success) with probability p, and a 0 (a failure) with probability $1 - p$, independently of what happens in other trials.

Of course, coin tossing is just a paradigm for a broad range of contexts involving a sequence of independent binary outcomes. For example, a Bernoulli process is often used to model systems involving arrivals of customers or jobs at service centers. Here, time is discretized into periods, and a "success" at the kth trial is associated with the arrival of at least one customer at the service center during the kth period. In fact, we will often use the term "arrival" in place of "success" when this is justified by the context.

In a more formal description, we define the Bernoulli process as a sequence $X_1, X_2, \ldots$ of **independent** Bernoulli random variables X_i with

$$\begin{aligned}\mathbf{P}(X_i = 1) &= \mathbf{P}(\text{success at the } i\text{th trial}) = p,\\ \mathbf{P}(X_i = 0) &= \mathbf{P}(\text{failure at the } i\text{th trial}) = 1 - p,\end{aligned}$$

for each i.†

Given an arrival process, one is often interested in random variables such as the number of arrivals within a certain time period, or the time until the first arrival. For the case of a Bernoulli process, some answers are already available from earlier chapters. Here is a summary of the main facts.

† Generalizing from the case of a finite number of random variables, the independence of an *infinite* sequence of random variables X_i is defined by the requirement that the random variables $X_1, \ldots, X_n$ be independent for any finite n. Intuitively, knowing the experimental values of any finite subset of the random variables does not provide any new probabilistic information on the remaining random variables, and the conditional distribution of the latter stays the same as the unconditional one.

Some Random Variables Associated with the Bernoulli Process and their Properties

- **The binomial with parameters p and n.** This is the number S of successes in n independent trials. Its PMF, mean, and variance are

$$p_S(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

$$\mathbf{E}[S] = np, \qquad \mathrm{var}(S) = np(1-p).$$

- **The geometric with parameter p.** This is the number T of trials up to (and including) the first success. Its PMF, mean, and variance are

$$p_T(t) = (1-p)^{t-1} p, \qquad t = 1, 2, \ldots,$$

$$\mathbf{E}[T] = \frac{1}{p}, \qquad \mathrm{var}(T) = \frac{1-p}{p^2}.$$

Independence and Memorylessness

The independence assumption underlying the Bernoulli process has important implications, including a memorylessness property (whatever has happened in past trials provides no information on the outcomes of future trials). An appreciation and intuitive understanding of such properties is very useful, and allows for the quick solution of many problems that would be difficult with a more formal approach. In this subsection, we aim at developing the necessary intuition.

Let us start by considering random variables that are defined in terms of what happened in a certain set of trials. For example, the random variable $Z = (X_1 + X_3)X_6X_7$ is defined in terms of the first, third, sixth, and seventh trial. If we have two random variables of this type and if the two sets of trials that define them have no common element, then these random variables are independent. This is a generalization of a fact first seen in Chapter 2: if two random variables U and V are independent, then any two functions of them, $g(U)$ and $h(V)$, are also independent.

Example 5.1.

(a) Let U be the number of successes in trials 1 to 5. Let V be the number of successes in trials 6 to 10. Then, U and V are independent. This is because $U = X_1 + \cdots + X_5$, $V = X_6 + \cdots + X_{10}$, and the two collections $\{X_1, \ldots, X_5\}$, $\{X_6, \ldots, X_{10}\}$ have no common elements.

(b) Let U (respectively, V) be the first odd (respectively, even) time i in which we have a success. Then, U is determined by the odd-time sequence $X_1, X_3, \ldots$, whereas V is determined by the even-time sequence $X_2, X_4, \ldots$. Since these two sequences have no common elements, U and V are independent.

Suppose now that a Bernoulli process has been running for n time steps, and that we have observed the experimental values of $X_1, X_2, \ldots, X_n$. We notice that the sequence of future trials $X_{n+1}, X_{n+2}, \ldots$ are independent Bernoulli trials and therefore form a Bernoulli process. In addition, these future trials are independent from the past ones. We conclude that starting from any given point in time, the future is also modeled by a Bernoulli process, which is independent of the past. We refer to this as the **fresh-start** property of the Bernoulli process.

Let us now recall that the time T until the first success is a geometric random variable. Suppose that we have been watching the process for n time steps and no success has been recorded. What can we say about the number $T-n$ of remaining trials until the first success? Since the future of the process (after time n) is independent of the past and constitutes a fresh-starting Bernoulli process, the number of future trials until the first success is described by the same geometric PMF. Mathematically, we have

$$\mathbf{P}(T-n=t \,|\, T>n) = (1-p)^{t-1}p = \mathbf{P}(T=t), \qquad t=1,2,\ldots.$$

This **memorylessness** property can also be derived algebraically, using the definition of conditional probabilities, but the argument given here is certainly more intuitive.

Memorylessness and the Fresh-Start Property of the Bernoulli Process

- The number $T-n$ of trials until the first success after time n has a geometric distribution with parameter p, and is independent of the past.
- For any given time n, the sequence of random variables $X_{n+1}, X_{n+2}, \ldots$ (the future of the process) is also a Bernoulli process, and is independent from $X_1, \ldots, X_n$ (the past of the process).

The next example deals with an extension of the fresh-start property, in which we start looking at the process at a *random* time, determined by the past history of the process.

Example 5.2. Let N be the first time in which we have a success immediately following a previous success. (That is, N is the first i for which $X_{i-1} = X_i = 1$.) What is the probability $\mathbf{P}(X_{N+1} = X_{N+2} = 0)$ that there are no successes in the two trials that follow?

Intuitively, once the condition $X_{N-1} = X_N = 1$ is satisfied, from then on, the future of the process still consists of independent Bernoulli trials. Therefore the probability of an event that refers to the future of the process is the same as in a fresh-starting Bernoulli process, so that $\mathbf{P}(X_{N+1} = X_{N+2} = 0) = (1-p)^2$.

To make this argument precise, we argue that the time N is a random variable, and by conditioning on the possible values of N, we have

$$\begin{aligned}\mathbf{P}(X_{N+1} = X_{N+2} = 0) &= \sum_n \mathbf{P}(N = n)\mathbf{P}(X_{N+1} = X_{N+2} = 0 \,|\, N = n)\\ &= \sum_n \mathbf{P}(N = n)\mathbf{P}(X_{n+1} = X_{n+2} = 0 \,|\, N = n)\end{aligned}$$

Because of the way that N was defined, the event $\{N = n\}$ occurs if and only if the experimental values of $X_1, \ldots, X_n$ satisfy a certain condition. But the latter random variables are independent of X_{n+1} and X_{n+2}. Therefore,

$$\mathbf{P}(X_{n+1} = X_{n+2} = 0 \,|\, N = n) = \mathbf{P}(X_{n+1} = X_{n+2} = 0) = (1-p)^2,$$

which leads to

$$\mathbf{P}(X_{N+1} = X_{N+2} = 0) = \sum_n \mathbf{P}(N = n)(1-p)^2 = (1-p)^2.$$

Interarrival Times

An important random variable associated with the Bernoulli process is the time of the kth success, which we denote by Y_k. A related random variable is the kth interarrival time, denoted by T_k. It is defined by

$$T_1 = Y_1, \qquad T_k = Y_k - Y_{k-1}, \qquad k = 2, 3, \ldots$$

and represents the number of trials following the $k-1$st success until the next success. See Fig. 5.1 for an illustration, and also note that

$$Y_k = T_1 + T_2 + \cdots + T_k.$$

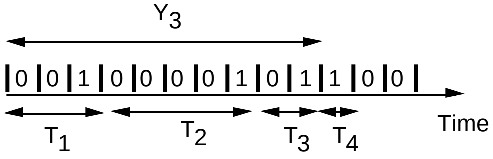


Figure 5.1: Illustration of interarrival times. In this example, $T_1 = 3$, $T_2 = 5$, $T_3 = 2$, $T_4 = 1$. Furthermore, $Y_1 = 3$, $Y_2 = 8$, $Y_3 = 10$, $Y_4 = 11$.

We have already seen that the time T_1 until the first success is a geometric random variable with parameter p. Having had a success at time T_1, the future is a fresh-starting Bernoulli process. Thus, the number of trials T_2 until the next success has the same geometric PMF. Furthermore, past trials (up to and including time T_1) are independent of future trials (from time $T_1 + 1$ onward). Since T_2 is determined exclusively by what happens in these future trials, we see that T_2 is independent of T_1. Continuing similarly, we conclude that the random variables $T_1, T_2, T_3, \ldots$ are independent and all have the same geometric distribution.

This important observation leads to an alternative, but equivalent way of describing the Bernoulli process, which is sometimes more convenient to work with.

Alternative Description of the Bernoulli Process

1. Start with a sequence of independent geometric random variables T_1, $T_2, \ldots$, with common parameter p, and let these stand for the interarrival times.
2. Record a success (or arrival) at times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, etc.

Example 5.3. A computer executes two types of tasks, priority and nonpriority, and operates in discrete time units (*slots*). A priority task arises with probability p at the beginning of each slot, independently of other slots, and requires one full slot to complete. A nonpriority task is executed at a given slot only if no priority task is available. In this context, it may be important to know the probabilistic properties of the time intervals available for nonpriority tasks.

With this in mind, let us call a slot *busy* if within this slot, the computer executes a priority task, and otherwise let us call it *idle*. We call a string of idle (or busy) slots, flanked by busy (or idle, respectively) slots, an *idle period* (or *busy period*, respectively). Let us derive the PMF, mean, and variance of the following random variables (cf. Fig. 5.2):

(a) T = the time index of the first idle slot;

(b) B = the length (number of slots) of the first busy period;

(c) I = the length of the first idle period.

We recognize T as a geometrically distributed random variable with parameter $1 - p$. Its PMF is

$$p_T(k) = p^{k-1}(1-p), \qquad k = 1, 2, \ldots.$$

Its mean and variance are

$$\mathbf{E}[T] = \frac{1}{1-p}, \qquad \mathrm{var}(T) = \frac{p}{(1-p)^2}.$$

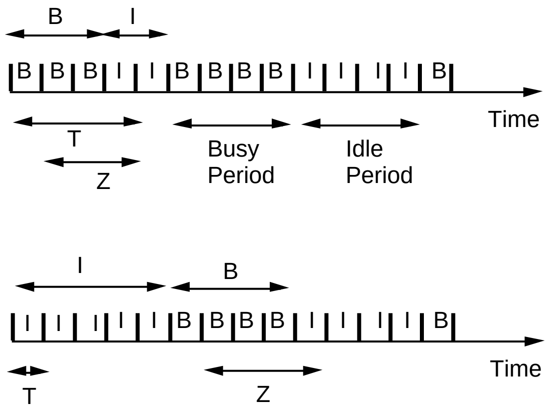


Figure 5.2: Illustration of busy (B) and idle (I) periods in Example 5.3. In the top diagram, $T = 4$, $B = 3$, and $I = 2$. In the bottom diagram, $T = 1$, $I = 5$, and $B = 4$.

Let us now consider the first busy period. It starts with the first busy slot, call it slot L. (In the top diagram in Fig. 5.2, $L = 1$; in the bottom diagram, $L = 6$.) The number Z of subsequent slots until (and including) the first subsequent idle slot has the same distribution as T, because the Bernoulli process starts fresh at time $L + 1$. We then notice that $Z = B$ and conclude that B has the same PMF as T.

If we reverse the roles of idle and busy slots, and interchange p with $1 - p$, we see that the length I of the first idle period has the same PMF as the time index of the first busy slot, so that

$$p_I(k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots, \qquad \mathbf{E}[I] = \frac{1}{p}, \quad \text{var}(I) = \frac{1-p}{p^2}.$$

We finally note that the argument given here also works for the second, third, etc. busy (or idle) period. Thus the PMFs calculated above apply to the ith busy and idle period, for any i.

The kth Arrival Time

The time Y_k of the kth success is equal to the sum $Y_k = T_1 + T_2 + \cdots + T_k$ of k independent identically distributed geometric random variables. This allows us to derive formulas for the mean, variance, and PMF of Y_k, which are given in the table that follows.

Properties of the kth Arrival Time

- The kth arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \cdots + T_k,$$

and the latter are independent geometric random variables with common parameter p.

- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \cdots + \mathbf{E}[T_k] = \frac{k}{p},$$

$$\text{var}(Y_k) = \text{var}(T_1) + \cdots + \text{var}(T_k) = \frac{k(1-p)}{p^2}.$$

- The PMF of Y_k is given by

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \qquad t = k, k+1, \ldots,$$

and is known as the **Pascal PMF of order** k.

To verify the formula for the PMF of Y_k, we first note that Y_k cannot be smaller than k. For $t \geq k$, we observe that the event $\{Y_k = t\}$ (the kth success comes at time t) will occur if and only if both of the following two events A and B occur:

(a) event A: trial t is a success;

(b) event B: exactly $k-1$ successes occur in the first $t-1$ trials.

The probabilities of these two events are

$$\mathbf{P}(A) = p$$

and

$$\mathbf{P}(B) = \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k},$$

respectively. In addition, these two events are independent (whether trial t is a success or not is independent of what happened in the first $t-1$ trials). Therefore,

$$p_{Y_k}(t) = \mathbf{P}(Y_k = t) = \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) = \binom{t-1}{k-1} p^k (1-p)^{t-k},$$

as claimed.

Example 5.4. In each minute of basketball play, Alice commits a single foul with probability p and no foul with probability $1-p$. The number of fouls in different minutes are assumed to be independent. Alice will foul out of the game once she commits her sixth foul, and will play 30 minutes if she does not foul out. What is the PMF of Alice's playing time?

We model fouls as a Bernoulli process with parameter p. Alice's playing time Z is equal to Y_6, the time until the sixth foul, except if Y_6 is larger than 30, in which case, her playing time is 30, the duration of the game; that is, $Z = \min\{Y_6, 30\}$. The random variable Y_6 has a Pascal PMF of order 6, which is given by

$$p_{Y_6}(t) = \binom{t-1}{5} p^6 (1-p)^{t-6}, \qquad t = 6, 7, \ldots$$

To determine the PMF $p_Z(z)$ of Z, we first consider the case where z is between 6 and 29. For z in this range, we have

$$p_Z(z) = \mathbf{P}(Z = z) = \mathbf{P}(Y_6 = z) = \binom{z-1}{5} p^6 (1-p)^{z-6}, \qquad z = 6, 7, \ldots, 29.$$

The probability that $Z = 30$ is then determined from

$$p_Z(30) = 1 - \sum_{z=6}^{29} p_Z(z).$$

Splitting and Merging of Bernoulli Processes

Starting with a Bernoulli process in which there is a probability p of an arrival at each time, consider **splitting** it as follows. Whenever there is an arrival, we choose to either keep it (with probability q), or to discard it (with probability $1-q$); see Fig. 5.3. Assume that the decisions to keep or discard are independent for different arrivals. If we focus on the process of arrivals that are kept, we see that it is a Bernoulli process: in each time slot, there is a probability pq of a kept arrival, independently of what happens in other slots. For the same reason, the process of discarded arrivals is also a Bernoulli process, with a probability of a discarded arrival at each time slot equal to $p(1-q)$.

In a reverse situation, we start with two *independent* Bernoulli processes (with parameters p and q, respectively) and **merge** them into a single process, as follows. An arrival is recorded in the merged process if and only if there is an arrival in at least one of the two original processes, which happens with probability $p + q - pq$ [one minus the probability $(1-p)(1-q)$ of no arrival in either process.] Since different time slots in either of the original processes are independent, different slots in the merged process are also independent. Thus, the merged process is Bernoulli, with success probability $p+q-pq$ at each time step; see Fig. 5.4.

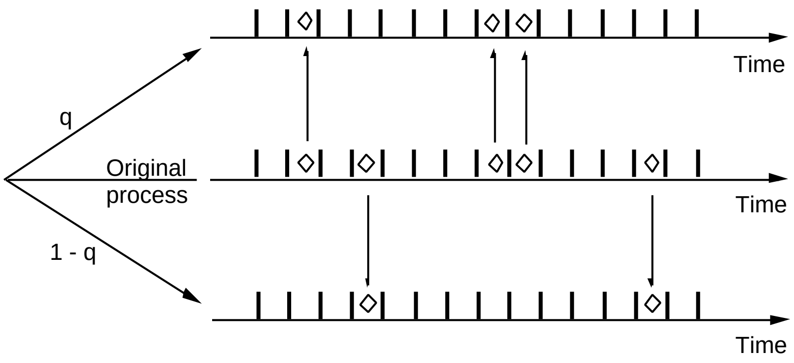


Figure 5.3: Splitting of a Bernoulli process.

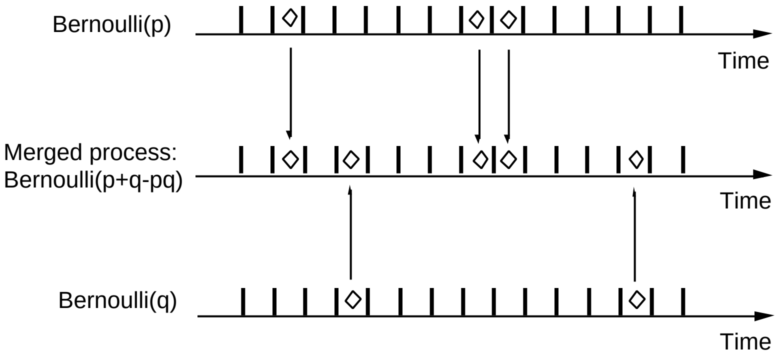


Figure 5.4: Merging of independent Bernoulli process.

Splitting and merging of Bernoulli (or other) arrival processes arises in many contexts. For example, a two-machine work center may see a stream of arriving parts to be processed and split them by sending each part to a randomly chosen machine. Conversely, a machine may be faced with arrivals of different types that can be merged into a single arrival stream.

The Poisson Approximation to the Binomial

The number of successes in n independent Bernoulli trials is a binomial random variable with parameters n and p, and its mean is np. In this subsection, we concentrate on the special case where n is large but p is small, so that the mean np has a moderate value. A situation of this type arises when one passes from discrete to continuous time, a theme to be picked up in the next section. For some more examples, think of the number of airplane accidents on any given day:

there is a large number of trials (airplane flights), but each one has a very small probability of being involved in an accident. Or think of counting the number of typos in a book: there is a large number n of words, but a very small probability of misspelling each one.

Mathematically, we can address situations of this kind, by letting n grow while simultaneously decreasing p, in a manner that keeps the product np at a constant value λ. In the limit, it turns out that the formula for the binomial PMF simplifies to the Poisson PMF. A precise statement is provided next, together with a reminder of some of the properties of the Poisson PMF that were derived in earlier chapters.

Poisson Approximation to the Binomial

- A Poisson random variable Z with parameter λ takes nonnegative integer values and is described by the PMF

$$p_Z(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

  Its mean and variance are given by

$$\mathbf{E}[Z] = \lambda, \qquad \mathrm{var}(Z) = \lambda.$$

- For any fixed nonnegative integer k, the binomial probability

$$p_S(k) = \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}$$

  converges to $p_Z(k)$, when we take the limit as $n \to \infty$ and $p = \lambda/n$, while keeping λ constant.

- In general, the Poisson PMF is a good approximation to the binomial as long as $\lambda = np$, n is very large, and p is very small.

The verification of the limiting behavior of the binomial probabilities was given in Chapter 2 as as an end-of-chapter problem, and is replicated here for convenience. We let $p = \lambda/n$ and note that

$$\begin{aligned} p_S(k) &= \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}\cdot\frac{\lambda^k}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}. \end{aligned}$$

$$= \frac{n}{n} \cdot \frac{(n-1)}{n} \cdots \frac{(n-k+1)}{n} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Let us focus on a fixed k and let $n \to \infty$. Each one of the ratios $(n-1)/n$, $(n-2)/n, \ldots, (n-k+1)/n$ converges to 1. Furthermore,†

$$\left(1 - \frac{\lambda}{n}\right)^{-k} \to 1, \qquad \left(1 - \frac{\lambda}{n}\right)^{n} \to e^{-\lambda}.$$

We conclude that for each fixed k, and as $n \to \infty$, we have

$$p_S(k) \to e^{-\lambda} \frac{\lambda^k}{k!}.$$

Example 5.5. As a rule of thumb, the Poisson/binomial approximation

$$e^{-\lambda} \frac{\lambda^k}{k!} \approx \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

is valid to several decimal places if $n \geq 100$, $p \leq 0.01$, and $\lambda = np$. To check this, consider the following.

Gary Kasparov, the world chess champion (as of 1999) plays against 100 amateurs in a large simultaneous exhibition. It has been estimated from past experience that Kasparov wins in such exhibitions 99% of his games on the average (in precise probabilistic terms, we assume that he wins each game with probability 0.99, independently of other games). What are the probabilities that he will win 100 games, 98 games, 95 games, and 90 games?

We model the number of games X that Kasparov does *not* win as a binomial random variable with parameters $n = 100$ and $p = 0.01$. Thus the probabilities that he will win 100 games, 98, 95 games, and 90 games are

$$\begin{aligned} p_X(0) &= (1 - 0.01)^{100} = 0.366, \\ p_X(2) &= \frac{100!}{98!2!} 0.01^2 (1 - 0.01)^{98} = 0.185, \\ p_X(5) &= \frac{100!}{95!5!} 0.01^5 (1 - 0.01)^{95} = 0.00290, \\ p_X(10) &= \frac{100!}{90!10!} 0.01^{10} (1 - 0.01)^{90} = 7.006 \times 10^{-8}, \end{aligned}$$

† We are using here, the well known formula $\lim_{x\to\infty}(1 - \frac{1}{x})^x = e^{-1}$. Letting $x = n/\lambda$, we have $\lim_{n\to\infty}(1 - \frac{\lambda}{n})^{n/\lambda} = e^{-1}$, from which it follows that $\lim_{n\to\infty}(1 - \frac{\lambda}{n})^n = e^{-\lambda}$.

respectively. Now let us check the corresponding Poisson approximations with $\lambda = 100 \cdot 0.01 = 1$. They are:

$$p_Z(0) = e^{-1}\frac{1}{0!} = 0.368,$$

$$p_Z(2) = e^{-1}\frac{1}{2!} = 0.184,$$

$$p_Z(5) = e^{-1}\frac{1}{5!} = 0.00306,$$

$$p_Z(10) = e^{-1}\frac{1}{10!} = 1.001 \times 10^{-8}.$$

By comparing the binomial PMF values $p_X(k)$ with their Poisson approximations $p_Z(k)$, we see that there is close agreement.

Suppose now that Kasparov plays simultaneously just 5 opponents, who are, however, stronger so that his probability of a win per game is 0.9. Here are the binomial probabilities $p_X(k)$ for $n = 5$ and $p = 0.1$, and the corresponding Poisson approximations $p_Z(k)$ for $\lambda = np = 0.5$,

$$\begin{aligned} p_X(0) &= 0.590, & p_Z(0) &= 0.605,\\ p_X(1) &= 0.328, & p_Z(1) &= 0.303,\\ p_X(2) &= 0.0729, & p_Z(2) &= 0.0758,\\ p_X(3) &= 0.0081, & p_Z(3) &= 0.0126,\\ p_X(4) &= 0.00045, & p_Z(4) &= 0.0016,\\ p_X(5) &= 0.00001, & p_Z(5) &= 0.00016. \end{aligned}$$

We see that the approximation, while not poor, is considerably less accurate than in the case where $n = 100$ and $p = 0.01$.

Example 5.6. A packet consisting of a string of n symbols is transmitted over a noisy channel. Each symbol has probability $p = 0.0001$ of being transmitted in error, independently of errors in the other symbols. How small should n be in order for the probability of incorrect transmission (at least one symbol in error) to be less than 0.001?

Each symbol transmission is viewed as an independent Bernoulli trial. Thus, the probability of a positive number S of errors in the packet is

$$1 - \mathbf{P}(S = 0) = 1 - (1 - p)^n.$$

For this probability to be less than 0.001, we must have $1 - (1 - 0.0001)^n < 0.001$ or

$$n < \frac{\ln 0.999}{\ln 0.9999} = 10.0045.$$

We can also use the Poisson approximation for $\mathbf{P}(S = 0)$, which is $e^{-\lambda}$ with $\lambda = np = 0.0001 \cdot n$, and obtain the condition $1 - e^{-0.0001 \cdot n} < 0.001$, which leads to

$$n < \frac{-\ln 0.999}{0.0001} = 10.005.$$

Given that n must be integer, both methods lead to the same conclusion that n can be at most 10.

5.2 THE POISSON PROCESS

The Poisson process can be viewed as a continuous-time analog of the Bernoulli process and applies to situations where there is no natural way of dividing time into discrete periods.

To see the need for a continuous-time version of the Bernoulli process, let us consider a possible model of traffic accidents within a city. We can start by discretizing time into one-minute periods and record a "success" during every minute in which there is at least one traffic accident. Assuming the traffic intensity to be constant over time, the probability of an accident should be the same during each period. Under the additional (and quite plausible) assumption that different time periods are independent, the sequence of successes becomes a Bernoulli process. Note that in real life, two or more accidents during the same one-minute interval are certainly possible, but the Bernoulli process model does not keep track of the exact number of accidents. In particular, it does not allow us to calculate the expected number of accidents within a given period.

One way around this difficulty is to choose the length of a time period to be very small, so that the probability of two or more accidents becomes negligible. But how small should it be? A second? A millisecond? Instead of answering this question, it is preferable to consider a limiting situation where the length of the time period becomes zero, and work with a continuous time model.

We consider an arrival process that evolves in continuous time, in the sense that any real number t is a possible arrival time. We define

$$P(k,\tau) = \mathbf{P}(\text{there are exactly } k \text{ arrivals during an interval of length } \tau),$$

and assume that this probability is the same for all intervals of the same length τ. We also introduce a positive parameter λ to be referred to as the **arrival rate** or **intensity** of the process, for reasons that will soon be apparent.

Definition of the Poisson Process

An arrival process is called a Poisson process with rate λ if it has the following properties:

(a) **(Time-homogeneity.)** The probability $P(k,\tau)$ of k arrivals is the same for all intervals of the same length τ.

(b) **(Independence.)** The number of arrivals during a particular interval is independent of the history of arrivals outside this interval.

(c) **(Small interval probabilities.)** The probabilities $P(k,\tau)$ satisfy

$$P(0,\tau) = 1 - \lambda\tau + o(\tau),$$
$$P(1,\tau) = \lambda\tau + o_1(\tau).$$

Here, $o(\tau)$ and $o_1(\tau)$ are functions of τ that satisfy

$$\lim_{\tau\to 0}\frac{o(\tau)}{\tau}=0, \qquad \lim_{\tau\to 0}\frac{o_1(\tau)}{\tau}=0.$$

The first property states that arrivals are "equally likely" at all times. The arrivals during any time interval of length τ are statistically the same, in the sense that they obey the same probability law. This is a counterpart of the assumption that the success probability p in a Bernoulli process is constant over time.

To interpret the second property, consider a particular interval $[t, t']$, of length $t'-t$. The unconditional probability of k arrivals during that interval is $P(k, t'-t)$. Suppose now that we are given complete or partial information on the arrivals outside this interval. Property (b) states that this information is irrelevant: the conditional probability of k arrivals during $[t, t']$ remains equal to the unconditional probability $P(k, t'-t)$. This property is analogous to the independence of trials in a Bernoulli process.

The third property is critical. The $o(\tau)$ and $o_1(\tau)$ terms are meant to be negligible in comparison to τ, when the interval length τ is very small. They can be thought of as the $O(\tau^2)$ terms in a Taylor series expansion of $P(k,\tau)$. Thus, for small τ, the probability of a single arrival is roughly $\lambda\tau$, plus a negligible term. Similarly, for small τ, the probability of zero arrivals is roughly $1-\lambda\tau$. Note that the probability of two or more arrivals is

$$1-P(0,\tau)-P(1,\tau)=-o(\tau)-o_1(\tau),$$

and is negligible in comparison to $P(1,\tau)$ as τ gets smaller and smaller.

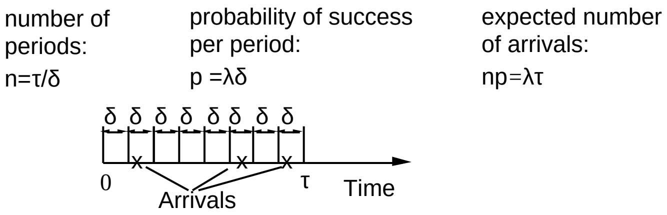


Figure 5.5: Bernoulli approximation of the Poisson process.

Let us now start with a fixed time interval of length τ and partition it into τ/δ periods of length δ, where δ is a very small number; see Fig. 5.5. The probability of more than two arrivals during any period can be neglected, because of property (c) and the preceding discussion. Different periods are independent, by property (b). Furthermore, each period has one arrival with probability approximately equal to $\lambda\delta$, or zero arrivals with probability approximately equal to $1 - \lambda\delta$. Therefore, the process being studied can be approximated by a Bernoulli process, with the approximation becoming more and more accurate the smaller δ is chosen. Thus the probability $P(k, \tau)$ of k arrivals in time τ, is approximately the same as the (binomial) probability of k successes in $n = \tau/\delta$ independent Bernoulli trials with success probability $p = \lambda\delta$ at each trial. While keeping the length τ of the interval fixed, we let the period length δ decrease to zero. We then note that the number n of periods goes to infinity, while the product np remains constant and equal to $\lambda\tau$. Under these circumstances, we saw in the previous section that the binomial PMF converges to a Poisson PMF with parameter $\lambda\tau$. We are then led to the important conclusion that

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

Note that a Taylor series expansion of $e^{-\lambda\tau}$, yields

$$\begin{aligned}
&P(0, \tau) = e^{-\lambda\tau} = 1 - \lambda\tau + O(\tau^2) \\
&P(1, \tau) = \lambda\tau e^{-\lambda\tau} = \lambda\tau - \lambda^2\tau^2 + O(\tau^3) = \lambda\tau + O(\tau^2),
\end{aligned}$$

consistent with property (c).

Using our earlier formulas for the mean and variance of the Poisson PMF, we obtain

$$\mathbf{E}[N_\tau] = \lambda\tau, \qquad \mathrm{var}(N_\tau) = \lambda\tau,$$

where N_τ stands for the number of arrivals during a time interval of length τ. These formulas are hardly surprising, since we are dealing with the limit of a binomial PMF with parameters $n = \tau/\delta$, $p = \lambda\delta$, mean $np = \lambda\tau$, and variance $np(1-p) \approx np = \lambda\tau$.

Let us now derive the probability law for the time T of the first arrival, assuming that the process starts at time zero. Note that we have $T > t$ if and only if there are no arrivals during the interval $[0, t]$. Therefore,

$$F_T(t) = \mathbf{P}(T \leq t) = 1 - \mathbf{P}(T > t) = 1 - P(0, t) = 1 - e^{-\lambda t}, \qquad t \geq 0.$$

We then differentiate the CDF $F_T(t)$ of T, and obtain the PDF formula

$$f_T(t) = \lambda e^{-\lambda t}, \qquad t \geq 0,$$

which shows that the time until the first arrival is exponentially distributed with parameter λ. We summarize this discussion in the table that follows. See also Fig. 5.6.

Random Variables Associated with the Poisson Process and their Properties

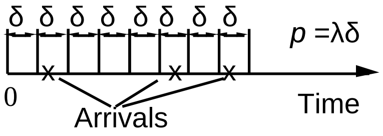
- **The Poisson with parameter $\lambda\tau$.** This is the number N_τ of arrivals in a Poisson process with rate λ, over an interval of length τ. Its PMF, mean, and variance are

$$p_{N_\tau}(k) = P(k,\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \ldots .$$

$$\mathbf{E}[N_\tau] = \lambda\tau, \qquad \mathrm{var}(N_\tau) = \lambda\tau.$$

- **The exponential with parameter λ.** This is the time T until the first arrival. Its PDF, mean, and variance are

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \qquad \mathbf{E}[T] = \frac{1}{\lambda}, \qquad \mathrm{var}(T) = \frac{1}{\lambda^2}.$$



	POISSON	**BERNOULLI**
Times of Arrival	Continuous	Discrete
PMF of # of Arrivals	Poisson	Binomial
Interarrival Time CDF	Exponential	Geometric
Arrival Rate	λ/unit time	p/per trial

Figure 5.6: View of the Bernoulli process as the discrete-time version of the Poisson. We discretize time in small intervals δ and associate each interval with a Bernoulli trial whose parameter is $p = \lambda\delta$. The table summarizes some of the basic correspondences.

Example 5.7. You get email according to a Poisson process at a rate of $\lambda = 0.2$ messages per hour. You check your email every hour. What is the probability of finding 0 and 1 new messages?

These probabilities can be found using the Poisson PMF $(\lambda\tau)^k e^{-\lambda\tau}/k!$, with $\tau = 1$, and $k = 0$ or $k = 1$:

$$\mathbf{P}(0,1) = e^{-0.2} = 0.819, \qquad \mathbf{P}(1,1) = 0.2 \cdot e^{-0.2} = 0.164$$

Suppose that you have not checked your email for a whole day. What is the probability of finding no new messages? We use again the Poisson PMF and obtain

$$\mathbf{P}(0,24) = e^{-0.2\cdot 24} = 0.008294.$$

Alternatively, we can argue that the event of no messages in a 24-hour period is the intersection of the events of no messages during each of 24 hours. These latter events are independent and the probability of each is $\mathbf{P}(0,1) = e^{-0.2}$, so

$$\mathbf{P}(0,24) = \big(\mathbf{P}(0,1)\big)^{24} = \big(e^{-0.2}\big)^{24} = 0.008294,$$

which is consistent with the preceding calculation method.

Example 5.8. Sum of Independent Poisson Random Variables. Arrivals of customers at the local supermarket are modeled by a Poisson process with a rate of $\lambda = 10$ customers per minute. Let M be the number of customers arriving between 9:00 and 9:10. Also, let N be the number of customers arriving between 9:30 and 9:35. What is the distribution of $M + N$?

We notice that M is Poisson with parameter $\mu = 10\cdot 10 = 100$ and N is Poisson with parameter $\nu = 10 \cdot 5 = 50$. Furthermore, M and N are independent. As shown in Section 4.1, using transforms, $M + N$ is Poisson with parameter $\mu + \nu = 150$. We will now proceed to derive the same result in a more direct and intuitive manner.

Let $\tilde{N}$ be the number of customers that arrive between 9:10 and 9:15. Note that $\tilde{N}$ has the same distribution as N (Poisson with parameter 50). Furthermore, $\tilde{N}$ is also independent of N. Thus, the distribution of $M + N$ is the same as the distribution of $M + \tilde{N}$. But $M + \tilde{N}$ is the number of arrivals during an interval of length 15, and has therefore a Poisson distribution with parameter $10 \cdot 15 = 150$.

This example makes a point that is valid in general. The probability of k arrivals during a *set* of times of total length τ is always given by $P(k, \tau)$, even if that set is not an interval. (In this example, we dealt with the set $[9:00, 9:10] \cup [9:30, 9:35]$, of total length 15.)

Example 5.9. During rush hour, from 8 am to 9 am, traffic accidents occur according to a Poisson process with a rate μ of 5 accidents per hour. Between 9 am and 11 am, they occur as an independent Poisson process with a rate ν of 3 accidents per hour. What is the PMF of the total number of accidents between 8 am and 11 am?

This is the sum of two independent Poisson random variables with parameters 5 and $3\cdot 2 = 6$, respectively. Since the sum of independent Poisson random variables is also Poisson, the total number of accidents has a Poisson PMF with parameter 5+6=11.

Independence and Memorylessness

The Poisson process has several properties that parallel those of the Bernoulli process, including the independence of nonoverlapping time sets, a fresh-start property, and the memorylessness of the interarrival time distribution. Given that the Poisson process can be viewed as a limiting case of a Bernoulli process, the fact that it inherits the qualitative properties of the latter should be hardly surprising.

(a) **Independence of nonoverlapping sets of times.** Consider two disjoint sets of times A and B, such as $A = [0, 1] \cup [4, \infty)$ and $B = [1.5, 3.6]$, for example. If U and V are random variables that are completely determined by what happens during A (respectively, B), then U and V are independent. This is a consequence of the second defining property of the Poisson process.

(b) **Fresh-start property.** As a special case of the preceding observation, we notice that the history of the process until a particular time t is independent from the future of the process. Furthermore, if we focus on that portion of the Poisson process that starts at time t, we observe that it inherits the defining properties of the original process. For this reason, *the portion of the Poisson process that starts at any particular time $t > 0$ is a probabilistic replica of the Poisson process starting at time 0, and is independent of the portion of the process prior to time t.* Thus, we can say that the Poisson process *starts afresh* at each time instant.

(c) **Memoryless interarrival time distribution.** We have already seen that the geometric PMF (interarrival time in the Bernoulli process) is memoryless: the number of *remaining trials* until the first future arrival does not depend on the past. The exponential PDF (interarrival time in the Poisson process) has a similar property: given the current time t and the past history, the future is a fresh-starting Poisson process, hence the *remaining time* until the next arrival has the same exponential distribution. In particular, if T is the time of the first arrival and if we are told that $T > t$, then the remaining time $T - t$ is exponentially distributed, with the same parameter λ. For an algebraic derivation of this latter fact, we first use the exponential CDF to obtain $\mathbf{P}(T > t) = e^{-\lambda t}$. We then note that for all positive scalars s and t, we have

$$\begin{aligned}\mathbf{P}(T > t+s \,|\, T > t) &= \frac{\mathbf{P}(T > t+s,\, T > t)}{\mathbf{P}(T > t)}\\ &= \frac{\mathbf{P}(T > t+s)}{\mathbf{P}(T > t)}\\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}}\\ &= e^{-\lambda s}.\end{aligned}$$

Here are some examples of reasoning based on the memoryless property.

Example 5.10. You and your partner go to a tennis court, and have to wait until the players occupying the court finish playing. Assume (somewhat unrealistically) that their playing time has an exponential PDF. Then the PDF of your waiting time (equivalently, their remaining playing time) also has the same exponential PDF, regardless of when they started playing.

Example 5.11. When you enter the bank, you find that all three tellers are busy serving other customers, and there are no other customers in queue. Assume that the service times for you and for each of the customers being served are independent identically distributed exponential random variables. What is the probability that you will be the last to leave?

The answer is 1/3. To see this, focus at the moment when you start service with one of the tellers. Then, the remaining time of each of the other two customers being served, as well as your own remaining time, have the same PDF. Therefore, you and the other two customers have equal probability 1/3 of being the last to leave.

Interarrival Times

An important random variable associated with a Poisson process that starts at time 0, is the time of the kth arrival, which we denote by Y_k. A related random variable is the kth interarrival time, denoted by T_k. It is defined by

$$T_1 = Y_1, \qquad T_k = Y_k - Y_{k-1}, \qquad k = 2, 3, \ldots$$

and represents the amount of time between the $k-1$st and the kth arrival. Note that

$$Y_k = T_1 + T_2 + \cdots + T_k.$$

We have already seen that the time T_1 until the first arrival is an exponential random variable with parameter λ. Starting from the time T_1 of the first arrival, the future is a fresh-starting Poisson process. Thus, the time until the next arrival has the same exponential PDF. Furthermore, the past of the process (up to time T_1) is independent of the future (after time T_1). Since T_2 is determined exclusively by what happens in the future, we see that T_2 is independent of T_1. Continuing similarly, we conclude that the random variables $T_1, T_2, T_3, \ldots$ are independent and all have the same exponential distribution.

This important observation leads to an alternative, but equivalent, way of describing the Poisson process.†

Alternative Description of the Poisson Process

1. Start with a sequence of independent exponential random variables $T_1, T_2, \ldots$, with common parameter λ, and let these stand for the interarrival times.
2. Record an arrival at times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, etc.

The kth Arrival Time

The time Y_k of the kth arrival is equal to the sum $Y_k = T_1 + T_2 + \cdots + T_k$ of k independent identically distributed exponential random variables. This allows us to derive formulas for the mean, variance, and PMF of Y_k, which are given in the table that follows.

Properties of the kth Arrival Time

- The kth arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \cdots + T_k,$$

and the latter are independent exponential random variables with common parameter λ.

† In our original definition, a process was called Poisson if it possessed certain properties. However, the astute reader may have noticed that we have not so far established that there exists a process with the required properties. In an alternative line of development, we could have defined the Poisson process by the alternative description given here, and such a process is clearly well-defined: we start with a sequence of independent interarrival times, from which the arrival times are completely determined. Starting with this definition, it is then possible to establish that the process satisfies all of the properties that were postulated in our original definition.

- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \cdots + \mathbf{E}[T_k] = \frac{k}{\lambda},$$

$$\text{var}(Y_k) = \text{var}(T_1) + \cdots + \text{var}(T_k) = \frac{k}{\lambda^2}.$$

- The PDF of Y_k is given by

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

  and is known as the **Erlang PDF of order** k.

To evaluate the PDF f_{Y_k} of Y_k, we can argue that for a small δ, the product $\delta \cdot f_{Y_k}(y)$ is the probability that the kth arrival occurs between times y and $y+\delta$.† When δ is very small, the probability of more than one arrival during the interval $[y, y+\delta]$ is negligible. Thus, the kth arrival occurs between y and $y+\delta$ if and only if the following two events A and B occur:

(a) event A: there is an arrival during the interval $[y, y+\delta]$;

(b) event B: there are exactly $k-1$ arrivals before time y.

The probabilities of these two events are

$$\mathbf{P}(A) \approx \lambda\delta, \qquad \text{and} \qquad \mathbf{P}(B) = P(k-1, y) = \frac{\lambda^{k-1} y^{k-1} e^{-\lambda y}}{(k-1)!}.$$

† For an alternative derivation that does not rely on approximation arguments, note that for a given $y \geq 0$, the event $\big\{Y_k \leq y\big\}$ is the same as the event

$$\{\text{number of arrivals in the interval } [0, y] \geq k\}.$$

Thus the CDF of Y_k is given by

$$F_{Y_k}(y) = \mathbf{P}\big(Y_k \leq y\big) = \sum_{n=k}^{\infty} P(n, y) = 1 - \sum_{n=0}^{k-1} P(n, y) = 1 - \sum_{n=0}^{k-1} \frac{(\lambda y)^n e^{-\lambda y}}{n!}.$$

The PDF of Y_k can be obtained by differentiating the above expression, which by straightforward calculation yields the Erlang PDF formula

$$f_{Y_k}(y) = \frac{d}{dy} F_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}.$$

Since A and B are independent, we have

$$\delta f_{Y_k}(y) \approx \mathbf{P}(y \leq Y_k \leq y+\delta) \approx \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \approx \lambda\delta \frac{\lambda^{k-1} y^{k-1} e^{-\lambda y}}{(k-1)!},$$

from which we obtain

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \qquad y \geq 0.$$

Example 5.12. You call the IRS hotline and you are told that you are the 56th person in line, excluding the person currently being served. Callers depart according to a Poisson process with a rate of $\lambda = 2$ per minute. How long will you have to wait on the average until your service starts, and what is the probability you will have to wait for more than an hour?

By the memoryless property, the remaining service time of the person currently being served is exponentially distributed with parameter 2. The service times of the 55 persons ahead of you are also exponential with the same parameter, and all of these random variables are independent. Thus, your waiting time Y is Erlang of order 56, and

$$\mathbf{E}[Y] = \frac{56}{\lambda} = 28.$$

The probability that you have to wait for more than an hour is given by the formula

$$\mathbf{P}(Y \geq 60) = \int_{60}^{\infty} \frac{\lambda^{56} y^{55} e^{-\lambda y}}{55!}\, dy.$$

Computing this probability is quite tedious. In Chapter 7, we will discuss a much easier way to compute approximately this probability. This is done using the central limit theorem, which allows us to approximate the CDF of the sum of a large number of random variables with a normal CDF and then to calculate various probabilities of interest by using the normal tables.

Splitting and Merging of Poisson Processes

Similar to the case of a Bernoulli process, we can start with a Poisson process with rate λ and split it, as follows: each arrival is kept with probability p and discarded with probability $1-p$, independently of what happens to other arrivals. In the Bernoulli case, we saw that the result of the splitting was also a Bernoulli process. In the present context, the result of the splitting turns out to be a Poisson process with rate λp.

Alternatively, we can start with two independent Poisson processes, with rates λ_1 and λ_2, and merge them by recording an arrival whenever an arrival occurs in either process. It turns out that the merged process is also Poisson with rate $\lambda_1 + \lambda_2$. Furthermore, any particular arrival of the merged process has probability $\lambda_1/(\lambda_1 + \lambda_2)$ of originating from the first process and probability $\lambda_2/(\lambda_1 + \lambda_2)$ of originating from the second, independently of all other arrivals and their origins.

We discuss these properties in the context of some examples, and at the same time provide a few different arguments to establish their validity.

Example 5.13. Splitting of Poisson Processes. A packet that arrives at a node of a data network is either a local packet which is destined for that node (this happens with probability p), or else it is a transit packet that must be relayed to another node (this happens with probability $1 - p$). Packets arrive according to a Poisson process with rate λ, and each one is a local or transit packet independently of other packets and of the arrival times. As stated above, the process of *local* packet arrivals is Poisson with rate λp. Let us see why.

We verify that the process of local packet arrivals satisfies the defining properties of a Poisson process. Since λ and p are constant (do not change with time), the first property (time homogeneity) clearly holds. Furthermore, there is no dependence between what happens in disjoint time intervals, verifying the second property. Finally, if we focus on an interval of small length δ, the probability of a local arrival is approximately the probability that there is a packet arrival, and that this turns out to be a local one, i.e., $\lambda\delta \cdot p$. In addition, the probability of two or more local arrivals is negligible in comparison to δ, and this verifies the third property. We conclude that local packet arrivals form a Poisson process and, in particular, the number L_τ of such arrivals during an interval of length τ has a Poisson PMF with parameter $p\lambda\tau$.

Let us now rederive the Poisson PMF of L_τ using transforms. The total number of packets N_τ during an interval of length τ is Poisson with parameter $\lambda\tau$. For $i = 1, \ldots, N_\tau$, let X_i be a Bernoulli random variable which is 1 if the ith packet is local, and 0 if not. Then, the random variables $X_1, X_2, \ldots$ form a Bernoulli process with success probability p. The number of local packets is the number of "successes," i.e.,

$$L_\tau = X_1 + \cdots + X_{N_\tau}.$$

We are dealing here with the sum of a random number of independent random variables. As discussed in Section 4.4, the transform associated with L_τ is found by starting with the transform associated with N_τ, which is

$$M_{N_\tau}(s) = e^{\lambda\tau(e^s-1)},$$

and replacing each occurrence of e^s by the transform associated with X_i, which is

$$M_X(s) = 1 - p + pe^s.$$

We obtain

$$M_{L_\tau}(s) = e^{\lambda\tau(1-p+pe^s-1)} = e^{\lambda\tau p(e^s-1)}.$$

We observe that this is the transform of a Poisson random variable with parameter $\lambda\tau p$, thus verifying our earlier statement for the PMF of L_τ.

We conclude with yet another method for establishing that the local packet process is Poisson. Let $T_1, T_2, \ldots$ be the interarrival times of packets of any type; these are independent exponential random variables with parameter λ. Let K be the total number of arrivals up to and including the first local packet arrival. In particular, the time S of the first local packet arrival is given by

$$S = T_1 + T_2 + \cdots + T_K.$$

Since each packet is a local one with probability p, independently of the others, and by viewing each packet as a trial which is successful with probability p, we recognize K as a geometric random variable with parameter p. Since the nature of the packets is independent of the arrival times, K is independent from the interarrival times. We are therefore dealing with a sum of a random (geometrically distributed) number of exponential random variables. We have seen in Chapter 4 (cf. Example 4.21) that such a sum is exponentially distributed with parameter λp. Since the interarrival times between successive local packets are clearly independent, it follows that the local packet arrival process is Poisson with rate λp.

Example 5.14. Merging of Poisson Processes. People with letters to mail arrive at the post office according to a Poisson process with rate λ_1, while people with packages to mail arrive according to an independent Poisson process with rate λ_2. As stated earlier the merged process, which includes arrivals of both types, is Poisson with rate $\lambda_1 + \lambda_2$. Let us see why.

First, it should be clear that the merged process satisfies the time-homogeneity property. Furthermore, since different intervals in each of the two arrival processes are independent, the same property holds for the merged process. Let us now focus on a small interval of length δ. Ignoring terms that are negligible compared to δ, we have

$$\mathbf{P}(0 \text{ arrivals in the merged process}) \approx (1 - \lambda_1\delta)(1 - \lambda_2\delta) \approx 1 - (\lambda_1 + \lambda_2)\delta,$$

$$\mathbf{P}(1 \text{ arrival in the merged process}) \approx \lambda_1\delta(1 - \lambda_2\delta) + (1 - \lambda_1\delta)\lambda_2\delta \approx (\lambda_1 + \lambda_2)\delta,$$

and the third property has been verified.

Given that an arrival has just been recorded, what is the probability that it is an arrival of a person with a letter to mail? We focus again on a small interval of length δ around the current time, and we seek the probability

$$\mathbf{P}(1 \text{ arrival of person with a letter} \mid 1 \text{ arrival}).$$

Using the definition of conditional probabilities, and ignoring the negligible probability of more than one arrival, this is

$$\frac{\mathbf{P}(1 \text{ arrival of person with a letter})}{\mathbf{P}(1 \text{ arrival})} \approx \frac{\lambda_1\delta}{(\lambda_1 + \lambda_2)\delta} = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Example 5.15. Competing Exponentials. Two light bulbs have independent and exponentially distributed lifetimes $T^{(1)}$ and $T^{(2)}$, with parameters λ_1 and λ_2, respectively. What is the distribution of the first time $Z = \min\{T^{(1)}, T^{(2)}\}$ at which a bulb burns out?

We can treat this as an exercise in derived distributions. For all $z \geq 0$, we have,

$$\begin{aligned} F_Z(z) &= \mathbf{P}\big(\min\{T^{(1)}, T^{(2)}\} \leq z\big) \\ &= 1 - \mathbf{P}\big(\min\{T^{(1)}, T^{(2)}\} > z\big) \\ &= 1 - \mathbf{P}(T^{(1)} > z, \, T^{(2)} > z) \\ &= 1 - \mathbf{P}(T^{(1)} > z)\mathbf{P}(T^{(2)} > z) \\ &= 1 - e^{-\lambda_1 z} e^{-\lambda_2 z} \\ &= 1 - e^{-(\lambda_1+\lambda_2)z}. \end{aligned}$$

This is recognized as the exponential CDF with parameter $\lambda_1 + \lambda_2$. Thus, the minimum of two independent exponentials with parameters λ_1 and λ_2 is an exponential with parameter $\lambda_1 + \lambda_2$.

For a more intuitive explanation of this fact, let us think of $T^{(1)}$ (respectively, $T^{(2)}$) as the times of the first arrival in two independent Poisson process with rate λ_1 (respectively, $T^{(2)}$). If we merge these two Poisson processes, the first arrival time will be $\min\{T^{(1)}, T^{(2)}\}$. But we already know that the merged process is Poisson with rate $\lambda_1 + \lambda_2$, and it follows that the first arrival time, $\min\{T^{(1)}, T^{(2)}\}$, is exponential with parameter $\lambda_1 + \lambda_2$.

The preceding discussion can be generalized to the case of more than two processes. Thus, the total arrival process obtained by merging the arrivals of n independent Poisson processes with arrival rates $\lambda_1, \ldots, \lambda_n$ is Poisson with arrival rate equal to the sum $\lambda_1 + \cdots + \lambda_n$.

Example 5.16. More on Competing Exponentials. Three light bulbs have independent exponentially distributed lifetimes with a common parameter λ. What is the expectation of the time until the last bulb burns out?

We think of the times when each bulb burns out as the first arrival times in independent Poisson processes. In the beginning, we have three bulbs, and the merged process has rate 3λ. Thus, the time T_1 of the first burnout is exponential with parameter 3λ, and mean $1/3\lambda$. Once a bulb burns out, and because of the memorylessness property of the exponential distribution, the remaining lifetimes of the other two lightbulbs are again independent exponential random variables with parameter λ. We thus have *two* Poisson processes running in parallel, and the remaining time T_2 until the first arrival in one of these two processes is now exponential with parameter 2λ and mean $1/2\lambda$. Finally, once a second bulb burns out, we are left with a single one. Using memorylessness once more, the remaining time T_3 until the last bulb burns out is exponential with parameter λ and mean $1/\lambda$. Thus, the expectation of the total time is

$$\mathbf{E}[T_1 + T_2 + T_3] = \frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda}.$$

Note that the random variables T_1, T_2, T_3 are independent, because of memorylessness. This also allows us to compute the variance of the total time:

$$\mathrm{var}(T_1 + T_2 + T_3) = \mathrm{var}(T_1) + \mathrm{var}(T_2) + \mathrm{var}(T_3) = \frac{1}{9\lambda^2} + \frac{1}{4\lambda^2} + \frac{1}{\lambda^2}.$$

We close by noting a related and quite deep fact, namely that the sum of a *large* number of (*not* necessarily Poisson) independent arrival processes, can be approximated by a Poisson process with arrival rate equal to the sum of the individual arrival rates. The component processes must have a small rate relative to the total (so that none of them imposes its probabilistic character on the total arrival process) and they must also satisfy some technical mathematical assumptions. Further discussion of this fact is beyond our scope, but we note that it is in large measure responsible for the abundance of Poisson-like processes in practice. For example, the telephone traffic originating in a city consists of many component processes, each of which characterizes the phone calls placed by individual residents. The component processes need not be Poisson; some people for example tend to make calls in batches, and (usually) while in the process of talking, cannot initiate or receive a second call. However, the total telephone traffic is well-modeled by a Poisson process. For the same reasons, the process of auto accidents in a city, customer arrivals at a store, particle emissions from radioactive material, etc., tend to have the character of the Poisson process.

The Random Incidence Paradox

The arrivals of a Poisson process partition the time axis into a sequence of interarrival intervals; each interarrival interval starts with an arrival and ends at the time of the next arrival. We have seen that the lengths of these interarrival intervals are independent exponential random variables with parameter λ and mean $1/\lambda$, where λ is the rate of the process. More precisely, for every k, the length of the kth interarrival interval has this exponential distribution. In this subsection, we look at these interarrival intervals from a different perspective.

Let us fix a time instant t^* and consider the length L of the interarrival interval to which it belongs. For a concrete context, think of a person who shows up at the bus station at some arbitrary time t^* and measures the time from the previous bus arrival until the next bus arrival. The arrival of this person is often referred to as a “random incidence,” but the reader should be aware that the term is misleading: t^* is just a particular time instance, not a random variable.

We assume that t^* is much larger than the starting time of the Poisson process so that we can be fairly certain that there has been an arrival prior to time t^*. To avoid the issue of determining how large a t^* is large enough, we can actually assume that the Poisson process has been running forever, so that we can be fully certain that there has been a prior arrival, and that L is well-defined. One might superficially argue that L is the length of a “typical” interarrival interval, and is exponentially distributed, but this turns out to be false. Instead, we will establish that L has an Erlang PDF of order two.

This is known as the *random incidence phenomenon or paradox*, and it can be explained with the help of Fig. 5.7. Let $[U, V]$ be the interarrival interval to which t^* belongs, so that $L = V - U$. In particular, U is the time of the first arrival prior to t^* and V is the time of the first arrival after t^*. We split L into two parts,

$$L = (t^* - U) + (V - t^*),$$

where $t^* - U$ is the elapsed time since the last arrival, and $V - t^*$ is the remaining time until the next arrival. Note that $t^* - U$ is determined by the past history of the process (before t^*), while $V - t^*$ is determined by the future of the process (after time t^*). By the independence properties of the Poisson process, the random variables $t^* - U$ and $V - t^*$ are independent. By the memorylessness property, the Poisson process starts fresh at time t^*, and therefore $V - t^*$ is exponential with parameter λ. The random variable $t^* - U$ is also exponential with parameter λ. The easiest way of seeing this is to realize that if we run a Poisson process backwards in time it remains Poisson; this is because the defining properties of a Poisson process make no reference to whether time moves forward or backward. A more formal argument is obtained by noting that

$$\mathbf{P}(t^* - U > x) = \mathbf{P}\big(\text{no arrivals during } [t^* - x, t^*]\big) = P(0, x) = e^{-\lambda x}, \qquad x \geq 0.$$

We have therefore established that L is the sum of two independent exponential random variables with parameter λ, i.e., Erlang of order two, with mean $2/\lambda$.

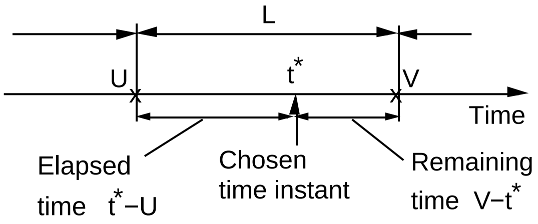


Figure 5.7: Illustration of the random incidence phenomenon. For a fixed time instant t^*, the corresponding interarrival interval $[U, V]$ consists of the elapsed time $t^* - U$ and the remaining time $V - t^*$. These two times are independent and are exponentially distributed with parameter λ, so the PDF of their sum is Erlang of order two.

Random incidence phenomena are often the source of misconceptions and errors, but these can be avoided with careful probabilistic modeling. The key issue is that even though interarrival intervals have length $1/\lambda$ on the average, an observer who arrives at an arbitrary time is more likely to fall in a large rather than a small interarrival interval. As a consequence the expected length seen by the observer is higher, $2/\lambda$ in this case. This point is amplified by the example that follows.

Example 5.17. Random incidence in a non-Poisson arrival process. Buses arrive at a station deterministically, on the hour, and fifteen minutes after the hour. Thus, the interarrival times alternate between 15 and 45 minutes. The average interarrival time is 30 minutes. A person shows up at the bus station at a "random" time. We interpret "random" to mean a time which is uniformly distributed within a particular hour. Such a person falls into an interarrival interval of length 15 with probability 1/4, and an interarrival interval of length 45 with probability 3/4. The expected value of the length of the chosen interarrival interval is

$$15 \cdot \frac{1}{4} + 45 \cdot \frac{3}{4} = 37.5,$$

which is considerably larger than 30, the average interarrival time.

SOLVED PROBLEMS

SECTION 5.1. The Bernoulli Process

Problem 1. Dave fails quizzes with probability 1/4, independently of other quizzes.

(a) What is the probability that that Dave fails exactly two of the next six quizzes?

(b) What is the expected number of quizzes that Dave will pass before he has failed three times?

(c) What is the probability that the second and third time Dave fails a quiz will occur when he takes his eighth and ninth quizzes, respectively?

(d) What is the probability that Dave fails two quizzes in a row before he passes two quizzes in a row?

Solution. (a) We have a Bernoulli process with parameter $p = 1/4$. The desired probability is given by the binomial formula:

$$\binom{6}{2} p^2 (1-p)^4 = \binom{6}{2} \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^4.$$

(b) The expected number of quizzes up to the third failure is the expected value of the third order Pascal with parameter 1/4, which is $3 \cdot 4 = 12$. Subtracting the number of failures, we have that the expected number of quizzes that Dave will pass is $12 - 3 = 9$.

(c) The event whose probability we want to find is the intersection of the following three independent events:

A: There is exactly one failure in the first seven quizzes.

B: Quiz eight is a failure.

C: Quiz nine is a failure.

We have

$$\mathbf{P}(A) = \binom{7}{1} \left(\frac{1}{4}\right) \left(\frac{3}{4}\right)^6, \qquad \mathbf{P}(B) = \mathbf{P}(C) = \frac{1}{4},$$

so the desired probability is

$$\mathbf{P}(A \cap B \cap C) = 7 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^6.$$

(d)

Problem 2. Each of n packages is loaded independently into either a red truck (with probability P) or into a green truck (with probability $1-P$). Let R be the total number of items selected for the red truck and let G be the total number of items selected for the green truck.

(a) Determine the PMF, expected value, and variance for random variable R.

(b) Evaluate $\mathbf{P}(A)$, the probability that the first item to be loaded ends up being the only one on its truck.

(c) Evaluate $\mathbf{P}(B)$, the probability that at least one truck ends up with a total of exactly one package.

(d) Evaluate the expectation and the variance for the difference, $D = R - G$.

(e) Assume $n \geq 2$. Given that both of the first two packages to be loaded go onto the red truck, find the conditional expectation, variance and probability law for random variable R.

Solution. (a) R is a binomial random variable with parameters p and N, hence

$$p_R(r) = \binom{n}{r}(1-p)^{n-r}p^r \quad \text{for } r = 0, 1, 2, \ldots, n.$$

Note that R can be written as a sum of n *i.i.d.* Bernoulli random variables, $R = X_1 + X_2 + \cdots + X_n$, where $X_i = 1$ if the i^{th} package is loaded onto the red truck and $X_i = 0$ otherwise. It is easy to verify that $\mathbf{E}[X] = p$ and $\text{var}(X) = p(1-p)$, thus

$$\mathbf{E}[R] = \mathbf{E}[X_1 + \cdots + X_n] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = np,$$

and

$$\text{var}(R) = \text{var}(X_1) + \cdots + \text{var}(X_n) = np(1-p).$$

(b) The event A is the union of two disjoint events:

1. the first item is placed in the red truck and the remaining $n-1$ are placed in the green truck, and
2. the first item is placed in the green truck and the remaining $n-1$ are placed in the red truck.

Thus, $\mathbf{P}(A) = p(1-p)^{n-1} + (1-p)p^{n-1}$.

(c) The event B occurs if exactly one or both of the trucks end up with exactly 1 package, so

$$\mathbf{P}(B) = \begin{cases} 1 & \text{if } n = 1 \\ 2p(1-p) & \text{if } n = 2 \\ \binom{n}{1}(1-p)^{n-1}p + \binom{n}{n-1}p^{n-1}(1-p) & \text{if } n = 3, 4, 5, \ldots \end{cases}$$

(d) $\mathbf{E}[D] = \mathbf{E}[R - G] = \mathbf{E}[R - (n - R)] = \mathbf{E}[2R - n] = 2\mathbf{E}[R] - n = 2np - n$. Since $D = 2R - n$, where n is a constant,

$$\text{var}(D) = 4\text{var}(R) = 4np(1-p).$$

(e) Let C be the event that each of the first 2 packages is loaded onto the red truck; then the random variable R given C becomes

$$2 + X_3 + X_4 + \cdots + X_n.$$

Hence,

$$\mathbf{E}[R \mid C] = \mathbf{E}[2 + X_3 + X_4 + \cdots + X_n] = 2 + (n-2)\mathbf{E}[X] = 2 + (n-2)p.$$

Similarly the conditional variance of R is:

$$\text{var}(R \mid C) = \text{var}(2 + X_3 + X_4 + \cdots + X_n) = (n-2)\text{var}(X_i) = (n-2)p(1-p).$$

Finally, given that the first two packages go to the red truck, the probability that a total of r packages are loaded onto the red truck is equal to the probability that $r-2$ of the remaining $n-2$ packages go to the red truck:

$$p_{R \mid C}(r \mid C) = \binom{n-2}{r-2}(1-p)^{(n-r)}p^{(r-2)} \text{ for } r = 2, \ldots, n.$$

Problem 3. * Consider the Bernoulli process with probability of success in each trial equal to p.

(a) Obtain a simple expression for the probability that the ith success occurs before the jth failure.

(b) Find the expected value and variance of the number of successes which occur before the jth failure.

Solution. (a) Let S_i and F_j be the times of the ith success and jth failure, respectively. The event $\{S_i < F_j\}$ is the same as the event of having greater or equal to i successes in the first $i+j-1$ trials. Thus, we have

$$\mathbf{P}(S_i < F_j) = \sum_{k=i}^{i+j-1} p^k (1-p)^{i+j-1-k}.$$

(b) Let k be the number of successes that occur before the jth failure. We have

$$\mathbf{E}[k] = \frac{pj}{1-p},$$

$$\text{var}(k) = \frac{pj}{(1-p)^2}.$$

Problem 4. * The PMF for the number of failures before the rth success in a Bernoulli process is sometimes called the **negative binomial** PMF. Relate the corresponding random variable to a Pascal random variable, and derive its PMF.

Solution. If X is the negative binomial random variable and Y is the Pascal random variable of order r, we have $X = Y - r$. Therefore $p_X(x) = p_Y(x+r)$, so that

$$p_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x.$$

Problem 5. * Random incidence in the Bernoulli process. A hotel receptionist with a passion for flipping a special coin that comes up a head with probability p is engaged in her favorite pastime. She flips her coin sequentially and independently, and does not raise her head to acknowledge a waiting customer until a head comes up. Suppose that a customer arrives as she flips a tail. What is the PMF of the length of the string of tails that starts with the tail following the first head prior to the customer's arrival and ends at the time she will acknowledge the customer.

Solution. Let T be the arrival time of the customer, let M be the time of the last head flip prior to T, and let N be the time of the first head flip after T. The random variable $X = N - T$ is geometrically distributed with parameter p. By symmetry and independence of the flips, the random variable $Y = T - M$ is also geometrically distributed with parameter p. The length of the string of tails between M and N is

$$L = N - M - 1 = X + Y - 1.$$

Thus $L + 1$ has a Pascal PMF of order two, and

$$p_L(l) = \binom{l}{1} p^2 (1-p)^{l-1}, \qquad l = 1, 2, \ldots$$

SECTION 5.2. The Poisson Process

Problem 6. A wombat in the San Diego zoo spends the day walking from a burrow to a food tray, eating, walking back to the burrow, resting, and repeating the cycle. The amount of time to walk from the burrow to the tray (and also from the tray to the burrow) is 20 secs. The amounts of time spent at the tray and resting are exponentially distributed with mean 30 secs. The wombat, with probability 1/3, will momentarily stand still (for a negligibly small time) during a walk to or from the tray, with all times being equally likely (and independently of what happened in the past). A photographer arrives at a random time and will take a picture at the first time the wombat will stand still. What is the expected value of the length of time the photographer has to wait to snap the wombat's picture?

Solution. We will calculate the expected length of the photographer's waiting time T conditioned on each of the two events: A, which is that the photographer arrives while the wombat is resting or eating, and A^c, which is that the photographer arrives while the wombat is walking. We will then use the total expectation theorem as follows:

$$\mathbf{E}[T] = \mathbf{P}(A)\mathbf{E}[T \,|\, A] + \mathbf{P}(A^c)\mathbf{E}[T \,|\, A^c].$$

The conditional expectation $\mathbf{E}[T \,|\, A]$ can be broken down in three components:

(1) The expected remaining time up to when the wombat starts its next walk; by the memoryless property, this time is exponentially distributed and its expected value is 30 secs.

(2) A number of walking and resting/eating intervals (each of expected length 50 secs) during which the wombat does not stop; if N is the number of these intervals,

then $N+1$ is geometrically distributed with parameter 1/3. Thus the expected length of these intervals is $(3-1)\cdot 50 = 100$ secs.

(3) The expected waiting time during the walking interval when the wombat stands still. This time is uniformly distributed between 0 and 20, so its expected value is 10 secs.

Collecting the above terms, we see that

$$\mathbf{E}[T \mid A] = 30 + 100 + 10 = 140.$$

The conditional expectation $\mathbf{E}[T \mid A^c]$ can be calculated using the total expectation theorem, by conditioning on three events: B_1, which is that the wombat does not stop during the photographer's arrival interval (probability 2/3); B_2, which is that the wombat stops during the photographer's arrival interval after the photographer arrives (probability 1/6); B_3, which is that the wombat stops during the photographer's arrival interval before the photographer arrives (probability 1/6). We have

$$\begin{aligned}\mathbf{E}[T \mid A^c, B_1] &= \mathbf{E}[\text{photographer's wait up to the end of the interval}] + \mathbf{E}[T \mid A]\\ &= 10 + 140 = 150.\end{aligned}$$

Also, it can be shown that if two points are randomly chosen in an interval of length l, the expected distance between the two points is $l/3$ (an end-of-chapter problem in Chapter 3), and using this fact, we have

$$\mathbf{E}[T \mid A^c, B_2] = \mathbf{E}[\text{photographer's wait up to the time when the wombat stops}] = \frac{20}{3}.$$

Similarly, it can be shown that if two points are randomly chosen in an interval of length l, the expected distance between each point and the nearest endpoint of the interval is $l/3$. Using this fact, we have

$$\begin{aligned}\mathbf{E}[T \mid A^c, B_3] &= \mathbf{E}[\text{photographer's wait up to the end of the interval}] + \mathbf{E}[T \mid A]\\ &= \frac{20}{3} + 140.\end{aligned}$$

Applying the total expectation theorem, we see that

$$\mathbf{E}[T \mid A^c] = \frac{2}{3}150 + \frac{1}{6}\frac{20}{3} + \frac{1}{6}\left(\frac{20}{3} + 140\right) = 125.55.$$

To apply the total expectation theorem for obtaining $\mathbf{E}[T]$, we need the probability

$$\mathbf{P}(A) = \int_0^\infty \frac{x}{20+x}\frac{1}{30}e^{-x/30}\,dx,$$

which must be calculated numerically. Once this is done all the quantities needed in the formula

$$\mathbf{E}[T] = \mathbf{P}(A)\mathbf{E}[T \mid A] + \big(1 - \mathbf{P}(A)\big)\mathbf{E}[T \mid A^c]$$

will have been calculated.

Problem 7. * Show that in the Poisson process, given that a single arrival occurred in a given interval $[a, b]$, the PDF of the time of arrival is uniform over $[a, b]$.

Problem 8. * Show that when a Poisson process is split into two Poisson processes by randomization, the two processes are independent.

Problem 9. * Let Y_k be the time of the kth arrival in a Poisson process with rate λ. Show that for all $y > 0$,

$$\sum_{k=1}^{\infty} f_{Y_k}(y) = \lambda.$$

Solution. We have

$$\begin{aligned}\sum_{k=1}^{\infty} f_{Y_k}(y) &= \sum_{k=1}^{\infty} \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!} \\ &= \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1} y^{k-1} e^{-\lambda y}}{(k-1)!} \qquad \text{(let } m = k-1\text{)} \\ &= \lambda \sum_{m=0}^{\infty} \frac{\lambda^m y^m e^{-\lambda y}}{m!} \\ &= \lambda.\end{aligned}$$

The last equality holds because the $\frac{\lambda^m y^m e^{-\lambda y}}{m!}$ terms are the values of a Poisson PMF with parameter λ and must therefore sum to 1.

For a more intuitive derivation, let δ be a small positive number and consider the following events:

A_k: the kth arrival occurs between y and $y + \delta$;

A: an arrival occurs between y and $y + \delta$.

Notice that the events $A_1, A_2, \ldots$ are disjoint and their union is A. Therefore,

$$\begin{aligned}\sum_{k=1}^{\infty} f_{Y_k}(y) \cdot \delta &\approx \sum_{k=1}^{\infty} \mathbf{P}(A_k) \\ &= \mathbf{P}(A) \\ &\approx \lambda\delta,\end{aligned}$$

and the desired result follows by canceling δ from both sides.

Problem 10. * Consider an experiment involving two independent Poisson processes with rates λ_1 and λ_2. Let $X_1(k)$ and $X_2(k)$ be the time of the kth arrival in the 1st and the 2nd process, respectively. Show that

$$\mathbf{P}\big(X_1(n) < X_2(m)\big) = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n+m-1-k}.$$

6

Markov Chains

Contents

The Bernoulli and Poisson processes studied in the preceding chapter are memoryless, in the sense that the future does not depend on the past: the occurrences of new "successes" or "arrivals" do not depend on the past history of the process. In this chapter, we consider processes where the future depends on and can be predicted to some extent by what has happened in the past.

We emphasize models where the effect of the past on the future is summarized by a **state**, which changes over time according to given probabilities. We restrict ourselves to models whose state can take a finite number of values and can change in discrete instants of time. We want to analyze the probabilistic properties of the sequence of state values.

The range of applications of the models of this chapter is truly vast. It includes just about any dynamical system whose evolution over time involves uncertainty, provided the state of the system is suitably defined. Such systems arise in a broad variety of fields, such as communications, automatic control, signal processing, manufacturing, economics, resource allocation, etc.

6.1 DISCRETE-TIME MARKOV CHAINS

We will first consider **discrete-time Markov chains**, in which the state changes at certain discrete time instants, indexed by an integer variable n. At each time step n, the Markov chain has a **state**, denoted by X_n, which belongs to a **finite** set $\mathcal{S}$ of possible states, called the **state space**. Without loss of generality, and unless there is a statement to the contrary, we will assume that $\mathcal{S} = \{1, \ldots, m\}$, for some positive integer m. The Markov chain is described in terms of its **transition probabilities** p_{ij}: whenever the state happens to be i, there is probability p_{ij} that the next state is equal to j. Mathematically,

$$p_{ij} = \mathbf{P}(X_{n+1} = j \mid X_n = i), \qquad i, j \in \mathcal{S}.$$

The key assumption underlying Markov processes is that the transition probabilities p_{ij} apply whenever state i is visited, no matter what happened in the past, and no matter how state i was reached. Mathematically, we assume the **Markov property**, which requires that

$$\begin{aligned}\mathbf{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) &= \mathbf{P}(X_{n+1} = j \mid X_n = i)\\ &= p_{ij},\end{aligned}$$

for all times n, all states $i, j \in \mathcal{S}$, and all possible sequences $i_0, \ldots, i_{n-1}$ of earlier states. Thus, the probability law of the next state X_{n+1} depends on the past only through the value of the present state X_n.

The transition probabilities p_{ij} must be of course nonnegative, and sum to one:

$$\sum_{j=1}^{m} p_{ij} = 1, \qquad \text{for all } i.$$

We will generally allow the probabilities p_{ii} to be positive, in which case it is possible for the next state to be the same as the current one. Even though the state does not change, we still view this as a state transition of a special type (a "self-transition").

Specification of Markov Models

- A Markov chain model is specified by identifying
  (a) the set of states $\mathcal{S} = \{1, \ldots, m\}$,
  (b) the set of possible transitions, namely, those pairs (i, j) for which $p_{ij} > 0$, and,
  (c) the numerical values of those p_{ij} that are positive.
- The Markov chain specified by this model is a sequence of random variables $X_0, X_1, X_2, \ldots$, that take values in $\mathcal{S}$ and which satisfy

$$\mathbf{P}(X_{n+1} = j \,|\, X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p_{ij},$$

  for all times n, all states $i, j \in \mathcal{S}$, and all possible sequences $i_0, \ldots, i_{n-1}$ of earlier states.

All of the elements of a Markov chain model can be encoded in a **transition probability matrix**, which is simply a two-dimensional array whose element at the ith row and jth column is p_{ij}:

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}.$$

It is also helpful to lay out the model in the so-called **transition probability graph**, whose nodes are the states and whose arcs are the possible transitions. By recording the numerical values of p_{ij} near the corresponding arcs, one can visualize the entire model in a way that can make some of its major properties readily apparent.

Example 6.1. Alice is taking a probability class and in each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in the given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). We assume that these probabilities do not depend on whether she was up-to-date or behind in previous weeks, so the problem has the typical Markov chain character (the future depends on the past only through the present).

Let us introduce states 1 and 2, and identify them with being up-to-date and behind, respectively. Then, the transition probabilities are

$$p_{11} = 0.8, \qquad p_{12} = 0.2, \qquad p_{21} = 0.6, \qquad p_{22} = 0.4,$$

and the transition probability matrix is

$$\begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}.$$

The transition probability graph is shown in Fig. 6.1.

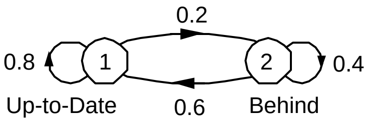


Figure 6.1: The transition probability graph in Example 6.1.

Example 6.2. A fly moves along a straight line in unit increments. At each time period, it moves one unit to the left with probability 0.3, one unit to the right with probability 0.3, and stays in place with probability 0.4, independently of the past history of movements. A spider is lurking at positions 1 and m: if the fly lands there, it is captured by the spider, and the process terminates. We want to construct a Markov chain model, assuming that the fly starts in one of the positions $2, \ldots, m-1$.

Let us introduce states $1, 2, \ldots, m$, and identify them with the corresponding positions of the fly. The nonzero transition probabilities are

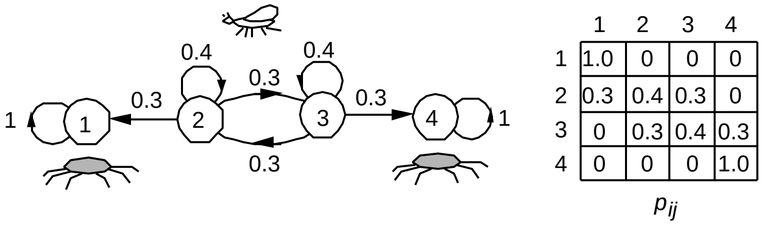
$$p_{11} = 1, \qquad p_{mm} = 1,$$

$$p_{ij} = \begin{cases} 0.3 & \text{if } j = i-1 \text{ or } j = i+1, \\ 0.4 & \text{if } j = i, \end{cases} \qquad \text{for } i = 2, \ldots, m-1.$$

The transition probability graph and matrix are shown in Fig. 6.2.

Given a Markov chain model, we can compute the probability of any particular sequence of future states. This is analogous to the use of the multiplication rule in sequential (tree) probability models. In particular, we have

$$\mathbf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{i_n} = i_n) = \mathbf{P}(X_0 = i_0) p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$



	1	2	3	4
1	1.0	0	0	0
2	0.3	0.4	0.3	0
3	0	0.3	0.4	0.3
4	0	0	0	1.0

p_{ij}

Figure 6.2: The transition probability graph and the transition probability matrix in Example 6.2, for the case where $m = 4$.

To verify this property, note that

$$\begin{aligned}\mathbf{P}(&X_0 = i_0, X_1 = i_1, \ldots, X_{i_n} = i_n)\\ &= \mathbf{P}(X_n = i_n \,|\, X_0 = i_0, \ldots, X_{n-1} = i_{n-1})\mathbf{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1})\\ &= p_{i_{n-1}i_n}\mathbf{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}),\end{aligned}$$

where the last equality made use of the Markov property. We then apply the same argument to the term $\mathbf{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1})$ and continue similarly, until we eventually obtain the desired expression. If the initial state X_0 is given and is known to be equal to some i_0, a similar argument yields

$$\mathbf{P}(X_1 = i_1, \ldots, X_{i_n} = i_n \,|\, X_0 = i_0) = p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

Graphically, a state sequence can be identified with a sequence of arcs in the transition probability graph, and the probability of such a path (given the initial state) is given by the product of the probabilities associated with the arcs traversed by the path.

Example 6.3. For the spider and fly example (Example 6.2), we have

$$\mathbf{P}(X_1 = 2, X_2 = 2, X_3 = 3, X_4 = 4 \,|\, X_0 = 2) = p_{22}p_{22}p_{23}p_{34} = (0.4)^2(0.3)^2.$$

We also have

$$\begin{aligned}\mathbf{P}(X_0 = 2, X_1 = 2, X_2 = 2, X_3 = 3, X_4 = 4) &= \mathbf{P}(X_0 = 2)p_{22}p_{22}p_{23}p_{34}\\ &= \mathbf{P}(X_0 = 2)(0.4)^2(0.3)^2.\end{aligned}$$

Note that in order to calculate a probability of this form, in which there is no conditioning on a fixed initial state, we need to specify a probability law for the initial state X_0.

n-Step Transition Probabilities

Many Markov chain problems require the calculation of the probability law of the state at some future time, conditioned on the current state. This probability law is captured by the **n-step transition probabilities**, defined by

$$r_{ij}(n) = \mathbf{P}(X_n = j \,|\, X_0 = i).$$

In words, $r_{ij}(n)$ is the probability that the state after n time periods will be j, given that the current state is i. It can be calculated using the following basic recursion, known as the **Chapman-Kolmogorov equation**.

Chapman-Kolmogorov Equation for the n-Step Transition Probabilities

The n-step transition probabilities can be generated by the recursive formula

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{kj}, \qquad \text{for } n > 1, \quad \text{and all } i,\ j,$$

starting with

$$r_{ij}(1) = p_{ij}.$$

To verify the formula, we apply the total probability theorem as follows:

$$\begin{aligned}\mathbf{P}(X_n = j \,|\, X_0 = i) &= \sum_{k=1}^{m} \mathbf{P}(X_{n-1} = k \,|\, X_0 = i) \cdot \mathbf{P}(X_n = j \,|\, X_{n-1} = k,\, X_0 = i) \\ &= \sum_{k=1}^{m} r_{ik}(n-1)p_{kj};\end{aligned}$$

see Fig. 6.3 for an illustration. We have used here the Markov property: once we condition on $X_{n-1} = k$, the conditioning on $X_0 = i$ does not affect the probability p_{kj} of reaching j at the next step.

We can view $r_{ij}(n)$ as the element at the ith row and jth column of a two-dimensional array, called the **n-step transition probability matrix**.† Figures

† Those readers familiar with matrix multiplication, may recognize that the Chapman-Kolmogorov equation can be expressed as follows: the matrix of n-step transition probabilities $r_{ij}(n)$ is obtained by multiplying the matrix of $(n-1)$-step transition probabilities $r_{ik}(n-1)$, with the one-step transition probability matrix. Thus, the n-step transition probability matrix is the nth power of the transition probability matrix.

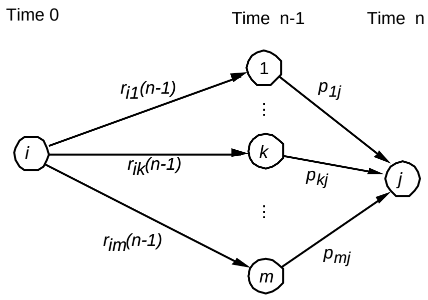
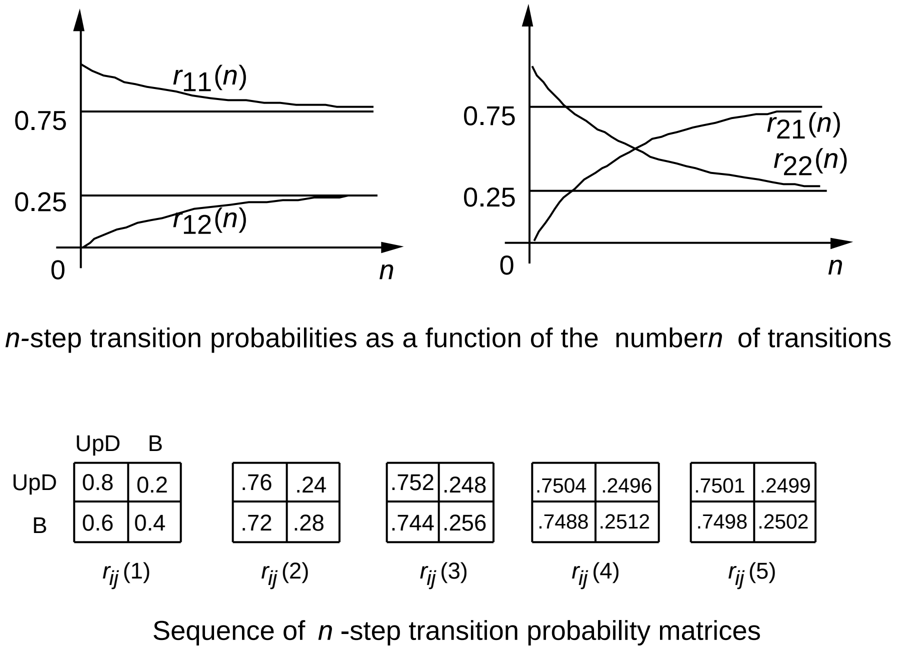


Figure 6.3: Derivation of the Chapman-Kolmogorov equation. The probability of being at state j at time n is the sum of the probabilities $r_{ik}(n-1)p_{kj}$ of the different ways of reaching j.



	UpD	B
UpD	0.8	0.2
B	0.6	0.4

.76	.24
.72	.28

.752	.248
.744	.256

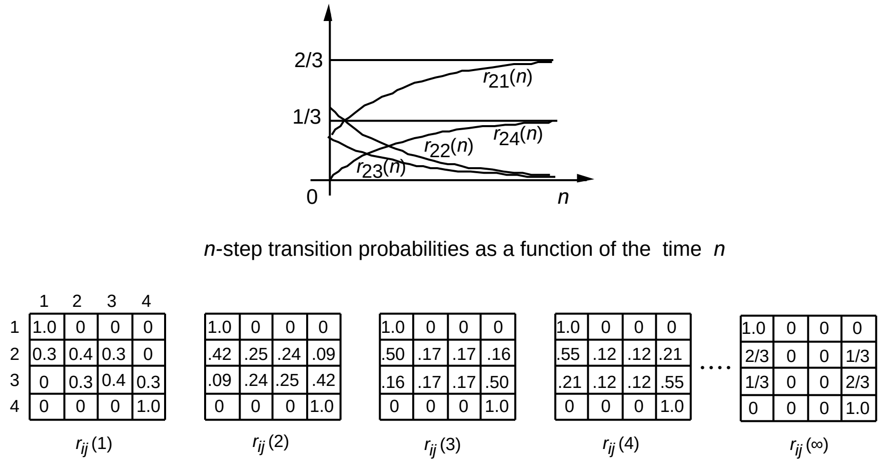
.7504	.2496
.7488	.2512

.7501	.2499
.7498	.2502

Figure 6.4: n-step transition probabilities for the "up-to-date/behind" Example 6.1. Observe that as $n \to \infty$, $r_{ij}(n)$ converges to a limit that does not depend on the initial state.

6.4 and 6.5 give the n-step transition probabilities $r_{ij}(n)$ for the cases of Examples 6.1 and 6.2, respectively. There are some interesting observations about the limiting behavior of $r_{ij}(n)$ in these two examples. In Fig. 6.4, we see that each $r_{ij}(n)$ converges to a limit, as $n \to \infty$, and this limit does not depend on the initial state. Thus, each state has a positive "steady-state" probability of being occupied at times far into the future. Furthermore, the probability $r_{ij}(n)$ depends on the initial state i when n is small, but over time this dependence diminishes. Many (but by no means all) probabilistic models that evolve over time have such a character: after a sufficiently long time, the effect of their initial condition becomes negligible.

In Fig. 6.5, we see a qualitatively different behavior: $r_{ij}(n)$ again converges, but the limit depends on the initial state, and can be zero for selected states. Here, we have two states that are "absorbing," in the sense that they are infinitely repeated, once reached. These are the states 1 and 4 that correspond to the capture of the fly by one of the two spiders. Given enough time, it is certain that some absorbing state will be reached. Accordingly, the probability of being at the non-absorbing states 2 and 3 diminishes to zero as time increases.



n-step transition probabilities as a function of the time n

	1	2	3	4
1	1.0	0	0	0
2	0.3	0.4	0.3	0
3	0	0.3	0.4	0.3
4	0	0	0	1.0

$r_{ij}(1)$

1.0	0	0	0
.42	.25	.24	.09
.09	.24	.25	.42
0	0	0	1.0

$r_{ij}(2)$

1.0	0	0	0
.50	.17	.17	.16
.16	.17	.17	.50
0	0	0	1.0

$r_{ij}(3)$

1.0	0	0	0
.55	.12	.12	.21
.21	.12	.12	.55
0	0	0	1.0

$r_{ij}(4)$

....

1.0	0	0	0
2/3	0	0	1/3
1/3	0	0	2/3
0	0	0	1.0

$r_{ij}(\infty)$

Sequence of transition probability matrices

Figure 6.5: n-step transition probabilities for the "spiders-and-fly" Example 6.2. Observe that $r_{ij}(n)$ converges to a limit that depends on the initial state.

These examples illustrate that there is a variety of types of states and asymptotic occupancy behavior in Markov chains. We are thus motivated to classify and analyze the various possibilities, and this is the subject of the next three sections.

6.2 CLASSIFICATION OF STATES

In the preceding section, we saw through examples several types of Markov chain states with qualitatively different characteristics. In particular, some states, after being visited once, are certain to be revisited again, while for some other states this may not be the case. In this section, we focus on the mechanism by which this occurs. In particular, we wish to classify the states of a Markov chain with a focus on the long-term frequency with which they are visited.

As a first step, we make the notion of revisiting a state precise. Let us say that a state j is **accessible** from a state i if for some n, the n-step transition probability $r_{ij}(n)$ is positive, i.e., if there is positive probability of reaching j, starting from i, after some number of time periods. An equivalent definition is that there is a possible state sequence $i, i_1, \ldots, i_{n-1}, j$, that starts at i and ends at j, in which the transitions $(i, i_1), (i_1, i_2), \ldots, (i_{n-2}, i_{n-1}), (i_{n-1}, j)$ all have positive probability. Let $A(i)$ be the set of states that are accessible from i. We say that i is **recurrent** if for every j that is accessible from i, i is also accessible from j; that is, for all j that belong to $A(i)$ we have that i belongs to $A(j)$.

When we start at a recurrent state i, we can only visit states $j \in A(i)$ from which i is accessible. Thus, from any future state, there is always some probability of returning to i and, given enough time, this is certain to happen. By repeating this argument, if a recurrent state is visited once, it will be revisited an infinite number of times.

A state is called **transient** if it is not recurrent. In particular, there are states $j \in A(i)$ such that i is not accessible from j. After each visit to state i, there is positive probability that the state enters such a j. Given enough time, this will happen, and state i cannot be visited after that. Thus, a transient state will only be visited a finite number of times.

Note that transience or recurrence is determined by the arcs of the transition probability graph [those pairs (i, j) for which $p_{ij} > 0$] and not by the numerical values of the p_{ij}. Figure 6.6 provides an example of a transition probability graph, and the corresponding recurrent and transient states.

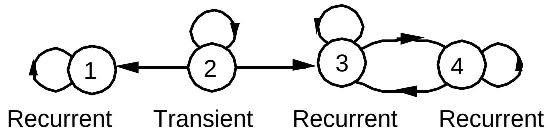


Figure 6.6: Classification of states given the transition probability graph. Starting from state 1, the only accessible state is itself, and so 1 is a recurrent state. States 1, 3, and 4 are accessible from 2, but 2 is not accessible from any of them, so state 2 is transient. States 3 and 4 are accessible only from each other (and themselves), and they are both recurrent.

If i is a recurrent state, the set of states $A(i)$ that are accessible from i

form a **recurrent class** (or simply **class**), meaning that states in $A(i)$ are all accessible from each other, and no state outside $A(i)$ is accessible from them. Mathematically, for a recurrent state i, we have $A(i) = A(j)$ for all j that belong to $A(i)$, as can be seen from the definition of recurrence. For example, in the graph of Fig. 6.6, states 3 and 4 form a class, and state 1 by itself also forms a class.

It can be seen that at least one recurrent state must be accessible from any given transient state. This is intuitively evident, and a more precise justification is given in the theoretical problems section. It follows that there must exist at least one recurrent state, and hence at least one class. Thus, we reach the following conclusion.

Markov Chain Decomposition

- A Markov chain can be decomposed into one or more recurrent classes, plus possibly some transient states.
- A recurrent state is accessible from all states in its class, but is not accessible from recurrent states in other classes.
- A transient state is not accessible from any recurrent state.
- At least one, possibly more, recurrent states are accessible from a given transient state.

Figure 6.7 provides examples of Markov chain decompositions. Decomposition provides a powerful conceptual tool for reasoning about Markov chains and visualizing the evolution of their state. In particular, we see that:

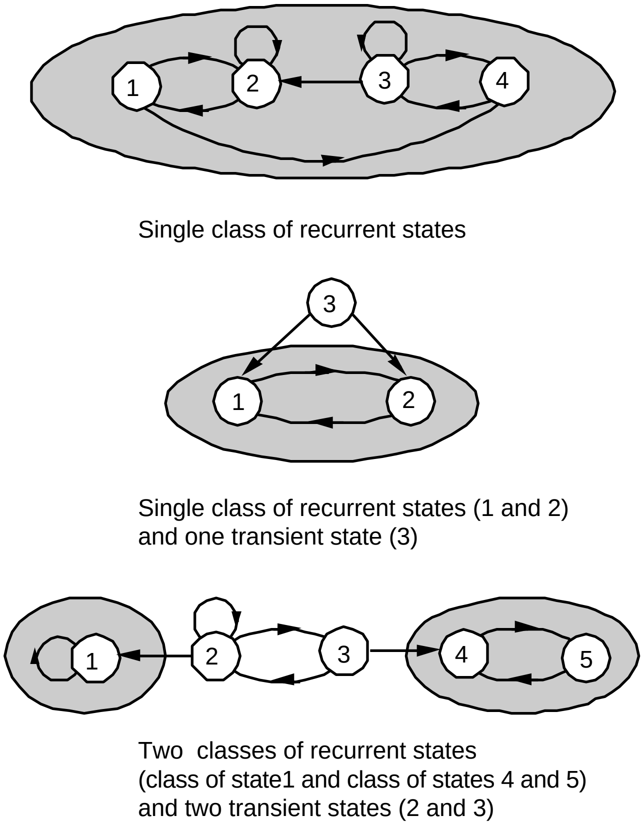
(a) once the state enters (or starts in) a class of recurrent states, it stays within that class; since all states in the class are accessible from each other, all states in the class will be visited an infinite number of times;

(b) if the initial state is transient, then the state trajectory contains an initial portion consisting of transient states and a final portion consisting of recurrent states from the same class.

For the purpose of understanding long-term behavior of Markov chains, it is important to analyze chains that consist of a single recurrent class. For the purpose of understanding short-term behavior, it is also important to analyze the mechanism by which any particular class of recurrent states is entered starting from a given transient state. These two issues, long-term and short-term behavior, are the focus of Sections 6.3 and 6.4, respectively.

Periodicity

One more characterization of a recurrent class is of special interest, and relates



Single class of recurrent states

Single class of recurrent states (1 and 2)
and one transient state (3)

Two classes of recurrent states
(class of state1 and class of states 4 and 5)
and two transient states (2 and 3)

Figure 6.7: Examples of Markov chain decompositions into recurrent classes and transient states.

to the presence or absence of a certain periodic pattern in the times that a state is visited. In particular, a recurrent class is said to be **periodic** if its states can be grouped in $d > 1$ disjoint subsets $S_1, \ldots, S_d$ so that all transitions from one subset lead to the next subset; see Fig. 6.8. More precisely,

$$\text{if} \quad i \in S_k \text{ and } p_{ij} > 0, \quad \text{then} \begin{cases} j \in S_{k+1}, & \text{if } k = 1, \ldots, d-1, \\ j \in S_1, & \text{if } k = d. \end{cases}$$

A recurrent class that is not periodic, is said to be **aperiodic**.

Thus, in a periodic recurrent class, we move through the sequence of subsets in order, and after d steps, we end up in the same subset. As an example, the recurrent class in the second chain of Fig. 6.7 (states 1 and 2) is periodic, and the same is true of the class consisting of states 4 and 5 in the third chain of Fig. 6.7. All other classes in the chains of this figure are aperiodic.

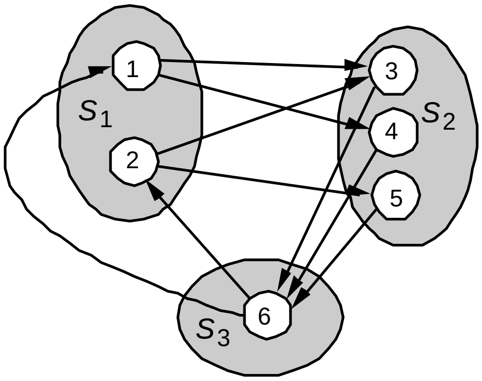


Figure 6.8: Structure of a periodic recurrent class.

Note that given a periodic recurrent class, a positive time n, and a state j in the class, there must exist some state i such that $r_{ij}(n) = 0$. The reason is that, from the definition of periodicity, the states are grouped in subsets $S_1, \ldots, S_d$, and the subset to which j belongs can be reached at time n from the states in only one of the subsets. Thus, a way to verify aperiodicity of a given recurrent class R, is to check whether there is a special time $\overline{n} \geq 1$ and a special state $s \in R$ that can be reached at time $\overline{n}$ from all initial states in R, i.e., $r_{is}(\overline{n}) > 0$ for all $i \in R$. As an example, consider the first chain in Fig. 6.7. State $s = 2$ can be reached at time $\overline{n} = 2$ starting from every state, so the unique recurrent class of that chain is aperiodic.

A converse statement, which we do not prove, also turns out to be true: if a recurrent class is not periodic, then a time $\overline{n}$ and a special state s with the above properties can always be found.

Periodicity

Consider a recurrent class R.

- The class is called **periodic** if its states can be grouped in $d > 1$ disjoint subsets $S_1, \ldots, S_d$, so that all transitions from S_k lead to S_{k+1} (or to S_1 if $k = d$).
- The class is **aperiodic** (not periodic) if and only if there exists a time $\overline{n}$ and a state s in the class, such that $p_{is}(\overline{n}) > 0$ for all $i \in R$.

6.3 STEADY-STATE BEHAVIOR

In Markov chain models, we are often interested in long-term state occupancy behavior, that is, in the n-step transition probabilities $r_{ij}(n)$ when n is very large. We have seen in the example of Fig. 6.4 that the $r_{ij}(n)$ may converge to steady-state values that are independent of the initial state, so to what extent is this behavior typical?

If there are two or more classes of recurrent states, it is clear that the limiting values of the $r_{ij}(n)$ must depend on the initial state (visiting j far into the future will depend on whether j is in the same class as the initial state i). We will, therefore, restrict attention to chains involving a single recurrent class, plus possibly some transient states. This is not as restrictive as it may seem, since we know that once the state enters a particular recurrent class, it will stay within that class. Thus, asymptotically, the presence of all classes except for one is immaterial.

Even for chains with a single recurrent class, the $r_{ij}(n)$ may fail to converge. To see this, consider a recurrent class with two states, 1 and 2, such that from state 1 we can only go to 2, and from 2 we can only go to 1 ($p_{12} = p_{21} = 1$). Then, starting at some state, we will be in that same state after any even number of transitions, and in the other state after any odd number of transitions. What is happening here is that the recurrent class is periodic, and for such a class, it can be seen that the $r_{ij}(n)$ generically oscillate.

We now assert that for every state j, the n-step transition probabilities $r_{ij}(n)$ approach a limiting value that is independent of i, provided we exclude the two situations discussed above (multiple recurrent classes and/or a periodic class). This limiting value, denoted by π_j, has the interpretation

$$\pi_j \approx \mathbf{P}(X_n = j), \qquad \text{when } n \text{ is large,}$$

and is called the **steady-state probability of** j. The following is an important theorem. Its proof is quite complicated and is outlined together with several other proofs in the theoretical problems section.

Steady-State Convergence Theorem

Consider a Markov chain with a single recurrent class, which is aperiodic. Then, the states j are associated with steady-state probabilities π_j that have the following properties.

(a) $$\lim_{n\to\infty} r_{ij}(n) = \pi_j, \qquad \text{for all } i,\ j.$$

(b) The π_j are the unique solution of the system of equations below:

$$\pi_j = \sum_{k=1}^{m} \pi_k p_{kj}, \qquad j = 1, \ldots, m,$$

$$1 = \sum_{k=1}^{m} \pi_k.$$

(c) We have

$$\pi_j = 0, \qquad \text{for all transient states } j,$$

$$\pi_j > 0, \qquad \text{for all recurrent states } j.$$

Since the steady-state probabilities π_j sum to 1, they form a probability distribution on the state space, called the **stationary distribution** of the chain. The reason for the name is that if the initial state is chosen according to this distribution, i.e., if

$$\mathbf{P}(X_0 = j) = \pi_j, \qquad j = 1, \ldots, m,$$

then, using the total probability theorem, we have

$$\mathbf{P}(X_1 = j) = \sum_{k=1}^{m} \mathbf{P}(X_0 = k) p_{kj} = \sum_{k=1}^{m} \pi_k p_{kj} = \pi_j,$$

where the last equality follows from part (b) of the steady-state convergence theorem. Similarly, we obtain $\mathbf{P}(X_n = j) = \pi_j$, for all n and j. Thus, if the initial state is chosen according to the stationary distribution, all subsequent states will have the same distribution.

The equations

$$\pi_j = \sum_{k=1}^{m} \pi_k p_{kj}, \qquad j = 1, \ldots, m,$$

are called the **balance equations**. They are a simple consequence of part (a) of the theorem and the Chapman-Kolmogorov equation. Indeed, once the convergence of $r_{ij}(n)$ to some π_j is taken for granted, we can consider the equation,

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1) p_{kj},$$

take the limit of both sides as $n \to \infty$, and recover the balance equations.† The balance equations are a linear system of equations that, together with $\sum_{k=1}^{m} \pi_k = 1$, can be solved to obtain the π_j. The following examples illustrate the solution process.

Example 6.4. Consider a two-state Markov chain with transition probabilities

$$p_{11} = 0.8, \qquad p_{12} = 0.2,$$

$$p_{21} = 0.6, \qquad p_{22} = 0.4.$$

[This is the same as the chain of Example 6.1 (cf. Fig. 6.1).] The balance equations take the form

$$\pi_1 = \pi_1 p_{11} + \pi_2 p_{21}, \qquad \pi_2 = \pi_1 p_{12} + \pi_2 p_{22},$$

or

$$\pi_1 = 0.8 \cdot \pi_1 + 0.6 \cdot \pi_2, \qquad \pi_2 = 0.2 \cdot \pi_1 + 0.4 \cdot \pi_2.$$

Note that the above two equations are dependent, since they are both equivalent to

$$\pi_1 = 3\pi_2.$$

This is a generic property, and in fact it can be shown that one of the balance equations depends on the remaining equations (see the theoretical problems). However, we know that the π_j satisfy the normalization equation

$$\pi_1 + \pi_2 = 1,$$

which supplements the balance equations and suffices to determine the π_j uniquely. Indeed, by substituting the equation $\pi_1 = 3\pi_2$ into the equation $\pi_1 + \pi_2 = 1$, we obtain $3\pi_2 + \pi_2 = 1$, or

$$\pi_2 = 0.25,$$

which using the equation $\pi_1 + \pi_2 = 1$, yields

$$\pi_1 = 0.75.$$

This is consistent with what we found earlier by iterating the Chapman-Kolmogorov equation (cf. Fig. 6.4).

Example 6.5. An absent-minded professor has two umbrellas that she uses when commuting from home to office and back. If it rains and an umbrella is available in

† According to a famous and important theorem from linear algebra (called the Perron-Frobenius theorem), the balance equations always have a nonnegative solution, for any Markov chain. What is special about a chain that has a single recurrent class, which is aperiodic, is that the solution is unique and is also equal to the limit of the n-step transition probabilities $r_{ij}(n)$.

her location, she takes it. If it is not raining, she always forgets to take an umbrella. Suppose that it rains with probability p each time she commutes, independently of other times. What is the steady-state probability that she gets wet on a given day?

We model this problem using a Markov chain with the following states:

State i: i umbrellas are available in her current location, $\quad i = 0, 1, 2.$

The transition probability graph is given in Fig. 6.9, and the transition probability matrix is

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{bmatrix}.$$

The chain has a single recurrent class that is aperiodic (assuming $0 < p < 1$), so the steady-state convergence theorem applies. The balance equations are

$$\pi_0 = (1-p)\pi_2, \quad \pi_1 = (1-p)\pi_1 + p\pi_2, \quad \pi_2 = \pi_0 + p\pi_1.$$

From the second equation, we obtain $\pi_1 = \pi_2$, which together with the first equation $\pi_0 = (1-p)\pi_2$ and the normalization equation $\pi_0 + \pi_1 + \pi_2 = 1$, yields

$$\pi_0 = \frac{1-p}{3-p}, \quad \pi_1 = \frac{1}{3-p}, \quad \pi_2 = \frac{1}{3-p}.$$

According to the steady-state convergence theorem, the steady-state probability that the professor finds herself in a place without an umbrella is π_0. The steady-state probability that she gets wet is π_0 times the probability of rain p.

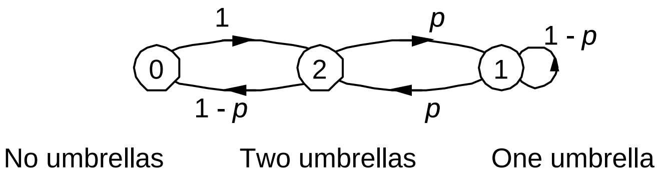


Figure 6.9: Transition probability graph for Example 6.5.

Example 6.6. A superstitious professor works in a circular building with m doors, where m is odd, and never uses the same door twice in a row. Instead he uses with probability p (or probability $1-p$) the door that is adjacent in the clockwise direction (or the counterclockwise direction, respectively) to the door he used last. What is the probability that a given door will be used on some particular day far into the future?

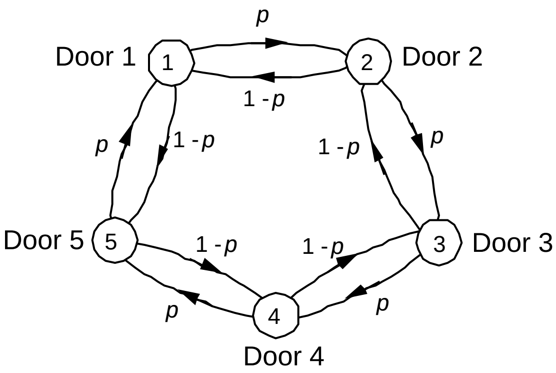


Figure 6.10: Transition probability graph in Example 6.6, for the case of $m = 5$ doors.

We introduce a Markov chain with the following m states:

State i: Last door used is door i, $\qquad i = 1, \ldots, m.$

The transition probability graph of the chain is given in Fig. 6.10, for the case $m = 5$. The transition probability matrix is

$$\begin{bmatrix} 0 & p & 0 & 0 & \ldots & 0 & 1-p \\ 1-p & 0 & p & 0 & \ldots & 0 & 0 \\ 0 & 1-p & 0 & p & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p & 0 & 0 & 0 & \ldots & 1-p & 0 \end{bmatrix}.$$

Assuming that $0 < p < 1$, the chain has a single recurrent class that is aperiodic. [To verify aperiodicity, argue by contradiction: if the class were periodic, there could be only two subsets of states such that transitions from one subset lead to the other, since it is possible to return to the starting state in two transitions. Thus, it cannot be possible to reach a state i from a state j in both an odd and an even number of transitions. However, if m is odd, this is true for states 1 and m – a contradiction (for example, see the case where $m = 5$ in Fig. 6.10, doors 1 and 5 can be reached from each other in 1 transition and also in 4 transitions).] The balance equations are

$$\begin{aligned} \pi_1 &= (1-p)\pi_2 + p\pi_m, \\ \pi_i &= p\pi_{i-1} + (1-p)\pi_{i+1}, \qquad i = 2, \ldots, m-1, \\ \pi_m &= (1-p)\pi_1 + p\pi_{m-1}. \end{aligned}$$

These equations are easily solved once we observe that by symmetry, all doors should have the same steady-state probability. This suggests the solution

$$\pi_j = \frac{1}{m}, \qquad j = 1, \ldots, m.$$

Indeed, we see that these π_j satisfy the balance equations as well as the normalization equation, so they must be the desired steady-state probabilities (by the uniqueness part of the steady-state convergence theorem).

Note that if either $p = 0$ or $p = 1$, the chain still has a single recurrent class but is periodic. In this case, the n-step transition probabilities $r_{ij}(n)$ do not converge to a limit, because the doors are used in a cyclic order. Similarly, if m is even, the recurrent class of the chain is periodic, since the states can be grouped into two subsets, the even and the odd numbered states, such that from each subset one can only go to the other subset.

Example 6.7. A machine can be either working or broken down on a given day. If it is working, it will break down in the next day with probability b, and will continue working with probability $1 - b$. If it breaks down on a given day, it will be repaired and be working in the next day with probability r, and will continue to be broken down with probability $1 - r$. What is the steady-state probability that the machine is working on a given day?

We introduce a Markov chain with the following two states:

State 1: Machine is working, State 2: Machine is broken down.

The transition probability graph of the chain is given in Fig. 6.11. The transition probability matrix is

$$\begin{bmatrix} 1-b & b \\ r & 1-r \end{bmatrix}.$$

This Markov chain has a single recurrent class that is aperiodic (assuming $0 < b < 1$ and $0 < r < 1$), and from the balance equations, we obtain

$$\pi_1 = (1-b)\pi_1 + r\pi_2, \qquad \pi_2 = b\pi_1 + (1-r)\pi_2,$$

or

$$b\pi_1 = r\pi_2.$$

This equation together with the normalization equation $\pi_1 + \pi_2 = 1$, yields the steady-state probabilities

$$\pi_1 = \frac{r}{b+r}, \qquad \pi_2 = \frac{b}{b+r}.$$

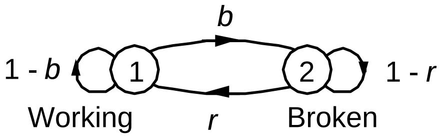


Figure 6.11: Transition probability graph for Example 6.7.

The situation considered in the previous example has evidently the Markov property, i.e., the state of the machine at the next day depends explicitly only on its state at the present day. However, it is possible to use a Markov chain model even if there is a dependence on the states at several past days. The general idea is to introduce some additional states which encode what has happened in preceding periods. Here is an illustration of this technique.

Example 6.8. Consider a variation of Example 6.7. If the machine remains broken for a given number of ℓ days, despite the repair efforts, it is replaced by a new working machine. To model this as a Markov chain, we replace the single state 2, corresponding to a broken down machine, with several states that indicate the number of days that the machine is broken. These states are

State $(2,i)$: The machine has been broken for i days, $i = 1, 2, \ldots, \ell$.

The transition probability graph is given in Fig. 6.12 for the case where $\ell = 4$. Again this Markov chain has a single recurrent class that is aperiodic. From the balance equations, we have

$$\begin{aligned}\pi_1 &= (1-b)\pi_1 + r(\pi_{(2,1)} + \cdots + \pi_{(2,\ell-1)}) + \pi_{(2,\ell)},\\ \pi_{(2,1)} &= b\pi_1,\\ \pi_{(2,i)} &= (1-r)\pi_{(2,i-1)}, \qquad i = 2, \ldots, \ell.\end{aligned}$$

The last two equations can be used to express $\pi_{(2,i)}$ in terms of π_1,

$$\pi_{(2,i)} = (1-r)^{i-1} b\pi_1, \qquad i = 1, \ldots, \ell.$$

Substituting into the normalization equation $\pi_1 + \sum_{i=1}^{\ell} \pi_{(2,i)} = 1$, we obtain

$$1 = \left(1 + b\sum_{i=1}^{\ell}(1-r)^{i-1}\right)\pi_1 = \left(1 + \frac{b\big(1-(1-r)^{\ell}\big)}{r}\right)\pi_1,$$

or

$$\pi_1 = \frac{r}{r + b\big(1-(1-r)^{\ell}\big)}.$$

Using the equation $\pi_{(2,i)} = (1-r)^{i-1} b\pi_1$, we can also obtain explicit formulas for the $\pi_{(2,i)}$.

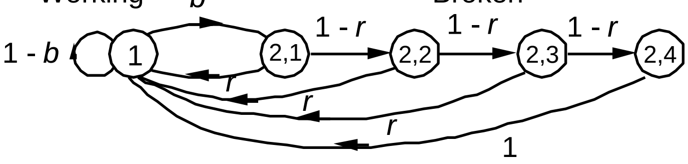


Figure 6.12: Transition probability graph for Example 6.8. A machine that has remained broken for $\ell = 4$ days is replaced by a new, working machine.

Long-Term Frequency Interpretations

Probabilities are often interpreted as relative frequencies in an infinitely long string of independent trials. The steady-state probabilities of a Markov chain admit a similar interpretation, despite the absence of independence.

Consider, for example, a Markov chain involving a machine, which at the end of any day can be in one of two states, working or broken-down. Each time it breaks down, it is immediately repaired at a cost of $1. How are we to model the long-term expected cost of repair **per day**? One possibility is to view it as the expected value of the repair cost on a randomly chosen day far into the future; this is just the steady-state probability of the broken down state. Alternatively, we can calculate the total expected repair cost in n days, where n is very large, and divide it by n. Intuition suggests that these two methods of calculation should give the same result. Theory supports this intuition, and in general we have the following interpretation of steady-state probabilities (a justification is given in the theoretical problems section).

Steady-State Probabilities as Expected State Frequencies

For a Markov chain with a single class that is aperiodic, the steady-state probabilities π_j satisfy

$$\pi_j = \lim_{n\to\infty} \frac{v_{ij}(n)}{n},$$

where $v_{ij}(n)$ is the expected value of the number of visits to state j within the first n transitions, starting from state i.

Based on this interpretation, π_j is the long-term expected fraction of time that the state is equal to j. Each time that state j is visited, there is probability p_{jk} that the next transition takes us to state k. We conclude that $\pi_j p_{jk}$ can be viewed as the long-term expected fraction of transitions that move the state from j to k.†

† In fact, some stronger statements are also true. Namely, whenever we carry out the probabilistic experiment and generate a trajectory of the Markov chain over an infinite time horizon, the observed long-term frequency with which state j is visited will be exactly equal to π_j, and the observed long-term frequency of transitions from j to k will be exactly equal to $\pi_j p_{jk}$. Even though the trajectory is random, these equalities hold with certainty, that is, with probability 1. The exact meaning of this statement will become more apparent in the next chapter, when we discuss concepts related to the limiting behavior of random processes.

Expected Frequency of a Particular Transition

Consider n transitions of a Markov chain with a single class that is aperiodic, starting from a given initial state. Let $q_{jk}(n)$ be the expected number of such transitions that take the state from j to k. Then, regardless of the initial state, we have

$$\lim_{n\to\infty} \frac{q_{jk}(n)}{n} = \pi_j p_{jk}.$$

The frequency interpretation of π_j and $\pi_j p_{jk}$ allows for a simple interpretation of the balance equations. The state is equal to j if and only if there is a transition that brings the state to j. Thus, the expected frequency π_j of visits to j is equal to the sum of the expected frequencies $\pi_k p_{kj}$ of transitions that lead to j, and

$$\pi_j = \sum_{k=1}^{m} \pi_k p_{kj};$$

see Fig. 6.13.

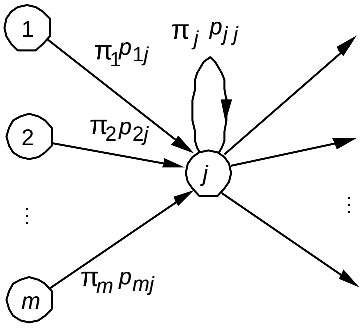


Figure 6.13: Interpretation of the balance equations in terms of frequencies. In a very large number of transitions, there will be a fraction $\pi_k p_{kj}$ that bring the state from k to j. (This also applies to transitions from j to itself, which occur with frequency $\pi_j p_{jj}$.) The sum of the frequencies of such transitions is the frequency π_j of being at state j.

Birth-Death Processes

A **birth-death** process is a Markov chain in which the states are linearly arranged and transitions can only occur to a neighboring state, or else leave the state unchanged. They arise in many contexts, especially in queueing theory.

Figure 6.14 shows the general structure of a birth-death process and also introduces some generic notation for the transition probabilities. In particular,

$$b_i = \mathbf{P}(X_{n+1} = i+1 \mid X_n = i), \qquad (\text{“birth” probability at state } i),$$
$$d_i = \mathbf{P}(X_{n+1} = i-1 \mid X_n = i), \qquad (\text{“death” probability at state } i).$$

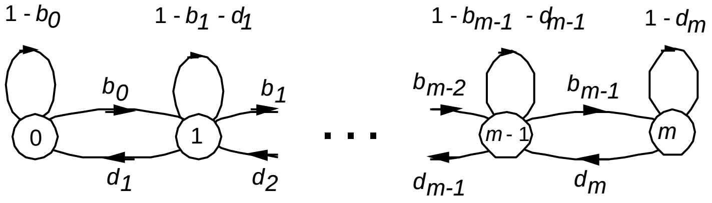


Figure 6.14: Transition probability graph for a birth-death process.

For a birth-death process, the balance equations can be substantially simplified. Let us focus on two neighboring states, say, i and $i+1$. In any trajectory of the Markov chain, a transition from i to $i+1$ has to be followed by a transition from $i+1$ to i, before another transition from i to $i+1$ can occur. Therefore, the frequency of transitions from i to $i+1$, which is $\pi_i b_i$, must be equal to the frequency of transitions from $i+1$ to i, which is $\pi_{i+1} d_{i+1}$. This leads to the **local balance** equations[†]

$$\pi_i b_i = \pi_{i+1} d_{i+1}, \qquad i = 0, 1, \ldots, m-1.$$

Using the local balance equations, we obtain

$$\pi_i = \pi_0 \frac{b_0 b_1 \cdots b_{i-1}}{d_1 d_2 \cdots d_i}, \qquad i = 1, \ldots, m.$$

Together with the normalization equation $\sum_i \pi_i = 1$, the steady-state probabilities π_i are easily computed.

Example 6.9. (Random Walk with Reflecting Barriers) A person walks along a straight line and, at each time period, takes a step to the right with probability b, and a step to the left with probability $1-b$. The person starts in one of

[†] A more formal derivation that does not rely on the frequency interpretation proceeds as follows. The balance equation at state 0 is $\pi_0(1-b_0) + \pi_1 d_1 = \pi_0$, which yields the first local balance equation $\pi_0 b_0 = \pi_1 d_1$.

The balance equation at state 1 is $\pi_0 b_0 + \pi_1(1-b_1-d_1) + \pi_2 d_2 = \pi_1$. Using the local balance equation $\pi_0 b_0 = \pi_1 d_1$ at the previous state, this is rewritten as $\pi_1 d_1 + \pi_1(1-b_1-d_1) + \pi_2 d_2 = \pi_1$, which simplifies to $\pi_1 b_1 = \pi_2 d_2$. We can then continue similarly to obtain the local balance states at all other states.

the positions $1, 2, \ldots, m$, but if he reaches position 0 (or position $m+1$), his step is instantly reflected back to position 1 (or position m, respectively). Equivalently, we may assume that when the person is in positions 1 or m. he will stay in that position with corresponding probability $1-b$ and b, respectively. We introduce a Markov chain model whose states are the positions $1, \ldots, m$. The transition probability graph of the chain is given in Fig. 6.15.

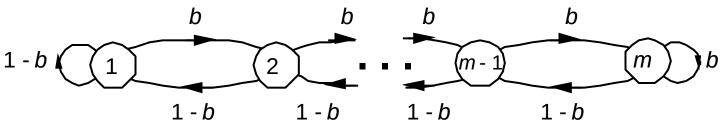


Figure 6.15: Transition probability graph for the random walk Example 6.9.

The local balance equations are

$$\pi_i b = \pi_{i+1}(1-b), \qquad i = 1, \ldots, m-1.$$

Thus, $\pi_{i+1} = \rho\pi_i$, where

$$\rho = \frac{b}{1-b},$$

and we can express all the π_j in terms of π_1, as

$$\pi_i = \rho^{i-1}\pi_1, \qquad i = 1, \ldots, m.$$

Using the normalization equation $1 = \pi_1 + \cdots + \pi_m$, we obtain

$$1 = \pi_1(1 + \rho + \cdots + \rho^{m-1})$$

which leads to

$$\pi_i = \frac{\rho^{i-1}}{1 + \rho + \cdots + \rho^{m-1}}, \qquad i = 1, \ldots, m.$$

Note that if $\rho = 1$, then $\pi_i = 1/m$ for all i.

Example 6.10. (Birth-Death Markov Chains – Queueing) Packets arrive at a node of a communication network, where they are stored in a buffer and then transmitted. The storage capacity of the buffer is m: if m packets are already present, any newly arriving packets are discarded. We discretize time in very small periods, and we assume that in each period, at most one event can happen that can change the number of packets stored in the node (an arrival of a new packet or a completion of the transmission of an existing packet). In particular, we assume that at each period, exactly one of the following occurs:

(a) one new packet arrives; this happens with a given probability $b > 0$;

(b) one existing packet completes transmission; this happens with a given probability $d > 0$ if there is at least one packet in the node, and with probability 0 otherwise;

(c) no new packet arrives and no existing packet completes transmission; this happens with a probability $1 - b - d$ if there is at least one packet in the node, and with probability $1 - b$ otherwise.

We introduce a Markov chain with states $0, 1, \ldots, m$, corresponding to the number of packets in the buffer. The transition probability graph is given in Fig. 6.16.

The local balance equations are

$$\pi_i b = \pi_{i+1} d, \qquad i = 0, 1, \ldots, m-1.$$

We define

$$\rho = \frac{b}{d},$$

and obtain $\pi_{i+1} = \rho\pi_i$, which leads to $\pi_i = \rho^i \pi_0$ for all i. By using the normalization equation $1 = \pi_0 + \pi_1 + \cdots + \pi_m$, we obtain

$$1 = \pi_0(1 + \rho + \cdots + \rho^m),$$

and

$$\pi_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{m+1}} & \text{if } \rho \neq 1, \\ \dfrac{1}{m+1} & \text{if } \rho = 1. \end{cases}$$

The steady-state probabilities are then given by

$$\pi_i = \begin{cases} \dfrac{\rho^i(1-\rho)}{1-\rho^{m+1}} & \text{if } \rho \neq 1, \\ \dfrac{1}{m+1} & \text{if } \rho = 1, \end{cases} \qquad i = 0, 1, \ldots, m.$$

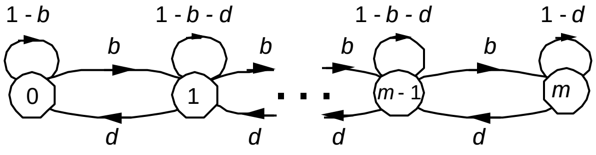


Figure 6.16: Transition probability graph in Example 6.10.

It is interesting to consider what happens when the buffer size m is so large that it can be considered as practically infinite. We distinguish two cases.

(a) Suppose that $b < d$, or $\rho < 1$. In this case, arrivals of new packets are less likely than departures of existing packets. This prevents the number of packets in the buffer from growing, and the steady-state probabilities π_i decrease with i. We observe that as $m \to \infty$, we have $1 - \rho^{m+1} \to 1$, and

$$\pi_i \to \rho^i(1-\rho), \qquad \text{for all } i.$$

We can view these as the steady-state probabilities in a system with an infinite buffer. [As a check, note that we have $\sum_{i=0}^{\infty} \rho^i(1-\rho) = 1$.]

(b) Suppose that $b > d$, or $\rho > 1$. In this case, arrivals of new packets are more likely than departures of existing packets. The number of packets in the buffer tends to increase, and the steady-state probabilities π_i increase with i. As we consider larger and larger buffer sizes m, the steady-state probability of any fixed state i decreases to zero:

$$\pi_i \to 0, \qquad \text{for all } i.$$

Were we to consider a system with an infinite buffer, we would have a Markov chain with a countably infinite number of states. Although we do not have the machinery to study such chains, the preceding calculation suggests that every state will have zero steady-state probability and will be "transient." The number of packets in queue will generally grow to infinity, and any particular state will be visited only a finite number of times.

6.4 ABSORPTION PROBABILITIES AND EXPECTED TIME TO ABSORPTION

In this section, we study the short-term behavior of Markov chains. We first consider the case where the Markov chain starts at a transient state. We are interested in the first recurrent state to be entered, as well as in the time until this happens.

When focusing on such questions, the subsequent behavior of the Markov chain (after a recurrent state is encountered) is immaterial. We can therefore assume, without loss of generality, that every recurrent state k is **absorbing**, i.e.,

$$p_{kk} = 1, \qquad p_{kj} = 0 \text{ for all } j \neq k.$$

If there is a unique absorbing state k, its steady-state probability is 1 (because all other states are transient and have zero steady-state probability), and will be reached with probability 1, starting from any initial state. If there are multiple absorbing states, the probability that one of them will be eventually reached is still 1, but the identity of the absorbing state to be entered is random and the associated probabilities may depend on the starting state. In the sequel, we fix a particular absorbing state, denoted by s, and consider the absorption probability a_i that s is eventually reached, starting from i:

$$a_i = \mathbf{P}(X_n \text{ eventually becomes equal to the absorbing state } s \,|\, X_0 = i).$$

Absorption probabilities can be obtained by solving a system of linear equations, as indicated below.

Absorption Probability Equations

Consider a Markov chain in which each state is either transient or absorbing. We fix a particular absorbing state s. Then, the probabilities a_i of eventually reaching state s, starting from i, are the unique solution of the equations

$$\begin{aligned} a_s &= 1, \\ a_i &= 0, \qquad \text{for all absorbing } i \neq s, \\ a_i &= \sum_{j=1}^{m} p_{ij} a_j, \qquad \text{for all transient } i. \end{aligned}$$

The equations $a_s = 1$, and $a_i = 0$, for all absorbing $i \neq s$, are evident from the definitions. To verify the remaining equations, we argue as follows. Let us consider a transient state i and let A be the event that state s is eventually reached. We have

$$\begin{aligned} a_i &= \mathbf{P}(A \,|\, X_0 = i) \\ &= \sum_{j=1}^{m} \mathbf{P}(A \,|\, X_0 = i, X_1 = j)\mathbf{P}(X_1 = j \,|\, X_0 = i) \qquad \text{(total probability thm.)} \\ &= \sum_{j=1}^{m} \mathbf{P}(A \,|\, X_1 = j) p_{ij} \qquad \text{(Markov property)} \\ &= \sum_{j=1}^{m} a_j p_{ij}. \end{aligned}$$

The uniqueness property of the solution of the absorption probability equations requires a separate argument, which is given in the theoretical problems section.

The next example illustrates how we can use the preceding method to calculate the probability of entering a given recurrent class (rather than a given absorbing state).

Example 6.11. Consider the Markov chain shown in Fig. 6.17(a). We would like to calculate the probability that the state eventually enters the recurrent class

$\{4, 5\}$ starting from one of the transient states. For the purposes of this problem, the possible transitions within the recurrent class $\{4, 5\}$ are immaterial. We can therefore lump the states in this recurrent class and treat them as a single absorbing state (call it state 6); see Fig. 6.17(b). It then suffices to compute the probability of eventually entering state 6 in this new chain.

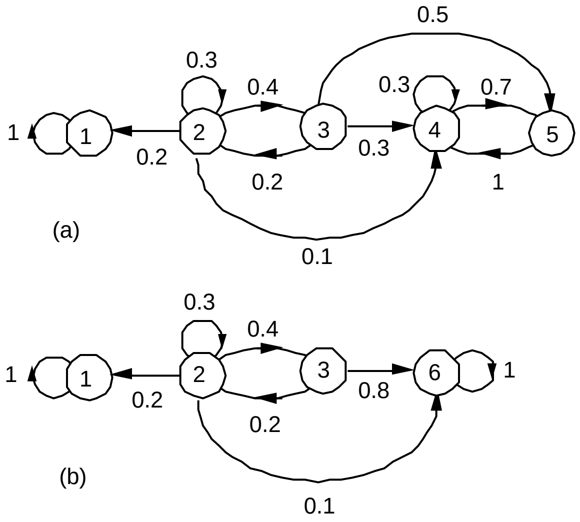


Figure 6.17: (a) Transition probability graph in Example 6.11. (b) A new graph in which states 4 and 5 have been lumped into the absorbing state $s = 6$.

The absorption probabilities a_i of eventually reaching state $s = 6$ starting from state i, satisfy the following equations:

$$a_2 = 0.2a_1 + 0.3a_2 + 0.4a_3 + 0.1a_6,$$
$$a_3 = 0.2a_2 + 0.8a_6.$$

Using the facts $a_1 = 0$ and $a_6 = 1$, we obtain

$$a_2 = 0.3a_2 + 0.4a_3 + 0.1,$$
$$a_3 = 0.2a_2 + 0.8.$$

This is a system of two equations in the two unknowns a_2 and a_3, which can be readily solved to yield $a_2 = 21/31$ and $a_3 = 29/31$.

Example 6.12. (Gambler's Ruin) A gambler wins \$1 at each round, with probability p, and loses \$1, with probability $1 - p$. Different rounds are assumed independent. The gambler plays continuously until he either accumulates a target amount of \$$m$, or loses all his money. What is the probability of eventually accumulating the target amount (winning) or of losing his fortune?

We introduce the Markov chain shown in Fig. 6.18 whose state i represents the gambler's wealth at the beginning of a round. The states $i = 0$ and $i = m$ correspond to losing and winning, respectively.

All states are transient, except for the winning and losing states which are absorbing. Thus, the problem amounts to finding the probabilities of absorption at each one of these two absorbing states. Of course, these absorption probabilities depend on the initial state i.

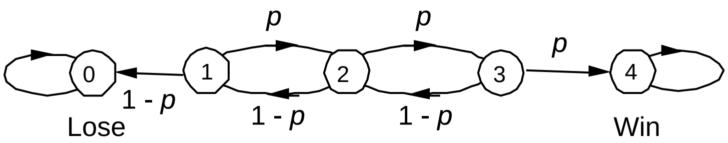


Figure 6.18: Transition probability graph for the gambler's ruin problem (Example 6.12). Here $m = 4$.

Let us set $s = 0$ in which case the absorption probability a_i is the probability of losing, starting from state i. These probabilities satisfy

$$\begin{aligned} a_0 &= 1, \\ a_i &= (1-p)a_{i-1} + pa_{i+1}, \qquad i = 1, \ldots, m-1, \\ a_m &= 0. \end{aligned}$$

These equations can be solved in a variety of ways. It turns out there is an elegant method that leads to a nice closed form solution.

Let us write the equations for the a_i as

$$(1-p)(a_{i-1} - a_i) = p(a_i - a_{i+1}), \qquad i = 1, \ldots, m-1.$$

Then, by denoting

$$\delta_i = a_i - a_{i+1}, \qquad i = 1, \ldots, m-1,$$

and

$$\rho = \frac{1-p}{p},$$

the equations are written as

$$\delta_i = \rho\delta_{i-1}, \qquad i = 1, \ldots, m-1,$$

from which we obtain

$$\delta_i = \rho^i\delta_0, \qquad i = 1, \ldots, m-1.$$

This, together with the equation $\delta_0 + \delta_1 + \cdots + \delta_{m-1} = a_0 - a_m = 1$, implies that

$$(1 + \rho + \cdots + \rho^{m-1})\delta_0 = 1.$$

Thus, we have

$$\delta_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^m} & \text{if } \rho \neq 1, \\ \dfrac{1}{m} & \text{if } \rho = 1, \end{cases}$$

and, more generally,

$$\delta_i = \begin{cases} \dfrac{\rho^i(1-\rho)}{1-\rho^m} & \text{if } \rho \neq 1, \\ \dfrac{1}{m} & \text{if } \rho = 1. \end{cases}$$

From this relation, we can calculate the probabilities a_i. If $\rho \neq 1$, we have

$$\begin{aligned} a_i &= a_0 - \delta_{i-1} - \cdots - \delta_0 \\ &= 1 - (\rho^{i-1} + \cdots + \rho + 1)\delta_0 \\ &= 1 - \frac{1-\rho^i}{1-\rho} \cdot \frac{1-\rho}{1-\rho^m}, \\ &= 1 - \frac{1-\rho^i}{1-\rho^m}, \end{aligned}$$

and finally the probability of losing, starting from a fortune i, is

$$a_i = \frac{\rho^i - \rho^m}{1-\rho^m}, \qquad i = 1, \ldots, m-1.$$

If $\rho = 1$, we similarly obtain

$$a_i = \frac{m-i}{m}.$$

The probability of winning, starting from a fortune i, is the complement $1 - a_i$, and is equal to

$$1 - a_i = \begin{cases} \dfrac{1-\rho^i}{1-\rho^m} & \text{if } \rho \neq 1, \\ \dfrac{i}{m} & \text{if } \rho = 1. \end{cases}$$

The solution reveals that if $\rho > 1$, which corresponds to $p < 1/2$ and unfavorable odds for the gambler, the probability of losing approaches 1 as $m \to \infty$ regardless of the size of the initial fortune. This suggests that if you aim for a large profit under unfavorable odds, financial ruin is almost certain.

Expected Time to Absorption

We now turn our attention to the expected number of steps until a recurrent state is entered (an event that we refer to as "absorption"), starting from a particular transient state. For any state i, we denote

$$\begin{aligned}\mu_i &= \mathbf{E}\big[\text{number of transitions until absorption, starting from } i\big]\\ &= \mathbf{E}\big[\min\{n \geq 0 \,|\, X_n \text{ is recurrent}\} \bigm| X_0 = i\big].\end{aligned}$$

If i is recurrent, this definition sets μ_i to zero.

We can derive equations for the μ_i by using the total expectation theorem. We argue that the time to absorption starting from a transient state i is equal to 1 plus the expected time to absorption starting from the next state, which is j with probability p_{ij}. This leads to a system of linear equations which is stated below. It turns out that these equations have a unique solution, but the argument for establishing this fact is beyond our scope.

Equations for the Expected Time to Absorption

The expected times μ_i to absorption, starting from state i are the unique solution of the equations

$$\mu_i = 0, \qquad \text{for all recurrent states } i,$$

$$\mu_i = 1 + \sum_{j=1}^{m} p_{ij}\mu_j, \qquad \text{for all transient states } i.$$

Example 6.13. (Spiders and Fly) Consider the spiders-and-fly model of Example 6.2. This corresponds to the Markov chain shown in Fig. 6.19. The states correspond to possible fly positions, and the absorbing states 1 and m correspond to capture by a spider.

Let us calculate the expected number of steps until the fly is captured. We have

$$\mu_1 = \mu_m = 0,$$

and

$$\mu_i = 1 + 0.3 \cdot \mu_{i-1} + 0.4 \cdot \mu_i + 0.3 \cdot \mu_{i+1}, \qquad \text{for } i = 2, \ldots, m-1.$$

We can solve these equations in a variety of ways, such as for example by successive substitution. As an illustration, let $m = 4$, in which case, the equations reduce to

$$\mu_2 = 1 + 0.4 \cdot \mu_2 + 0.3 \cdot \mu_3, \qquad \mu_3 = 1 + 0.3 \cdot \mu_2 + 0.4 \cdot \mu_3.$$

The first equation yields $\mu_2 = (1/0.6) + (1/2)\mu_3$, which we can substitute in the second equation and solve for μ_3. We obtain $\mu_3 = 10/3$ and by substitution again, $\mu_2 = 10/3$.

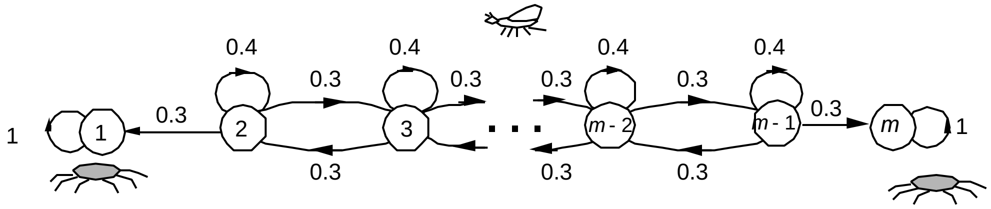


Figure 6.19: Transition probability graph in Example 6.13.

Mean First Passage Times

The same idea used to calculate the expected time to absorption can be used to calculate the expected time to reach a particular recurrent state, starting from any other state. Throughout this subsection, we consider a Markov chain with a single recurrent class. We focus on a special recurrent state s, and we denote by t_i the **mean first passage time from state** i **to state** s, defined by

$$\begin{aligned} t_i &= \mathbf{E}\big[\text{number of transitions to reach } s \text{ for the first time, starting from } i\big] \\ &= \mathbf{E}\big[\min\{n \geq 0 \,|\, X_n = s\} \;\big|\; X_0 = i\big]. \end{aligned}$$

The transitions out of state s are irrelevant to the calculation of the mean first passage times. We may thus consider a new Markov chain which is identical to the original, except that the special state s is converted into an absorbing state (by setting $p_{ss} = 1$, and $p_{sj} = 0$ for all $j \neq s$). We then compute t_i as the expected number of steps to absorption starting from i, using the formulas given earlier in this section. We have

$$\begin{aligned} t_i &= 1 + \sum_{j=1}^{m} p_{ij} t_j, \qquad \text{for all } i \neq s, \\ t_s &= 0. \end{aligned}$$

This system of linear equations can be solved for the unknowns t_i, and is known to have a unique solution.

The above equations give the expected time to reach the special state s starting from any other state. We may also want to calculate the **mean recurrence time** of the special state s, which is defined as

$$\begin{aligned} t_s^* &= \mathbf{E}[\text{number of transitions up to the first return to } s, \text{ starting from } s] \\ &= \mathbf{E}\big[\min\{n > 1 \,|\, X_n = s\} \;\big|\; X_0 = s\big]. \end{aligned}$$

We can obtain t_s^*, once we have the first passage times t_i, by using the equation

$$t_s^* = 1 + \sum_{j=1}^{m} p_{sj} t_j.$$

To justify this equation, we argue that the time to return to s, starting from s, is equal to 1 plus the expected time to reach s from the next state, which is j with probability p_{sj}. We then apply the total expectation theorem.

Example 6.14. Consider the "up-to-date"–"behind" model of Example 6.1. States 1 and 2 correspond to being up-to-date and being behind, respectively, and the transition probabilities are

$$p_{11} = 0.8, \qquad p_{12} = 0.2,$$
$$p_{21} = 0.6, \qquad p_{22} = 0.4.$$

Let us focus on state $s = 1$ and calculate the mean first passage time to state 1, starting from state 2. We have $t_1 = 0$ and

$$t_2 = 1 + p_{21}t_1 + p_{22}t_2 = 1 + 0.4 \cdot t_2,$$

from which

$$t_2 = \frac{1}{0.6} = \frac{5}{3}.$$

The mean recurrence time to state 1 is given by

$$t_1^* = 1 + p_{11}t_1 + p_{12}t_2 = 1 + 0 + 0.2 \cdot \frac{5}{3} = \frac{4}{3}.$$

Summary of Facts About Mean First Passage Times

Consider a Markov chain with a single recurrent class, and let s be a particular recurrent state.

- The mean first passage times t_i to reach state s starting from i, are the unique solution to the system of equations

$$t_s = 0, \qquad t_i = 1 + \sum_{j=1}^{m} p_{ij} t_j, \qquad \text{for all } i \neq s.$$

- The mean recurrence time t_s^* of state s is given by

$$t_s^* = 1 + \sum_{j=1}^{m} p_{sj} t_j.$$

6.5 MORE GENERAL MARKOV CHAINS

The discrete-time, finite-state Markov chain model that we have considered so far is the simplest example of an important Markov process. In this section, we briefly discuss some generalizations that involve either a countably infinite number of states or a continuous time, or both. A detailed theoretical development for these types of models is beyond our scope, so we just discuss their main underlying ideas, relying primarily on examples.

Chains with Countably Infinite Number of States

Consider a Markov process $\{X_1, X_2, \ldots\}$ whose state can take any positive integer value. The transition probabilities

$$p_{ij} = \mathbf{P}(X_{n+1} = j \,|\, X_n = i), \qquad i, j = 1, 2, \ldots$$

are given, and can be used to represent the process with a transition probability graph that has an infinite number of nodes, corresponding to the integers $1, 2, \ldots$

It is straightforward to verify, using the total probability theorem in a similar way as in Section 6.1, that the n-step transition probabilities

$$r_{ij}(n) = \mathbf{P}(X_n = j \,|\, X_0 = i), \qquad i, j = 1, 2, \ldots$$

satisfy the Chapman-Kolmogorov equations

$$r_{ij}(n+1) = \sum_{k=1}^{\infty} r_{ik}(n) p_{kj}, \qquad i, j = 1, 2, \ldots$$

Furthermore, if the $r_{ij}(n)$ converge to steady-state values π_j as $n \to \infty$, then by taking limit in the preceding equation, we obtain

$$\pi_j = \sum_{k=1}^{\infty} \pi_k p_{kj}, \qquad i, j = 1, 2, \ldots$$

These are the balance equations for a Markov chain with states $1, 2, \ldots$

It is important to have conditions guaranteeing that the $r_{ij}(n)$ indeed converge to steady-state values π_j as $n \to \infty$. As we can expect from the finite-state case, such conditions should include some analog of the requirement that there is a single recurrent class that is aperiodic. Indeed, we require that:

(a) each state is accessible from every other state;

(b) the set of all states is aperiodic in the sense that there is no $d > 1$ such that the states can be grouped in $d > 1$ disjoint subsets $S_1, \ldots, S_d$ so that all transitions from one subset lead to the next subset.

These conditions are sufficient to guarantee the convergence to a steady-state

$$\lim_{n\to\infty} r_{ij}(n) = \pi_j, \qquad i, j = 1, 2, \ldots$$

but something peculiar may also happen here, which is not possible if the number of states is finite: the limits π_j may not add to 1, so that $(\pi_1, \pi_2, \ldots)$ may not be a probability distribution. In fact, we can prove the following theorem (the proof is beyond our scope).

Steady-State Convergence Theorem

Under the above accessibility and aperiodicity assumptions (a) and (b), there are only two possibilities:

(1) The $r_{ij}(n)$ converge to a steady state probability distribution $(\pi_1, \pi_2, \ldots)$. In this case the π_j uniquely solve the balance equations together with the normalization equation $\pi_1 + \pi_2 + \cdots = 1$. Furthermore, the π_j have an expected frequency interpretation:

$$\pi_j = \lim_{n\to\infty} \frac{v_{ij}(n)}{n},$$

where $v_{ij}(n)$ is the expected number of visits to state j within the first n transitions, starting from state i.

(2) All the $r_{ij}(n)$ converge to 0 as $n \to \infty$ and the balance equations have no solution, other than $\pi_j = 0$ for all j.

For an example of possibility (2) above, consider the packet queueing system of Example 6.10 for the case where the probability b of a packet arrival in each period is larger than the probability d of a departure. Then, as we saw in that example, as the buffer size m increases, the size of the queue will tend to increase without bound, and the steady-state probability of any one state will tend to 0 as $m \to \infty$. In effect, with infinite buffer space, the system is "unstable" when $b > d$, and all states are "transient."

An important consequence of the steady-state convergence theorem is that if we can find a probability distribution $(\pi_1, \pi_2, \ldots)$ that solves the balance equations, then we can be sure that it is the steady-state distribution. This line of argument is very useful in queueing systems as illustrated in the following two examples.

Example 6.15. (Queueing with Infinite Buffer Space) Consider, as in Example 6.10, a communication node, where packets arrive and are stored in a buffer before getting transmitted. We assume that the node can store an infinite number of packets. We discretize time in very small periods, and we assume that in each period, one of the following occurs:

(a) one new packet arrives; this happens with a given probability $b > 0$;

(b) one existing packet completes transmission; this happens with a given probability $d > 0$ if there is at least one packet in the node, and with probability 0 otherwise;

(c) no new packet arrives and no existing packet completes transmission; this happens with a probability $1 - b - d$ if there is at least one packet in the node, and with probability $1 - b$ otherwise.

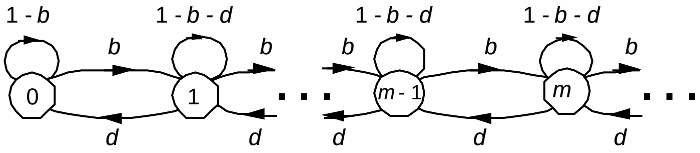


Figure 6.20: Transition probability graph in Example 6.15.

We introduce a Markov chain with states are $0, 1, \ldots$, corresponding to the number of packets in the buffer. The transition probability graph is given in Fig. 6.20. As in the case of a finite number of states, the local balance equations are

$$\pi_i b = \pi_{i+1} d, \qquad i = 0, 1, \ldots,$$

and we obtain $\pi_{i+1} = \rho \pi_i$, where $\rho = b/d$. Thus, we have $\pi_i = \rho^i \pi_0$ for all i. If $\rho < 1$, the normalization equation $1 = \sum_{i=0}^{\infty} \pi_i$ yields

$$1 = \pi_0 \sum_{i=0}^{\infty} \rho^i = \frac{\pi_0}{1 - \rho},$$

in which case $\pi_0 = 1 - \rho$, and the steady-state probabilities are

$$\pi_i = \rho^i (1 - \rho), \qquad i = 0, 1, \ldots$$

If $\rho \geq 1$, which corresponds to the case where the arrival probability b is no less than the departure probability d, the normalization equation $1 = \pi_0(1 + \rho + \rho^2 + \cdots)$ implies that $\pi_0 = 0$, and also $\pi_i = \rho^i \pi_0 = 0$ for all i.

Example 6.16. (The $M/G/1$ Queue) Packets arrive at a node of a communication network, where they are stored at an infinite capacity buffer and are then transmitted one at a time. The arrival process of the packets is Poisson with rate

λ, and the transmission time of a packet has a given CDF. Furthermore, the transmission times of different packets are independent and are also independent from all the interarrival times of the arrival process.

This queueing system is known as the $M/G/1$ system. With changes in terminology, it applies to many different practical contexts where "service" is provided to "arriving customers," such as in communication, transportation, and manufacturing, among others. The name $M/G/1$ is an example of shorthand terminology from queueing theory, whereby the first letter (M in this case) characterizes the customer arrival process (Poisson in this case), the second letter (G in this case) characterizes the distribution of the service time of the queue (general in this case), and the number (1 in this case) characterizes the number of customers that can be simultaneously served.

To model this system as a discrete-time Markov chain, we focus on the time instants when a packet completes transmission and departs from the system. We denote by X_n the number of packets in the system just after the nth customer's departure. We have

$$X_{n+1} = \begin{cases} X_n - 1 + S_n & \text{if } X_n > 0, \\ S_n & \text{if } X_n = 0, \end{cases}$$

where S_n is the number of packet arrivals during the $(n+1)$st packet's transmission. In view of the Poisson assumption, the random variables $S_1, S_2, \ldots$ are independent and their PMF can be calculated using the given CDF of the transmission time, and the fact that in an interval of length r, the number of packet arrivals is Poisson-distributed with parameter λr. In particular, let us denote

$$\alpha_k = \mathbf{P}(S_n = k), \qquad k = 0, 1, \ldots,$$

and let us assume that if the transmission time R of a packet is a discrete random variable taking the values $r_1, \ldots, r_m$ with probabilities $p_1, \ldots, p_m$. Then, we have for all $k \geq 0$,

$$\alpha_k = \sum_{j=1}^{m} p_j \frac{e^{-\lambda r_j}(\lambda r_j)^k}{k!},$$

while if R is a continuous random variable with PDF $f_R(r)$, we have for all $k \geq 0$,

$$\alpha_k = \int_{r=0}^{\infty} \mathbf{P}(S_n = k \,|\, R = r) f_R(r)\,dr = \int_{r=0}^{\infty} \frac{e^{-\lambda r}(\lambda r)^k}{k!} f_R(r)\,dr.$$

The probabilities α_k define in turn the transition probabilities of the Markov chain $\{X_n\}$, as follows (see Fig. 6.21):

$$p_{ij} = \begin{cases} \alpha_j & \text{if } i = 0 \text{ and } j > 0, \\ \alpha_{j-i+1} & \text{if } i > 0 \text{ and } j \geq i - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, this Markov chain satisfies the accessibility and aperiodicity conditions that guarantee steady-state convergence. There are two possibilities: either $(\pi_0, \pi_1, \ldots)$ form a probability distribution, or else $\pi_j > 0$ for all j. We will clarify

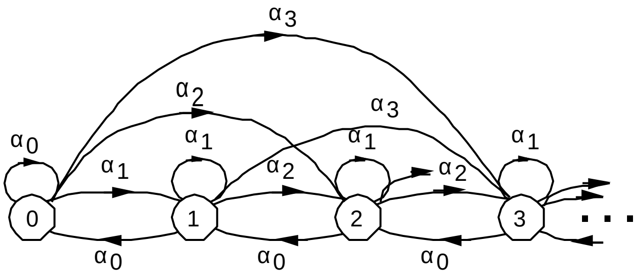


Figure 6.21: Transition probability graph for the number of packets left behind by a packet completing transmission in the $M/G/1$ queue (Example 6.16).

the conditions under which each of these cases holds, and we will also calculate the transform $M(s)$ (when it exists) of the steady-state distribution $(\pi_0, \pi_1, \ldots)$:

$$M(s) = \sum_{j=0}^{\infty} \pi_j e^{sj}.$$

For this purpose, we will use the transform of the PMF $\{\alpha_k\}$:

$$A(s) = \sum_{j=0}^{\infty} \alpha_j e^{sj}.$$

Indeed, let us multiply the balance equations

$$\pi_j = \pi_0 \alpha_j + \sum_{i=1}^{j+1} \pi_i \alpha_{j-i+1},$$

with e^{sj} and add over all j. We obtain

$$\begin{aligned} M(s) &= \sum_{j=0}^{\infty} \pi_0 \alpha_j e^{sj} + \sum_{j=0}^{\infty} \left(\sum_{i=1}^{j+1} \pi_i \alpha_{j-i+1} \right) e^{sj} \\ &= A(s) + \sum_{i=1}^{\infty} \pi_i e^{s(i-1)} \sum_{j=i-1}^{\infty} \alpha_{j-i+1} e^{s(j-i+1)} \\ &= A(s) + \frac{A(s)}{e^s} \sum_{i=1}^{\infty} \pi_i e^{si} \\ &= A(s) + \frac{A(s)\big(M(s) - \pi_0\big)}{e^s}, \end{aligned}$$

or

$$M(s) = \frac{(e^s - 1)\pi_0 A(s)}{e^s - A(s)}.$$

To calculate π_0, we take the limit as $s \to 0$ in the above formula, and we use the fact $M(0) = 1$ when $\{\pi_j\}$ is a probability distribution. We obtain, using the fact $A(0) = 1$ and L'Hospital's rule,

$$1 = \lim_{s \to 0} \frac{(e^s - 1)\pi_0 A(s)}{e^s - A(s)} = \frac{\pi_0}{1 - \big(dA(s)/ds\big)\big|_{s=0}} = \frac{\pi_0}{1 - \mathbf{E}[N]},$$

where $\mathbf{E}[N] = \sum_{j=0}^{\infty} j\alpha_j$ is the expected value of the number N of packet arrivals within a packet's transmission time. Using the iterated expectations formula, we have

$$\mathbf{E}[N] = \lambda \mathbf{E}[R],$$

where $\mathbf{E}[R]$ is the expected value of the transmission time. Thus,

$$\pi_0 = 1 - \lambda \mathbf{E}[R],$$

and the transform of the steady-state distribution $\{\pi_j\}$ is

$$M(s) = \frac{(e^s - 1)\big(1 - \lambda \mathbf{E}[R]\big)A(s)}{e^s - A(s)}.$$

For the above calculation to be correct, we must have $\mathbf{E}[N] < 1$, i.e., packets should arrive at a rate that is smaller than the transmission rate of the node. If this is not true, the system is not "stable" and there is no steady-state distribution, i.e., the only solution of the balance equations is $\pi_j = 0$ for all j.

Let us finally note that we have introduced the π_j as the steady-state probability that j packets are left behind in the system by a packet upon completing transmission. However, it turns out that π_j is also equal to the steady-state probability of j packets found in the system by an observer that looks at the system at a "typical" time far into the future. This is discussed in the theoretical problems, but to get an idea of the underlying reason, note that for each time the number of packets in the system increases from n to $n + 1$ due to an arrival, there will be a corresponding future decrease from $n + 1$ to n due to a departure. Therefore, in the long run, the frequency of transitions from n to $n + 1$ is equal to the frequency of transitions from $n + 1$ to n. Therefore, in steady-state, the system appears statistically identical to an arriving and to a departing packet. Now, because the packet interarrival times are independent and exponentially distributed, the times of packet arrivals are "typical" and do not depend on the number of packets in the system. With some care this argument can be made precise, and shows that at the times when packets complete their transmissions and depart, the system is "typically loaded."

Continuous-Time Markov Chains

We have implicitly assumed so far that the transitions between states take unit time. When the time between transitions takes values from a continuous range, some new questions arise. For example, what is the proportion of time that the system spends at a particular state (as opposed to the frequency of visits into the state)?

Let the states be denoted by $1, 2, \ldots$, and let us assume that state transitions occur at discrete times, but the time from one transition to the next is random. In particular, we assume that:

(a) If the current state is i, the next state will be j with a given probability p_{ij}.

(b) The time interval Δ_i between the transition to state i and the transition to the next state is exponentially distributed with a given parameter ν_i:

$$\mathbf{P}(\Delta_i \leq \delta \,|\, \text{current state is } i) \leq 1 - e^{-\nu_i \delta}.$$

Furthermore, Δ_i is independent of earlier transition times and states.

The parameter ν_i is referred to as the *transition rate associated with state* i. Since the expected transition time is

$$\mathbf{E}[\Delta_i] = \int_0^\infty \delta \nu_i e^{-\nu_i \delta} d\delta = \frac{1}{\nu_i},$$

we can interpret ν_i as the average number of transitions per unit time. We may also view

$$q_{ij} = p_{ij} \nu_i$$

as the rate at which the process makes a transition to j when at state i. Consequently, we call q_{ij} the *transition rate from i to j*. Note that given the transition rates q_{ij}, one can obtain the node transition rates using the formula $\nu_i = \sum_{j=1}^\infty q_{ij}$.

The state of the chain at time $t \geq 0$ is denoted by $X(t)$, and stays constant between transitions. Let us recall the memoryless property of the exponential distribution, which in our context implies that, for any time t between the kth and $(k+1)$st transition times t_k and t_{k+1}, the additional time $t_{k+1} - t$ needed to effect the next transition is independent of the time $t - t_k$ that the system has been in the current state. This implies the Markov character of the process, i.e., that at any time $\overline{t}$, the future of the process, [the random variables $X(t)$ for $t > \overline{t}$] depend on the past of the process [the values of the random variables $X(t)$ for $t \leq \overline{t}$] only through the present value of $X(\overline{t})$.

Example 6.17. (The M/M/1 Queue) Packets arrive at a node of a communication network according to a Poisson process with rate λ. The packets are stored at an infinite capacity buffer and are then transmitted one at a time. The transmission time of a packet is exponentially distributed with parameter μ, and the transmission times of different packets are independent and are also independent from all the interarrival times of the arrival process. Thus, this queueing system is identical to the special case of the $M/G/1$ system, where the transmission times are exponentially distributed (this is indicated by the second M in the $M/M/1$ name).

We will model this system using a continuous-time process with state $X(t)$ equal to the number of packets in the system at time t [if $X(t) > 0$, then $X(t) - 1$ packets are waiting in the queue and one packet is under transmission]. The state increases by one when a new packet arrives and decreases by one when an existing packet departs. To show that this process is a continuous-time Markov chain, let us identify the transition rates ν_i and q_{ij} at each state i.

Consider first the case where at some time $\overline{t}$, the system becomes empty, i.e., the state becomes equal to 0. Then the next transition will occur at the next arrival, which will happen in time that is exponentially distributed with parameter λ. Thus at state 0, we have the transition rates

$$q_{0j} = \begin{cases} \lambda & \text{if } j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Consider next the case of a positive state i, and suppose that a transition occurs at some time $\overline{t}$ to $X(\overline{t}) = i$. If the next transition occurs at time $\overline{t}+\Delta_i$, then Δ_i is the minimum of two exponentially distributed random variables: the time to the next arrival, call it Y, which has parameter λ, and the time to the next departure, call it Z, which has parameter μ. (We are again using here the memoryless property of the exponential distribution.) Thus according to Example 5.15, which deals with “competing exponentials,” the time Δ_i is exponentially distributed with parameter $\nu_i = \lambda + \mu$. Furthermore, the probability that the next transition corresponds to an arrival is

$$\begin{aligned} \mathbf{P}(Y \le Z) &= \int_{y \le z} \lambda e^{-\lambda y} \cdot \mu e^{\mu z} \, dy \, dz \\ &= \lambda\mu \int_0^\infty e^{-\lambda y} \left(\int_y^\infty e^{\mu z} \, dz \right) dy \\ &= \lambda\mu \int_0^\infty e^{-\lambda y} \left(\frac{e^{-\mu y}}{\mu} \right) dy \\ &= \lambda \int_0^\infty e^{-(\lambda+\mu) y} \, dy \\ &= \frac{\lambda}{\lambda + \mu}. \end{aligned}$$

We thus have for $i > 0$, $q_{i,i+1} = \nu_i \mathbf{P}(Y \le Z) = (\lambda + \mu)\big(\lambda/(\lambda + \mu)\big) = \lambda$. Similarly, we obtain that the probability that the next transition corresponds to a departure is $\mu/(\lambda + \mu)$, and we have $q_{i,i-1} = \nu_i \mathbf{P}(Y \ge Z) = (\lambda + \mu)\big(\mu/(\lambda + \mu)\big) = \mu$. Thus

$$q_{ij} = \begin{cases} \lambda & \text{if } j = i + 1, \\ \mu & \text{if } j = i - 1, \\ 0 & \text{otherwise.} \end{cases}$$

The positive transition rates q_{ij} are recorded next to the arcs (i, j) of the transition diagram, as in Fig. 6.22.

We will be interested in chains for which the discrete-time Markov chain corresponding to the transition probabilities p_{ij} satisfies the accessibility and

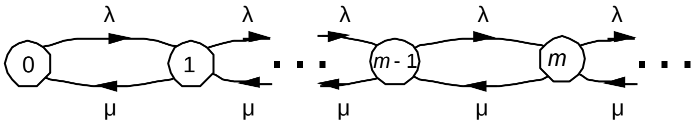


Figure 6.22: Transition graph for the $M/M/1$ queue (Example 6.17).

aperiodicity assumptions of the preceding section. We also require a technical condition, namely that the number of transitions in any finite length of time is finite with probability one. Almost all models of practical use satisfy this condition, although it is possible to construct examples that do not.

Under the preceding conditions, it can be shown that the limit

$$\pi_j = \lim_{t\to\infty} P\big(X(t) = j \,|\, X(0) = i\big)$$

exists and is independent of the initial state i. We refer to π_j as the steady-state probability of state j. It can be shown that if $T_j(t)$ is the expected value of the time spent in state j up to time t, then, regardless of the initial state, we have

$$\pi_j = \lim_{t\to\infty} \frac{T_j(t)}{t}$$

that is, π_j can be viewed as the long-term proportion of time the process spends in state j.

The balance equations for a continuous-time Markov chain take the form

$$p_j \sum_{i=0}^{\infty} q_{ji} = \sum_{i=0}^{\infty} p_i q_{ij}, \qquad j = 0, 1, \ldots$$

Similar to discrete-time Markov chains, it can be shown that there are two possibilities:

(1) The steady-state probabilities are all positive and solve uniquely the balance equations together with the normalization equation $\pi_1 + \pi_2 + \cdots = 1$.

(2) The steady-state probabilities are all zero.

To interpret the balance equations, we note that since π_i is the proportion of time the process spends in state i, it follows that $\pi_i q_{ij}$ can be viewed as frequency of transitions from i to j (expected number of transitions from i to j per unit time). It is seen therefore that the balance equations express the intuitive fact that the frequency of transitions out of state j (the left side term $\pi_j \sum_{i=1}^{\infty} q_{ji}$) is equal to the frequency of transitions into state j (the right side term $\sum_{i=0}^{\infty} \pi_i q_{ij}$).

The continuous-time analog of the local balance equations for discrete-time chains is

$$\pi_j q_{ji} = \pi_i q_{ij}, \qquad i, j = 1, 2, \ldots$$

These equations hold in birth-death systems where $q_{ij} = 0$ for $|i-j| > 1$, but need not hold in other types of Markov chains. They express the fact that the frequencies of transitions from i to j and from j to i are equal.

To understand the relationship between the balance equations for continuous-time chains and the balance equations for discrete-time chains, consider any $\delta > 0$, and the discrete-time Markov chain $\{Z_n \,|\, n \geq 0\}$, where

$$Z_n = X(n\delta), \qquad n = 0, 1, \ldots$$

The steady-state distribution of $\{Z_n\}$ is clearly $\{\pi_j \,|\, j \geq 0\}$, the steady-state distribution of the continuous chain. The transition probabilities of $\{Z_n \,|\, n \geq 0\}$ can be derived by using the properties of the exponential distribution. We obtain

$$\overline{p}_{ij} = \delta q_{ij} + o(\delta), \qquad i \neq j,$$
$$\overline{p}_{jj} = 1 - \delta \sum_{\substack{i=0 \\ i \neq j}}^{\infty} q_{ji} + o(\delta)$$

Using these expressions, the balance equations

$$\pi_j = \sum_{i=0}^{\infty} \pi_i \, \overline{p}_{ij} \qquad j \geq 0$$

for the discrete-time chain $\{Z_n\}$, we obtain

$$\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij} = p_j \big(1 - \delta \sum_{\substack{i=0 \\ i \neq j}}^{\infty} q_{ji} + o(\delta)\big) + \sum_{\substack{i=0 \\ i \neq j}} p_i \big(\delta q_{ij} + o(\delta)\big).$$

Taking the limit as $\delta \to 0$, we obtain the balance equations for the continuous-time chain.

Example 6.18. (The $M/M/1$ Queue – Continued) As in the case of a finite number of states, the local balance equations are

$$\pi_i \lambda = \pi_{i+1} \mu, \qquad i = 0, 1, \ldots,$$

and we obtain $\pi_{i+1} = \rho \pi_i$, where $\rho = \lambda/\mu$. Thus, we have $\pi_i = \rho^i \pi_0$ for all i. If $\rho < 1$, the normalization equation $1 = \sum_{i=0}^{\infty} \pi_i$ yields

$$1 = \pi_0 \sum_{i=0}^{\infty} \rho^i = \frac{\pi_0}{1-\rho},$$

in which case $\pi_0 = 1 - \rho$, and the steady-state probabilities are

$$\pi_i = \rho^i (1 - \rho), \qquad i = 0, 1, \ldots$$

If $\rho \geq 1$, which corresponds to the case where the arrival probability b is no less than the departure probability d, the normalization equation $1 = \pi_0(1+\rho+\rho^2+\cdots)$ implies that $\pi_0 = 0$, and also $\pi_i = \rho^i \pi_0 = 0$ for all i.

Example 6.19. (The $M/M/m$ and $M/M/\infty$ Queues) The $M/M/m$ queueing system is identical to the $M/M/1$ system except that m packets can be simultaneously transmitted (i.e., the transmission line of the node has m transmission channels). A packet at the head of the queue is routed to any channel that is available. The corresponding state transition diagram is shown in Fig. 6.24.

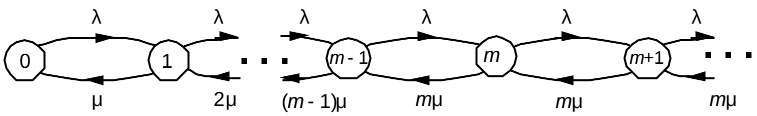


Figure 6.24: Transition graph for the $M/M/m$ queue (Example 6.19).

By writing down the local balance equations for the steady-state probabilities π_n, we obtain

$$\lambda \pi_{n-1} = \begin{cases} n\mu\pi_n & \text{if } n \leq m, \\ m\mu\pi_n & \text{if } n > m. \end{cases}$$

From these equations, we obtain

$$\pi_n = \begin{cases} p_0 \dfrac{(m\rho)^n}{n!}, & n \leq m \\ \\ p_0 \dfrac{m^m \rho^n}{m!}, & n > m \end{cases}$$

where ρ is given by

$$\rho = \frac{\lambda}{m\mu}.$$

Assuming $\rho < 1$, we can calculate π_0 using the above equations and the condition $\sum_{n=0}^{\infty} \pi_n = 1$. We obtain

$$\pi_0 = \left(1 + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + \sum_{n=m}^{\infty} \frac{(m\rho)^n}{m!} \frac{1}{m^{n-m}}\right)^{-1}$$

and, finally,

$$\pi_0 = \left(\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1-\rho)}\right)^{-1}.$$

In the limiting case where $m = \infty$ in the $M/M/m$ system (which is called the $M/M/\infty$ system), the local balance equations become

$$\lambda \pi_{n-1} = n\mu\pi_n, \qquad n = 1, 2, \ldots$$

so

$$\pi_n = \pi_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}, \qquad n = 1, 2, \ldots$$

From the condition $\sum_{n=0}^{\infty} \pi_n = 1$, we obtain

$$\pi_0 = \left(1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right)^{-1} = e^{-\lambda/\mu},$$

so, finally,

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \frac{e^{-\lambda/\mu}}{n!}, \qquad n = 0, 1, \ldots$$

Therefore, in steady-state, the number in the system is Poisson distributed with parameter λ/μ.

SOLVED PROBLEMS

SECTION 6.1. Discrete-Time Markov Chains

SECTION 6.2. Classification of States

Problem 1. * Existence of a Recurrent State. Show that in a Markov chain at least one recurrent state must be accessible from any given state, i.e., for any i, there is at least one recurrent j in the set $A(i)$ of accessible states from i.

Solution. Start with a state i. If $A(i)$ contains a recurrent state we are done; otherwise choose a transient state $i_1 \in A(i)$ such that $i \notin A(i_1)$ and hence also $i_1 \neq i$ [such a state must exist for otherwise i and all the states in $A(i)$ would form a recurrent class]. If $A(i_1)$ contains a recurrent state we are done; otherwise choose a transient state $i_2 \in A(i_1)$ such that $i_1 \notin A(i_2)$ and hence also $i_1 \neq i_2$ [such a state must exist for otherwise i_1 and all the states in $A(i_1)$ would form a recurrent class]. Note that we must have $i_2 \neq i$ for otherwise we would be able to come back to i from i_1 through i_2. Continuing this process, at the kth step, we will either obtain a recurrent state i_k which is accessible from i, or else we will obtain a transient state i_k which is different than all the preceding states $i, i_1, \ldots, i_{k-1}$. Since there is only a finite number of states, a recurrent state must ultimately be obtained.

Problem 2. * Periodic Classes.

(a) Show that a recurrent class is periodic if and only if there exists an integer $d > 1$ such that for all states i in the class, if n is not an integer multiple of d, then $r_{ii}(n) = 0$.

(b) Show that a recurrent class is not periodic if and only if we can find a time $\overline{n}$ and a special state s in the class such that $r_{is}(n) > 0$ for all i in the class and $n \geq \overline{n}$.

SECTION 6.3. Steady-State Behavior

Problem 3. The book pile problem. A professor has m probability books, which she keeps in a pile. Occasionally she picks out a book from the pile (independently of past choices) and after using it, she places it at the top of the pile. Show that if the ith book is picked up with probability β_i and $\beta_i > 0$ for all i, then the steady-state probability that this book is at the top of the pile is also β_i.

Solution. Let the states be $1, \ldots, m$, where state j corresponds to the jth book being at the top. The transition probabilities are

$$p_{ij} = \beta_j, \qquad i, j = 1, \ldots, m.$$

The chain has a single recurrent class which is aperiodic, so steady-state probabilities π_j and solve uniquely the normalization equation $\pi_1 + \cdots + \pi_m = 1$ together with the balance equations

$$\pi_j = \sum_{i=1}^{m} \pi_i p_{ij}, \qquad j = 1, \ldots, m.$$

It is seen that $\pi_j = \beta_j$ solve these equations, so they are the unique steady-state probabilities.

Problem 4. A particular professor gives tests that are either hard, medium, or easy. Fortunately, she never gives two hard tests in a row. If she gives a hard test, her next test is equally likely to be either medium or easy. However, if she gives a medium or easy test, there is a 50% chance that her next test will be of the same difficulty, and a 25% chance that it will be either of the two other difficulties. Formulate the appropriate Markov Chain and find the steady state probabilities.

Solution. We use a Markov chain model with 3 states, where the state is the difficulty of the most recent exam

$$\{S_1 = H, S_2 = M, S_3 = E\}.$$

We are given the transition probabilities

$$\begin{pmatrix} r_{HH} & r_{HM} & r_{HE} \\ r_{MH} & r_{MM} & r_{ME} \\ r_{EH} & r_{EM} & r_{EE} \end{pmatrix} = \begin{pmatrix} 0 & .5 & .5 \\ .25 & .5 & .25 \\ .25 & .25 & .5 \end{pmatrix}.$$

It is easy to see that our Markov chain has a single, aperiodic recurrent class. Thus we can use the steady-state convergence theorem which tells us that if we can find $\{\pi_i\}$ that satisfy $\pi_j = \sum_i \pi_i p_{ij}$ and $\sum_i \pi_i = 1$ then the $\{\pi_i\}$ are in fact the steady state probabilities. Therefore we have $\pi_j = \sum_i \pi_i p_{ij}$, i.e.,

$$\frac{1}{4}(\pi_2 + \pi_3) = \pi_1$$
$$\frac{1}{2}(\pi_1 + \pi_2) + \frac{1}{4}\pi_3 = \pi_2$$
$$\frac{1}{2}(\pi_1 + \pi_3) + \frac{1}{4}\pi_2 = \pi_3$$

and solving these with the constraint $\sum_i \pi_i = 1$ gives

$$\pi_1 = \frac{1}{5}, \pi_2 = \pi_3 = \frac{2}{5}.$$

Problem 5. Alvin likes to sail each Saturday to his cottage on a nearby island off the coast. Alvin is an avid fisherman, and enjoys fishing off his boat on the way to and from the island, as long as the weather is good. Unfortunately, the weather is good on the way to or from the island with probability p, independently of what the weather was on any past trip (so the weather could be nice on the way to the island, but poor on the way back). Now, if the weather is nice, Alvin will take one of his n fishing rods for the trip, but if the weather is bad, he will not bring a fishing rod with him. We want to find the probability that on a given leg of the trip to or from the island the weather will be nice, but Alvin will not fish because all his fishing rods are at his other home.

(a) Formulate an appropriate Markov chain model with $n + 1$ states and find the steady-state probabilities.

(b) What is the probability that on a given trip, Alvin sails with nice weather but without a fishing rod?

(c) If Alvin owns 4 fishing rods, find p such that the time he wants to, but cannot fish is maximized.

Solution. (a) We need to know the number of fishing rods on and off the island. It is enough to know the number of fishing rods off the island. Therefore the states of our chain will be the number of fishing rods off the island. We will consider the state of the chain after a round trip to and from the island. This is a birth-death Markov chain, since the only states that are adjacent are states S_i, S_{i+1} for appropriate i. Note that because of this, the number of transitions from state i to state $i+1$ must be within 1 of the number of transitions from state $i+1$ to state i. Therefore since we are seeking the steady state probabilities, we can equate the two. Before we can solve these equations for the limiting probabilities, we need to find the transition probabilities

$$p_{ii} = \begin{cases} (1-p)^2 + p^2 & \text{for } i \geq 1, \\ (1-p) & \text{for } i = 0. \end{cases}$$

$$p_{i,i+1} = \begin{cases} (1-p)p & \text{for } 1 \leq i < n, \\ p & \text{for } i = 0. \end{cases}$$

$$p_{i,i-1} = \begin{cases} (1-p)p & \text{for } 1 \leq i. \\ 0 & \text{for } i = 0. \end{cases}$$

Thus we have

$$\pi_0 p_{01} = \pi_1 p_{10},$$

implying

$$\pi_1 = \frac{\pi_0}{1-p}$$

and similarly,

$$\pi_n = \cdots = \pi_2 = \pi_1 = \frac{\pi_0}{1-p}$$

thus we have

$$\sum_i \pi_1 = \pi_0(1 + \frac{n}{1-p}) = 1,$$

thus yielding

$$\pi_0 = \frac{1-p}{n+1-p}, \qquad \pi_i = \frac{1}{n+1-p}, \quad \text{for all } i > 0.$$

(b) Let A denote the event that the weather is nice but Alvin has no fishing rods with him. Then

$$P(A) = \pi_n(1-p)p + \pi_0(p) = \frac{2p - 2p^2}{n+1-p}.$$

(c) If $n = 4$ we have

$$P(A) = \frac{2p - 2p^2}{5-p}$$

and we can maximize this by taking a derivative with respect to p and setting equal to zero. Solving, we find that the "optimal" value of p is

$$p = 5 - 2\sqrt{5}.$$

Problem 6. Bernoulli-Laplace model of diffusion. Each of two urns contains m balls. Out of the total of the $2m$ balls, m are white and m are black. A ball is simultaneously selected from each urn and moved to the other urn, an the process is indefinitely repeated, What is the steady-state distribution of the number of balls in each urn?

Solution. Let $j = 0, 1 \ldots, m$ be the states, with state j corresponding to the first urn containing j white balls. The nonzero transition probabilities are

$$p_{j,j-1} = \left(\frac{j}{m}\right)^2, \qquad p_{j,j+1} = \left(\frac{m-j}{m}\right)^2, \qquad p_{jj} = \frac{2j(m-j)}{m^2}.$$

The chain has a single recurrent class that is aperiodic. From the balance equations, the steady-state probabilities can be shown to be

$$\pi_j = \frac{\binom{m}{j}^2}{\binom{2m}{m}}, \qquad j = 0, 1, \ldots, m.$$

Problem 7. * Consider a Markov chain with two states denoted 1 and 2, and transition probabilities

$$p_{11} = 1 - \alpha, \qquad p_{12} = \alpha,$$
$$p_{21} = \beta, \qquad p_{22} = 1 - \beta,$$

where α and β are such that either $0 < \alpha < 1$ or $0 < \beta < 1$ (or both).

(a) Show that the two states of the chain form a recurrent and aperiodic class.

(b) Use induction to show that for all n, we have

$$r_{11}(n) = \frac{\beta}{\alpha+\beta} + \frac{\alpha(1-\alpha-\beta)^n}{\alpha+\beta}, \quad r_{12}(n) = \frac{\alpha}{\alpha+\beta} - \frac{\alpha(1-\alpha-\beta)^n}{\alpha+\beta},$$

$$r_{21}(n) = \frac{\beta}{\alpha+\beta} - \frac{\beta(1-\alpha-\beta)^n}{\alpha+\beta}, \quad r_{22}(n) = \frac{\alpha}{\alpha+\beta} + \frac{\beta(1-\alpha-\beta)^n}{\alpha+\beta}.$$

(c) What are the steady-state probabilities π_1 and π_2?

Solution. (a) The states form a recurrent class since all possible transitions have positive probability. The class is aperiodic since state 1 (as well as state 2) can be reached in one step starting from any state.

(b) The Chapman-Kolmogorov equations are

$$r_{ij}(n) = \sum_{k=1}^{2} r_{ik}(n-1)p_{kj}, \qquad \text{for } n > 1, \quad \text{and } i, j = 1, 2,$$

starting with $r_{ij}(1) = p_{ij}$, so they have the form

$$r_{11}(n) = r_{11}(n-1)(1-\alpha) + r_{12}(n-1)\beta, \qquad r_{12}(n) = r_{11}(n-1)\alpha + r_{12}(n-1)(1-\beta),$$

$$r_{21}(n) = r_{21}(n-1)(1-\alpha) + r_{22}(n-1)\beta, \qquad r_{22}(n) = r_{21}(n-1)\alpha + r_{22}(n-1)(1-\beta).$$

If the $r_{ij}(n-1)$ have the form given, it is easily verified by substitution in the Chapman-Kolmogorov equations that the $r_{ij}(n)$ also have the form given.

(c) The steady-state probabilities π_1 and π_2 are obtained by taking the limit of $r_{i1}(n)$ and $r_{i2}(n)$, respectively, as $n \to \infty$. Thus, we have

$$\pi_1 = \frac{\beta}{\alpha+\beta}, \qquad \pi_2 = \frac{\alpha}{\alpha+\beta}.$$

Problem 8. * The parking garage at MIT has installed a card operated gate, which, unfortunately, is vulnerable to absent-minded faculty and staff. In particular, in each day a car crashes the gate with probability p, in which case a new gate must be installed. Also a gate that has survived for m days must be replaced as a matter of periodic maintenance. What is the long-term expected frequency of gate replacements?

Solution. Let the state be the number of days that the gate has survived. The balance equations are

$$\pi_0 = \pi_0 p + \pi_1 p + \cdots + \pi_{m-1} p + \pi_m,$$

$$\pi_1 = \pi_0(1-p),$$

$$\pi_2 = \pi_1(1-p) = \pi_0(1-p)^2,$$

and similarly

$$\pi_i = \pi_0(1-p)^i, \qquad i = 1, \ldots, m.$$

We have using the normalization equation

$$1 = \pi_0 + \sum_{i=1}^{m} \pi_i = \pi_0 \left(1 + \sum_{i=1}^{m} (1-p)^i\right),$$

so

$$\pi_0 = \frac{p}{1-(1-p)^{m+1}}.$$

The long-term expected frequency of gate replacements is equal to the long-term expected frequency of visits to state 0, which is π_0.

Problem 9. * Doubly Stochastic Matrices. Consider an irreducible and aperiodic Markov chain whose transition probability matrix is **doubly stochastic**, i.e., it has the property that all its columns (as well as all its rows) add to unity:

$$\sum_{i=1}^{n} p_{ij} = 1, \qquad \text{for all } j = 1, \ldots, m.$$

(a) Show that the transition probability matrix of the chain in Example 6.6 is doubly stochastic.

(b) Show that the steady-state probabilities are

$$\pi_j = \frac{1}{m}, \qquad j = 1, \ldots, m.$$

(c) Suppose that the recurrent class of the chain is instead periodic. Use the results of Problem 5 to show that $\pi_j = \frac{1}{m}$ for all j are the unique solution of the balance equations. Discuss your answer in the context of Example 6.6 for the case where m is even.

Solution. (a) Indeed the rows and the columns of the transition probability matrix in this example all add to 1.

(b) It is seen that the given probabilities $\pi_j = 1/m$ indeed satisfy the balance equations.

(c) If m is even in Example 6.6, the chain is periodic with period 2. Despite this fact $\pi_j = 1/m$ solve uniquely the balance equations.

Problem 10. * Consider the queueing Example 6.10, but assume that the probabilities of a packet arrival and a packet transmission depend on the state of the queue. In particular, in each period where there are i packets in the node, one of the following occurs:

(1) One new packet arrives; this happens with a given probability $b_i > 0$.

(2) One existing packet completes transmission; this happens with a given probability $d_i > 0$ if $i \geq 1$, and with probability 0 otherwise.

(3) No new packet arrives and no existing packet completes transmission; this happens with a probability $1 - b_i - d_i$ if $i \geq 1$, and with probability $1 - b_i$ otherwise.

Calculate the steady-state probabilities of the corresponding Markov chain.

Solution. We introduce a Markov chain where the states are $0, 1, \ldots, m$ and correspond to the number of packets currently stored at the node. The transition probability graph is given in Fig. 6.25.

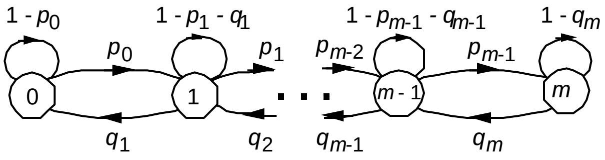


Figure 6.25: Transition probability graph for the queueing problem.

Similar to Example 6.10, the balance equations are easily solved by applying the formula

$$\text{Frequency of transitions into } S \;=\; \text{Frequency of transitions out of } S$$

for

$$S = \{0, 1, \ldots, i\}, \qquad i = 0, 1, \ldots, m - 1.$$

We obtain

$$\pi_i b_i = \pi_{i+1} d_{i+1}, \qquad i = 0, 1, \ldots, m-1.$$

Thus we have $\pi_{i+1} = \rho_i \pi_i$ where

$$\rho_i = \frac{b_i}{d_{i+1}}.$$

Hence $\pi_i = (\rho_0 \cdots \rho_{i-1})\pi_0$ for all $i = 1, \ldots, m$. By using the normalization equation $1 = \pi_0 + \pi_1 + \cdots + \pi_m$, we obtain

$$1 = \pi_0(1 + \rho_0 + \cdots + \rho_0 \cdots \rho_{m-1}),$$

from which

$$\pi_0 = \frac{1}{1 + \rho_0 + \cdots + \rho_0 \cdots \rho_{m-1}}.$$

The remaining steady-state probabilities are

$$\pi_i = \frac{\rho_0 \cdots \rho_{i-1}}{1 + \rho_0 + \cdots + \rho_0 \cdots \rho_{m-1}}, \qquad i = 1, \ldots, m.$$

Problem 11. * Dependence of the Balance Equations. Show that if we add the first $m-1$ balance equations $\pi_j = \sum_{k=1}^m \pi_k p_{kj}$, $j = 1, \ldots, m-1$, we obtain the last equation $\pi_m = \sum_{k=1}^m \pi_k p_{km}$.

Solution. By adding the first $m-1$ balance equations, we obtain

$$\begin{aligned}
\sum_{j=1}^{m-1} \pi_j &= \sum_{j=1}^{m-1} \sum_{k=1}^{m} \pi_k p_{kj} \\
&= \sum_{k=1}^{m} \pi_k \sum_{j=1}^{m-1} p_{kj} \\
&= \sum_{k=1}^{m} \pi_k (1 - p_{km}) \\
&= \pi_m + \sum_{k=1}^{m-1} \pi_k - \sum_{k=1}^{m} \pi_k p_{km}.
\end{aligned}$$

This equation is equivalent to the last balance equation $\pi_m = \sum_{k=1}^m \pi_k p_{km}$.

Problem 12. * Local Balance Equations. Given a Markov chain with transition probabilities p_{ij}, $i, j = 1, \ldots, m$, and a single class that is aperiodic, suppose that we have found a solution $(\pi_1, \ldots, \pi_m)$ to the following system of local balance equations

$$\pi_i p_{ij} = \pi_j p_{ji}, \qquad i, j = 1, \ldots, m,$$

$$\sum_{i=1}^{m} \pi_i = 1, \qquad \pi_i \geq 0, \quad i = 1, \ldots, m.$$

(a) Show that the π_j are the unique steady-state probabilities. *Hint*: Given a state j, add the local balance equations corresponding to the pairs of the form (i,j).

(b) What is the interpretation of the equations $\pi_i p_{ij} = \pi_j p_{ji}$ in terms of steady-state expected frequencies of transition between i and j?

(c) Show that the local balance equations hold for any birth-death Markov chain.

(d) Construct an example of a Markov chain with a single class that is aperiodic, where the local balance equations are not satisfied by the steady-state probabilities.

Solution. (a) By adding the local balance equations $\pi_i p_{ij} = \pi_j p_{ji}$ over i, we obtain

$$\sum_{i=1}^{m} \pi_i p_{ij} = \sum_{i=1}^{m} \pi_j p_{ji} = \pi_j,$$

so the π_j also satisfy the balance equations. Therefore, they are equal to the steady-state probabilities.

(b) We know that $\pi_i p_{ij}$ can be interpreted as the expected long-term frequency of transitions from i to j, so the local balance equations imply that the expected long-term frequency of any transition is equal to the expected long-term frequency of the reverse transition. (This property is also known as *time reversibility* of the chain.)

(c) In any Markov chain, given a subset of states S, the expected long-term frequency of transitions form states in S to states in the complementary set $\overline{S}$, is equal to the expected long-term frequency of transitions form states in $\overline{S}$ to states in S. Applying this property to a birth-death Markov chain to the sets $S = \{1, \ldots, i\}$ and $\overline{S} = \{i+1, \ldots, m\}$, we obtain the local balance equations.

(d) We need a minimum of three states for such an example. Let the states be $1, 2, 3$, and let $p_{12} > 0$, $p_{13} > 0$, $p_{21} > 0$, $p_{32} > 0$, with all other transition probabilities being 0. The chain has a single recurrent aperiodic class. The local balance equations do not hold because the expected frequency of transitions from 1 to 3 is positive, but the expected frequency of reverse transitions is 0.

Problem 13. * Sampled Markov chains. Consider a Markov chain $\{X_1, X_2, \ldots\}$ with transition probabilities p_{ij}, $i, j = 1, \ldots, m$, and let $r_{ij}(n)$ be the n-step transition probabilities.

(a) Show that for all $n \geq 1$ and $l \geq 1$, we have

$$r_{ij}(n+l) = \sum_{k=1}^{m} r_{ik}(n) r_{kj}(l).$$

(b) Let the Markov chain have a single recurrent class that is aperiodic. Suppose that the process is sampled every l transitions, thus generating the process $\{Y_1, Y_2, \ldots\}$, where $Y_n = X_{ln}$. Show that the sampled process can be modeled by a Markov chain with a single aperiodic recurrent class and transition probabilities $r_{ij}(l)$.

(c) Use part (a) to show that the sampled process of part (b) has the same steady-state probabilities as the original process.

Problem 14. * Given an irreducible and aperiodic Markov chain $\{X_n\}$, consider the Markov chain whose state at time n is $\{X_{n-1}, X_n\}$ (thus the state in the new chain can be associated with the last transition in the original chain).

(a) Show that the steady-probabilities are

$$\eta_{ij} = \pi_i p_{ij}.$$

(b) Generalize part (a) to the case of the Markov chain $\{X_n, X_{n+1}, \ldots, X_{n+k}\}$, whose state can be associated with the last k transitions of the original chain.

(c) Consider an infinite sequence of independent coin tosses, where the probability of a head is p. Each time the coin comes up heads twice in a row you get \$1 (so the sequence $THHHHT$ gets you \$3). What is the (approximate) expected amount of money that you will get per coin toss, averaged over a large number of coin tosses.

SECTION 6.4. Absorption Probabilities and Expected Time to Absorption

Problem 15. There are m classes offered by a particular department, and each year, the students rank each class from 1 to m, in order of difficulty. Unfortunately, the ranking is completely arbitrary. In fact, any given class is equally likely to receive any given rank on a given year (two classes may not receive the same rank). A certain professor chooses to remember only the highest ranking his class has ever gotten.

(a) Find the transition probabilities of the Markov chain that models the ranking that the professor remembers.

(b) Find the recurrent and the transient states.

(c) Find the expected number of years for the professor to achieve the highest ranking given that in the first year he achieved the ith ranking.

Solution. (a) For $i > j$, we have $p_{ij} = 0$. Since the professor will continue to remember the highest ranking, even if he gets a lower ranking in a subsequent year, we have $p_{ii} = i/m$. Finally, for $j > i$, we have $p_{ij} = 1/m$, since the class is equally likely to receive any given rating.

(b) There is a positive probability that on any given year, the professor will receive the highest ranking, namely $1/m$. Therefore the probability that he does not receive this ranking in n years is $\left(\frac{m-1}{m}\right)^n$ and this probability goes to 0 as $n \to \infty$. As a result, he is guaranteed of eventually receiving the highest ranking. Now note that this state is absorbing. It is the only recurrent state in the Markov chain. All other states are transient.

(c) This question can be answered by finding the mean first passage time to the absorbing state m starting from i. It is simpler though to argue as follows: since the probability of achieving the highest ranking in a given year is $1/m$, independently of the current state, the required expected number of years is the expected number of trials to the first success in a Bernoulli process with success probability $1/m$. Thus, the expected number of years is m.

Problem 16. * Mean First Passage Times. Consider a Markov chain with a single recurrent class, and let s be a fixed recurrent state.

(a) Show that for some scalars c and γ, with $c > 0$ and $0 < \gamma < 1$, we have

$$\mathbf{P}(X_1 \neq s, \ldots, X_n \neq s \,|\, X_0 = i) \leq c\gamma^n$$

for all i and $n \geq 1$.

(b) Show that for all $i \neq s$, we have

$$\mu_i \leq \frac{c}{(1-\gamma)^2},$$

where μ_i is the mean first passage time to go from a state $i \neq s$ to state s.

(c) Show that the μ_i are the unique solution of the system of equations

$$\mu_i = 1 + \sum_{\substack{j=1\\ j\neq s}}^{m} p_{ij}\mu_j, \qquad \text{for all } i \neq s.$$

Solution. (a) For notational convenience, denote

$$q_i(n) = \mathbf{P}(X_1 \neq s, \ldots, X_n \neq s \,|\, X_0 = i).$$

Note that $q_i(n)$ is monotonically nonincreasing with n, and that $q_i(m) < 1$, since by assumption, state s is accessible from each state i. Let

$$\beta = \max_{i=1,\ldots,m} q_i(m).$$

We have for all i, $q_i(m) \leq \beta < 1$, and also

$$\begin{aligned} q_i(2m) &= \mathbf{P}(X_1 \neq s, \ldots, X_m \neq s, X_{m+1} \neq s, \ldots, X_{2m} \neq s \,|\, X_0 = i) \\ &= \mathbf{P}(X_1 \neq s, \ldots, X_m \neq s \,|\, X_0 = i) \\ &\quad \cdot \mathbf{P}(X_{m+1} \neq s, \ldots, X_{2m} \neq s \,|\, X_0 = i, X_1 \neq s, \ldots, X_m \neq s) \\ &\leq \beta^2, \end{aligned}$$

where the last relation follows from the calculation

$$\begin{aligned} &\mathbf{P}(X_{m+1} \neq s, \ldots, X_{2m} \neq s \,|\, X_0 = i, X_1 \neq s, \ldots, X_m \neq s) \\ &\quad = \sum_{j\neq s} \mathbf{P}(X_m = j \,|\, X_0 = i, X_1 \neq s, \ldots, X_{m-1} \neq s) \\ &\qquad\quad \cdot \mathbf{P}(X_{m+1} \neq s, \ldots, X_{2m} \neq s \,|\, X_0 = i, X_1 \neq s, \ldots, X_{m-1} \neq s, X_m = j) \\ &\quad = \sum_{j\neq s} \mathbf{P}(X_m = j \,|\, X_0 = i, X_1 \neq s, \ldots, X_{m-1} \neq s) \\ &\qquad\quad \cdot \mathbf{P}(X_{m+1} \neq s, \ldots, X_{2m} \neq s \,|\, X_m = j) \\ &\quad \leq \sum_{j\neq s} \mathbf{P}(X_m = j \,|\, X_0 = i, X_1 \neq s, \ldots, X_{m-1} \neq s) \cdot \beta \\ &\quad \leq \beta. \end{aligned}$$

Similarly, we show that for all i and k,

$$q_i(km) \le \beta^k.$$

Let n be any positive integer, and let k be the integer such that $km \le n < (k+1)m$. Then, we have

$$q_i(n) \le q_i(km) \le \beta^k = \beta^{-1}\left(\beta^{1/m}\right)^{(k+1)m} \le \beta^{-1}\left(\beta^{1/m}\right)^n.$$

Thus, the desired relation holds with $c = \beta^{-1}$ and $\gamma = \beta^{1/m}$.

(b) We have

$$\begin{aligned}
\mu_i &= \sum_{n=1}^{\infty} n\mathbf{P}(X_1 \ne s, \ldots, X_{n-1} \ne s, X_n = s \mid X_0 = i) \\
&\le \sum_{n=1}^{\infty} nq_i(n-1) \\
&\le \sum_{n=1}^{\infty} nc\gamma^{n-1} \\
&= \frac{c}{1-\gamma}\sum_{n=1}^{\infty} n(1-\gamma)\gamma^{n-1} \\
&= \frac{c}{(1-\gamma)^2},
\end{aligned}$$

where the last relation holds using the expression for the mean of the geometric distribution.

(c) That the μ_i are a solution to the given equations follows by applying the total expectation theorem. To show uniqueness, let $\overline{\mu}_i$ be another solution. Then we have for all $i \ne s$

$$\mu_i = 1 + \sum_{j \ne s} p_{ij}\mu_j, \qquad \overline{\mu}_i = 1 + \sum_{j \ne s} p_{ij}\overline{\mu}_j,$$

and by subtraction, we obtain

$$\delta_i = \sum_{j \ne s} p_{ij}\delta_j,$$

where $\delta_i = \overline{\mu}_i - \mu_i$. By applying m successive times this relation, if follows that

$$\delta_i = \sum_{j_1 \ne s} p_{ij_1} \sum_{j_2 \ne s} p_{j_1 j_2} \cdots \sum_{j_m \ne s} p_{j_{m-1} j_m} \cdot \delta_{j_m}.$$

Hence, we have for all $i \ne s$

$$\begin{aligned}
|\delta_i| &\le \left| \sum_{j_1 \ne s} p_{ij_1} \sum_{j_2 \ne s} p_{j_1 j_2} \cdots \sum_{j_m \ne s} p_{j_{m-1} j_m} \right| \cdot |\delta_{j_m}| \\
&= q_i(m) \cdot |\delta_{j_m}| \\
&\le \beta \cdot \max_{j \ne s} |\delta_j|.
\end{aligned}$$

The above relation holds for all $i \neq s$, so we obtain

$$\max_{j \neq s} |\delta_j| \leq \beta \cdot \max_{j \neq s} |\delta_j|,$$

which implies that $\max_{j \neq s} |\delta_j| = 0$ or $\mu_j = \overline{\mu}_j$ for all $j \neq s$.

Problem 17. * Balance Equations and Mean Recurrence Times. Consider a Markov chain with a single recurrent class. For any two states i and j, with j being recurrent, let

$$\rho_{ij} = \mathbf{E}\big[\text{Number of visits to } i \text{ between two successive visits to } j\big],$$

where by convention, $\rho_{jj} = 1$.

(a) Fix a recurrent state s. Show that the scalars

$$\pi_i = \frac{\rho_{is}}{t_s^*}, \qquad i = 1, \ldots, m,$$

satisfy the balance equations, where t_s^* is the mean recurrence time of s (th expected number of transitions up to the first return to s, starting from s.

(b) Show that if $(\pi_1, \ldots, \pi_m)$ satisfy the balance equations, then

$$\pi_i = \begin{cases} \dfrac{1}{\mu_i} & \text{if } i \text{ is recurrent}, \\ 0 & \text{if } i \text{ is transient}, \end{cases}$$

where μ_i is the mean recurrence time of i.

(c) Show that the distribution of part (a) is the unique probability distribution that satisfies the balance equations.

Solution. (a) We first assert that

$$\rho_{is} = \sum_{n=1}^{\infty} \mathbf{P}(X_1 \neq s, \ldots, X_{n-1} \neq s, X_n = i \,|\, X_0 = s).$$

To see this, note that

$$\rho_{is} = \mathbf{E}\left[\sum_{n=1}^{\infty} I_n \;\middle|\; X_0 = s\right].$$

where I_n is the random variable that takes the value 1 if $X_1 \neq s, \ldots, X_{n-1} \neq s$, and $X_n = i$, and the value 0 otherwise, so that

$$\mathbf{E}\big[I_n \,|\, X_0 = s\big] = \mathbf{P}(X_1 \neq s, \ldots, X_{n-1} \neq s, X_n = i \neq s \,|\, X_0 = s).$$

We next use the total probability theorem to write for $n \geq 2$

$$\begin{aligned}\mathbf{P}&(X_1 \neq s, \ldots, X_{n-1} \neq s, X_n = i \,|\, X_0 = s) \\ &= \sum_{k \neq s} \mathbf{P}(X_1 \neq s, \ldots, X_{n-2} \neq s, X_{n-1} = k \,|\, X_0 = s) p_{ki}.\end{aligned}$$

We thus obtain

$$\begin{aligned}
\rho_{is} &= \mathbf{E}\left[\sum_{n=1}^{\infty} I_n \,\middle|\, X_0 = s\right] \\
&= p_{si} + \sum_{n=2}^{\infty} \mathbf{P}(X_1 \neq s, \ldots, X_{n-1} \neq s, X_n = i \,|\, X_0 = s) \\
&= p_{si} + \sum_{n=2}^{\infty} \sum_{k \neq s} \mathbf{P}(X_1 \neq s, \ldots, X_{n-2} \neq s, X_{n-1} = k \,|\, X_0 = s) p_{ki} \\
&= p_{si} + \sum_{k \neq s} p_{ki} \sum_{n=2}^{\infty} \mathbf{P}(X_1 \neq s, \ldots, X_{n-2} \neq s, X_{n-1} = k \,|\, X_0 = s) \\
&= \rho_{ss} p_{si} + \sum_{k \neq s} p_{ki} \rho_{ks} \\
&= \sum_{k=1}^{m} \rho_{ks} p_{ki}.
\end{aligned}$$

Dividing both sides of this relation by t_s^*, we obtain

$$\pi_i = \sum_{k=1}^{m} \pi_k p_{ki},$$

where $\pi_i = \rho_{is}/t_s^*$. Thus, the π_i solve the balance equations. Furthermore, the π_i are nonnegative, and we clearly have $\sum_{i=1}^{m} \rho_{is} = t_s^*$ or $\sum_{i=1}^{m} \pi_i = 1$. Hence, $(\pi_1, \ldots, \pi_m)$ is a probability distribution.

(b) Consider a probability distribution $(\pi_1, \ldots, \pi_m)$ that satisfies the balance equations. Fix a recurrent state s, let t_s^* be the mean recurrence time of s, and let t_i be the mean first passage time from a state $i \neq s$ to state s. We will show that $\pi_s t_s^* = 1$. Indeed, we have

$$t_s^* = 1 + \sum_{j \neq s} p_{sj} t_j,$$

$$t_i = 1 + \sum_{j \neq s} p_{ij} t_j, \qquad \text{for all } i \neq s.$$

Multiplying these equations with π_s and π_i, respectively, and adding, we obtain

$$\pi_s t_s^* + \sum_{i \neq s} \pi_i t_i = 1 + \sum_{i=1}^{m} \pi_i \sum_{j \neq s} p_{ij} t_j.$$

By using the balance equations, the right-hand side is equal to

$$1 + \sum_{i=1}^{m} \pi_i \sum_{j \neq s} p_{ij} t_j = 1 + \sum_{j \neq s} t_j \sum_{i=1}^{m} \pi_i p_{ij} = 1 + \sum_{j \neq s} t_j \pi_j.$$

By combining the last two equations, we obtain $\pi_s t_s^* = 1$.

Since the probability distribution $(\pi_1, \ldots, \pi_m)$ satisfies the balance equations, if the initial state X_0 is chosen according to this distribution, all subsequent states X_n have the same distribution. If we start at a transient state i, the probability of being at that state at time n diminishes to 0 as $n \to \infty$. It follows that we must have $\pi_i = 0$.

(c) Part (a) shows that there exists at least one stationary probability distribution. Part (b) shows that there can be only one stationary probability distribution.

Problem 18. * Steady-State Convergence. Consider a Markov chain with a single recurrent class, and assume that there exists a special recurrent state s and a time $\overline{n}$ such that s can be reached in $\overline{n}$ steps starting from any state:

$$r_{is}(\overline{n}) > 0, \qquad i = 1, \ldots, m.$$

(a) Show that there exists a time $\hat{n}$ such that for all $i = 1, \ldots, m$, all recurrent j, and all $n \geq \hat{n}$, we have

$$r_{ij}(n) > 0.$$

Furthermore, the recurrent class of the chain is aperiodic.

(b) Show that for all i and j, we have

$$\lim_{n\to\infty} r_{ij}(n) = \pi_j,$$

where the π_j are the unique solution of the balance equations (cf. the preceding problem). In particular, there exist scalars c and γ with $c > 0$ and $0 < \gamma < 1$ such that for all n, i, and j, we have

$$\bigl|r_{ij}(n) - \pi_j\bigr| \leq c\gamma^n.$$

Solution. (a) We first claim that for all i, we have

$$r_{is}(n) > 0, \qquad \text{for all } n \geq \overline{n}.$$

This follows by induction. In particular, if $r_{is}(n) > 0$ for all i, we have

$$r_{is}(n+1) = \sum_{j=1}^{m} p_{ij} r_{js}(n) \geq \min_{i=1,\ldots,m} r_{js}(n) > 0, \qquad \text{for all } i.$$

We next note that for all j that are reachable from s in k steps, we have $r_{ij}(n+k) > 0$ for all i, as long as $r_{is}(n) > 0$ for all i. Hence, we have $r_{ij}(n) > 0$ for all i and all $n \geq \overline{n} + k$. Since all recurrent states j are reachable from s within $m-1$ steps, we have $r_{ij}(n) > 0$ for all i and j, and all $n \geq \overline{n} + m - 1$. This also shows that the chain is aperiodic.

(b) For all states i and all transient states j, we have

$$r_{ij}(n) = \mathbf{P}(X_n = j \,|\, X_0 = i) \leq \mathbf{P}(X_1 : \text{transient}, \ldots, X_n : \text{transient} \,|\, X_0 = i).$$

Hence $r_{ij}(m) < 1$, and by using an argument similar to the one used to prove finiteness of the first passage time in chains with a single recurrent class [cf. Problem 3(a)], we obtain $r_{ij}(n) \leq c\gamma^n$ for all n, where c and γ are some scalars with $c > 0$ and $0 < \gamma < 1$. Thus, $|r_{ij}(n) - \pi_j| \leq c\gamma^n$ for all n.

We introduce a copy of the given Markov chain and denote its state at time n by Y_n. We start the given chain at some fixed initial state $X_0 = i$ and we select the initial state Y_0 according to its unique solution of the balance equations. Let T be the first time n for which we have $X_n = Y_n$:

$$T = \min\{n \,|\, X_n = Y_n\}.$$

Fix a recurrent state j. We have

$$r_{ij}(n) = \mathbf{P}(X_n = j) = \mathbf{P}(X_n = j, n \geq T) + \mathbf{P}(X_n = j, n < T),$$

$$\pi_j = \mathbf{P}(Y_n = j) = \mathbf{P}(Y_n = j, n \geq T) + \mathbf{P}(Y_n = j, n < T).$$

By subtracting these two equations and taking the absolute value of both sides,

$$\begin{aligned} |r_{ij}(n) - \pi_j| &\leq \big|\mathbf{P}(X_n = j, n \geq T) - \mathbf{P}(Y_n = j, n \geq T)\big| \\ &\quad + \big|\mathbf{P}(X_n = j, n < T) - \mathbf{P}(Y_n = j, n < T)\big| \\ &\leq \big|\mathbf{P}(X_n = j, n < T) - \mathbf{P}(Y_n = j, n < T)\big| \\ &\leq \mathbf{P}(n < T), \end{aligned}$$

where the last inequality follows using the generic relation

$$|\mathbf{P}(B \cap A) - \mathbf{P}(C \cap A)| \leq \mathbf{P}(A), \qquad \text{for any events } A, B, C.$$

We now introduce for every n the probability $q_i(n)$ that X_k and Y_k are not simultaneously equal to state j for any $k \leq n$:

$$q_i(n) = \mathbf{P}\left(\cap_{k=1}^{n}\big(\{X_k \neq j\} \cup \{Y_k \neq j\}\big)\right).$$

We have $q_i(\overline{n}) < 1$ since by part (a), j is reachable from all states in at most $\hat{n}$ steps, and by using again an argument similar to the one used in Problem 3(a), we obtain $q_i(n) \leq c\gamma^n$ for all n, where c and γ are some scalars with $c > 0$ and $0 < \gamma < 1$. On the other hand, we have for all n,

$$|r_{ij}(n) - \pi_j| \leq \mathbf{P}(n < T) \leq q_i(n),$$

so $|r_{ij}(n) - \pi_j| \leq c\gamma^n$ for all n.

Problem 19. * Expected Long-Term Frequency Interpretation. Consider a Markov chain with a single recurrent class that satisfies the condition of the preceding problem. Show that

$$\pi_j = \lim_{n\to\infty} \frac{v_{ij}(n)}{n}, \qquad \text{for all } i,\, j = 1, \ldots, m,$$

where π_j are the steady-state probabilities, and $v_{ij}(n)$ is the expected value of the number of visits to state j within the first n transitions, starting from state i.

Solution. We first assert that for all n, i, and j, we have

$$v_{ij}(n) = \sum_{k=1}^{n} r_{ij}(k).$$

To see this, note that

$$v_{ij}(n) = \mathbf{E}\left[\sum_{k=1}^{n} I_k \;\Big|\; X_0 = i\right],$$

where I_k is the random variable that takes the value 1 if $X_k = j$, and the value 0 otherwise, so that

$$\mathbf{E}\big[I_k \,|\, X_0 = i\big] = r_{ij}(k).$$

From Problem 6(b), we have that there exist scalars c and γ, with $c > 0$ and $0 < \gamma < 1$, such that for all n, i, and j, we have

$$\big|r_{ij}(n) - \pi_j\big| \leq c\gamma^n.$$

Hence

$$\begin{aligned}\left|\frac{v_{ij}(n)}{n} - \pi_j\right| &= \left|\frac{\sum_{k=1}^{n}\big(r_{ij}(k) - \pi_j\big)}{n}\right| \\ &\leq c\left|\frac{\sum_{k=1}^{n}\gamma^k}{n}\right| \\ &\leq \frac{c\gamma}{n(1-\gamma)}.\end{aligned}$$

This proves the desired result.

Problem 20. * Absorption Probabilities. Consider a Markov chain where each state is either transient or absorbing. Fix an absorbing state s. Show that the probabilities a_i of eventually reaching s starting from a state i are the unique solution of the equations

$$a_s = 1,$$

$$a_i = 0, \quad \text{for all absorbing } i \neq s,$$

$$a_i = \sum_{j=1}^{m} p_{ij} a_j, \qquad \text{for all transient } i.$$

Solution. By the total probability theorem, the a_i are a solution to the given equations. To show uniqueness, let $\overline{a}_i$ be another solution, and let $\delta_i = \overline{a}_i - a_i$. Denoting by A the set of absorbing states and using the fact $\delta_j = 0$ for all $j \in A$, we have

$$\delta_i = \sum_{j=1}^{m} p_{ij}\delta_j = \sum_{j \notin A} p_{ij}\delta_j, \qquad \text{for all transient } i.$$

Applying this relation m successive times, we obtain

$$\delta_i = \sum_{j_1 \notin A} p_{ij_1} \sum_{j_2 \notin A} p_{j_1 j_2} \cdots \sum_{j_m \notin A} p_{j_{m-1} j_m} \cdot \delta_{j_m}.$$

Hence

$$\begin{aligned} |\delta_i| &\le \left| \sum_{j_1 \neq s} p_{ij_1} \sum_{j_2 \notin A} p_{j_1 j_2} \cdots \sum_{j_m \notin A} p_{j_{m-1} j_m} \right| \cdot |\delta_{j_m}| \\ &= \mathbf{P}(X_1 \notin A, \ldots, X_m \notin A \,|\, X_0 = i) \cdot |\delta_{j_m}| \\ &\le \mathbf{P}(X_1 \notin A, \ldots, X_m \notin A \,|\, X_0 = i) \cdot \max_{j \notin A} |\delta_j|. \end{aligned}$$

The above relation holds for all $i \neq s$, so we obtain

$$\max_{j \notin A} |\delta_j| \le \beta \cdot \max_{j \notin A} |\delta_j|,$$

where

$$\beta = \mathbf{P}(X_1 \notin A, \ldots, X_m \notin A \,|\, X_0 = i).$$

Since $\beta < 1$, it follows that $\max_{j \notin A} |\delta_j| = 0$ or $a_i = \overline{a}_i$ for all i that are not absorbing. We also have $a_j = \overline{a}_j$ for all absorbing j, so $a_i = \overline{a}_i$ for all i.

Problem 21. * Multiple Recurrent Classes. Consider a Markov chain that has more that one recurrent classes, as well as some transient states. Assume that all the recurrent classes are aperiodic.

(a) For any transient state i, let $a_i(k)$ be the probability that starting from i we will reach a state in the kth recurrent class. Derive a system of equations whose solution are the $a_i(k)$.

(b) Show that each n-step transition probabilities $r_{ij}(n)$ converges to a limit denoted π_{ij}. Derive a system of equations whose solution are the π_{ij}.

Solution. (a) We introduce a new Markov chain that has only transient and absorbing states. The transient states correspond to the transient states of the original, while the absorbing states correspond to the recurrent classes of the original. The transition probabilities $\hat{p}_{ij}$ of the new chain are as follows: if i and j are transient, $\hat{p}_{ij} = p_{ij}$; if i is a transient state and j corresponds to a recurrent class, $\hat{p}_{ij}$ is the sum of the transition probabilities from i to states in the recurrent class in the original Markov chain.

The desired probabilities $a_i(k)$ are the absorption probabilities in the new Markov chain and are given by the corresponding formulas:

$$a_i(k) = \hat{p}_{ik} + \sum_{j:\text{transient}} \hat{p}_{ij} a_j(k), \qquad \text{for all transient } i.$$

(b) If i and j are recurrent but belong to different classes, again $r_{ij}(n)$ converges to 0. If i and j are recurrent but belong to the same class, $r_{ij}(n)$ converges to the steady-state probability of j conditioned on the initial state being in the class of j. If j is transient, $r_{ij}(n)$ converges to 0. Finally, if i is transient and j is recurrent, then $r_{ij}(n)$

converges to the product of two probabilities: (1) the probability that starting from i we will reach a state in the recurrent class of j, and (2) the steady-state probability of j conditioned on the initial state being in the class of j.

SECTION 6.5. More General Markov Chains

Problem 22. Persons arrive at a taxi stand with room for five taxis according to a Poisson process with rate one per minute. A person boards a taxi upon arrival if one is available and otherwise waits in a line. Taxis arrive at the stand according to a Poisson process with rate two per minute. An arriving taxi that finds the stand full departs immediately; otherwise, it picks up a customer if at least one is waiting, or else joins the queue of waiting taxis. What is the steady-state probability distribution of the taxi queue size?

Solution. Consider a continuous-time Markov chain with state

$$n = \text{Number of people waiting} + \text{number of empty taxi positions}.$$

Then the state goes from n to $n+1$ each time a person arrives and goes from n to $n-1$ (if $n \geq 1$) when a taxi arrives. Thus the system behaves like an $M/M/1$ queue with arrival rate 1 per min and departure rate 2 per min. Therefore the occupancy distribution is

$$\pi_n = \frac{(1-\rho)}{\rho^n},$$

where $\rho = 1/2$. State n, for $0 \leq n \leq 4$ corresponds to 5, 4, 3, 2, 1 taxis waiting while $n > 4$ corresponds to no taxi waiting. Therefore

$$\mathbf{P}(5 \text{ taxis waiting}) = 1/2,$$

$$\mathbf{P}(4 \text{ taxis waiting}) = 1/4,$$

$$\mathbf{P}(3 \text{ taxis waiting}) = 1/8,$$

$$\mathbf{P}(2 \text{ taxis waiting}) = 1/16,$$

$$\mathbf{P}(1 \text{ taxi waiting}) = 1/32.$$

and $\mathbf{P}(\text{no taxi waiting})$ is obtained by subtracting the sum of the probabilities above from unity. This gives $\mathbf{P}(\text{no taxi waiting}) = 1/32$.

Problem 23. Empty taxis pass by a street corner at a Poisson rate of two per minute and pick up a passenger if one is waiting there. Passengers arrive at the street corner at a Poisson rate of one per minute and wait for a taxi only if there are less than four persons waiting; otherwise they leave and never return. Find the expected waiting time of a passenger that joins the queue.

Solution. We consider a continuous-time Markov chain with state $n = 0, 1, \ldots, 4$, where

$$n = \text{Number of people waiting}.$$

The transitions from n to $n+1$ have rate 1, and the transitions from $n+1$ to n have rate 2. The balance equations are

$$\pi_n = \frac{\pi_{n-1}}{2}, \qquad n = 1, \ldots, 4.$$

Solving these equations together with the normalization equation $\sum_{i=0}^{4} \pi_i = 1$, we find

$$\pi_n = \frac{2^{4-n}}{3!}, \qquad n = 0, 1, \ldots, 4.$$

A customer that joins the queue will find n customers ahead of him where $n = 0, 1, 2, 3$ with steady-state probabilities $\pi_n/(\pi_0 + \pi_1 + \pi_2 + \pi_3)$. The expected number of customers found by an arriving customer who joins the queue is

$$\mathbf{E}[N] = \frac{\pi_1 + 2\pi_2 + 3\pi_3}{\pi_0 + \pi_1 + \pi_2 + \pi_3}.$$

Since the expected waiting time for a new taxi is 1/2 minute, the expected waiting time, by the law of iterated expectations is

$$\mathbf{E}[T] = \mathbf{E}[N]\frac{1}{2}.$$

Problem 24. An athletic facility has 5 tennis courts. Players arrive at the courts according to a Poisson process with rate of one pair per 10 minutes, and use a court for an exponentially distributed time with mean 40 minutes. Suppose a pair of players arrives and finds all courts busy and k other pairs waiting in queue. What is the expected waiting time to get a court?

Solution. When all the courts are busy, the expected time between two departures is $40/5 = 8$ minutes. If a pair sees k pairs waiting in the queue, there must be exactly $k+1$ departures from the system before they get a court. Since all the courts would be busy during this whole time, the expected waiting time required before $k+1$ departures is $8(k+1)$ minutes.

Problem 25. A facility of m identical machines is sharing a single repairperson. The time to repair a failed machine is exponentially distributed with mean $1/\lambda$. A machine once operational, fails after a time that is exponentially distributed with mean $1/\mu$. All failure and repair times are independent. What is the steady-state proportion of time where there is no operational machine?

Solution. Define the state to be the number of operational machines. This results in a continuous-time Markov chain, which is the same as an $M/M/1$ queue with arrival rate λ and service rate μ, except that there is storage for any m customers. The required probability is simply π_0 for this queue.

7

Limit Theorems

Contents

Consider a sequence $X_1, X_2, \ldots$ of independent identically distributed random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots + X_n$$

be the sum of the first n of them. Limit theorems are mostly concerned with the properties of S_n and related random variables, as n becomes very large.

Because of independence, we have

$$\text{var}(S_n) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n\sigma^2.$$

Thus, the distribution of S_n spreads out as n increases, and does not have a meaningful limit. The situation is different if we consider the **sample mean**

$$M_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

A quick calculation yields

$$\mathbf{E}[M_n] = \mu, \qquad \text{var}(M_n) = \frac{\sigma^2}{n}.$$

In particular, the variance of M_n decreases to zero as n increases, and the bulk of its distribution must be very close to the mean μ. This phenomenon is the subject of certain laws of large numbers, which generally assert that the sample mean M_n (a random variable) converges to the true mean μ (a number), in a precise sense. These laws provide a mathematical basis for the loose interpretation of an expectation $\mathbf{E}[X] = \mu$ as the average of a large number of independent samples drawn from the distribution of X.

We will also consider a quantity which is intermediate between S_n and M_n. We first subtract $n\mu$ from S_n, to obtain the zero-mean random variable $S_n - n\mu$ and then divide by $\sigma\sqrt{n}$, to obtain

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

It can be verified (see Section 7.4) that

$$\mathbf{E}[Z_n] = 0, \qquad \text{var}(Z_n) = 1.$$

Since the mean and the variance of Z_n remain unchanged as n increases, its distribution neither spreads, nor shrinks to a point. The **central limit theorem** is concerned with the asymptotic shape of the distribution of Z_n and asserts that it becomes the standard normal distribution.

Limit theorems are useful for several reasons:

(a) Conceptually, they provide an interpretation of expectations (as well as probabilities) in terms of a long sequence of identical independent experiments.

(b) They allow for an approximate analysis of the properties of random variables such as S_n. This is to be contrasted with an exact analysis which would require a formula for the PMF or PDF of S_n, a complicated and tedious task when n is large.

7.1 SOME USEFUL INEQUALITIES

In this section, we derive some important inequalities. These inequalities use the mean, and possibly the variance, of a random variable to draw conclusions on the probabilities of certain events. They are primarily useful in situations where the mean and variance of a random variable X are easily computable, but the distribution of X is either unavailable or hard to calculate.

We first present the **Markov inequality**. Loosely speaking it asserts that if a *nonnegative* random variable has a small mean, then the probability that it takes a large value must also be small.

Markov Inequality

If a random variable X can only take nonnegative values, then

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}, \qquad \text{for all } a > 0.$$

To justify the Markov inequality, let us fix a positive number a and consider the random variable Y_a defined by

$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a. \end{cases}$$

It is seen that the relation

$$Y_a \leq X$$

always holds and therefore,

$$\mathbf{E}[Y_a] \leq \mathbf{E}[X].$$

On the other hand,

$$\mathbf{E}[Y_a] = a\mathbf{P}(Y_a = a) = a\mathbf{P}(X \geq a),$$

from which we obtain

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

Example 7.1. Let X be uniformly distributed on the interval $[0, 4]$ and note that $\mathbf{E}[X] = 2$. Then, the Markov inequality asserts that

$$\mathbf{P}(X \geq 2) \leq \frac{2}{2} = 1, \qquad \mathbf{P}(X \geq 3) \leq \frac{2}{3} = 0.67, \qquad \mathbf{P}(X \geq 4) \leq \frac{2}{4} = 0.5.$$

By comparing with the exact probabilities

$$\mathbf{P}(X \geq 2) = 0.5, \qquad \mathbf{P}(X \geq 3) = 0.25, \qquad \mathbf{P}(X \geq 4) = 0,$$

we see that the bounds provided by the Markov inequality can be quite loose.

We continue with the **Chebyshev inequality**. Loosely speaking, it asserts that if the variance of a random variable is small, then the probability that it takes a value far from its mean is also small. Note that the Chebyshev inequality does not require the random variable to be nonnegative.

Chebyshev Inequality

If X is a random variable with mean μ and variance σ^2, then

$$\mathbf{P}\big(|X-\mu| \geq c\big) \leq \frac{\sigma^2}{c^2}, \qquad \text{for all } c > 0.$$

To justify the Chebyshev inequality, we consider the nonnegative random variable $(X-\mu)^2$ and apply the Markov inequality with $a = c^2$. We obtain

$$\mathbf{P}\big((X-\mu)^2 \geq c^2\big) \leq \frac{\mathbf{E}\big[(X-\mu)^2\big]}{c^2} = \frac{\sigma^2}{c^2}.$$

The derivation is completed by observing that the event $(X-\mu)^2 \geq c^2$ is identical to the event $|X-\mu| \geq c$ and

$$\mathbf{P}\big(|X-\mu| \geq c\big) = \mathbf{P}\big((X-\mu)^2 \geq c^2\big) \leq \frac{\sigma^2}{c^2}.$$

An alternative form of the Chebyshev inequality is obtained by letting $c = k\sigma$, where k is positive, which yields

$$\mathbf{P}\big(|X-\mu| \geq k\sigma\big) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Thus, the probability that a random variable takes a value more than k standard deviations away from its mean is at most $1/k^2$.

The Chebyshev inequality is generally more powerful than the Markov inequality (the bounds that it provides are more accurate), because it also makes use of information on the variance of X. Still, the mean and the variance of a random variable are only a rough summary of the properties of its distribution, and we cannot expect the bounds to be close approximations of the exact probabilities.

Example 7.2. As in Example 7.1, let X be uniformly distributed on $[0, 4]$. Let us use the Chebyshev inequality to bound the probability that $|X - 2| \geq 1$. We have $\sigma^2 = 16/12 = 4/3$, and

$$\mathbf{P}\big(|X-2| \geq 1\big) \leq \frac{4}{3},$$

which is not particularly informative.

For another example, let X be exponentially distributed with parameter $\lambda = 1$, so that $\mathbf{E}[X] = \mathrm{var}(X) = 1$. For $c > 1$, using Chebyshev's inequality, we obtain

$$\mathbf{P}(X \geq c) = \mathbf{P}(X - 1 \geq c - 1) \leq \mathbf{P}\big(|X - 1| \geq c - 1\big) \leq \frac{1}{(c-1)^2}.$$

This is again conservative compared to the exact answer $\mathbf{P}(X \geq c) = e^{-c}$.

7.2 THE WEAK LAW OF LARGE NUMBERS

The weak law of large numbers asserts that the sample mean of a large number of independent identically distributed random variables is very close to the true mean, with high probability.

As in the introduction to this chapter, we consider a sequence $X_1, X_2, \ldots$ of independent identically distributed random variables with mean μ and variance σ^2, and define the sample mean by

$$M_n = \frac{X_1 + \cdots + X_n}{n}.$$

We have

$$\mathbf{E}[M_n] = \frac{\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}{n} = \frac{n\mu}{n} = \mu,$$

and, using independence,

$$\mathrm{var}(M_n) = \frac{\mathrm{var}(X_1 + \cdots + X_n)}{n^2} = \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

We apply Chebyshev's inequality and obtain

$$\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) \leq \frac{\sigma^2}{n\epsilon^2}, \qquad \text{for any } \epsilon > 0.$$

We observe that for any fixed $\epsilon > 0$, the right-hand side of this inequality goes to zero as n increases. As a consequence, we obtain the weak law of large numbers, which is stated below. It turns out that this law remains true even if the X_i

have infinite variance, but a much more elaborate argument is needed, which we omit. The only assumption needed is that $\mathbf{E}[X_i]$ is well-defined and finite.

The Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots$ be independent identically distributed random variables with mean μ. For every $\epsilon > 0$, we have

$$\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) = \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0, \qquad \text{as } n \to \infty.$$

The WLLN states that for large n, the "bulk" of the distribution of M_n is concentrated near μ. That is, if we consider a positive length interval $[\mu-\epsilon, \mu+\epsilon]$ around μ, then there is high probability that M_n will fall in that interval; as $n \to \infty$, this probability converges to 1. Of course, if ϵ is very small, we may have to wait longer (i.e., need a larger value of n) before we can assert that M_n is highly likely to fall in that interval.

Example 7.3. Probabilities and Frequencies. Consider an event A defined in the context of some probabilistic experiment. Let $p = \mathbf{P}(A)$ be the probability of that event. We consider n independent repetitions of the experiment, and let M_n be the fraction of time that event A occurred; in this context, M_n is often called the **empirical frequency** of A. Note that

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

where X_i is 1 whenever A occurs, and 0 otherwise; in particular, $\mathbf{E}[X_i] = p$. The weak law applies and shows that when n is large, the empirical frequency is most likely to be within ϵ of p. Loosely speaking, this allows us to say that empirical frequencies are faithful estimates of p. Alternatively, this is a step towards interpreting the probability p as the frequency of occurrence of A.

Example 7.4. Polling. Let p be the fraction of voters who support a particular candidate for office. We interview n "randomly selected" voters and record the fraction M_n of them that support the candidate. We view M_n as our estimate of p and would like to investigate its properties.

We interpret "randomly selected" to mean that the n voters are chosen independently and uniformly from the given population. Thus, the reply of each person interviewed can be viewed as an independent Bernoulli trial X_i with success probability p and variance $\sigma^2 = p(1-p)$. The Chebyshev inequality yields

$$\mathbf{P}\big(|M_n - p| \geq \epsilon\big) \leq \frac{p(1-p)}{n\epsilon^2}.$$

The true value of the parameter p is assumed to be unknown. On the other hand, it is easily verified that $p(1-p) \leq 1/4$, which yields

$$\mathbf{P}\big(|M_n - p| \geq \epsilon\big) \leq \frac{1}{4n\epsilon^2}.$$

For example, if $\epsilon = 0.1$ and $n = 100$, we obtain

$$\mathbf{P}\big(|M_{100} - p| \geq 0.1\big) \leq \frac{1}{4 \cdot 100 \cdot (0.1)^2} = 0.25.$$

In words, with a sample size of $n = 100$, the probability that our estimate is wrong by more than 0.1 is no larger than 0.25.

Suppose now that we impose some tight specifications on our poll. We would like to have high confidence (probability at least 95%) that our estimate will be very accurate (within .01 of p). How many voters should be sampled?

The only guarantee that we have at this point is the inequality

$$\mathbf{P}\big(|M_n - p| \geq 0.01\big) \leq \frac{1}{4n(0.01)^2}.$$

We will be sure to satisfy the above specifications if we choose n large enough so that

$$\frac{1}{4n(0.01)^2} \leq 1 - 0.95 = 0.05,$$

which yields $n \geq 50,000$. This choice of n has the specified properties but is actually fairly conservative, because it is based on the rather loose Chebyshev inequality. A refinement will be considered in Section 7.4.

7.3 CONVERGENCE IN PROBABILITY

We can interpret the WLLN as stating that "M_n converges to μ." However, since $M_1, M_2, \ldots$ is a sequence of random variables, not a sequence of numbers, the meaning of convergence has to be made precise. A particular definition is provided below. To facilitate the comparison with the ordinary notion of convergence, we also include the definition of the latter.

Convergence of a Deterministic Sequence

Let $a_1, a_2, \ldots$ be a sequence of real numbers, and let a be another real number. We say that the sequence a_n converges to a, or $\lim_{n\to\infty} a_n = a$, if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \leq \epsilon, \qquad \text{for all } n \geq n_0.$$

Intuitively, for any given accuracy level ϵ, a_n must be within ϵ of a, when n is large enough.

Convergence in Probability

Let $Y_1, Y_2, \ldots$ be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence Y_n **converges to a in probability**, if for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbf{P}\big(|Y_n - a| \geq \epsilon\big) = 0.$$

Given this definition, the WLLN simply says that the sample mean converges in probability to the true mean μ.

If the random variables $Y_1, Y_2, \ldots$ have a PMF or a PDF and converge in probability to a, then according to the above definition, "almost all" of the PMF or PDF of Y_n is concentrated to within a an ϵ-interval around a for large values of n. It is also instructive to rephrase the above definition as follows: for every $\epsilon > 0$, and for every $\delta > 0$, there exists some n_0 such that

$$\mathbf{P}\big(|Y_n - a| \geq \epsilon\big) \leq \delta, \qquad \text{for all } n \geq n_0.$$

If we refer to ϵ as the *accuracy* level, and δ as the *confidence* level, the definition takes the following intuitive form: for any given level of accuracy and confidence, Y_n will be equal to a, within these levels of accuracy and confidence, provided that n is large enough.

Example 7.5. Consider a sequence of independent random variables X_n that are uniformly distributed over the interval $[0, 1]$, and let

$$Y_n = \min\{X_1, \ldots, X_n\}.$$

The sequence of values of Y_n cannot increase as n increases, and it will occasionally decrease (when a value of X_n that is smaller than the preceding values is obtained). Thus, we intuitively expect that Y_n converges to zero. Indeed, for $\epsilon > 0$, we have using the independence of the X_n,

$$\begin{aligned}\mathbf{P}\big(|Y_n - 0| \geq \epsilon\big) &= \mathbf{P}(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon)\\ &= \mathbf{P}(X_1 \geq \epsilon) \cdots \mathbf{P}(X_n \geq \epsilon)\\ &= (1-\epsilon)^n.\end{aligned}$$

Since this is true for every $\epsilon > 0$, we conclude that Y_n converges to zero, in probability.

Example 7.6. Let Y be an exponentially distributed random variable with parameter $\lambda = 1$. For any positive integer n, let $Y_n = Y/n$. (Note that these random variables are dependent.) We wish to investigate whether the sequence Y_n converges to zero.

For $\epsilon > 0$, we have

$$\mathbf{P}\big(|Y_n - 0| \geq \epsilon\big) = \mathbf{P}(Y_n \geq \epsilon) = \mathbf{P}(Y \geq n\epsilon) = e^{-n\epsilon}.$$

In particular,

$$\lim_{n\to\infty} \mathbf{P}\big(|Y_n - 0| \geq \epsilon\big) = \lim_{n\to\infty} e^{-n\epsilon} = 0.$$

Since this is the case for every $\epsilon > 0$, Y_n converges to zero, in probability.

One might be tempted to believe that if a sequence Y_n converges to a number a, then $\mathbf{E}[Y_n]$ must also converge to a. The following example shows that this need not be the case.

Example 7.7. Consider a sequence of discrete random variables Y_n with the following distribution:

$$\mathbf{P}(Y_n = y) = \begin{cases} 1 - \dfrac{1}{n}, & \text{for } y = 0, \\ \dfrac{1}{n}, & \text{for } y = n^2, \\ 0, & \text{elsewhere.} \end{cases}$$

For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbf{P}\big(|Y_n| \geq \epsilon\big) = \lim_{n\to\infty} \frac{1}{n} = 0,$$

and Y_n converges to zero in probability. On the other hand, $\mathbf{E}[Y_n] = n^2/n = n$, which goes to infinity as n increases.

7.4 THE CENTRAL LIMIT THEOREM

According to the weak law of large numbers, the distribution of the sample mean M_n is increasingly concentrated in the near vicinity of the true mean μ. In particular, its variance tends to zero. On the other hand, the variance of the sum $S_n = X_1 + \cdots + X_n = nM_n$ increases to infinity, and the distribution of S_n cannot be said to converge to anything meaningful. An intermediate view is obtained by considering the deviation $S_n - n\mu$ of S_n from its mean $n\mu$, and scaling it by a factor proportional to $1/\sqrt{n}$. What is special about this particular scaling is that it keeps the variance at a constant level. The central limit theorem asserts that the distribution of this scaled random variable approaches a normal distribution.

More specifically, let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2. We define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

An easy calculation yields

$$\mathbf{E}[Z_n] = \frac{\mathbf{E}[X_1 + \cdots + X_n] - n\mu}{\sigma\sqrt{n}} = 0,$$

and

$$\text{var}(Z_n) = \frac{\text{var}(X_1 + \cdots + X_n)}{\sigma^2 n} = \frac{\text{var}(X_1) + \cdots + \text{var}(X_n)}{\sigma^2 n} = \frac{n\sigma^2}{n\sigma^2} = 1.$$

The Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with common mean μ and variance σ^2, and define

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then, the CDF of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\,dx,$$

in the sense that

$$\lim_{n\to\infty} \mathbf{P}(Z_n \le z) = \Phi(z), \qquad \text{for every } z.$$

The central limit theorem is surprisingly general. Besides independence, and the implicit assumption that the mean and variance are well-defined and finite, it places no other requirement on the distribution of the X_i, which could be discrete, continuous, or mixed random variables. It is of tremendous importance for several reasons, both conceptual, as well as practical. On the conceptual side, it indicates that the sum of a large number of independent random variables is approximately normal. As such, it applies to many situations in which a random effect is the sum of a large number of small but independent random factors. Noise in many natural or engineered systems has this property. In a wide array of contexts, it has been found empirically that the statistics of noise are well-described by normal distributions, and the central limit theorem provides a convincing explanation for this phenomenon.

On the practical side, the central limit theorem eliminates the need for detailed probabilistic models and for tedious manipulations of PMFs and PDFs. Rather, it allows the calculation of certain probabilities by simply referring to the normal CDF table. Furthermore, these calculations only require the knowledge of means and variances.

Approximations Based on the Central Limit Theorem

The central limit theorem allows us to calculate probabilities related to Z_n as if Z_n were normal. Since normality is preserved under linear transformations, this is equivalent to treating S_n as a normal random variable with mean $n\mu$ and variance $n\sigma^2$.

Normal Approximation Based on the Central Limit Theorem

Let $S_n = X_1 + \cdots + X_n$, where the X_i are independent identically distributed random variables with mean μ and variance σ^2. If n is large, the probability $\mathbf{P}(S_n \leq c)$ can be approximated by treating S_n as if it were normal, according to the following procedure.

1. Calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n.
2. Calculate the normalized value $z = (c - n\mu)/\sigma\sqrt{n}$.
3. Use the approximation

$$\mathbf{P}(S_n \leq c) \approx \Phi(z),$$

where $\Phi(z)$ is available from standard normal CDF tables.

Example 7.8. We load on a plane 100 packages whose weights are independent random variables that are uniformly distributed between 5 and 50 pounds. What is the probability that the total weight will exceed 3000 pounds? It is not easy to calculate the CDF of the total weight and the desired probability, but an approximate answer can be quickly obtained using the central limit theorem.

We want to calculate $\mathbf{P}(S_{100} > 3000)$, where S_{100} is the sum of the 100 packages. The mean and the variance of the weight of a single package are

$$\mu = \frac{5 + 50}{2} = 27.5, \qquad \sigma^2 = \frac{(50 - 5)^2}{12} = 168.75,$$

based on the formulas for the mean and variance of the uniform PDF. We thus calculate the normalized value

$$z = \frac{3000 - 100 \cdot 27.5}{\sqrt{168.75 \cdot 100}} = \frac{250}{129.9} = 1.92,$$

and use the standard normal tables to obtain the approximation

$$\mathbf{P}(S_{100} \le 3000) \approx \Phi(1.92) = 0.9726.$$

Thus the desired probability is

$$\mathbf{P}(S_{100} > 3000) = 1 - \mathbf{P}(S_{100} \le 3000) \approx 1 - 0.9726 = 0.0274.$$

Example 7.9. A machine processes parts, one at a time. The processing times of different parts are independent random variables, uniformly distributed on $[1, 5]$. We wish to approximate the probability that the number of parts processed within 320 time units is at least 100.

Let us call N_{320} this number. We want to calculate $\mathbf{P}(N_{320} \ge 100)$. There is no obvious way of expressing the random variable N_{320} as the sum of independent random variables, but we can proceed differently. Let X_i be the processing time of the ith part, and let $S_{100} = X_1 + \cdots + X_{100}$ be the total processing time of the first 100 parts. The event $\{N_{320} \ge 100\}$ is the same as the event $\{S_{100} \le 320\}$, and we can now use a normal approximation to the distribution of S_{100}. Note that $\mu = \mathbf{E}[X_i] = 3$ and $\sigma^2 = \text{var}(X_i) = 16/12 = 4/3$. We calculate the normalized value

$$z = \frac{320 - n\mu}{\sigma\sqrt{n}} = \frac{320 - 300}{\sqrt{100 \cdot 4/3}} = 1.73,$$

and use the approximation

$$\mathbf{P}(S_{100} \le 320) \approx \Phi(1.73) = 0.9582.$$

If the variance of the X_i is unknown, but an upper bound is available, the normal approximation can be used to obtain bounds on the probabilities of interest.

Example 7.10. Let us revisit the polling problem in Example 7.4. We poll n voters and record the fraction M_n of those polled who are in favor of a particular candidate. If p is the fraction of the entire voter population that supports this candidate, then

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

where the X_i are independent Bernoulli random variables with parameter p. In particular, M_n has mean p and variance $p(1-p)/n$. By the normal approximation,

$X_1 + \cdots + X_n$ is approximately normal, and therefore M_n is also approximately normal.

We are interested in the probability $\mathbf{P}\big(|M_n - p| \geq \epsilon\big)$ that the polling error is larger than some desired accuracy ϵ. Because of the symmetry of the normal PDF around the mean, we have

$$\mathbf{P}\big(|M_n - p| \geq \epsilon\big) \approx 2\mathbf{P}(M_n - p \geq \epsilon).$$

The variance $p(1-p)/n$ of $M_n - p$ depends on p and is therefore unknown. We note that the probability of a large deviation from the mean increases with the variance. Thus, we can obtain an upper bound on $\mathbf{P}\big(M_n - p \geq \epsilon\big)$ by assuming that $M_n - p$ has the largest possible variance, namely, $1/4n$. To calculate this upper bound, we evaluate the standardized value

$$z = \frac{\epsilon}{1/(2\sqrt{n})},$$

and use the normal approximation

$$\mathbf{P}\big(M_n - p \geq \epsilon\big) \leq 1 - \Phi(z) = 1 - \Phi\big(2\epsilon\sqrt{n}\big).$$

For instance, consider the case where $n = 100$ and $\epsilon = 0.1$. Assuming the worst-case variance, we obtain

$$\begin{aligned}\mathbf{P}\big(|M_{100} - p| \geq 0.1\big) &\approx 2\mathbf{P}(M_n - p \geq 0.1)\\ &\leq 2 - 2\Phi\big(2 \cdot 0.1 \cdot \sqrt{100}\big) = 2 - 2\Phi(2) = 2 - 2 \cdot 0.977 = 0.046.\end{aligned}$$

This is much smaller (more accurate) than the estimate that was obtained in Example 7.4 using the Chebyshev inequality.

We now consider a reverse problem. How large a sample size n is needed if we wish our estimate M_n to be within 0.01 of p with probability at least 0.95? Assuming again the worst possible variance, we are led to the condition

$$2 - 2\Phi\big(2 \cdot 0.01 \cdot \sqrt{n}\big) \leq 0.05,$$

or

$$\Phi\big(2 \cdot 0.01 \cdot \sqrt{n}\big) \geq 0.975.$$

From the normal tables, we see that $\Phi(1.96) = 0.975$, which leads to

$$2 \cdot 0.01 \cdot \sqrt{n} \geq 1.96,$$

or

$$n \geq \frac{(1.96)^2}{4 \cdot (0.01)^2} = 9604.$$

This is significantly better than the sample size of 50,000 that we found using Chebyshev's inequality.

The normal approximation is increasingly accurate as n tends to infinity, but in practice we are generally faced with specific and finite values of n. It would be useful to know how large an n is needed before the approximation can be trusted, but there are no simple and general guidelines. Much depends on whether the distribution of the X_i is close to normal to start with and, in particular, whether it is symmetric. For example, if the X_i are uniform, then S_8 is already very close to normal. But if the X_i are, say, exponential, a significantly larger n will be needed before the distribution of S_n is close to a normal one. Furthermore, the normal approximation to $\mathbf{P}(S_n \le c)$ is generally more faithful when c is in the vicinity of the mean of S_n.

The De Moivre – Laplace Approximation to the Binomial

A binomial random variable S_n with parameters n and p can be viewed as the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$, with common parameter p:

$$S_n = X_1 + \cdots + X_n.$$

Recall that

$$\mu = \mathbf{E}[X_i] = p, \qquad \sigma = \sqrt{\mathrm{var}(X_i)} = \sqrt{p(1-p)},$$

We will now use the approximation suggested by the central limit theorem to provide an approximation for the probability of the event $\{k \le S_n \le \ell\}$, where k and ℓ are given integers. We express the event of interest in terms of a standardized random variable, using the equivalence

$$k \le S_n \le \ell \qquad \Longleftrightarrow \qquad \frac{k-np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{\ell - np}{\sqrt{np(1-p)}}.$$

By the central limit theorem, $(S_n - np)/\sqrt{np(1-p)}$ has approximately a standard normal distribution, and we obtain

$$\begin{aligned}\mathbf{P}(k \le S_n \le \ell) =& \mathbf{P}\left(\frac{k-np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{\ell - np}{\sqrt{np(1-p)}}\right)\\ \approx& \Phi\left(\frac{\ell-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right).\end{aligned}$$

An approximation of this form is equivalent to treating S_n as a normal random variable with mean np and variance $np(1-p)$. Figure 7.1 provides an illustration and indicates that a more accurate approximation may be possible if we replace k and ℓ by $k - \frac{1}{2}$ and $\ell + \frac{1}{2}$, respectively. The corresponding formula is given below.

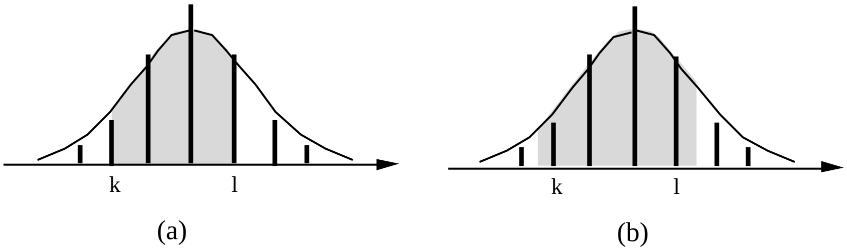


Figure 7.1: The central limit approximation treats a binomial random variable S_n as if it were normal with mean np and variance $np(1-p)$. This figure shows a binomial PMF together with the approximating normal PDF. (a) A first approximation of a binomial probability $\mathbf{P}(k \le S_n \le \ell)$ is obtained by integrating the area under the normal PDF from k to ℓ, which is the shaded area in the figure. (b) With the approach in (a), if we have $k = \ell$, the probability $\mathbf{P}(S_n = k)$ would be approximated by zero. A potential remedy would be to use the normal probability between $k - \frac{1}{2}$ and $k + \frac{1}{2}$ to approximate $\mathbf{P}(S_n = k)$. By extending this idea, $\mathbf{P}(k \le S_n \le \ell)$ can be approximated by using the area under the normal PDF from $k - \frac{1}{2}$ to $\ell + \frac{1}{2}$, which corresponds to the shaded area.

De Moivre – Laplace Approximation to the Binomial

If S_n is a binomial random variable with parameters n and p, n is large, and k, ℓ are nonnegative integers, then

$$\mathbf{P}(k \le S_n \le \ell) \approx \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

Example 7.11. Let S_n be a binomial random variable with parameters $n = 36$ and $p = 0.5$. An exact calculation yields

$$\mathbf{P}(S_n \le 21) = \sum_{k=0}^{21} \binom{36}{k} (0.5)^{36} = 0.8785.$$

The central limit approximation, without the above discussed refinement, yields

$$\mathbf{P}(S_n \le 21) \approx \Phi\left(\frac{21 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21 - 18}{3}\right) = \Phi(1) = 0.8413.$$

Using the proposed refinement, we have

$$\mathbf{P}(S_n \le 21) \approx \Phi\left(\frac{21.5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21.5 - 18}{3}\right) = \Phi(1.17) = 0.879,$$

which is much closer to the exact value.

The de Moivre – Laplace formula also allows us to approximate the probability of a single value. For example,

$$\mathbf{P}(S_n = 19) \approx \Phi\left(\frac{19.5 - 18}{3}\right) - \Phi\left(\frac{18.5 - 18}{3}\right) = 0.6915 - 05675 = 0.124.$$

This is very close to the exact value which is

$$\binom{36}{19}(0.5)^{36} = 0.1251.$$

7.5 THE STRONG LAW OF LARGE NUMBERS

The strong law of large numbers is similar to the weak law in that it also deals with the convergence of the sample mean to the true mean. It is different, however, because it refers to another type of convergence.

The Strong Law of Large Numbers (SLLN)

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with mean μ. Then, the sequence of sample means $M_n = (X_1 + \cdots + X_n)/n$ converges to μ, *with probability 1*, in the sense that

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

In order to interpret the SSLN, we need to go back to our original description of probabilistic models in terms of sample spaces. The contemplated experiment is infinitely long and generates experimental values for each one of the random variables in the sequence $X_1, X_2, \ldots$. Thus, it is best to think of the sample space Ω as a set of infinite sequences $\omega = (x_1, x_2, \ldots)$ of real numbers: any such sequence is a possible outcome of the experiment. Let us now define the subset A of Ω consisting of those sequences $(x_1, x_2, \ldots)$ whose long-term average is μ, i.e.,

$$(x_1, x_2, \ldots) \in A \qquad \Longleftrightarrow \qquad \lim_{n\to\infty} \frac{x_1 + \cdots + x_n}{n} = \mu.$$

The SLLN states that all of the probability is concentrated on this particular subset of Ω. Equivalently, the collection of outcomes that do not belong to A (infinite sequences whose long-term average is not μ) has probability zero.

The difference between the weak and the strong law is subtle and deserves close scrutiny. The weak law states that the probability $\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big)$ of a significant deviation of M_n from μ goes to zero as $n \to \infty$. Still, for any finite n, this probability can be positive and it is conceivable that once in a while, even if infrequently, M_n deviates significantly from μ. The weak law provides no conclusive information on the number of such deviations, but the strong law does. According to the strong law, and with probability 1, M_n converges to μ. This implies that for any given $\epsilon > 0$, the difference $|M_n - \mu|$ will exceed ϵ only a finite number of times.

Example 7.12. Probabilities and Frequencies. As in Example 7.3, consider an event A defined in terms of some probabilistic experiment. We consider a sequence of independent repetitions of the same experiment, and let M_n be the fraction of the first n trials in which A occurs. The strong law of large numbers asserts that M_n converges to $\mathbf{P}(A)$, with probability 1.

We have often talked intuitively about the probability of an event A as the frequency with which it occurs in an infinitely long sequence of independent trials. The strong law backs this intuition and establishes that the long-term frequency of occurrence of A is indeed equal to $\mathbf{P}(A)$, with certainty (the probability of this happening is 1).

Convergence with Probability 1

The convergence concept behind the strong law is different than the notion employed in the weak law. We provide here a definition and some discussion of this new convergence concept.

Convergence with Probability 1

Let $Y_1, Y_2, \ldots$ be a sequence of random variables (not necessarily independent) associated with the same probability model. Let c be a real number. We say that Y_n converges to c **with probability 1** (or **almost surely**) if

$$\mathbf{P}\left(\lim_{n\to\infty} Y_n = c\right) = 1.$$

Similar to our earlier discussion, the right way of interpreting this type of convergence is in terms of a sample space consisting of infinite sequences: all of the probability is concentrated on those sequences that converge to c. This does not mean that other sequences are impossible, only that they are extremely unlikely, in the sense that their total probability is zero.

The example below illustrates the difference between convergence in probability and convergence with probability 1.

Example 7.13. Consider a discrete-time arrival process. The set of times is partitioned into consecutive intervals of the form $I_k = \{2^k, 2^k + 1, \ldots, 2^{k+1} - 1\}$. Note that the length of I_k is 2^k, which increases with k. During each interval I_k, there is exactly one arrival, and all times within an interval are equally likely. The arrival times within different intervals are assumed to be independent. Let us define $Y_n = 1$ if there is an arrival at time n, and $Y_n = 0$ if there is no arrival.

We have $\mathbf{P}(Y_n \neq 0) = 1/2^k$, if $n \in I_k$. Note that as n increases, it belongs to intervals I_k with increasingly large indices k. Consequently,

$$\lim_{n\to\infty} \mathbf{P}(Y_n \neq 0) = \lim_{k\to\infty} \frac{1}{2^k} = 0,$$

and we conclude that Y_n converges to 0 in probability. However, when we carry out the experiment, the total number of arrivals is infinite (one arrival during each interval I_k). Therefore, Y_n is unity for infinitely many values of n, the event $\{\lim_{n\to\infty} Y_n = 0\}$ has zero probability, and we do not have convergence with probability 1.

Intuitively, the following is happening. At any given time, there is a small (and diminishing with n) probability of a substantial deviation from 0 (convergence in probability). On the other hand, given enough time, a substantial deviation from 0 is certain to occur, and for this reason, we do not have convergence with probability 1.

Example 7.14. Let $X_1, X_2, \ldots$ be a sequence of independent random variables that are uniformly distributed on $[0, 1]$, and let $Y_n = \min\{X_1, \ldots, X_n\}$. We wish to show that Y_n converges to 0, with probability 1.

In any execution of the experiment, the sequence Y_n is nonincreasing, i.e., $Y_{n+1} \leq Y_n$ for all n. Since this sequence is bounded below by zero, it must have a limit, which we denote by Y. Let us fix some $\epsilon > 0$. If $Y \geq \epsilon$, then $X_i \geq \epsilon$ for all i, which implies that

$$\mathbf{P}(Y \geq \epsilon) \leq \mathbf{P}(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) = (1 - \epsilon)^n.$$

Since this is true for all n, we must have

$$\mathbf{P}(Y \geq \epsilon) \leq \lim_{n\to\infty} (1 - \epsilon)^n = 0.$$

This shows that $\mathbf{P}(Y \geq \epsilon) = 0$, for any positive ϵ. We conclude that $\mathbf{P}(Y > 0) = 0$, which implies that $\mathbf{P}(Y = 0) = 1$. Since Y is the limit of Y_n, we see that Y_n converges to zero with probability 1.