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

These notes are copyright-protected but may be freely distributed for instructional nonprofit pruposes.

Contents

1. Sample Space and Probability . . . . . .

1.1. Sets . . . . . .
1.2. Probabilistic Models . . . . . .
1.3. Conditional Probability . . . . . .
1.4. Independence . . . . . .
1.5. Total Probability Theorem and Bayes' Rule . . . . . .
1.6. Counting . . . . . .
1.7. Summary and Discussion . . . . . .

2. Discrete Random Variables . . . . . .

2.1. Basic Concepts . . . . . .
2.2. Probability Mass Functions . . . . . .
2.3. Functions of Random Variables . . . . . .
2.4. Expectation, Mean, and Variance . . . . . .
2.5. Joint PMFs of Multiple Random Variables . . . . . .
2.6. Conditioning . . . . . .
2.7. Independence . . . . . .
2.8. Summary and Discussion . . . . . .

3. General Random Variables . . . . . .

3.1. Continuous Random Variables and PDFs . . . . . .
3.2. Cumulative Distribution Functions . . . . . .
3.3. Normal Random Variables . . . . . .
3.4. Conditioning on an Event . . . . . .
3.5. Multiple Continuous Random Variables . . . . . .
3.6. Derived Distributions . . . . . .
3.7. Summary and Discussion . . . . . .

4. Further Topics on Random Variables and Expectations . . . . . .

4.1. Transforms . . . . . .
4.2. Sums of Independent Random Variables - Convolutions . . . . . .

4.3. Conditional Expectation as a Random Variable . . . . . .
4.4. Sum of a Random Number of Independent Random Variables . . . .
4.5. Covariance and Correlation . . . . . .
4.6. Least Squares Estimation . . . . . .
4.7. The Bivariate Normal Distribution . . . . . .

5. The Bernoulli and Poisson Processes . . . . . .

5.1. The Bernoulli Process . . . . . .
5.2. The Poisson Process . . . . . .

6. Markov Chains . . . . . .

6.1. Discrete-Time Markov Chains . . . . . .
6.2. Classification of States . . . . . .
6.3. Steady-State Behavior . . . . . .
6.4. Absorption Probabilities and Expected Time to Absorption . . . . .
6.5. More General Markov Chains . . . . . .

7. Limit Theorems . . . . . .

7.1. Some Useful Inequalities . . . . . .
7.2. The Weak Law of Large Numbers . . . . . .
7.3. Convergence in Probability . . . . . .
7.4. The Central Limit Theorem . . . . . .
7.5. The Strong Law of Large Numbers . . . . . .

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

1.1. Sets . . . . . p. 3
1.2. Probabilistic Models . . . . . p. 6
1.3. Conditional Probability . . . . . p. 16
1.4. Total Probability Theorem and Bayes' Rule . . . . . p. 25
1.5. Independence . . . . . p. 31
1.6. Counting . . . . . p. 41
1.7. Summary and Discussion . . . . . p. 48

"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

$$\{x \mid x \text{ satisfies } P\}.$$

(The symbol "|" is to be read as "such that.") For example the set of even integers can be written as $\{k \mid k/2 \text{ is integer}\}$. Similarly, the set of all scalars x in the interval $[0, 1]$ can be written as $\{x \mid 0 \leq x \leq 1\}$. 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 theoretical 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 $\{x \in \Omega \mid x \notin S\}$ 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,

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

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

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

and

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

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.

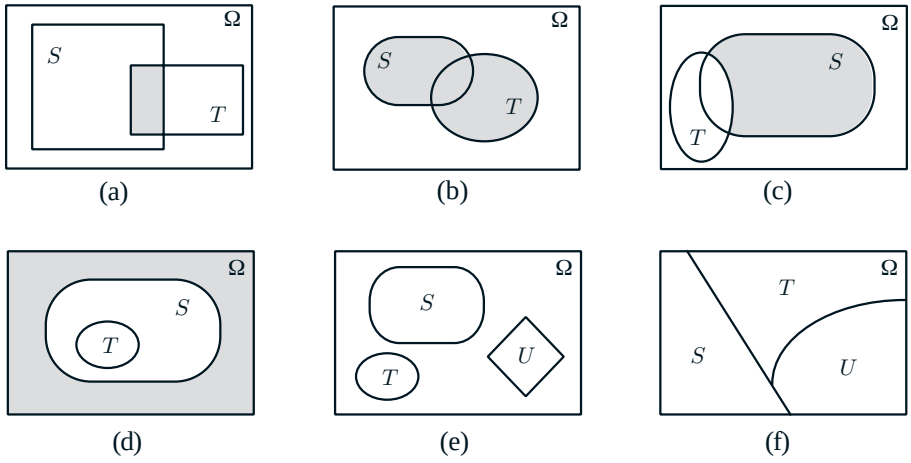


Figure 1.1: Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subset S$. The shaded region is the complement of S. (e) The sets S, T, and U are disjoint. (f) The sets S, T, and U form a partition of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

The Algebra of Sets

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

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

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.

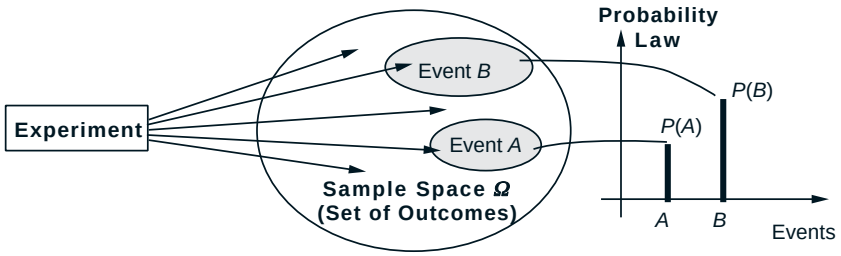


Figure 1.2: The main ingredients of a probabilistic model.

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.

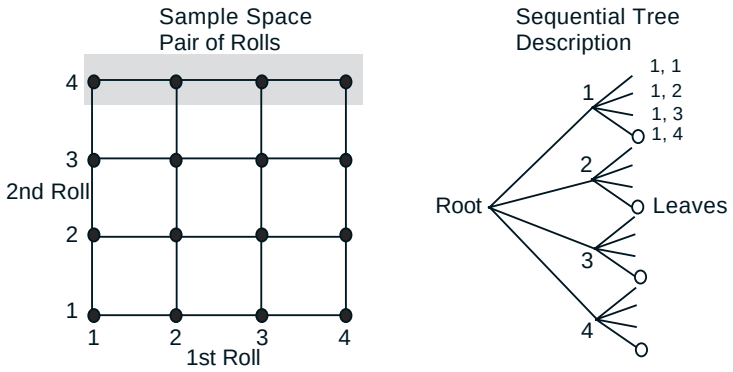


Figure 1.3: Two equivalent descriptions of the sample space of an experiment involving two rolls of a 4-sided die. The possible outcomes are all the ordered pairs of the form (i, j), where i is the result of the first roll, and j is the result of the second. These outcomes can be arranged in a 2-dimensional 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 $\mathbf{P}\big(\{s_1\}\big), \ldots, \mathbf{P}\big(\{s_n\}\big)$ 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:

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

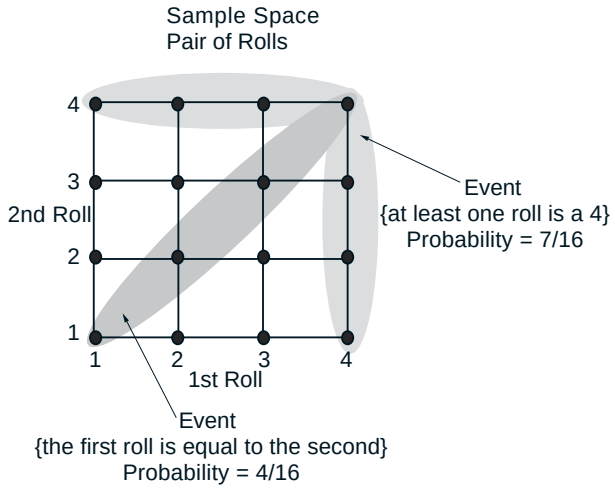


Figure 1.4: Various events in the experiment of rolling a pair of 4-sided dice, and their probabilities, calculated according to the discrete uniform law.

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.

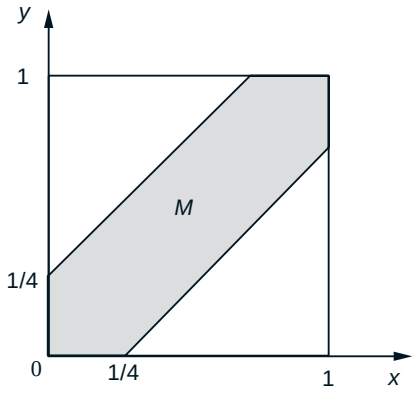


Figure 1.5: The event M that Romeo and Juliet will arrive within 15 minutes of each other (cf. Example 1.5) is

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

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

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

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

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

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

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

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

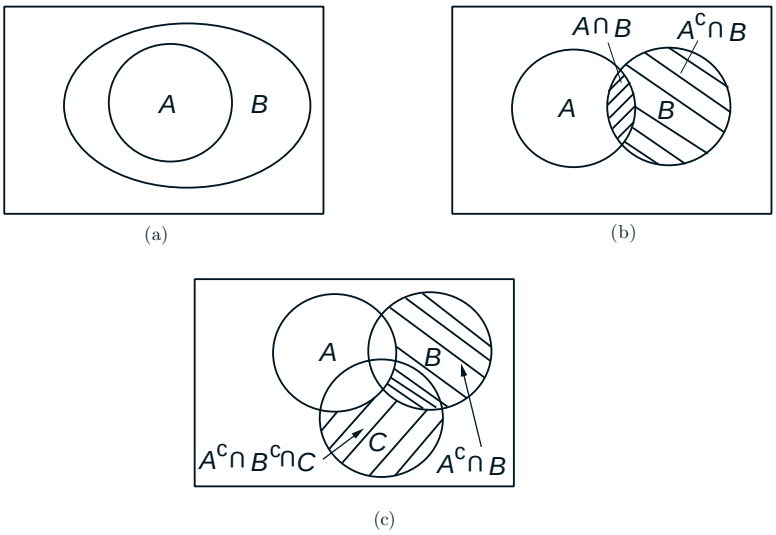


Figure 1.6: Visualization and verification of various properties of probability laws using Venn diagrams. If $A \subset B$, then B is the union of the two disjoint events A and $A^c \cap B$; see diagram (a). Therefore, by the additivity axiom, we have

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

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 $\mathbf{P}(A \mid B)$.

We would like the conditional probabilities $\mathbf{P}(A \mid B)$ 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

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

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

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

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

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

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, $\mathbf{P}(A \mid B)$ 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 $\mathbf{P}(A \mid B)$ form a legitimate probability law that satisfies the three axioms. Indeed, nonnegativity is clear. Furthermore,

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

and the normalization axiom is also satisfied. In fact, since we have $\mathbf{P}(B \mid B) = \mathbf{P}(B)/\mathbf{P}(B) = 1$, 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 $\mathbf{P}(A \cup C) \leq \mathbf{P}(A) + \mathbf{P}(C)$ translates to the new fact

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

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 $\mathbf{P}(A \mid B)$ 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 $\mathbf{P}(A \mid B)$ is

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

Because all possible outcomes are equally likely here, we can also compute $\mathbf{P}(A \mid B)$ 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 $\mathbf{P}(A \mid B)$ 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.

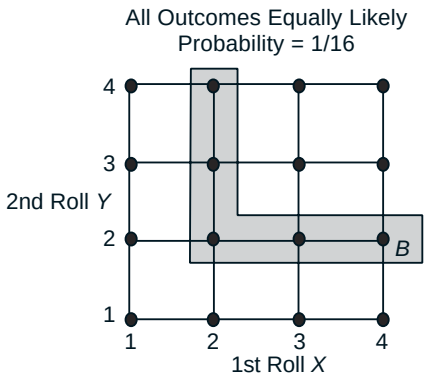


Figure 1.7: Sample space of an experiment involving two rolls of a 4-sided die. (cf. Example 1.7). The conditioning event $B = \{\min(X, Y) = 2\}$ consists of the 5-element shaded set. The set $A = \{\max(X, Y) = m\}$ shares with B two elements if $m = 3$ or $m = 4$, one element if $m = 2$, and no element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

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

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.

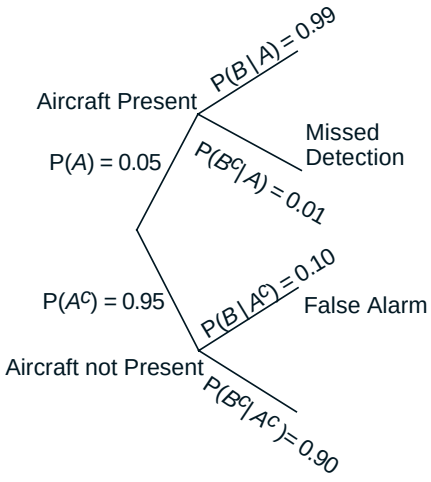


Figure 1.8: Sequential description of the sample space for the radar detection problem in Example 1.9

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

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

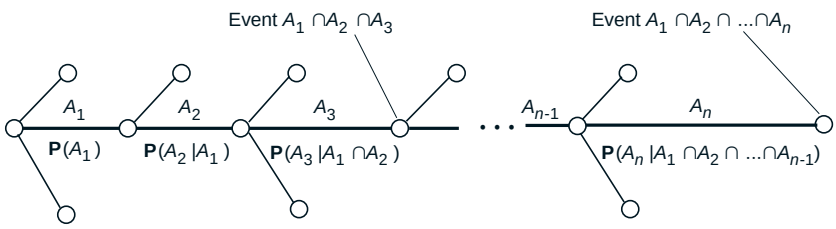


Figure 1.9: Visualization of the total probability theorem. The intersection event $A = A_1 \cap A_2 \cap \cdots \cap A_n$ is associated with a path on the tree of a sequential description of the experiment. We associate the branches of this path with the events $A_1, \ldots, A_n$, and we record next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

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

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

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

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,

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

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,

$$\begin{aligned}
\mathbf{P}(\text{1st is not a heart and 2nd is a heart}) &= \frac{39}{52} \cdot \frac{13}{51}, \\
\mathbf{P}(\text{1st two are not hearts and 3rd is a heart}) &= \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{13}{50}.
\end{aligned}$$

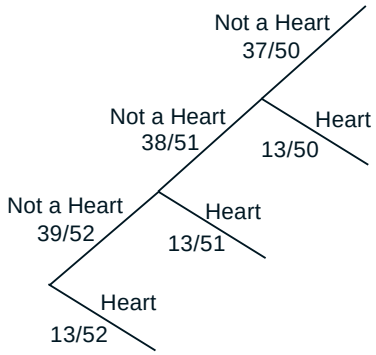


Figure 1.10: Sequential description of the sample space of the 3-card selection problem in Example 1.10.

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

$$A_1 = \{\text{students 1 and 2 are in different groups}\},$$

$$A_2 = \{\text{students 1, 2, and 3 are in different groups}\},$$

$$A_3 = \{\text{students 1, 2, 3, and 4 are in different groups}\}.$$

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.

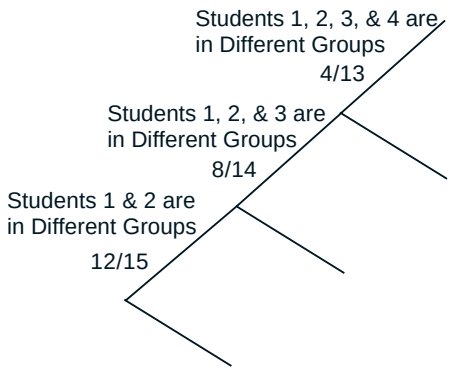


Figure 1.11: Sequential description of the sample space of the student problem in Example 1.11.

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

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

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 $\mathbf{P}(B \mid A_i)$ 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

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

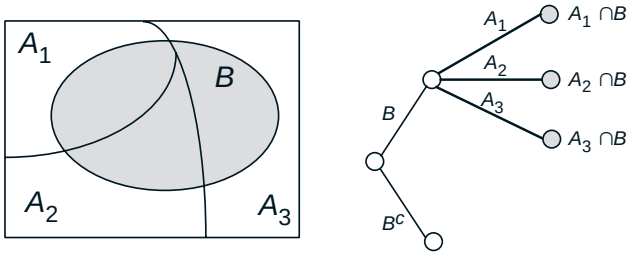


Figure 1.12: Visualization and verification of the total probability theorem. The events $A_1, \ldots, A_n$ form a partition of the sample space, so the event B can be decomposed into the disjoint union of its intersections $A_i \cap B$ with the sets A_i, i.e.,

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

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 caclulation 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 $\mathbf{P}(A \,|\, B)$ with conditional probabilities of the form $\mathbf{P}(B \,|\, A)$, 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

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

To verify Bayes' rule, note that $\mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i)$ and $\mathbf{P}(A_i \,|\, B)\mathbf{P}(B)$ 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.

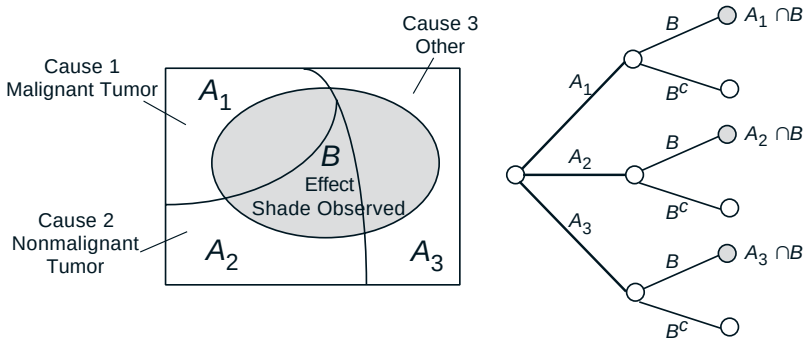


Figure 1.13: An example of the inference context that is implicit in Bayes' rule. We observe a shade in a person's X-ray (this is event B, the "effect") and we want to estimate the likelihood of three mutually exclusive and collectively exhaustive potential causes: cause 1 (event A_1) is that there is 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

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

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

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

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

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

so that $\mathbf{P}(H_1 \cap H_2 \mid D) \neq \mathbf{P}(H_1 \mid D)\mathbf{P}(H_2 \mid D)$, 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,

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

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

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

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 theoretical 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:

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

We have

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

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 theoretical 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

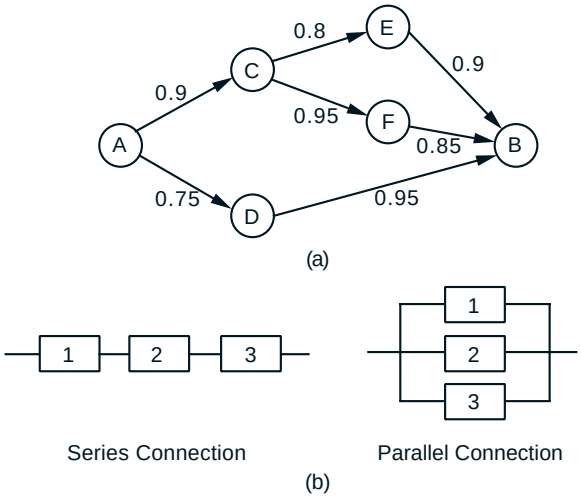


Figure 1.14: (a) Network for Example 1.22. The number next to each link (i, j) indicates the probability that the link is up. (b) Series and parallel connections of three components in a reliability problem.

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 $X \to Y$ 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

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

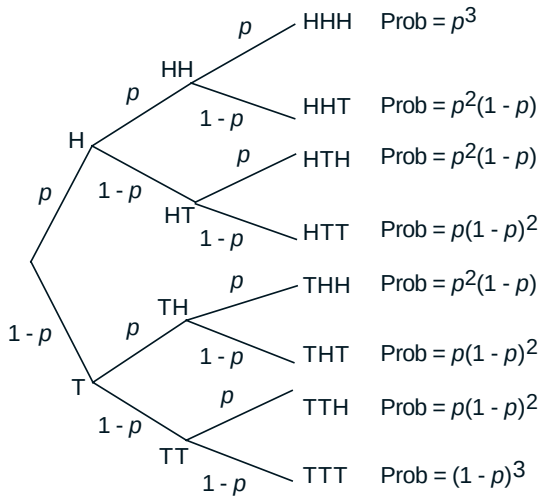


Figure 1.15: Sequential description of the sample space of an experiment involving three independent tosses of a biased coin. Along the branches of the tree, we record the corresponding conditional probabilities, and by the multiplication rule, the probability of obtaining a particular 3-toss sequence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

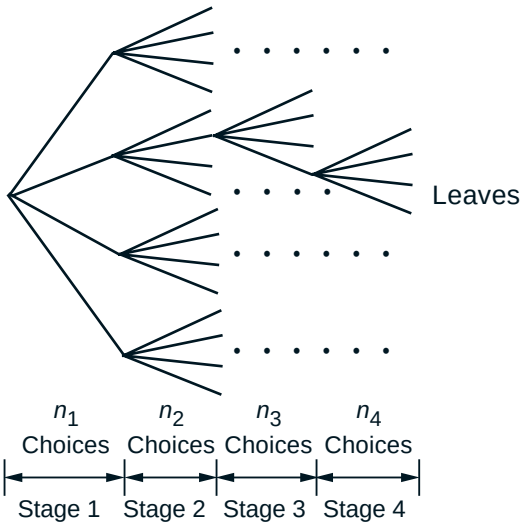


Figure 1.16: Illustration of the basic counting principle. The counting is carried out in r stages ($r = 4$ in the figure). The first stage has n_1 possible results. For every possible result of the first $i - 1$ stages, there are n_i possible results at the ith stage. The number of leaves is $n_1 n_2 \cdots n_r$. 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 $k \leq n$. 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 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 theoretical problems.) Let us rewrite TATTOO in the form $T_1AT_2T_3O_1O_2$ 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 T_1, T_1, and T_3, as well as any of the 2! possible permutations of O_1 and 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: $\dbinom{n}{k} = \dfrac{n!}{k!(n-k)!}$
- 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, be-

cause 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.

2

Discrete Random Variables

Contents

2.1. Basic Concepts . . . . . p. 2
2.2. Probability Mass Functions . . . . . p. 4
2.3. Functions of Random Variables . . . . . p. 9
2.4. Expectation, Mean, and Variance . . . . . p. 11
2.5. Joint PMFs of Multiple Random Variables . . . . . p. 22
2.6. Conditioning . . . . . p. 27
2.7. Independence . . . . . p. 36
2.8. Summary and Discussion . . . . . p. 42

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**.

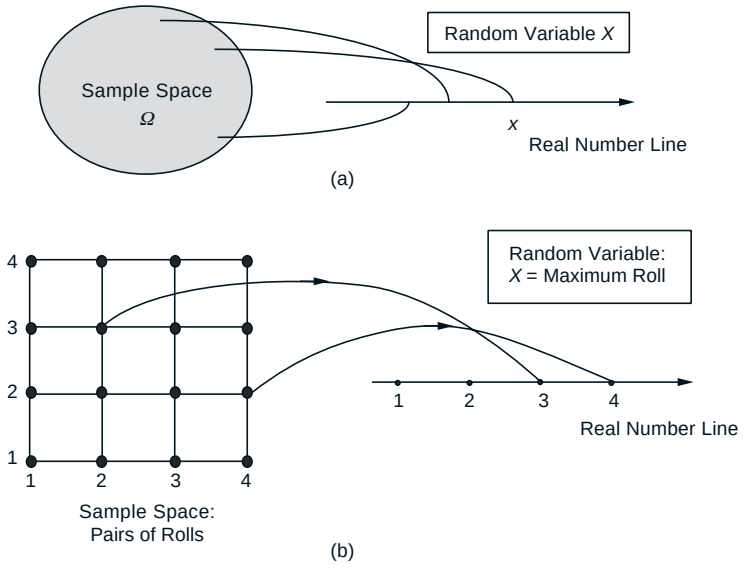


Figure 2.1: (a) Visualization of a random variable. It is a function that assigns a numerical value to each possible outcome of the experiment. (b) An example of a random variable. The experiment consists of two rolls of a 4-sided die, and the random variable is the maximum of the two rolls. If the outcome 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

$$\text{sgn}(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a < 0, \end{cases}$$

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

$$p_X(x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

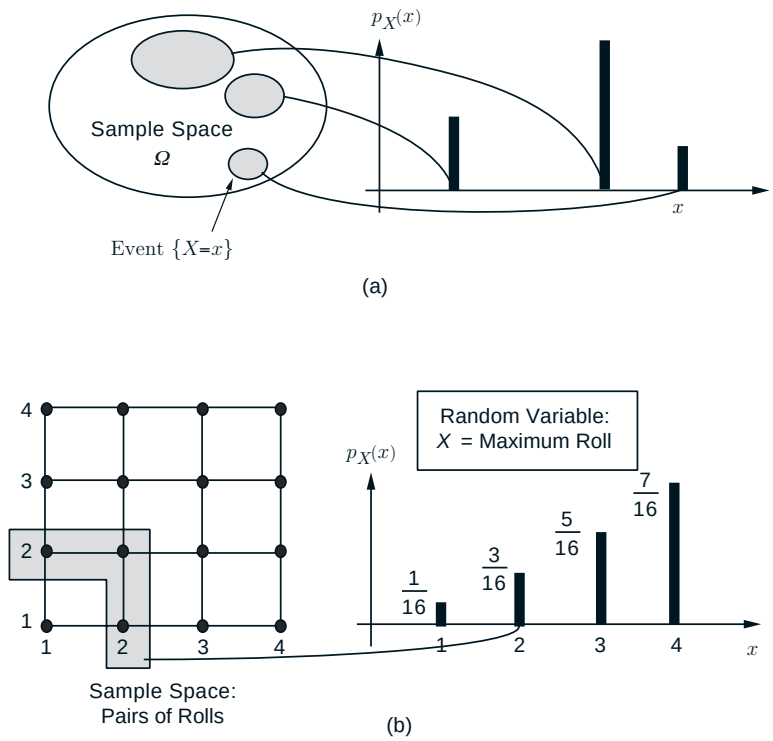


Figure 2.2: (a) Illustration of the method to calculate the PMF of a random variable X. For each possible value x, we collect all the outcomes that give rise to $X = x$ and add their probabilities to obtain $p_X(x)$. (b) Calculation of the PMF p_X of the random variable X = maximum roll in two independent 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.

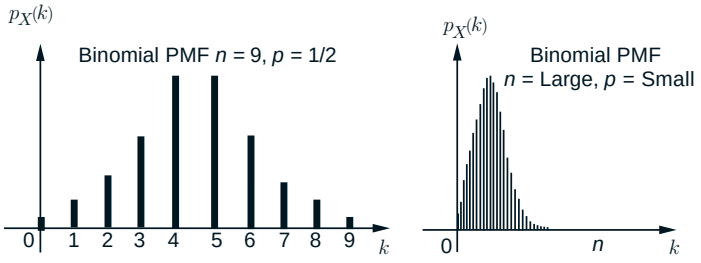


Figure 2.3: The PMF of a binomial random variable. If $p = 1/2$, the PMF is symmetric around $n/2$. Otherwise, the PMF is skewed towards 0 if $p < 1/2$, and towards n if $p > 1/2$.

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.

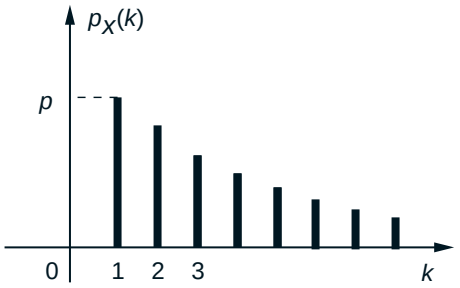


Figure 2.4: The PMF

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots,$$

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.

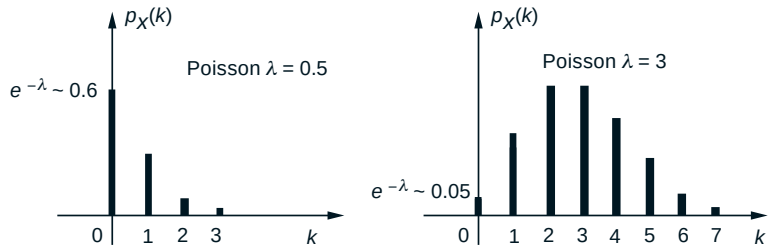


Figure 2.5: The PMF $e^{-\lambda}\frac{\lambda^k}{k!}$ of the Poisson random variable for different values of λ. Note that if $\lambda < 1$, then the PMF is monotonically decreasing, while if $\lambda > 1$, the PMF first increases and then decreases as the value of k increases (this is shown in the end-of-chapter 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 $g(x) = y$:

$$p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x).$$

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 $p_Z(z) = \sum_{\{x \mid x^2=z\}} p_X(x)$ or the formula $p_Z(z) = \sum_{\{y \mid y^2=z\}} p_Y(y)$, we obtain

$$p_Z(z) = \begin{cases} 2/9 & \text{if } z = 1, 4, 9, 16, \\ 1/9 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

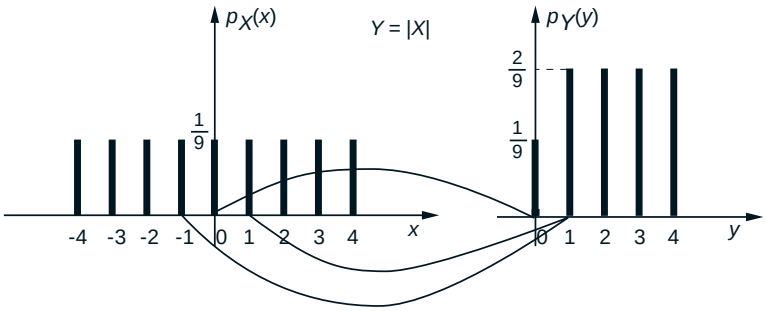


Figure 2.7: The PMFs of X and $Y = |X|$ in Example 2.1.

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.

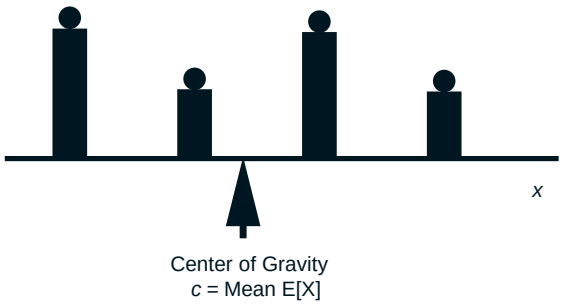


Figure 2.8: Interpretation of the mean as a center of gravity. Given a bar with a weight $p_X(x)$ placed at each point x with $p_X(x) > 0$, the center of gravity c is the point at which the sum of the torques from the weights to its left are equal to the sum of the torques from the weights to its right, 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 2**nd moment** of the random variable X as the expected value of the random variable X^2. More generally, we define the n**th 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 $\text{var}(X)$, 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 $p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x)$ derived in the preceding section, we have

$$\begin{aligned}
\mathbf{E}\big[g(X)\big] &= \mathbf{E}[Y] \\
&= \sum_y y p_Y(y) \\
&= \sum_y y \sum_{\{x \mid g(x)=y\}} p_X(x) \\
&= \sum_y \sum_{\{x \mid g(x)=y\}} y p_X(x) \\
&= \sum_y \sum_{\{x \mid g(x)=y\}} g(x) p_X(x) \\
&= \sum_x g(x) p_X(x).
\end{aligned}$$

Using the expected value rule, we can write the variance of X as

$$\operatorname{var}(X) = \mathbf{E}\left[\big(X - \mathbf{E}[X]\big)^2\right] = \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x).$$

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

$$\begin{aligned}
\operatorname{var}(X) &= \mathbf{E}\left[\big(X - \mathbf{E}[X]\big)^2\right] \\
&= \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x) \\
&= \frac{1}{9} \sum_{x=-4}^{4} x^2 \qquad \text{since } \mathbf{E}[X] = 0 \\
&= \frac{1}{9}(16 + 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16) \\
&= \frac{60}{9},
\end{aligned}$$

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

$$\begin{aligned}\mathrm{var}(X) &= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2\\ &= \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2\\ &= \frac{1}{12}(n+1)(4n+2-3n-3)\\ &= \frac{n^2-1}{12}.\end{aligned}$$

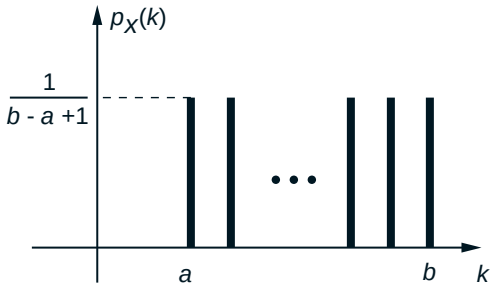


Figure 2.9: PMF of the discrete random variable that is uniformly distributed between two integers a and b. Its mean and variance are

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \text{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

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

$$\text{var}(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{(b-a)(b-a+2)}{12}.$$

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:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{the } k=0 \text{ term is zero} \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^m}{m!} \qquad \text{let } m = k-1 \\ &= \lambda. \end{aligned}$$

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):

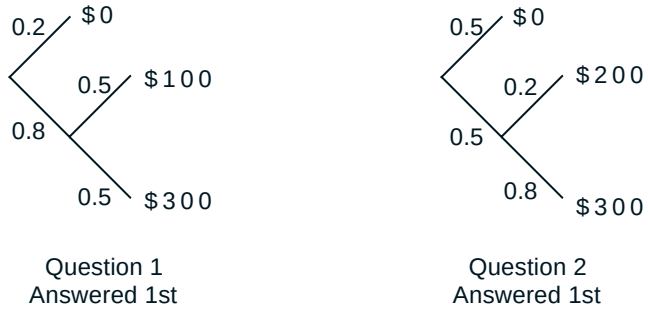


Figure 2.10: Sequential description of the sample space of the quiz problem for the two cases where we answer question 1 or question 2 first.

(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

$$T = \frac{2}{V}, \qquad \text{and} \quad \mathbf{E}[T] = \mathbf{E}\left[\frac{2}{V}\right] \neq \frac{2}{\mathbf{E}[V]}.$$

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

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

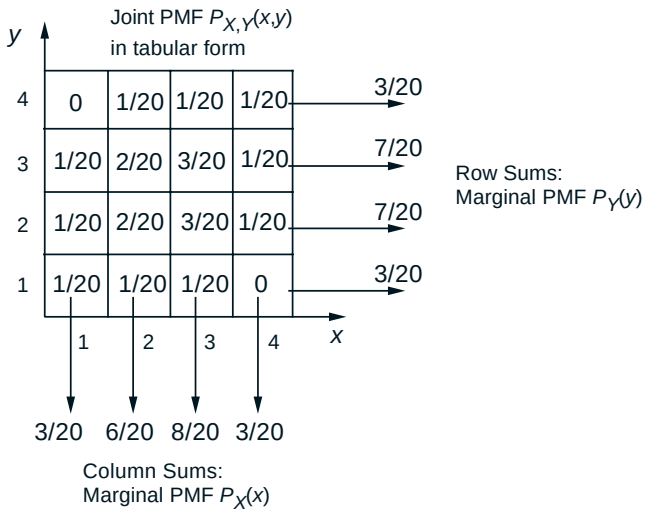


Figure 2.11: Illustration of the tabular method for calculating marginal PMFs from joint PMFs. The joint PMF is represented by a table, where the number in each square (x, y) gives the value of $p_{X,Y}(x, y)$. To calculate the marginal PMF $p_X(x)$ for a given value of x, we add the numbers in the column 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).

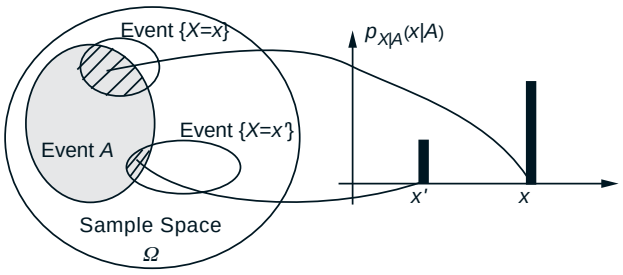


Figure 2.12: Visualization and calculation of the conditional PMF $p_{X|A}(x)$. For each x, we add the probabilities of the outcomes in the intersection $\{X = x\} \cap A$ and normalize by diving with $\mathbf{P}(A)$.

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.

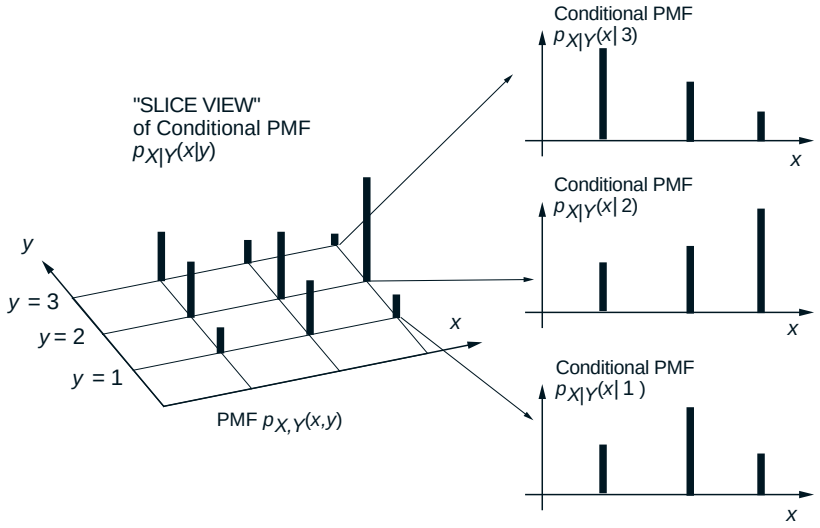


Figure 2.13: Visualization of the conditional PMF $p_{X|Y}(x \,|\, y)$. For each y, we view the joint PMF along the slice $Y = y$ and renormalize so that

$$\sum_x p_{X|Y}(x \,|\, y) = 1.$$

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

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y\,|\,x).$$

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

$$\begin{aligned}\mathbf{P}(\text{at least one wrong answer}) &= p_{X,Y}(1,1) + p_{X,Y}(2,1) + p_{X,Y}(2,2)\\ &= \frac{4}{48} + \frac{6}{48} + \frac{1}{48}.\end{aligned}$$

.

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

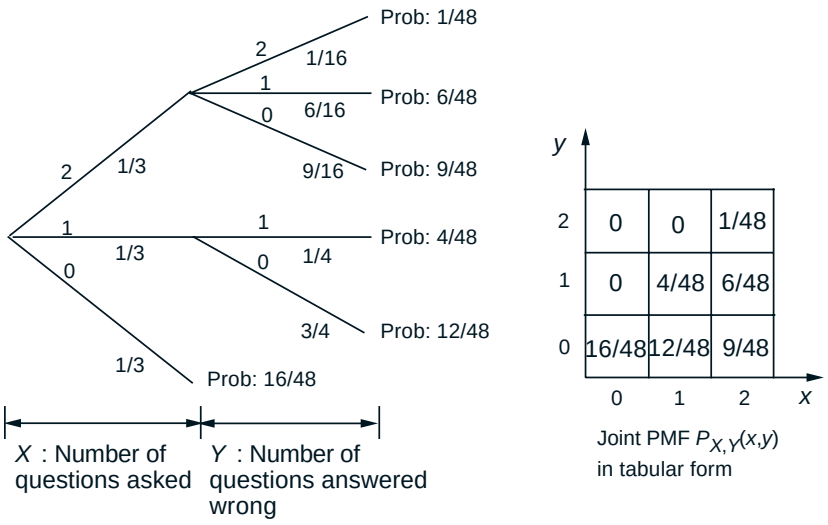


Figure 2.14: Calculation of the joint PMF $p_{X,Y}(x, y)$ in Example 2.11.

$1, 3, 4, 5, 6$ 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 $0 \le x + y \le 4$. 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 $0 \le x + y \le 4$. 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:

$$X : \text{the travel time of a given message}, \quad Y : \text{the length of the given message}.$$

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

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p, \qquad \text{var}(X) = \sum_{k=1}^{\infty}(k - \mathbf{E}[X])^2(1-p)^{k-1}p,$$

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

$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

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 $p_{X,Y}(x,y) = p_{X|Y}(x \mid y)p_Y(y)$ shows that independence is equivalent to the condition

$$p_{X|Y}(x \mid y) = p_X(x), \qquad \text{for all } y \text{ with } p_Y(y) > 0 \text{ and all } x.$$

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

$$\mathbf{P}(X = x, Y = y \mid A) = \mathbf{P}(X = x \mid A)\mathbf{P}(Y = y \mid A), \qquad \text{for all } x \text{ and } y,$$

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

$$p_{X|Y,A}(x \mid y) = p_{X|A}(x) \qquad \text{for all } x \text{ and } y \text{ such that } p_{Y|A}(y) > 0.$$

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:

$$\begin{aligned}
\mathbf{E}[XY] &= \sum_x \sum_y xy p_{X,Y}(x,y) \\
&= \sum_x \sum_y xy p_X(x) p_Y(y) \qquad \text{by independence} \\
&= \sum_x x p_X(x) \sum_y y p_Y(y) \\
&= \mathbf{E}[X]\,\mathbf{E}[Y].
\end{aligned}$$

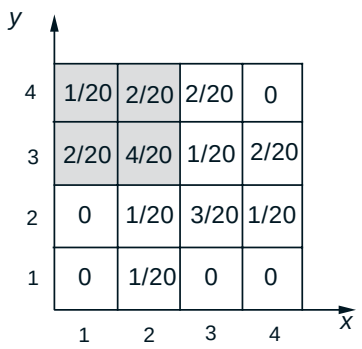


Figure 2.15: Example illustrating that conditional independence may not imply unconditional independence. For the PMF shown, the random variables X and Y are not independent. For example, we have

$$p_{X|Y}(1\,|\,1) = \mathbf{P}(X = 1\,|\,Y = 1) = 0 \neq \mathbf{P}(X = 1) = p_X(1).$$

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 $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$,

$$\begin{aligned}
\operatorname{var}(Z) &= \mathbf{E}\big[\big(X + Y - \mathbf{E}[X + Y]\big)^2\big] \\
&= \mathbf{E}\big[\big(X + Y - \mathbf{E}[X] - \mathbf{E}[Y]\big)^2\big] \\
&= \mathbf{E}\big[\big(\big(X - \mathbf{E}[X]\big) + \big(Y - \mathbf{E}[Y]\big)\big)^2\big] \\
&= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)^2\big] + \mathbf{E}\big[\big(Y - \mathbf{E}[Y]\big)^2\big] \\
&\quad + 2\mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] \\
&= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)^2\big] + \mathbf{E}\big[\big(Y - \mathbf{E}[Y]\big)^2\big].
\end{aligned}$$

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 $X_1, \ldots, X_n$,

$$\mathrm{var}(S_n) = \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{var}(X_i) = \frac{p(1-p)}{n}.$$

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

$$\mathrm{var}(S_n) = \frac{\mathrm{var}(X)}{n},$$

as long as the X_i are independent, with common mean $\mathbf{E}[X]$ and variance $\mathrm{var}(X)$. 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 $[a, b]$:

$$p_X(k) = \begin{cases} \dfrac{1}{b - a + 1} & \text{if } k = a, a + 1, \ldots, b, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a + b}{2}, \qquad \text{var}(X) = \frac{(b - a)(b - a + 2)}{12}.$$

Bernoulli with Parameter p: (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \qquad \text{var}(X) = p(1 - p).$$

Binomial with Parameters p and n: (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

$$\mathbf{E}[X] = np, \qquad \text{var}(X) = np(1-p).$$

Geometric with Parameter p: (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1} p, \qquad k = 1, 2, \ldots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \qquad \text{var}(X) = \frac{1-p}{p^2}.$$

Poisson with Parameter λ: (Approximates the binomial PMF when n is large, p is small, and $\lambda = np$.)

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, \ldots,$$

$$\mathbf{E}[X] = \lambda, \qquad \text{var}(X) = \lambda.$$

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:

$$\mathbf{E}[X] = \sum_y p_Y(y)\mathbf{E}[X \mid Y = y].$$

The concepts and methods of this chapter extend appropriately to general random variables (see the next chapter), and are fundamental for our subject.

3

General Random Variables

Contents

3.1. Continuous Random Variables and PDFs . . . . . p. 2
3.2. Cumulative Distribution Functions . . . . . p. 11
3.3. Normal Random Variables . . . . . p. 17
3.4. Conditioning on an Event . . . . . p. 21
3.5. Multiple Continuous Random Variables . . . . . p. 27
3.6. Derived Distributions . . . . . p. 39
3.7. Summary and Discussion . . . . . p. 51

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.

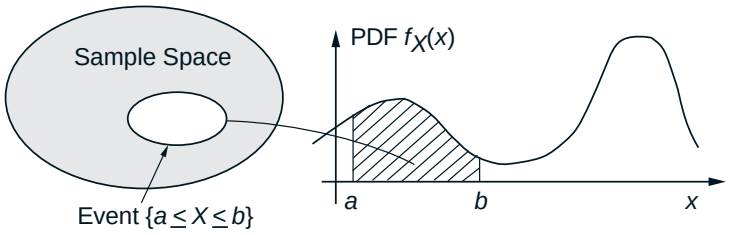


Figure 3.1: Illustration of a PDF. The probability that X takes value in an interval $[a, b]$ is $\int_a^b f_X(x)\,dx$, which is the shaded area in the figure.

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.

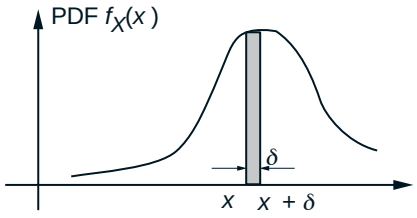


Figure 3.2: Interpretation of the PDF $f_X(x)$ as "probability mass per unit length" around x. If δ is very small, the probability that X takes value in the interval $[x, x + \delta]$ is the shaded area in the figure, which is approximately equal to $f_X(x) \cdot \delta$.

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

$$c = \frac{1}{b-a}.$$

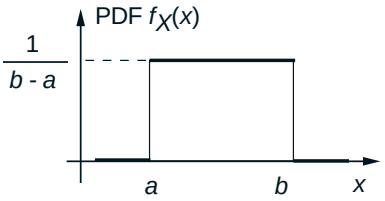


Figure 3.3: The PDF of a uniform random variable.

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:

$$1 = \int_{a_1}^{a_n} f_X(x)\,dx = \sum_{i=1}^{n-1} \int_{a_i}^{a_{i+1}} c_i\,dx = \sum_{i=1}^{n-1} c_i(a_{i+1} - a_i).$$

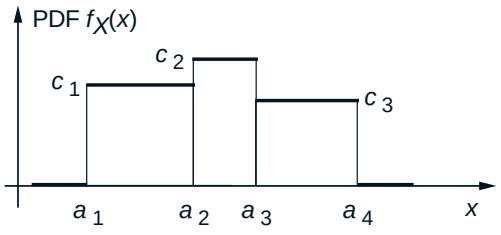


Figure 3.4: A piecewise constant PDF involving three intervals.

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
$$0 \leq \mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.$$
- 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.

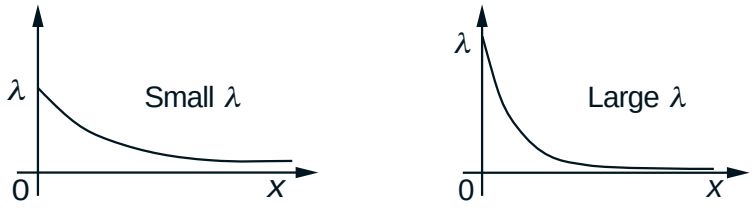


Figure 3.5: The PDF $\lambda e^{-\lambda x}$ of an exponential random variable.

The mean and the variance can be calculated to be

$$\mathbf{E}[X] = \frac{1}{\lambda}, \qquad \text{var}(X) = \frac{1}{\lambda^2}.$$

These formulas can be verified by straightforward calculation, as we now show. We have, using integration by parts,

$$\begin{aligned}
\mathbf{E}[X] &= \int_0^\infty x\lambda e^{-\lambda x}\,dx \\
&= (-xe^{-\lambda x})\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \\
&= 0 - \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty \\
&= \frac{1}{\lambda}.
\end{aligned}$$

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 $\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2$, we obtain

$$\text{var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

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

$$\mathbf{P}(1/4 \le X \le 3/4) = \mathbf{P}(X \ge 1/4) - \mathbf{P}(X > 3/4) = e^{-1/40} - e^{-3/40} = 0.0476,$$

where we have used the formula $\mathbf{P}(X \ge a) = \mathbf{P}(X > a) = e^{-\lambda a}$.

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 $k - (3/4) \le X \le k - (1/4)$. Since these events are disjoint, the probability of interest is

$$\sum_{k=1}^{\infty} \mathbf{P}\Big(k - \frac{3}{4} \le X \le k - \frac{1}{4}\Big) = \sum_{k=1}^{\infty} \Big(\mathbf{P}\Big(X \ge k - \frac{3}{4}\Big) - \mathbf{P}\Big(X > k - \frac{1}{4}\Big)\Big)$$
$$= \sum_{k=1}^{\infty} \left(e^{-(4k-3)/40} - e^{-(4k-1)/40}\right).$$

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 $\mathbf{P}(X \le x)$. In particular, for every x we have

$$F_X(x) = \mathbf{P}(X \le x) = \begin{cases} \displaystyle\sum_{k \le x} p_X(k) & X\text{: discrete,} \\ \displaystyle\int_{-\infty}^{x} f_X(t)\, dt & X\text{: continuous.} \end{cases}$$

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 $\{X \le x\}$ 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.

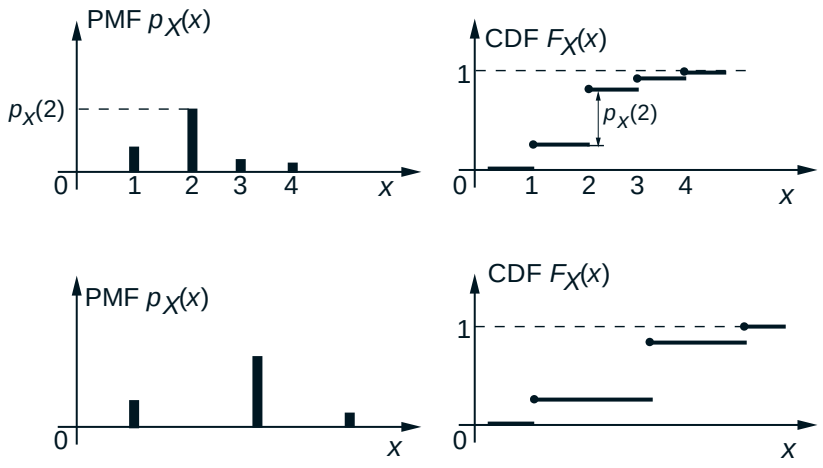


Figure 3.6: CDFs of some discrete random variables. The CDF is related to the PMF through the formula

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{k \leq x} p_X(k),$$

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 corresponding 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

$$F^{\text{exp}}(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x}, \qquad \text{for } x > 0.$$

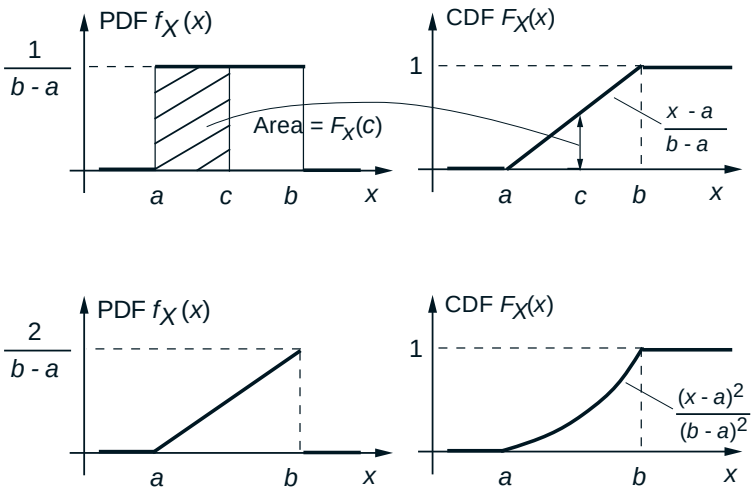


Figure 3.7: CDFs of some continuous random variables. The CDF is related to the PDF through the formula

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

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

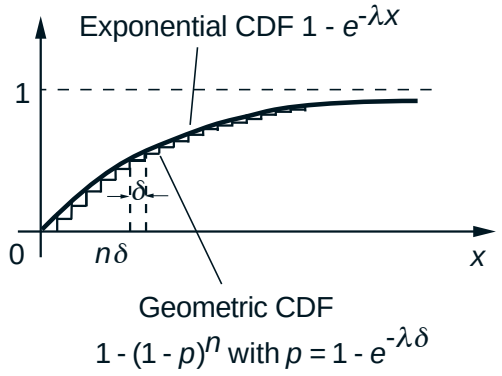


Figure 3.8: Relation of the geometric and the exponential CDFs. We have

$$F^{\exp}(n\delta) = F^{\text{geo}}(n), \qquad n = 1, 2, \ldots,$$

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 theoretical problems).

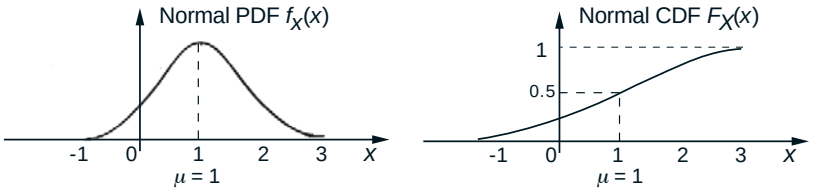


Figure 3.9: A normal PDF and CDF, with $\mu = 1$ and $\sigma^2 = 1$. We observe that the PDF is symmetric around its mean μ, and has a characteristic bell-shape. As x gets further from μ, the term $e^{-(x-\mu)^2/2\sigma^2}$ decreases very rapidly. In this figure, the PDF is very close to zero outside the interval $[-1, 3]$.

The mean and the variance can be calculated to be

$$\mathbf{E}[X] = \mu, \qquad \text{var}(X) = \sigma^2.$$

To see this, note that the PDF is symmetric around μ, so its mean must be μ. Furthermore, the variance is given by

$$\text{var}(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}\,dx.$$

Using the change of variables $y = (x-\mu)/\sigma$ and integration by parts, we have

$$\begin{aligned}
\text{var}(X) &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2}\,dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(-y e^{-y^2/2}\right) \Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\
&= \sigma^2.
\end{aligned}$$

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

$$\mathbf{E}[Y] = a\mu + b, \qquad \text{var}(Y) = a^2\sigma^2.$$

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,

$$\Phi(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(Y < y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2}\, dt.$$

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 $y \geq 0$, 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

$$\begin{aligned}\Phi(-0.5) &= \mathbf{P}(Y \leq -0.5) = \mathbf{P}(Y \geq 0.5) = 1 - \mathbf{P}(Y < 0.5) \\ &= 1 - \Phi(0.5) = 1 - .6915 = 0.3085.\end{aligned}$$

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.

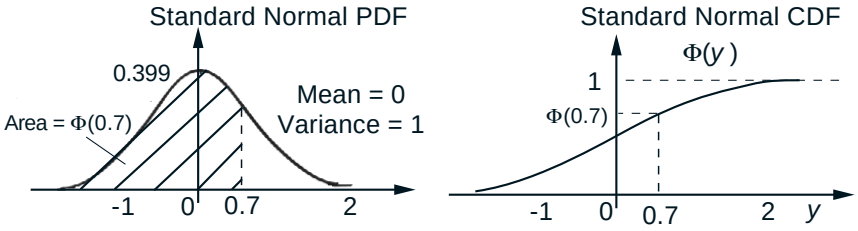


Figure 3.10: The PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

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

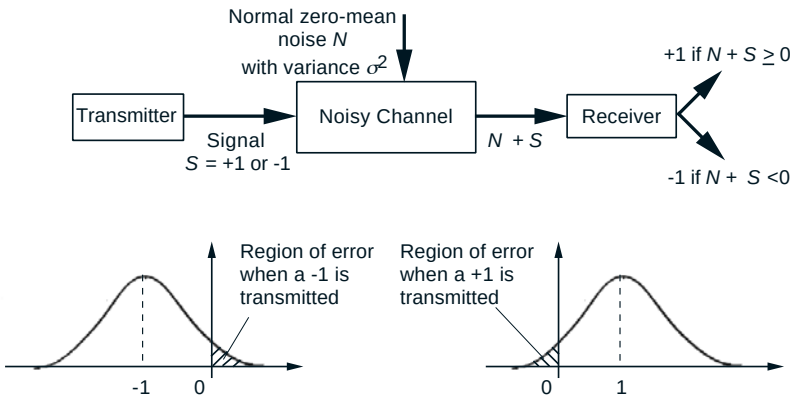


Figure 3.11: The signal detection scheme of Example 3.9. The area of the shaded region gives the probability of error in the two cases where -1 and $+1$ is transmitted.

than -1 so that $N + S = N + 1 < 0$. In the former case, the probability of error is

$$\mathbf{P}(N \geq 1) = 1 - \mathbf{P}(N < 1) = 1 - \mathbf{P}\left(\frac{N-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right)$$
$$= 1 - \Phi\left(\frac{1-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1}{\sigma}\right).$$

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

$$\mathbf{P}(X \in B \,|\, A) = \int_B f_{X|A}(x)\, dx,$$

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.

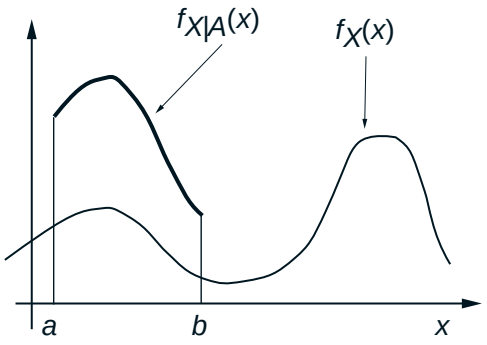


Figure 3.13: The unconditional PDF f_X and the conditional PDF $f_{X|A}$, where A is the interval $[a, b]$. Note that within the conditioning event A, $f_{X|A}$ retains the same shape as f_X, except that it is scaled along the vertical axis.

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

$$\begin{aligned}\mathbf{P}(X > x \,|\, A) &= \mathbf{P}(T > t + x \,|\, T > t) \\ &= \frac{\mathbf{P}(T > t + x \text{ and } T > t)}{\mathbf{P}(T > t)} \\ &= \frac{\mathbf{P}(T > t + x)}{\mathbf{P}(T > t)} \\ &= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} \\ &= e^{-\lambda x},\end{aligned}$$

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

$$\mathbf{P}(X \in B \,|\, A) = \int_B f_{X|A}(x)\, dx.$$

- 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

$$\mathbf{P}(X \in B \,|\, X \in A) = \int_B f_{X|A}(x)\, dx,$$

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

$$\mathbf{P}(X \leq x) = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{P}(X \leq x \mid A_i).$$

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

$$A_1 = \big\{X \text{ lies in the first interval } [0,1]\big\},$$
$$A_2 = \big\{X \text{ lies in the second interval } (1,2]\big\}.$$

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,

$$\mathbf{E}[X \mid A_1] = \frac{1}{2}, \qquad \mathbf{E}[X \mid A_2] = \frac{3}{2},$$
$$\mathbf{E}\big[X^2 \mid A_1\big] = \frac{1}{3}, \qquad \mathbf{E}\big[X^2 \mid A_2\big] = \frac{7}{3}.$$

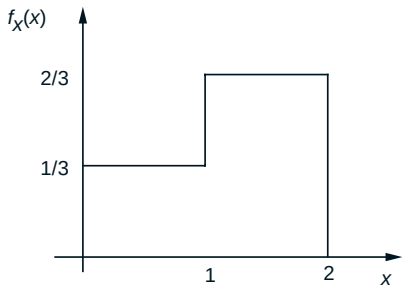


Figure 3.14: Piecewise constant PDF for Example 3.11.

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

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{15}{9} - \frac{49}{36} = \frac{11}{36}.$$

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?

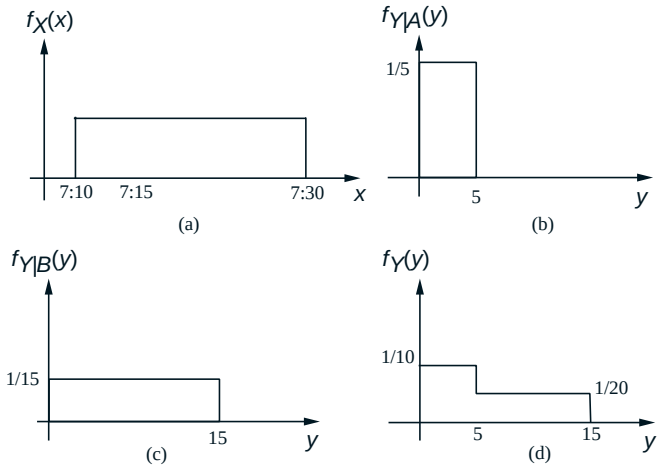


Figure 3.15: The PDFs f_X, $f_{Y|A}$, $f_{Y|B}$, and f_Y in Example 3.12.

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

$$A = \{\text{7:10} \le X \le \text{7:15}\} = \{\text{you board the 7:15 train}\},$$

$$B = \{7{:}15 < X \leq 7{:}30\} = \{\text{you board the 7:30 train}\}.$$

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,

$$f_Y(y) = \frac{1}{4} \cdot \frac{1}{5} + \frac{3}{4} \cdot \frac{1}{15} = \frac{1}{10}, \quad \text{for } 0 \leq y \leq 5,$$

and

$$f_Y(y) = \frac{1}{4} \cdot 0 + \frac{3}{4} \cdot \frac{1}{15} = \frac{1}{20}, \quad \text{for } 5 < y \leq 15.$$

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

$$\mathbf{P}(a \leq X \leq b,\, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)\, dx\, dy.$$

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

$$\mathbf{P}(a \le X \le a+\delta,\, c \le Y \le c+\delta) = \int_c^{c+\delta} \int_a^{a+\delta} f_{X,Y}(x,y)\,dx\,dy \approx f_{X,Y}(a,c)\cdot\delta^2,$$

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

$$f_{X,Y}(x,y) = \begin{cases} c & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

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.

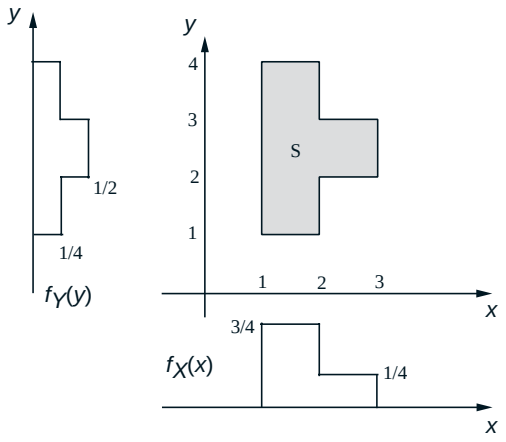


Figure 3.16: The joint PDF in Example 3.14 and the resulting marginal PDFs.

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?

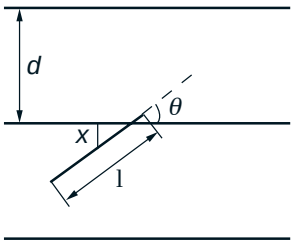


Figure 3.17: Buffon's needle. The length of the line segment between the midpoint of the needle and the point of intersection of the axis of the needle with the closest parallel line is $x/\sin\theta$. The needle will intersect the closest parallel line if and only if this length is less than $l/2$.

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

$$X \leq \frac{l}{2}\sin\Theta,$$

so the probability of intersection is

$$\begin{aligned} \mathbf{P}\big(X \leq (l/2)\sin\Theta\big) &= \int\int_{x \leq (l/2)\sin\theta} f_{X,\Theta}(x,\theta)\,dx\,d\theta \\ &= \frac{4}{\pi d}\int_0^{\pi/2}\int_0^{(l/2)\sin\theta} dx\,d\theta \\ &= \frac{4}{\pi d}\int_0^{\pi/2}\frac{l}{2}\sin\theta\,d\theta \\ &= \frac{2l}{\pi d}(-\cos\theta)\Big|_0^{\pi/2} \\ &= \frac{2l}{\pi d}. \end{aligned}$$

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.

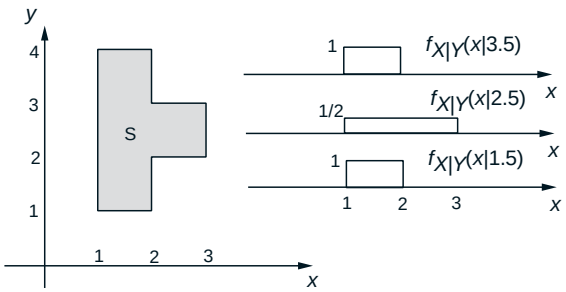


Figure 3.18: Visualization of the conditional PDF $f_{X|Y}(x\,|\,y)$. Let X, Y have a joint PDF which is uniform on the set S. For each fixed y, we consider the joint PDF along the slice $Y = y$ and normalize it so that it integrates to 1.

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

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\text{area of the circle}} & \text{if } (x,y) \text{ is in the circle,} \\ 0 & \text{otherwise,} \end{cases}$$

$$= \begin{cases} \dfrac{1}{\pi r^2} & \text{if } x^2 + y^2 \le r^2, \\ 0 & \text{otherwise.} \end{cases}$$

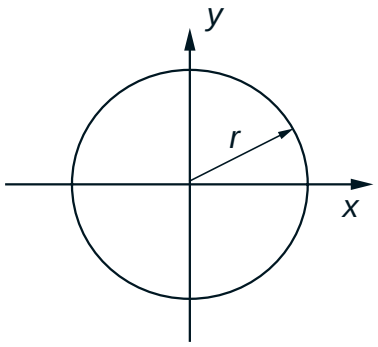


Figure 3.19: Circular target for Example 3.16.

To calculate the conditional PDF $f_{X|Y}(x \,|\, y)$, 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:

$$\begin{aligned}
\mathbf{E}[g(X)] &= \int g(x) f_X(x)\,dx,\\
\mathbf{E}\big[g(X,Y)\big] &= \int\int g(x,y) f_{X,Y}(x,y)\,dx\,dy,\\
\mathbf{E}\big[g(X) \,|\, Y = y\big] &= \int g(x) f_{X|Y}(x \,|\, y)\,dx,\\
\mathbf{E}\big[g(X,Y) \,|\, Y = y\big] &= \int g(x,y) f_{X|Y}(x \,|\, y)\,dx.
\end{aligned}$$

- We have the following versions of the total expectation theorem:

$$\begin{aligned}
\mathbf{E}[X] &= \int \mathbf{E}[X \,|\, Y = y] f_Y(y)\,dy,\\
\mathbf{E}\big[g(X)\big] &= \int \mathbf{E}\big[g(X) \,|\, Y = y\big] f_Y(y)\,dy,\\
\mathbf{E}\big[g(X,Y)\big] &= \int \mathbf{E}\big[g(X,Y) \,|\, Y = y\big] f_Y(y)\,dy.
\end{aligned}$$

To justify the first version of the total expectation theorem, we observe that

$$\begin{aligned}
\int \mathbf{E}[X \,|\, Y = y] f_Y(y)\,dy &= \int \left[\int x f_{X|Y}(x \,|\, y)\,dx\right] f_Y(y)\,dy\\
&= \int\int x f_{X|Y}(x \,|\, y) f_Y(y)\,dx\,dy\\
&= \int\int x f_{X,Y}(x,y)\,dx\,dy
\end{aligned}$$

$$\begin{aligned} &= \int x \left[\int f_{X,Y}(x,y)\, dy \right] dx \\ &= \int x f_X(x)\, dx \\ &= \mathbf{E}[X]. \end{aligned}$$

The other two versions are justified similarly.

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 theoretical 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
$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y).$$

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

$$F_Y(y) = \mathbf{P}\big(g(X) \le y\big) = \int_{\{x \mid g(x) \le y\}} f_X(x)\, dx.$$

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. His average speed 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,

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}), \qquad y \ge 0.$$

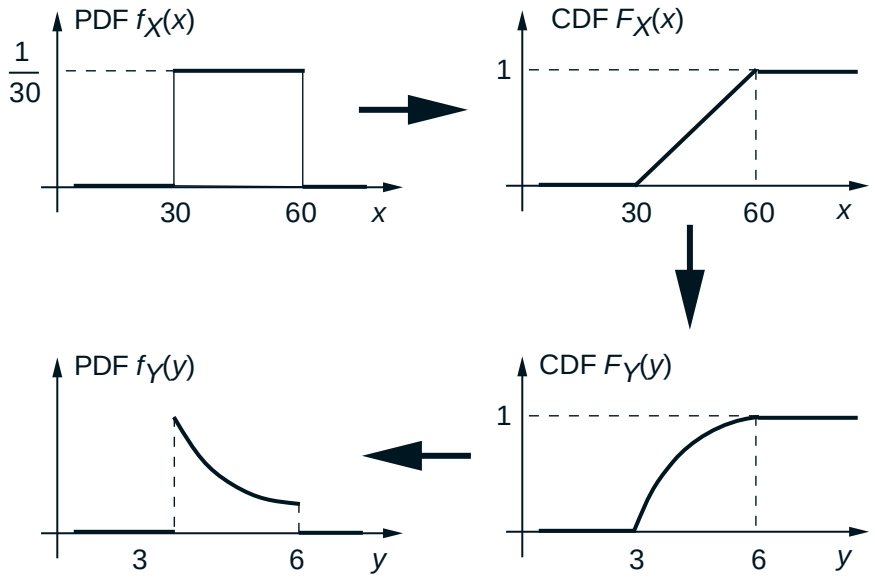


Figure 3.20: The calculation of the PDF of $Y = 180/X$ in Example 3.22. The arrows indicate the flow of the calculation.

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

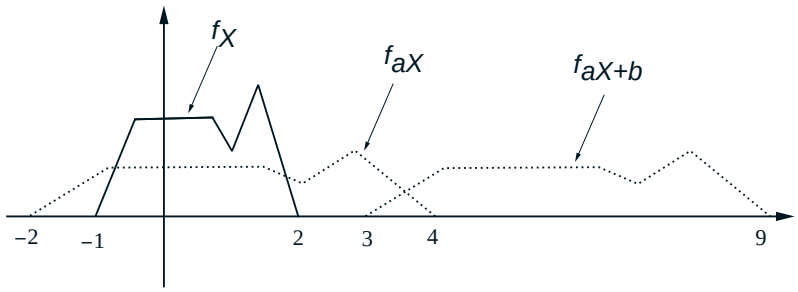


Figure 3.21: The PDF of $aX + b$ in terms of the PDF of X. In this figure, $a = 2$ and $b = 5$. As a first step, we obtain the PDF of aX. The range of Y is wider than the range of X, by a factor of a. Thus, the PDF f_X must be stretched (scaled horizontally) by this factor. But in order to keep the total 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.

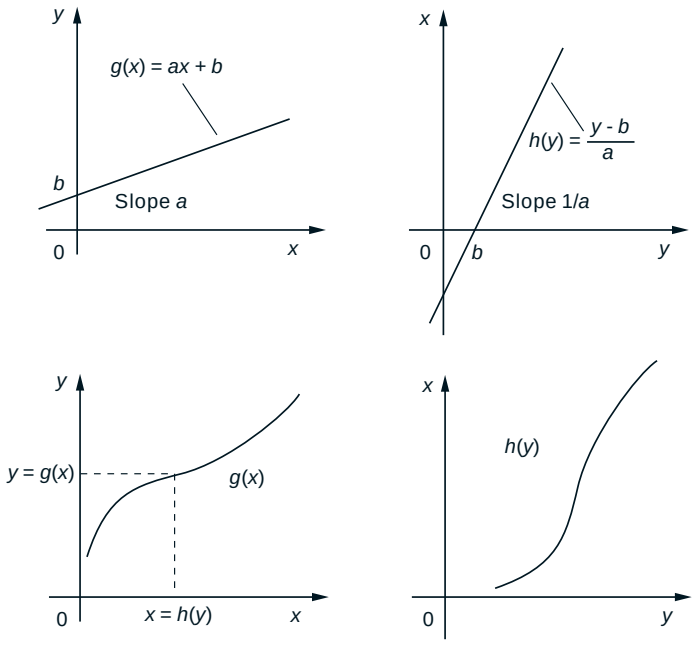


Figure 3.22: A monotonically increasing function g (on the left) and its inverse (on the right). Note that the graph of h has the same shape as the graph of g, except that it is rotated by 90 degrees and then reflected (this is the same as interchanging the x and y axes).

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

$$f_Y(y) = f_X(x) \Big/ \left| \frac{dg}{dx}(x) \right|.$$

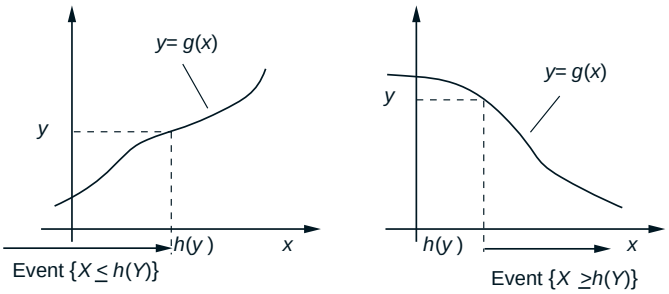


Figure 3.23: Calculating the probability $\mathbf{P}\big(g(X) \le y\big)$. When g is monotonically increasing (left figure), the event $\{g(X) \le y\}$ is the same as the event $\{X \le h(y)\}$. When g is monotonically decreasing (right figure), the event $\{g(X) \le y\}$ is the same as the event $\{X \ge h(y)\}$.

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 $Z = Y/X$?

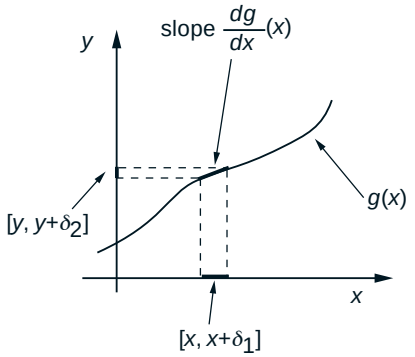


Figure 3.24: Illustration of the PDF formula for a monotonically increasing function g. Consider an interval $[x, x+\delta_1]$, where δ_1 is a small number. Under the mapping g, the image of this interval is another interval $[y, y+\delta_2]$. Since $(dg/dx)(x)$ is the slope of g, we have

$$\frac{\delta_2}{\delta_1} \approx \frac{dg}{dx}(x),$$

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

$$f_Z(z) = \begin{cases} 1/2 & \text{if } 0 \le z \le 1, \\ 1/(2z^2) & \text{if } z > 1, \\ 0 & \text{otherwise.} \end{cases}$$

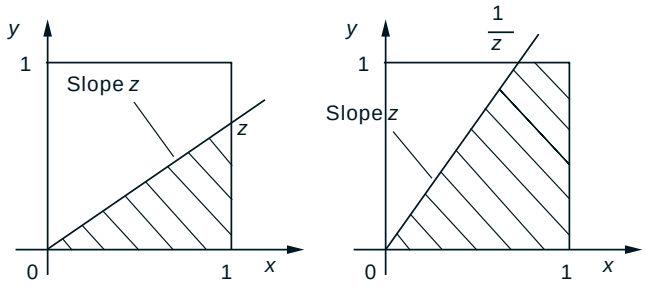


Figure 3.25: The calculation of the CDF of $Z = Y/X$ in Example 3.28. The value $\mathbf{P}(Y/X \le z)$ is equal to the shaded subarea of the unit square. The figure on the left deals with the case where $0 \le z \le 1$ and the figure on the right refers to the case where $z > 1$.

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)

$$\begin{aligned} F_Z(z) &= \mathbf{P}(X - Y \le z) \\ &= 1 - \mathbf{P}(X - Y > z) \\ &= 1 - \int_0^\infty \left(\int_{z+y}^\infty f_{X,Y}(x,y)\,dx \right) dy \\ &= 1 - \int_0^\infty \lambda e^{-\lambda y} \left(\int_{z+y}^\infty \lambda e^{-\lambda x}\,dx \right) dy \\ &= 1 - \int_0^\infty \lambda e^{-\lambda y} e^{-\lambda(z+y)}\,dy \\ &= 1 - e^{-\lambda z} \int_0^\infty \lambda e^{-2\lambda y}\,dy \\ &= 1 - \frac{1}{2} e^{-\lambda z}. \end{aligned}$$

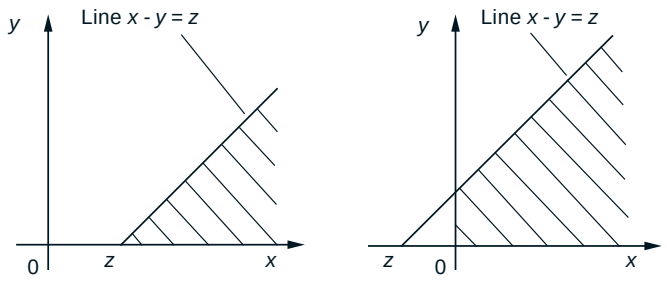


Figure 3.26: The calculation of the CDF of $Z = X - Y$ in Example 3.29. To obtain the value $\mathbf{P}(X - Y > z)$ we must integrate the joint PDF $f_{X,Y}(x, y)$ over the shaded area in the above figures, which correspond to $z \geq 0$ (left side) and $z < 0$ (right side).

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 $[a, b]$:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \text{var}(X) = \frac{(b-a)^2}{12}.$$

Exponential with Parameter λ:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0, \\ 0 & \text{otherwise,} \end{cases} \qquad F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{1}{\lambda}, \qquad \text{var}(X) = \frac{1}{\lambda^2}.$$

Normal with Parameters μ and σ^2:

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

$$\mathbf{E}[X] = \mu, \qquad \text{var}(X) = \sigma^2.$$

4

Further Topics on Random Variables and Expectations

Contents

4.1. Transforms . . . . . p. 2
4.2. Sums of Independent Random Variables - Convolutions . . . p. 13
4.3. Conditional Expectation as a Random Variable . . . . . p. 17
4.4. Sum of a Random Number of Independent Random Variables p. 25
4.5. Covariance and Correlation . . . . . p. 29
4.6. Least Squares Estimation . . . . . p. 32
4.7. The Bivariate Normal Distribution . . . . . p. 39

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).

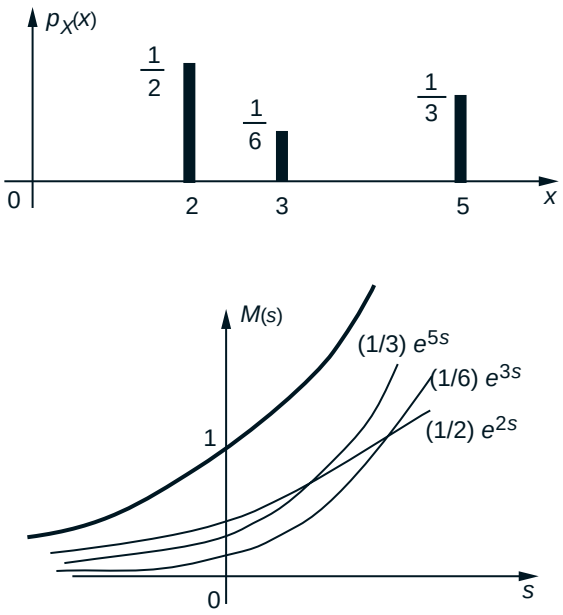


Figure 4.1: The PMF and the corresponding transform for Example 4.1. The transform $M(s)$ consists of the weighted sum of the three exponentials shown. Note that at $s = 0$, the transform takes the value 1. This is generically true since

$$M(0) = \sum_x e^{0\cdot x} p_X(x) = \sum_x p_X(x) = 1.$$

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,

$$\begin{aligned} M(s) =& \lambda\int_0^{\infty} e^{sx}e^{-\lambda x}\,dx \\ =& \lambda\int_0^{\infty} e^{(s-\lambda)x}\,dx \\ =& \lambda\frac{e^{(s-\lambda)x}}{s-\lambda}\bigg|_0^{\infty} \qquad \text{(if } s<\lambda\text{)} \\ =& \frac{\lambda}{\lambda-s}. \end{aligned}$$

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

$$M_Y(s) = \left(1 - p + pe^s\right)^n.$$

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

$$M_{X_1,\ldots,X_n}(s_1, \ldots, s_n) = \mathbf{E}\big[e^{sX_1 + \cdots + s_n X_n}\big].$$

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,

$$\begin{aligned} p_W(w) =& \mathbf{P}(X + Y = w) \\ =& \sum_{(x,y):\ x+y=w} \mathbf{P}(X = x \text{ and } Y = y) \\ =& \sum_x \mathbf{P}(X = x \text{ and } Y = w - x) \\ =& \sum_x p_X(x) p_Y(w - x). \end{aligned}$$

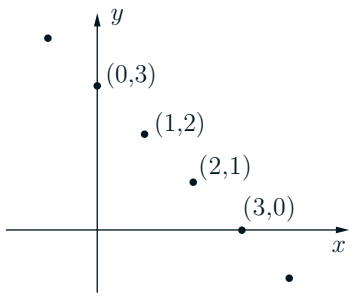


Figure 4.2: The probability $p_W(3)$ that $X+Y=3$ is the sum of the probabilities of all pairs (x,y) such that $x+y=3$, which are the points indicated in the figure. The probability of a generic such point is of the form $p_{X,Y}(x,3-x) = p_X(x)p_Y(3-x)$.

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,

$$f_W(w) = \begin{cases} \min\{1, w\} - \max\{0, w-1\}, & 0 \le w \le 2, \\ 0, & \text{otherwise}, \end{cases}$$

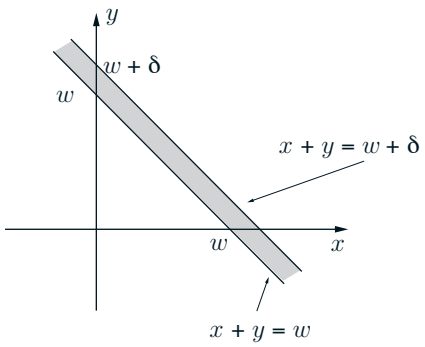


Figure 4.3: Illustration of the convolution formula for the case of continuous random variables (compare with Fig. 4.2). For small δ, the probability of the strip indicated in the figure is $\mathbf{P}(w \le X + Y \le w + \delta) \approx f_W(w) \cdot \delta$. Thus,

$$\begin{aligned} f_W(w) \cdot \delta =& \mathbf{P}(w \le X + Y \le w + \delta) \\ &= \int_{x=-\infty}^{\infty} \int_{y=w-x}^{w-x+\delta} f_X(x) f_Y(y)\, dy\, dx \\ &\approx \int_{x=-\infty}^{\infty} f_X(x) f_Y(w-x) \delta\, dx. \end{aligned}$$

The desired formula follows by canceling δ from both sides.

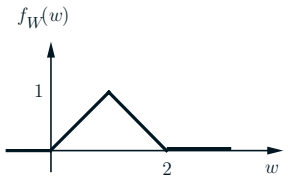


Figure 4.4: The PDF of the sum of two independent uniform random variables in $[0, 1]$.

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 $a \le t \le b$ and a PDF $f_Y(t)$ which is zero outside the range $c \le t \le d$. 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.

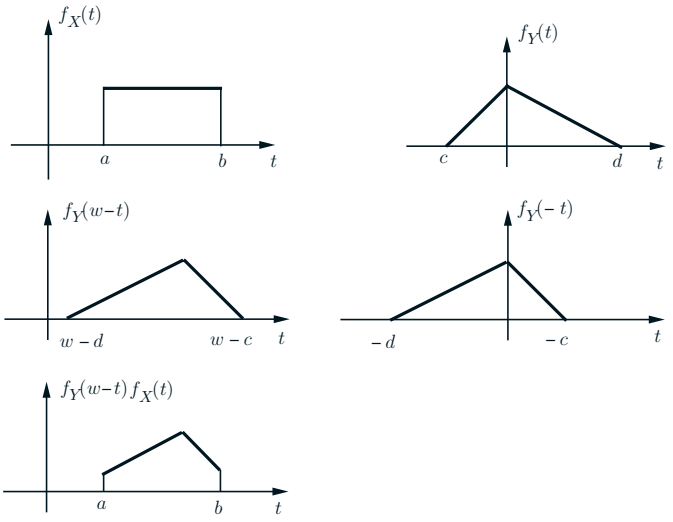


Figure 4.5: Illustration of the convolution calculation. For the value of w under consideration, $f_W(w)$ is equal to the integral of the function shown in the last plot.

4.3 CONDITIONAL EXPECTATION AS A RANDOM VARIABLE

The value of the conditional expectation $\mathbf{E}[X \,|\, Y = y]$ of a random variable X given another random variable Y depends on the realized experimental value y of Y. This makes $\mathbf{E}[X \,|\, Y]$ a function of Y, and therefore a random variable. In this section, we study the expectation and variance of $\mathbf{E}[X \,|\, Y]$. 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 $Y = y$.

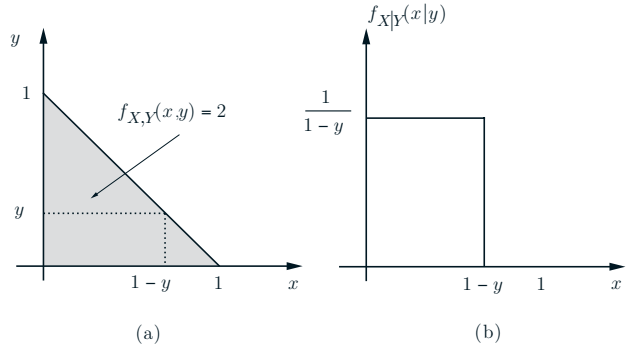


Figure 4.6: (a) The joint PDF in Example 4.15. (b) The conditional density of X.

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 $0 \leq y < 1$, the conditional mean $\mathbf{E}[X \,|\, Y = y]$ is the expectation of the uniform PDF in Fig. 4.6(b), and we have

$$\mathbf{E}[X \,|\, Y = y] = \frac{1 - y}{2}, \qquad 0 \leq y < 1.$$

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 $\mathbf{E}[X \,|\, Y] = (1 - Y)/2$ [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 $\mathbf{E}[X \,|\, Y] = Y/2$, 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,

$$\mathbf{E}[X] = \mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big] = \mathbf{E}\left[\frac{Y}{2}\right] = \frac{\mathbf{E}[Y]}{2} = \frac{\ell}{4}.$$

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,

$$\mathbf{E}[X \mid Y = s] = \frac{1}{n_s} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i = m_s.$$

A randomly selected student belongs to section s with probability n_s/n, i.e., $\mathbf{P}(Y = s) = n_s/n$. Hence,

$$\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \sum_{s=1}^{S} \mathbf{E}[X \mid Y = s]\mathbf{P}(Y = s) = \sum_{s=1}^{S} \frac{n_s}{n} m_s.$$

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 $\mathbf{E}[X \mid Y]$.

We view $\mathbf{E}[X \mid Y] - \mathbf{E}[X]$ as the forecast revision, in light of the mid-year information. The law of iterated expectations implies that

$$\mathbf{E}\big[\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big] = 0.$$

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 $\mathbf{E}[X \mid Y = y]$, 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 $Y = y$:

$$\operatorname{var}(X \mid Y = y) = \mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y = y]\big)^2 \mid Y = y\Big].$$

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 $\operatorname{var}(X \mid Y)$.

Arguing by analogy to the law of iterated expectations, we may conjecture that the expectation of the conditional variance $\operatorname{var}(X \mid Y)$ is related to the unconditional variance $\operatorname{var}(X)$. This is indeed the case, but the relation is more complex.

Law of Conditional Variances:

$$\operatorname{var}(X) = \mathbf{E}\big[\operatorname{var}(X \mid Y)\big] + \operatorname{var}\big(\mathbf{E}[X \mid Y]\big)$$

To verify the law of conditional variances, we start with the identity

$$X - \mathbf{E}[X] = \big(X - \mathbf{E}[X \mid Y]\big) + \big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big).$$

We square both sides and then take expectations to obtain

$$\begin{aligned}\operatorname{var}(X) &= \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] \\ &= \mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)^2\Big] + \mathbf{E}\Big[\big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big)^2\Big] \\ &\quad + 2\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)\big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big)\Big].\end{aligned}$$

Using the law of iterated expectations, the first term in the right-hand side of the above equation can be written as

$$\mathbf{E}\Big[\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)^2 \mid Y\Big]\Big],$$

which is the same as $\mathbf{E}\big[\operatorname{var}(X \mid Y)\big]$. The second term is equal to $\operatorname{var}\big(\mathbf{E}[X \mid Y]\big)$, since $\mathbf{E}[X]$ is the mean of $\mathbf{E}[X \mid Y]$. Finally, the third term is zero, as we now show. Indeed, if we define $h(Y) = 2\big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big)$, the third term is

$$\begin{aligned}\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)h(Y)\Big] &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[\mathbf{E}[X \mid Y]h(Y)\big] \\ &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\Big[\mathbf{E}\big[Xh(Y) \mid Y\big]\Big] \\ &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[Xh(Y)\big] \\ &= 0.\end{aligned}$$

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 $\mathbf{E}[X \,|\, Y] = Y/2$, so since Y is uniformly distributed between 0 and ℓ,

$$\mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big) = \mathrm{var}(Y/2) = \frac{1}{4}\mathrm{var}(Y) = \frac{1}{4} \cdot \frac{\ell^2}{12} = \frac{\ell^2}{48}.$$

Also, since X is uniformly distributed between 0 and Y, we have

$$\mathrm{var}(X \,|\, Y) = \frac{Y^2}{12}.$$

Thus, since Y is uniformly distributed between 0 and ℓ,

$$\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] = \frac{1}{12}\int_0^\ell \frac{1}{\ell} y^2 dy = \frac{1}{12}\frac{1}{3\ell} y^3 \Big|_0^\ell = \frac{\ell^2}{36}.$$

Using now the law of conditional variances, we obtain

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] + \mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big) = \frac{\ell^2}{48} + \frac{\ell^2}{36} = \frac{7\ell^2}{144}.$$

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

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] + \mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big).$$

In this context, $\mathrm{var}(X \,|\, Y = s)$ is the variance of the quiz scores within section s. Then, $\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big]$ is the average of the section variances. This latter expectation is an average over the probability distribution of Y, i.e.,

$$\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] = \sum_{s=1}^{S} \frac{n_s}{n} \mathrm{var}(X \,|\, Y = s).$$

Recall that $\mathbf{E}[X \,|\, Y = s]$ is the average score in section s. Then, $\mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big)$ 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 $\mathbf{E}\big[\mathrm{var}(X \mid Y)\big]$ *within* individual sections.

(b) The variability $\mathrm{var}\big(\mathbf{E}[X \mid Y]\big)$ *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, $\mathbf{E}[X \mid Y]$ takes the values $1/2$ and $3/2$, with probabilities $1/3$ and $2/3$, respectively. Thus, the mean of $\mathbf{E}[X \mid Y]$ is $7/6$. Therefore,

$$\mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = \frac{1}{3}\left(\frac{1}{2} - \frac{7}{6}\right)^2 + \frac{2}{3}\left(\frac{3}{2} - \frac{7}{6}\right)^2 = \frac{2}{9}.$$

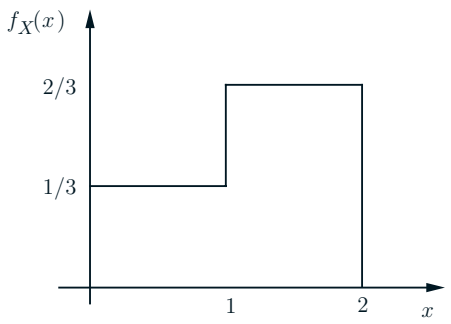


Figure 4.7: The PDF in Example 4.20.

Conditioned on either value of Y, X is uniformly distributed on a unit length interval. Therefore, $\mathrm{var}(X \mid Y = y) = 1/12$ for each of the two possible values of y, and $\mathbf{E}\big[\mathrm{var}(X \mid Y)\big] = 1/12$. Putting everything together, we obtain

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = \frac{1}{12} + \frac{2}{9} = \frac{11}{36}.$$

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).
- $\mathrm{var}(X \mid Y)$ is a random variable whose experimental value is $\mathrm{var}(X \mid Y = y)$, whenever the experimental value of Y is y.
- $\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big)$.

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,

$$\begin{aligned}\mathrm{var}(Y \mid N = n) &= \mathrm{var}(X_1 + \cdots + X_N \mid N = n) \\ &= \mathrm{var}(X_1 + \cdots + X_n) \\ &= n\sigma^2.\end{aligned}$$

Since this is true for every nonnegative integer n, the random variable $\mathrm{var}(Y \mid N)$ is equal to $N\sigma^2$. We now use the law of conditional variances to obtain

$$\begin{aligned}\mathrm{var}(Y) &= \mathbf{E}\big[\mathrm{var}(Y \mid N)\big] + \mathrm{var}\big(\mathbf{E}[Y \mid N]\big) \\ &= \mathbf{E}[N]\sigma^2 + \mathrm{var}(N\mu) \\ &= \mathbf{E}[N]\sigma^2 + \mu^2\mathrm{var}(N).\end{aligned}$$

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

$$\mathrm{var}(Y) = \mathbf{E}[N]\mathrm{var}(X) + (\mathbf{E}[X])^2\mathrm{var}(N) = \frac{1}{p} \cdot \frac{1}{\lambda^2} + \frac{1}{\lambda^2} \cdot \frac{1-p}{p^2} = \frac{1}{\lambda^2 p^2}.$$

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,

$$\text{cov}(X, Y) = \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] = \mathbf{E}[XY] = 0,$$

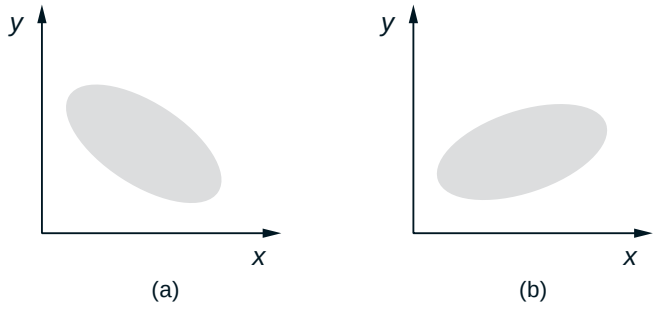


Figure 4.8: Examples of positively and negatively correlated random variables. Here X and Y are uniformly distributed over the ellipses shown. In case (a) the covariance $\text{cov}(X, Y)$ is negative, while in case (b) it is positive.

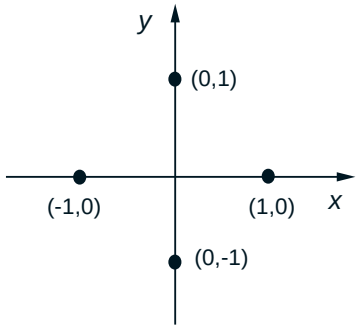


Figure 4.9: Joint PMF of X and Y for Example 4.21. Each of the four points shown has probability 1/4. Here X and Y are uncorrelated but not independent.

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

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

It may be viewed as a normalized version of the covariance $\text{cov}(X, Y)$, 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

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] \\ &= -\mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\ &= -\operatorname{var}(X).\end{aligned}$$

Hence, the correlation coefficient is

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} = \frac{-\operatorname{var}(X)}{\sqrt{\operatorname{var}(X)\operatorname{var}(X)}} = -1.$$

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

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{cov}(X_i, X_j).$$

This can be seen from the following calculation, where for brevity, we denote $\tilde{X}_i = X_i - \mathbf{E}[X_i]$:

$$\begin{aligned}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbf{E}\left[\left(\sum_{i=1}^{n} \tilde{X}_i\right)^2\right] \\ &= \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{X}_i \tilde{X}_j\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1} \mathbf{E}[\tilde{X}_i{}^2] + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{cov}(X_i, X_j).\end{aligned}$$

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).

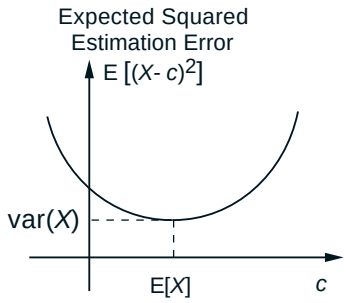


Figure 4.10: The mean squared error $\mathbf{E}\big[(X-c)^2\big]$, as a function of the estimate c, is a quadratic in c and is minimized when $c=\mathbf{E}[X]$. The minimum value of the mean squared error is $\mathrm{var}(X)$.

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.

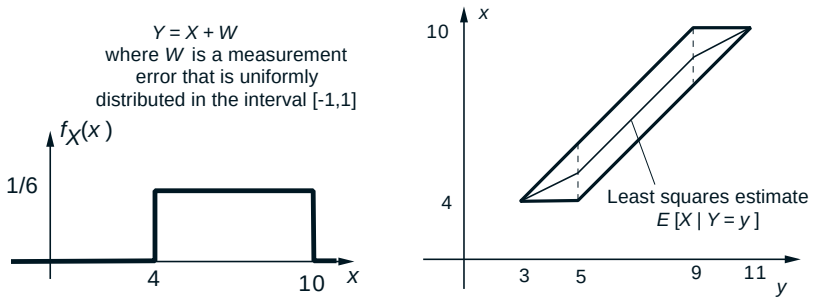


Figure 4.11: The PDFs in Example 4.27. The least squares estimate of X given the experimental value y of the random variable $Y = X + W$ depends on y and is represented by the piecewise linear function shown in the figure on the right.

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 $g(Y)$.



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

$$\begin{aligned}
\mathrm{var}(X) &= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)^2\big] \\
&= \mathbf{E}\Big[\big(\hat{X} - \mathbf{E}[X] + \tilde{X}\big)^2\Big] \\
&= \mathbf{E}\Big[\big(\hat{X} - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}[\tilde{X}^2] + 2\mathbf{E}\big[\big(\hat{X} - \mathbf{E}[X]\big)\tilde{X}\big] \\
&= \mathbf{E}\Big[\big(\hat{X} - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}[\tilde{X}^2] \\
&= \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X}).
\end{aligned}$$

(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 $\hat{X} = \mathbf{E}[X \,|\, Y] = \mathbf{E}[X]$, for every realization of Y.

If X and Y are independent, we have $\mathbf{E}[X \,|\, Y] = \mathbf{E}[X]$ and Y is indeed uninformative, which is quite intuitive. The converse, however, is not true. That is, it is possible for $\mathbf{E}[X \,|\, Y]$ 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,

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y_1, \ldots, Y_n]\big)^2\Big] \leq \mathbf{E}\Big[\big(X - g(Y_1, \ldots, Y_n)\big)^2\Big],$$

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 $\mathrm{cov}(X,Y)$ is the covariance of X and Y:

$$\mathrm{cov}(X,Y) = \mathbf{E}\big[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])\big].$$

This is a quadratic function of a, which is minimized at the point where its derivative is zero, that is, if

$$a = \frac{\mathrm{cov}(X,Y)}{\sigma_Y^2} = \frac{\rho\sigma_X\sigma_Y}{\sigma_Y^2} = \rho\frac{\sigma_X}{\sigma_Y},$$

where

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$$

is the correlation coefficient. With this choice of a, the mean squared estimation error is given by

$$\begin{aligned}\sigma_X^2 + a^2\sigma_Y^2 - 2a\cdot\mathrm{cov}(X,Y) =& \sigma_X^2 + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}\sigma_Y^2 - 2\rho\frac{\sigma_X}{\sigma_y}\rho\sigma_X\sigma_Y \\ =& (1-\rho^2)\sigma_X^2.\end{aligned}$$

Linear Least Mean Squares Estimation Formulas

The least mean squares linear estimator of X based on Y is

$$\mathbf{E}[X] + \frac{\mathrm{cov}(X,Y)}{\sigma_Y^2}\big(Y-\mathbf{E}[Y]\big).$$

The resulting mean squared estimation error is equal to

$$(1-\rho^2)\mathrm{var}(X).$$

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,

$$\operatorname{cov}(Y, \tilde{X}) = \mathbf{E}[Y\tilde{X}] = \mathbf{E}[YX] - \mathbf{E}[Y\hat{X}] = \mathbf{E}[YX] - \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}\mathbf{E}[Y^2] = 0.$$

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 $\mathbf{E}[X \mid Y]$ and variance
$$\sigma_{\hat{X}}^2 = (1 - \rho^2)\sigma_X^2.$$

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.

5

Stochastic Processes

Contents

5.1. The Bernoulli Process . . . . . p. 3
5.2. The Poisson Process . . . . . p. 15

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

$$\mathbf{P}(X_i = 1) = \mathbf{P}(\text{success at the } i\text{th trial}) = p,$$
$$\mathbf{P}(X_i = 0) = \mathbf{P}(\text{failure at the } i\text{th trial}) = 1 - p,$$

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

$$p_S(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

$$\mathbf{E}[S] = np, \qquad \text{var}(S) = np(1-p).$$

- **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

$$p_T(t) = (1-p)^{t-1} p, \qquad t = 1, 2, \ldots,$$

$$\mathbf{E}[T] = \frac{1}{p}, \qquad \text{var}(T) = \frac{1-p}{p^2}.$$

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

$$Y_k = T_1 + T_2 + \cdots + T_k.$$

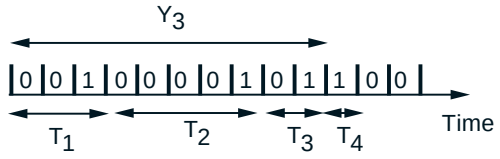


Figure 5.1: Illustration of interarrival times. In this example, $T_1 = 3$, $T_2 = 5$, $T_3 = 2$, $T_4 = 1$. Furthermore, $Y_1 = 3$, $Y_2 = 8$, $Y_3 = 10$, $Y_4 = 11$.

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

$$\mathbf{E}[T] = \frac{1}{1-p}, \qquad \text{var}(T) = \frac{p}{(1-p)^2}.$$

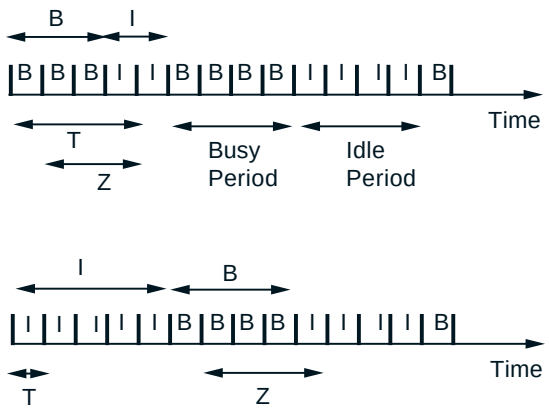


Figure 5.2: Illustration of busy (B) and idle (I) periods in Example 5.3. In the top diagram, $T = 4$, $B = 3$, and $I = 2$. In the bottom diagram, $T = 1$, $I = 5$, and $B = 4$.

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.

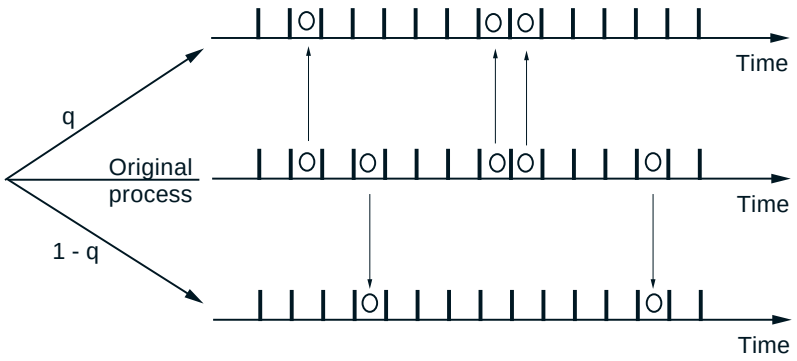


Figure 5.3: Splitting of a Bernoulli process.

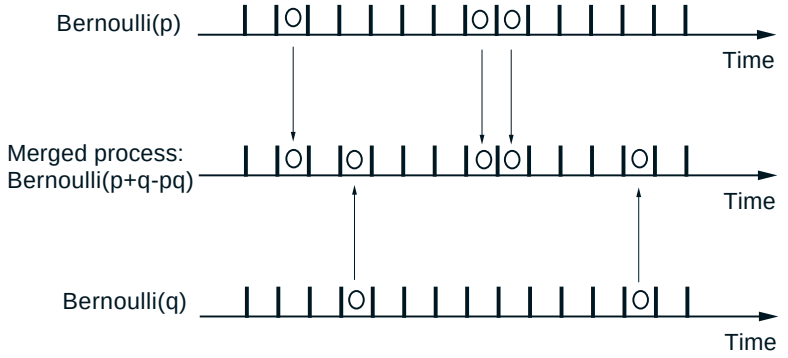


Figure 5.4: Merging of independent Bernoulli process.

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

$$\mathbf{E}[Z] = \lambda, \qquad \text{var}(Z) = \lambda.$$

- 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 $\lambda = np = 0.5$,

$$p_X(0) = 0.590, \qquad p_Z(0) = 0.605,$$
$$p_X(1) = 0.328, \qquad p_Z(1) = 0.303,$$
$$p_X(2) = 0.0729, \qquad p_Z(2) = 0.0758,$$
$$p_X(3) = 0.0081, \qquad p_Z(3) = 0.0126,$$
$$p_X(4) = 0.00045, \qquad p_Z(4) = 0.0016,$$
$$p_X(5) = 0.00001, \qquad p_Z(5) = 0.00016.$$

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.

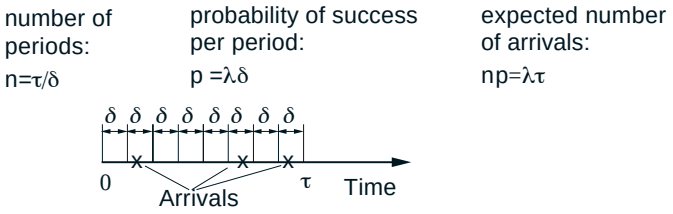


Figure 5.5: Bernoulli approximation of the Poisson process.

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

$$P(0, \tau) = e^{-\lambda\tau} = 1 - \lambda\tau + O(\tau^2)$$
$$P(1, \tau) = \lambda\tau e^{-\lambda\tau} = \lambda\tau - \lambda^2\tau^2 + O(\tau^3) = \lambda\tau + O(\tau^2),$$

consistent with property (c).

Using our earlier formulas for the mean and variance of the Poisson PMF, we obtain

$$\mathbf{E}[N_\tau] = \lambda\tau, \qquad \text{var}(N_\tau) = \lambda\tau,$$

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

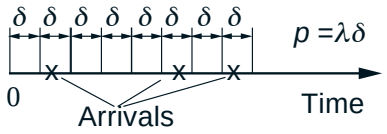
- **The Poisson with parameter $\lambda\tau$.** This is the number N_τ of arrivals in a Poisson process with rate λ, over an interval of length τ. Its PMF, mean, and variance are

$$p_{N_\tau}(k) = P(k,\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \ldots.$$

$$\mathbf{E}[N_\tau] = \lambda\tau, \qquad \text{var}(N_\tau) = \lambda\tau.$$

- **The exponential with parameter λ.** This is the time T until the first arrival. Its PDF, mean, and variance are

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \qquad \mathbf{E}[T] = \frac{1}{\lambda}, \qquad \text{var}(T) = \frac{1}{\lambda^2}.$$



	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 $\{Y_k \leq y\}$ 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 **s**plit 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

$$\mathbf{P}(\text{0 arrivals in the merged process}) \approx (1 - \lambda_1\delta)(1 - \lambda_2\delta) \approx 1 - (\lambda_1 + \lambda_2)\delta,$$

$$\mathbf{P}(\text{1 arrival in the merged process}) \approx \lambda_1\delta(1 - \lambda_2\delta) + (1 - \lambda_1\delta)\lambda_2\delta \approx (\lambda_1 + \lambda_2)\delta,$$

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

$$\mathbf{P}(\text{1 arrival of person with a letter} \,|\, \text{1 arrival}).$$

Using the definition of conditional probabilities, and ignoring the negligible probability of more than one arrival, this is

$$\frac{\mathbf{P}(\text{1 arrival of person with a letter})}{\mathbf{P}(\text{1 arrival})} \approx \frac{\lambda_1\delta}{(\lambda_1 + \lambda_2)\delta} = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

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:

$$\text{var}(T_1 + T_2 + T_3) = \text{var}(T_1) + \text{var}(T_2) + \text{var}(T_3) = \frac{1}{9\lambda^2} + \frac{1}{4\lambda^2} + \frac{1}{\lambda^2}.$$

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 $2/\lambda$.

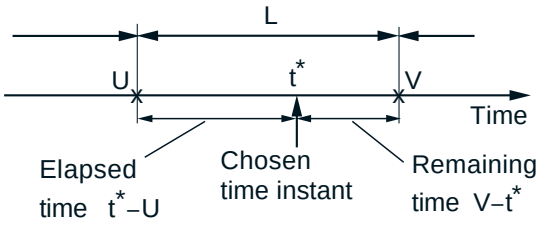


Figure 5.7: Illustration of the random incidence phenomenon. For a fixed time instant t^*, the corresponding interarrival interval $[U, V]$ consists of the elapsed time $t^* - U$ and the remaining time $V - t^*$. These two times are independent and are exponentially distributed with parameter λ, so the 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.

6

Markov Chains

Contents

6.1. Discrete-Time Markov Chains . . . . . p. 2
6.2. Classification of States . . . . . p. 9
6.3. Steady-State Behavior . . . . . p. 13
6.4. Absorption Probabilities and Expected Time to Absorption . p. 25
6.5. More General Markov Chains . . . . . p. 33

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,

$$p_{ij} = \mathbf{P}(X_{n+1} = j \,|\, X_n = i), \qquad i,\, j \in \mathcal{S}.$$

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

$$\begin{aligned}\mathbf{P}(X_{n+1} = j \,|\, X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) &= \mathbf{P}(X_{n+1} = j \,|\, X_n = i) \\ &= p_{ij},\end{aligned}$$

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.

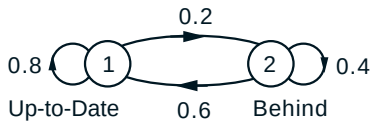


Figure 6.1: The transition probability graph in Example 6.1.

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

$$\mathbf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{i_n} = i_n) = \mathbf{P}(X_0 = i_0) p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

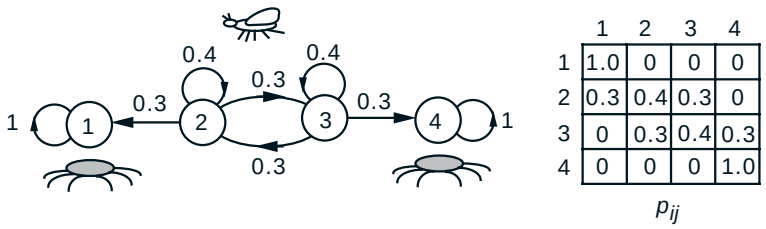


Figure 6.2: The transition probability graph and the transition probability matrix in Example 6.2, for the case where $m = 4$.

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.

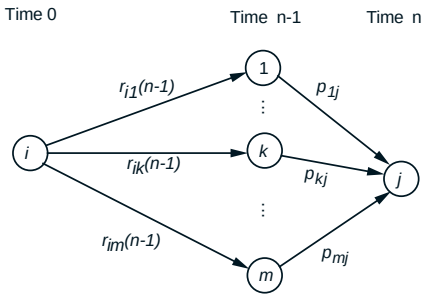
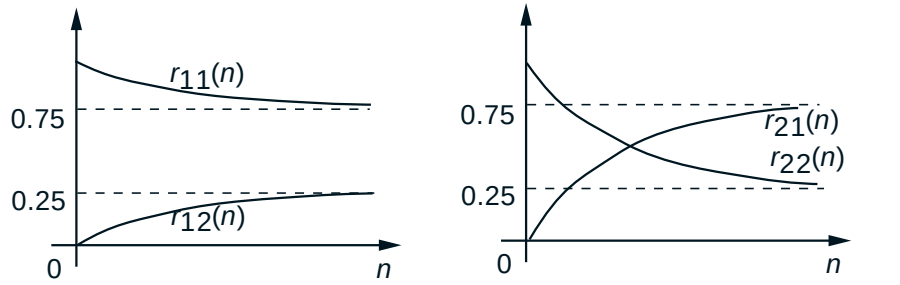


Figure 6.3: Derivation of the Chapman-Kolmogorov equation. The probability of being at state j at time n is the sum of the probabilities $r_{ik}(n-1)p_{kj}$ of the different ways of reaching j.



n-step transition probabilities as a function of the number*n* of transitions

	UpD	B
UpD	0.8	0.2
B	0.6	0.4

$r_{ij}(1)$

.76	.24
.72	.28

$r_{ij}(2)$

.752	.248
.744	.256

$r_{ij}(3)$

.7504	.2496
.7488	.2512

$r_{ij}(4)$

.7501	.2499
.7498	.2502

$r_{ij}(5)$

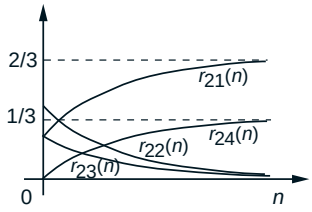
Sequence of *n* -step transition probability matrices

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.

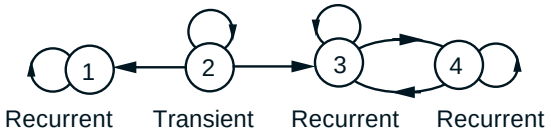


Figure 6.6: Classification of states given the transition probability graph. Starting from state 1, the only accessible state is itself, and so 1 is a recurrent state. States 1, 3, and 4 are accessible from 2, but 2 is not accessible from any of them, so state 2 is transient. States 3 and 4 are accessible 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:

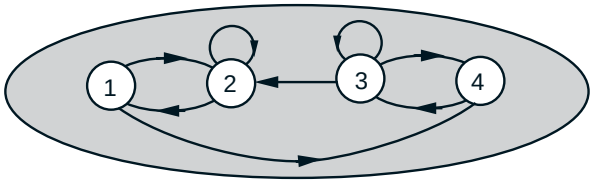
(a) once the state enters (or starts in) a class of recurrent states, it stays within that class; since all states in the class are accessible from each other, all states in the class will be visited an infinite number of times;

(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.

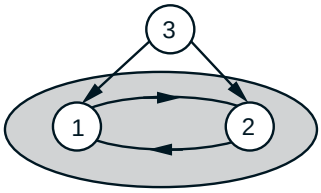
For the purpose of understanding long-term behavior of Markov chains, it is important to analyze chains that consist of a single recurrent class. For the purpose of understanding short-term behavior, it is also important to analyze the mechanism by which any particular class of recurrent states is entered 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

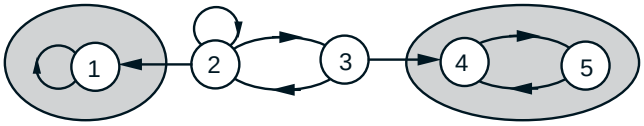
One more characterization of a recurrent class is of special interest, and relates



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.

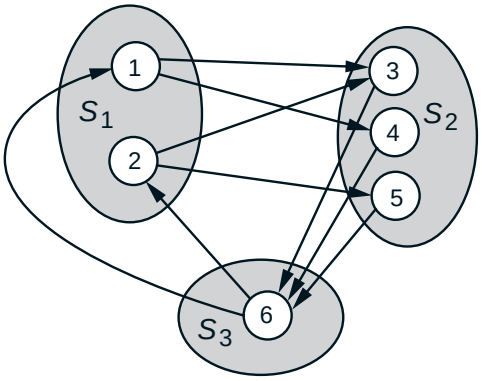


Figure 6.8: Structure of a periodic recurrent class.

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.

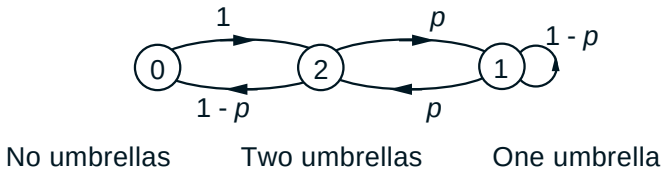


Figure 6.9: Transition probability graph for Example 6.5.

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?

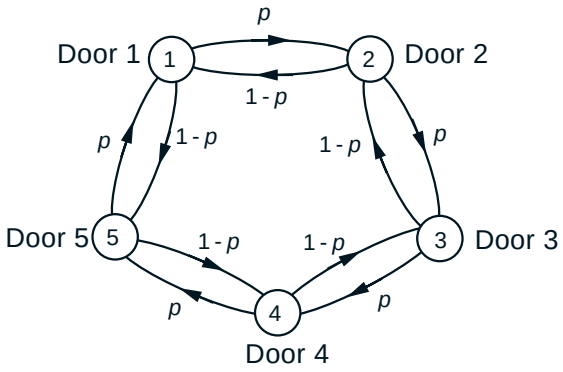


Figure 6.10: Transition probability graph in Example 6.6, for the case of $m = 5$ doors.

We introduce a Markov chain with the following m states:

$$\text{State } i\text{: Last door used is door } i, \qquad i = 1, \ldots, m.$$

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 uniquenes 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

$$\pi_1 = \frac{r}{b+r}, \qquad \pi_2 = \frac{b}{b+r}.$$

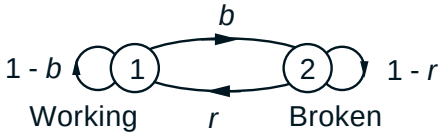


Figure 6.11: Transition probability graph for Example 6.7.

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 $\pi_{(2,i)}$.

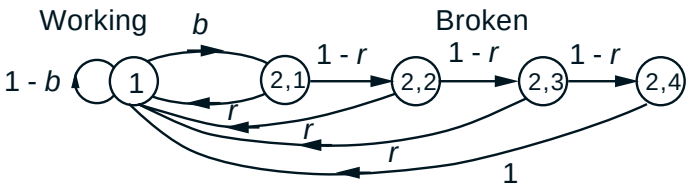


Figure 6.12: Transition probability graph for Example 6.8. A machine that has remained broken for $\ell = 4$ days is replaced by a new, working machine.

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.

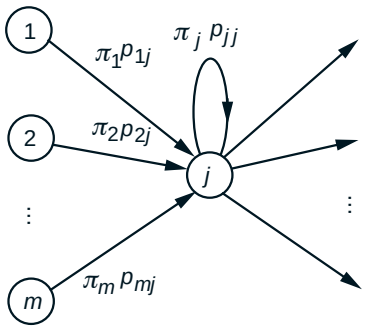


Figure 6.13: Interpretation of the balance equations in terms of frequencies. In a very large number of transitions, there will be a fraction $\pi_k p_{kj}$ that bring the state from k to j. (This also applies to transitions from j to itself, which occur with frequency $\pi_j p_{jj}$.) The sum of the 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,

$$b_i = \mathbf{P}(X_{n+1} = i+1 \,|\, X_n = i), \qquad \text{("birth" probability at state } i\text{)},$$
$$d_i = \mathbf{P}(X_{n+1} = i-1 \,|\, X_n = i), \qquad \text{("death" probability at state } i\text{)}.$$

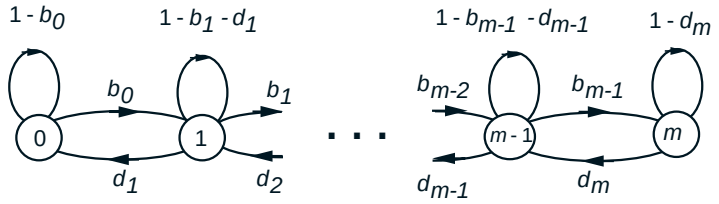


Figure 6.14: Transition probability graph for a birth-death process.

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.

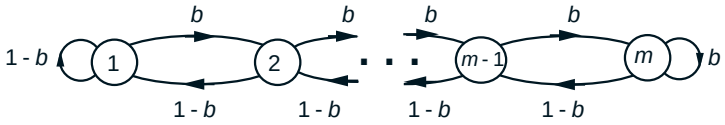


Figure 6.15: Transition probability graph for the random walk Example 6.9.

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

$$\pi_i = \begin{cases} \dfrac{\rho^i (1-\rho)}{1-\rho^{m+1}} & \text{if } \rho \neq 1, \\ \dfrac{1}{m+1} & \text{if } \rho = 1, \end{cases} \qquad i = 0, 1, \ldots, m.$$

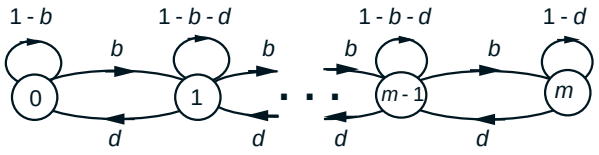


Figure 6.16: Transition probability graph in Example 6.10.

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.

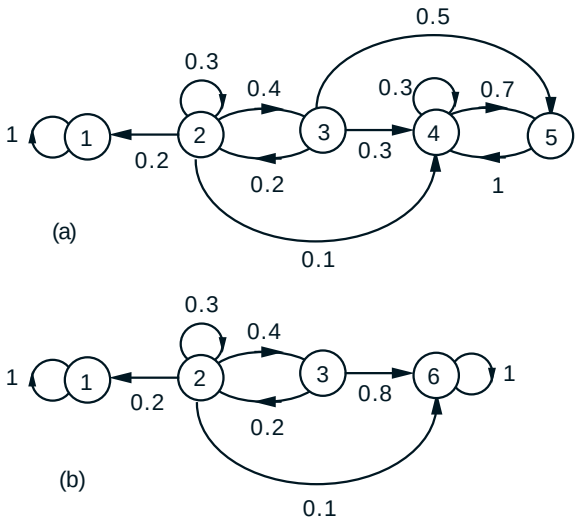


Figure 6.17: (a) Transition probability graph in Example 6.11. (b) A new graph in which states 4 and 5 have been lumped into the absorbing state $s = 6$.

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.

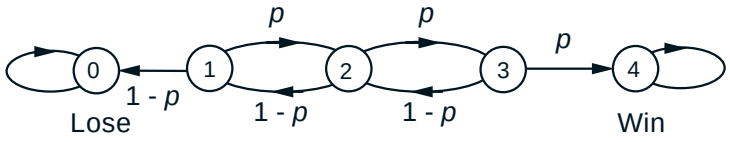


Figure 6.18: Transition probability graph for the gambler's ruin problem (Example 6.12). Here $m = 4$.

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

$$\begin{aligned}\mu_i &= \mathbf{E}\big[\text{number of transitions until absorption, starting from } i\big] \\ &= \mathbf{E}\big[\,\min\{n \geq 0 \,|\, X_n \text{ is recurrent}\} \;\big|\; X_0 = i\big].\end{aligned}$$

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, $\mu_2 = 10/3$.

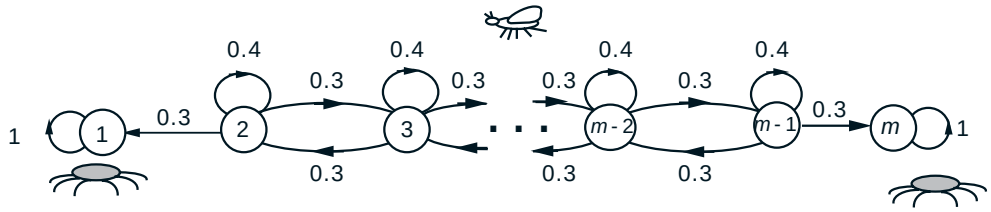


Figure 6.19: Transition probability graph in Example 6.13.

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.

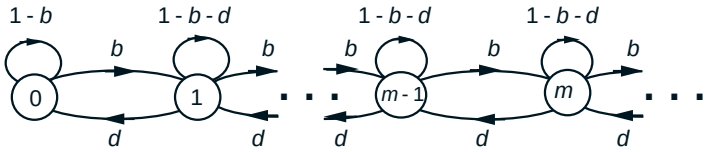


Figure 6.20: Transition probability graph in Example 6.15.

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 Poissson 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

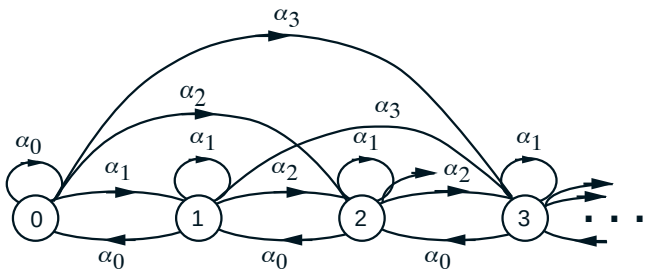


Figure 6.21: Transition probability graph for the number of packets left behind by a packet completing transmission in the $M/G/1$ queue (Example 6.16).

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 Poissson 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 inthe $M/M/1$ 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

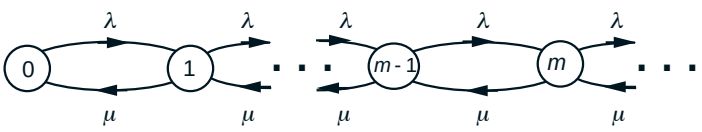


Figure 6.22: Transition graph for the $M/M/1$ queue (Example 6.17).

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.

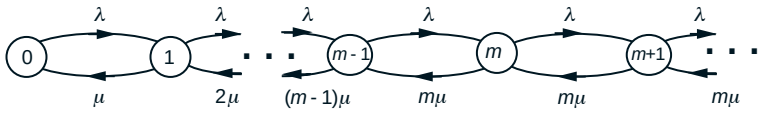


Figure 6.24: Transition graph for the $M/M/m$ queue (Example 6.19).

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.

7

Limit Theorems

Contents

7.1. Some Useful Inequalities . . . . . p. 3
7.2. The Weak Law of Large Numbers . . . . . p. 5
7.3. Convergence in Probability . . . . . p. 7
7.4. The Central Limit Theorem . . . . . p. 9
7.5. The Strong Law of Large Numbers . . . . . p. 16

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 $\mathbf{E}[X] = \text{var}(X) = 1$. 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,

$$\text{var}(M_n) = \frac{\text{var}(X_1 + \cdots + X_n)}{n^2} = \frac{\text{var}(X_1) + \cdots + \text{var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

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

$$\mathrm{var}(Z_n) = \frac{\mathrm{var}(X_1 + \cdots + X_n)}{\sigma^2 n} = \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)}{\sigma^2 n} = \frac{n\sigma^2}{n\sigma^2} = 1.$$

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

$$\lim_{n\to\infty} \mathbf{P}(Z_n \leq z) = \Phi(z), \qquad \text{for every } z.$$

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

$$\mathbf{P}(S_{100} \leq 3000) \approx \Phi(1.92) = 0.9726.$$

Thus the desired probability is

$$\mathbf{P}(S_{100} > 3000) = 1 - \mathbf{P}(S_{100} \leq 3000) \approx 1 - 0.9726 = 0.0274.$$

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 $\mathbf{P}(N_{320} \geq 100)$. 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 $\{N_{320} \geq 100\}$ is the same as the event $\{S_{100} \leq 320\}$, 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

$$\mathbf{P}(S_{100} \leq 320) \approx \Phi(1.73) = 0.9582.$$

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 $\mathbf{P}(S_n \leq c)$ 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 $\{k \leq S_n \leq \ell\}$, where k and ℓ are given integers. We express the event of interest in terms of a standardized random variable, using the equivalence

$$k \leq S_n \leq \ell \qquad \Longleftrightarrow \qquad \frac{k-np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}}.$$

By the central limit theorem, $(S_n - np)/\sqrt{np(1-p)}$ has approximately a standard normal distribution, and we obtain

$$\begin{aligned} \mathbf{P}(k \leq S_n \leq \ell) =& \mathbf{P}\left(\frac{k-np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}} \right) \\ \approx& \Phi\left(\frac{\ell - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}} \right). \end{aligned}$$

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.

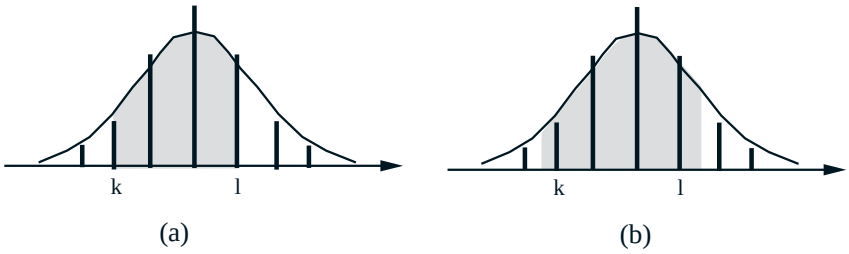


Figure 7.1: The central limit approximation treats a binomial random variable S_n as if it were normal with mean np and variance $np(1-p)$. This figure shows a binomial PMF together with the approximating normal PDF. (a) A first approximation of a binomial probability $\mathbf{P}(k \leq S_n \leq \ell)$ 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, $\mathbf{P}(k \leq S_n \leq \ell)$ 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

$$\mathbf{P}(k \leq S_n \leq \ell) \approx \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

Example 7.11. Let S_n be a binomial random variable with parameters $n = 36$ and $p = 0.5$. An exact calculation yields

$$\mathbf{P}(S_n \leq 21) = \sum_{k=0}^{21} \binom{36}{k} (0.5)^{36} = 0.8785.$$

The central limit approximation, without the above discussed refinement, yields

$$\mathbf{P}(S_n \leq 21) \approx \Phi\left(\frac{21 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21 - 18}{3}\right) = \Phi(1) = 0.8413.$$

Using the proposed refinement, we have

$$\mathbf{P}(S_n \leq 21) \approx \Phi\left(\frac{21.5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21.5 - 18}{3}\right) = \Phi(1.17) = 0.879,$$

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.