Introduction to Probability

J.R. Baxter

November 17, 2024

Contents

List of Figures

Preface

This is an introduction to probability theory, designed for self-study. It covers the same topics as the one-semester introductory courses which I taught at the University of Minnesota, with some extra discussion for reading on your own. The reasons which underlie the rules of probability are emphasized.

Probability theory is certainly useful. But how does it feel to study it? Well, like other areas of mathematics, probability theory contains elegant concepts, and it gives you a chance to exercise your ingenuity, which is often fun. But in addition, randomness and probability are part of our experience in the real world, present everywhere and yet still somewhat mysterious. This gives the subject of probability a special interest.

With self-study in mind, detailed solutions are given for all the exercises here. The exercises are mixed in with the exposition, and you are encouraged to solve them (on paper) as you read the theory. To get the benefit of an exercise, please work it out, or attempt it seriously, before reading the solution. Tackling at least some of the exercises is essential for learning.

Many facts are stated as numbered lemmas or remarks, often with descriptive names. This adds some noise, but should help in following the train of thought on your own. If a proof is given, the purpose is to clarify concepts, and all details are explained. Proofs are always optional in this book, but readers are encouraged to work at them, since proofs are one of the ways in which we internalize mathematical ideas. Internalizing ideas means making them part of our thinking, rather than leaving them as recipes from an outside source. Solving problems, working through examples, and thinking about the physical meaning of concepts are other ways of internalizing mathematics.

When reading this book on a computer (which is the intended way) you can use links to hop back and forth between exercises and solutions, as well as to follow references to equations and theorems. There is a large table

of contents and a large index with links to topics and definitions. (Most pdf viewers can return from following a link, coming back to the previous spot. This saves time. There may be a button to return, or a keystroke like "ctrl-left-arrow" or "alt-left-arrow".)

The order of the chapters is fairly logical, but a different order might be just as natural. Later chapters assume knowledge of calculus. Depending on your interests, some chapters can be omitted, or read quickly.

Learning mathematics always requires some "intense solitary thought", but it is also a human activity. If you have an opportunity to discuss your work and share ideas with others, try to do that. There are many good textbooks on probability theory, and dipping into another book can be very stimulating, especially if you find a different approach to a topic.

It is just possible that there are a few misprints. Corrections and suggestions will be gratefully received at probabilitybook@gmail.com. I particularly wish to thank Larry Susanka, who contributed many insightful comments on probability.

This book is dedicated to all the participants in my probability classes. Thanks for listening!

John Baxter

September 2023

Chapter 1

Probability and Events

In this chapter we try to explain the real-world background for probability. The discussion does not make much use of mathematics, and can be read quite rapidly. After working through later chapters, readers may find it worthwhile to look over this introduction again, and compare it with the precise statements of mathematical probability.

1.1 Common sense probability

Events in the real world are often unpredictable, and happen without clear causes. Such events are said to be *random*. To deal with randomness we all use "common sense probability", and we do so with little or no use of mathematics. For example, no one needs to study probability theory to decide whether it is safe to cross the road. But the concepts in probability are of interest, and mathematical probability theory is used widely in science and industry. In this book we are studying mathematical probability theory. We will build upon our understanding of common sense probability.

Can we give a simple definition of the concept of probability? A simple definition of a concept would be one that is expressed in terms of other concepts. But some concepts seem to be so basic that they cannot be given this sort of definition. For example: we all have some understanding of the geometrical concept of three-dimensional space, but if someone asked you to give a simple definition of space, what would you say? We seem to build up our intellectual understanding of space gradually, not through a simple definition, but through the use of this concept as we experience the world

around us.

Is probability like that, or not? Certainly probability is a very different property from space or time. But let's try an experiment.

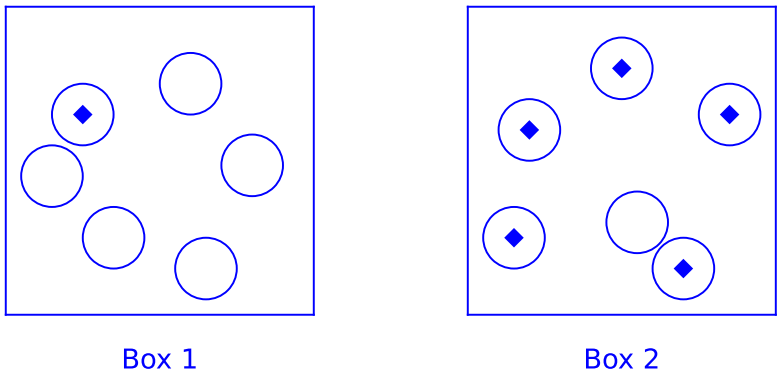


Figure 1.1: Box 1 and Box 2

Example 1.1 (The two boxes). Imagine that someone presents you with a choice of two boxes, Box 1 and Box 2. You cannot see inside either box, but you are allowed to choose one of the two boxes, and then reach into that box and take out one object. You must select an object in the box without looking, so you have no control over which object you obtain from the box.

You know that Box 1 contains six objects. One is a valuable diamond, and the other five objects are merely stones from the road. And you also know that Box 2 contains six objects. Five of these objects are valuable diamonds, and the remaining one is a stone without value. See Figure 1.1.

Remember, you must choose either Box 1 or Box 2 before you make your selection from the box. After you make your selection, you will be holding one object in your hand, either a valuable diamond or a worthless stone. Assuming that you wish to get rich, which box should you choose?

The unanimous answer is surely "Box 2". That is an example of common sense probability. But now comes the challenge: explain why you would choose Box 2, without using the word "probability", and without using any synonym, such as "chance" or "odds" or "likelihood".

An answer to this challenge might help in formulating a definition of probability. However, in Example 1.1 we didn't state exactly what constitutes an explanation, so someone might respond by simply saying "Box 2 gives you more ways to win". Such an answer certainly identifies a difference, but doesn't explain why this difference matters. So one could debate whether this is a sufficient explanation. But it doesn't give us a definition.

At any rate, as far as your author knows, no one has ever given a simple definition of probability. And that's ok! In this book we will build up our understanding of probability through examples and mathematical properties, drawing on our experience with probability in the real world.

Exercise 1.1. Consider a more complicated version of Example 1.1. Keep Box 2 the same, but change Box 1 to have 10 diamonds and 90 stones. In that case Box 1 certainly gives you "more ways to win". Is Box 2 still a better choice?

[Solution]

1.2 Probability as belief?

One can regard probability as a way of measuring what might be called "degree of belief". To a possible future event, we assign a number between 0 and 1, called the probability, which expresses our confidence that the event will happen.

Probability 1 for any event means we are certain the event will happen, probability 0 means we think it is impossible. Probability values which are between 0 and 1 mean we are not sure.

Our common sense probability judgments are based on knowledge. Your knowledge might be different from mine, and as a result we might assign very different likelihood to the same possible event. So it is natural to try to describe probability as a belief inside your head, i.e. something subjective. Is this a sensible definition?

Defining probability as "degree of belief" turns out to be an elegant way to think about the formulas of probability theory. And it is not wrong, just insufficient. We must still try to connect those probabilities inside our heads with the external world, and explain the brutal fact that correct assessments

of probability tend to keep you alive, and incorrect assessments of probability tend to kill you. A practical connection between probability and the real world will be stated as Probability Fact 1.1, after some discussion of concepts.

1.3 Experiments

We will use the phrase "experimental situation" as a convenient general term to describe a situation in which you know the setting but may have incomplete information. For brevity we might also just say "experiment" to describe this situation.

For example, perhaps someone will take an action, or has taken an action, and the result of this action is unknown to you, although you may learn the result later. We are calling the situation and the action an experimental situation, even though it need not arise from something you do in a scientific laboratory. It might just be tossing a coin, and indeed a coin toss is one of our standard examples.

The result of the experiment will often be called the *outcome*.

A real experiment takes place at a definite place and time, is carried out by particular people, and so on. Most of those details are irrelevant when calculating a probability.

When we talk about the outcome of the experiment, we usually only mean the features which are essential for our purposes. So for a coin toss we tersely say that the outcome is either a head or a tail.

1.4 Repeated coin tosses

Think more about tossing a coin.

We are not surprised that the result of tossing a coin is unpredictable. It seems that small changes, even ones that are too small to notice, can have an effect on the result of the toss. The coin is usually spinning in the air, and if it spins just a little faster, or we toss it just a little higher, that can change the result. Even if we try to toss the coin the same way each time, for most people there seems to be some kind of "shakiness" in the motions of their arms and hands. Perhaps that leads to unpredictability.

Suppose someone asserts, in everyday language, that a particular coin has probability .55 of coming up heads when tossed. This number .55 does

not help very much in predicting what will happen the next time we toss the coin! What is such a probability value good for?

It is perhaps surprising that probability does tell us something useful in this situation, provided that we are willing to toss the coin many times. Given that the probability of a head is .55, we expect that if we toss the coin 10000 times, it is likely that approximately 5500 of the results will be heads, although it is unlikely that *exactly* 5500 heads will be obtained. Please note that there are two vague words in the previous sentence: "likely" and "approximately". And yet, despite the vagueness, this is a key insight about the world.

The concept of the *frequency* gives us a convenient way to express what probability tells us. Here's the definition of frequency. We'll state it for the coin-tossing situation, but it applies to any experimental situation.

Definition 1.2 (Frequencies). When the coin is tossed N times, and heads occur on M of the tosses, we say that the frequency f of heads is given by

$$f = \frac{M}{N}. \tag{1.1}$$

Thus the frequency of heads is the *fraction of times* that a head is obtained.

Our interpretation of the probability .55 is: if we toss the coin many times, we are confident that the frequency of heads will be approximately .55. This is an example of the "Frequency Interpretation of Probability". The general statement is given below in Probability Fact 1.1.

Readers will be familiar with this way of thinking about probability. We expect that a baseball player with a high batting average is more likely to get a hit than someone with a low average, and so on.

But perhaps we should try to be surprised, just for a moment! Suppose we toss a coin $10,000$ times, and get 5439 heads. If we toss the coin another $10,000$ times, we certainly don't know what will happen on any particular toss. And yet, even if no one told us the probability of a head with this coin, we feel confident that the *total* number of heads the next time will not be too different from what was obtained the first time! So in this limited sense we can predict the future, and that is still enormously helpful.

Try to imagine a world in which the frequency in one series of tosses told us nothing about the frequency in the next series of tosses. That world would be far more chaotic than the one we live in. Planning and decision-making might be so difficult that we could not survive. And the concept of probability would not exist.

1.5 Selecting from the box

Return to the experiment described in Example 1.1. Imagine that you are able to repeat this experiment many times.

Suppose that on each repetition you choose Box 2, and then remove one object from Box 2, which must be either a diamond or a stone. What do you think will happen?

Each repetition of this experiment is supposed to be a fresh start, with no connection to the results of the previous repetitions. Box 2 contains five diamonds and one stone, each time. We can picture the box as being shaken vigorously each time before the object is selected, so we have no idea of the positions of objects inside. And we should assume that the diamonds and the stone are indistinguishable by touch, so we have no control at all over which object is selected.

In a long series of repetitions of this experiment, very likely you will obtain a diamond from the box in approximately 5/6 of the repetitions.

Of course, if Box 1 were chosen for each repetition of the experiment, we would expect that approximately 1/6 of the time a diamond would be obtained. If we define "success" to mean that a diamond is obtained, we can say that Box 2 is a better choice than Box 1 because it gives a larger frequency of success.

To express our thoughts more concisely, we can use *probability language* instead of frequency language.

Thus we would say that when selecting an object from Box 2, the *probability* of success is 5/6, and when selecting an object from Box 1, the probability of success is only 1/6.

In conversations about practical situations, most people people seem well aware of the connection between probability and frequency. In theoretical discussions this connection is often called the "frequency interpretation of probability".

1.6 The frequency interpretation

We will be talking about an interpretation for the probability of an event. The word "event" is used in ordinary speech, but let's define a slightly more precise usage here.

Definition 1.3 (Events). We will use the term "event" to describe the occurrence or non-occurrence of a property of the outcome of an experiment. We often denote such an event by a letter, so for example we might speak of the event A.

Remember that the concept of probability has not been given a precise definition, although we've talked about common sense probability, and we've talked about probability as a degree of belief. In a particular situation, one may estimate the probability of an event by means of careful observation, or, less precisely, from general experience. Once we have decided on the probability of an event, the general laws of probability will then determine the probabilities of other events. We don't have a neat definition of probability, but thinking about frequencies will help us to use probability correctly.

Here's a convenient standard notation: for any event A, let us write $\mathbf{P}(A)$ to denote the probability that we assign to an event A.

If the event is defined for a particular experiment, imagine carrying out the experiment repeatedly, for a total of N repetitions. Sometimes each experiment in the sequence of repetitions is called a "trial". The repeated experiments are distinct actions, but are supposed to take place in *similar* settings.

What does it mean to say that settings are "similar"? Settings which look similar may have subtle differences that influence the outcomes which we observe. This means that we must think hard when applying probability to real-world settings, and use our practical experience as well as mathematical theory. But we won't worry about that right now.

In Section 1.4 we talked about the frequency with which a head is obtained in coin-tossing. In the general, the frequency with which an event occurs is defined in the same way: it is the fraction of the trials for which the event actually occurs.

Any physically meaningful probability value must be consistent with the following.

Probability Fact 1.1 (The frequency interpretation of probability). For an event A, the observed frequency of occurrence of A, in any sufficiently long sequence of repetitions of similar experimental situations, will likely be close to $\mathbf{P}(A)$.

In applications, we can use the frequency interpretation to find a probability that we don't know, and to predict a frequency from a probability that we do know.

If the frequency of an event in repeated experiments does not match the probability that we have assigned to the event, that indicates an error.

Remark 1.4 (Do we have a definition of probability here?). The answer to this question depends on your standards for definitions. However, it must be noted that the frequency interpretation of probability cannot provide a *precise* definition of probability. Since the word "likely" is used, a definition based on the frequency interpretation would be **circular**, since you would have to already know at least something about the meaning of probability, in order to understand its definition! Furthermore, the statement is vague. Look at those weasel-words "sufficiently long" and "close", in the statement. If you want the observed frequency to be, say, within 1% of the probability, how long is "sufficiently long"?

And yet, despite its theoretical deficiencies, the frequency interpretation is the most important practical statement we can make about the connection between mathematical probability statements and physical probability statements. Whatever assumptions we make later about mathematical probabilities must be consistent with the frequency interpretation.

As in our discussion of choices from Box 1 and Box 2, in general we can use probability language as a convenient way to express frequencies of events. In some practical situations frequency language may seem more informative, and either formulation is correct.

Example 1.5 (Events that are certain and events that are impossible). For some experiment, let A be an event which is certain to occur, and let C be an event which is impossible. Then we say that $\mathbf{P}(A) = 1$, and $\mathbf{P}(C) = 0$.

Let's take a moment to think about a question: are these definitions forced upon us by the frequency interpretation?

The frequency interpretation says that if we repeat the experiment many times, the measured frequency of A is likely to be close to $\mathbf{P}(A)$.

Consider N repetitions of the experiment. Since A is certain, it will occur N times, giving an experimental frequency $f = N/N = 1$. The frequency tells us the probability, so the frequency interpretation of probability does seem to require that $\mathbf{P}(A) = 1$.

Of course, if we want to be fussy about our logic, we might remember that the frequency interpretation of probability does not say that the frequency is *equal* to the probability, it says that when an experiment is repeated many times, the frequency is *likely* to be *close* to the probability. So let's take a further moment here to give a more careful argument.

The frequency interpretation of probability says that when N is sufficiently large, the difference between the frequency f and $\mathbf{P}(A)$ is likely to be small. So, for example, if the number of repetitions is sufficiently large then we will likely have $|f - \mathbf{P}(A)| < .01$, i.e. $|1 - \mathbf{P}(A)| < .01$.

Using .01 to estimate the difference was just an example. If we perform an even larger number of repetitions, then with enough repetitions the frequency interpretation says that we will likely have $|1 - \mathbf{P}(A)| < .001$. And so on!

In the real world, our research budget will not cover endless repetitions of the experiment. But in our minds we can imagine longer and longer sequences of repetitions, for which the likely difference between 1 and $\mathbf{P}(A)$ becomes smaller and smaller, as small as we wish.

That can only be true if $\mathbf{P}(A) = 1$, so yes, the frequency interpretation of probability forces us to conclude that the value of $\mathbf{P}(A)$ must be equal to 1.

In the same way, the frequency interpretation requires that $\mathbf{P}(C) = 0$.

1.7 Adding up probabilities

Suppose you are working in a big office in Chicago, it's 2:30 pm, and the phone rings.

You know that the phones where you work only receive calls from the branch offices. There are branch offices in exactly five cities: New York City, Baltimore, Miami, San Francisco and Los Angeles.

Like most people, you are familiar with the concept of probability as it is used in practical situations. Based on the experience of people working in your office, it is believed that at this time of day, the probability that the

call is from New York is .20, the probability that the call is from Baltimore is .17, the probability that the call is from Miami is .20, the probability that the call is from San Francisco is .18, and the probability that the call is from Los Angeles is .25.

Suppose you would like to know the probability that this particular call is from the east coast. Is that an easy number to find?

It is easy. We simply add up the probabilities of calls from the cities on the east coast: New York, Baltimore and Miami. So:

$$\text{probability call is from the east coast} = .20 + .17 + .20 = .57 \tag{1.2}$$

But why do we add the probabilities for the separate cities? Can we justify this calculation?

If we think of probabilities simply as degrees of belief, it's not clear why adding is ok. Feelings are not numbers. So let's think about frequencies instead.

Think of each incoming call as an experiment. Suppose that, typically, your phones get a total of 100 calls per day, during the time period from 2:00 pm to 3:00 pm. The probabilities stated earlier suggest that you will likely get around 20 calls from New York, 17 calls from Baltimore, and 20 calls from Miami. So you will get approximately 57 calls from the east coast, out of a total of 100 calls.

Since the frequency of east coast calls is 57/100, the probability of a call being from the east coast should be around .57, and that is what you get by adding the probabilities.

What do you think of this argument? It is a bit careless, because the frequency interpretation applies to a large number of repetitions of the experiment, and 100 calls is not a large number of repetitions. But the idea is sound. To argue more carefully, think about a longer period, as long as you like, say 30 days. Then the total number of calls to the office during that time of day will be roughly $30 \times 100 = 3000$. Call that number N. Since N is large, we feel reasonably confident that approximately $N \times .20$ calls will come from New York, $N \times .17$ calls will come from Baltimore, and $N \times .20$ calls will come from Miami. Thus approximately $N \times (.20 + .17 + .20)$ calls will come from the east coast. And so:

$$\text{frequency of calls from the east coast} \approx \frac{N \times (.20 + .17 + .20)}{N} = .20{+}.17{+}.20,$$

where we write $\approx$ to mean "approximately equal". By the frequency interpretation, the sum $.20 + .17 + .20$ is the correct probability.

The same argument could be carried out in general, of course! So we have an important general rule, stated next. We will state this rule rather formally. It has to be stated that way, because it is a general rule, which is supposed to apply to many different situations. And we need to be careful in what we say, because we want our theoretical arguments to be reliable. Thinking theoretically is a lot less work that carrying out experiments, but it has to be right!

Probability Fact 1.2 (Probabilities are additive over cases). Let $D_1, \ldots, D_k$ be events for a some experiment.

Suppose that events $D_1, \ldots, D_k$ are mutually exclusive, meaning that **at most one** of the events D_i can occur. Let A be the event that one of $D_1, \ldots, D_k$ occurs. This means that if any of the events $D_1, \ldots, D_k$ occurs, by definition A occurs. Then:

$$\mathbf{P}(A) = \mathbf{P}(D_1) + \ldots + \mathbf{P}(D_k). \tag{1.3}$$

In the situation with the office phone call, we could let D_1 denote the event that that call came from New York, D_2 denote the event that the call came from Baltimore, and D_3 denote the event that the call came from Miami. A would be the event that the call came from the east coast. Then equation (1.3) and equation (1.2) say the same thing.

An important special case of Probability Fact 1.2 is the situation in which $D_1, \ldots, D_k$ cover *all* possibilities, meaning that one of these events *always* occurs. In that case we have:

$$\mathbf{P}(D_1) + \ldots + \mathbf{P}(D_k) = 1. \tag{1.4}$$

Why does equation (1.4) follow from equation (1.3)? Well, since $D_1, \ldots, D_k$ cover all possibilities, A always happens! (It's a really boring event.) Just as in Example 1.5 we then conclude immediately that $\mathbf{P}(A) = 1$, so equation (1.3) turns into equation (1.4).

1.8 Back to the boxes!

Return again to the problem of choosing from one of two boxes (Example 1.1). In Section 1.5 we stated that the frequency of success using Box 2 was 5/6.

We didn't really justify this statement, although it certainly seemed plausible. Let's give a more careful analysis now, to practice using Probability Fact 1.2.

Think about the six objects in Box 2. Our practical experience tells us that each of these objects would be chosen approximately one-sixth of the time. Since five of these objects are diamonds, the combined frequency of obtaining a diamond is 5/6. In probability language, we would say that each object has probability one-sixth of being chosen, and then say that by Probability Fact 1.2,

$$\text{the probability that a diamond is chosen} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6}.$$

That's more or less the whole story. But we might say a bit more.

Why do we think that each object in the box has probability one-sixth to be chosen? There are six objects, and we think that each one has the same chance of being chosen, don't we? That's true, but we should realize that we are building in our real-world experience when we assert that.

It is related to the comment made at the end of in Section 1.4. We said that if we toss a coin many times, and then perform a second sequence of tosses with the same coin, we expect that the frequency of heads in the second series of tosses will be roughly consistent with the frequency of heads in the first series.

Now we are considering a situation involving the six different objects in Box 2, rather than a single coin. However, the six objects are the same in any way which affects the results of the experiment. For that reason, we expect that in a long series of trials, each object will be selected with roughly the same frequency. In the language of probability, we think that each of the six objects has the same probability of being selected. If this assumption turns out to be false, we will conclude that we did not understand the experiment.

If we accept that each of the six objects has the same probability of being chosen, call this probability p. By equation (1.4),

$$p + p + p + p + p + p = 1,$$

so yep, $p = 1/6$.

1.9 Some simple examples

Example 1.6 (One coin toss). For a coin toss there seem to be only two interesting events, the event H that the result is a head, and the event T

that the result is a tail.

A coin is said to be fair if the probability of obtaining a head is equal to the probability of obtaining a tail. Gamblers are typically expected to use fair coins in their games.

A real coin may be fair or unfair. For any coin,

$$\mathbf{P}(H) + \mathbf{P}(T) = 1. \tag{1.5}$$

Exercise 1.2. How is equation (1.5) related to Probability Fact 1.2?

[Solution]

Example 1.7 (Rolling Dice). Instead of thinking about tossing a coin, let's consider rolling a die. Most people have played games in which a die is rolled, or perhaps two dice are rolled. The die is a cube, so it has six faces. Rolling the die a single time is an experiment with six possible outcomes. The outcome of the experiment is the number of dots on the uppermost face of the die when it settles. The possible outcomes are $1, 2, 3, 4, 5, 6$.

A die is said to be fair if all the outcomes $1, 2, 3, 4, 5, 6$ have the same probability.

One possible event when rolling a die is the event that the outcome is 5. We might call this event A. The event A only occurs when the outcome is 5.

Another possible event is the event that the outcome is an odd number. We might call this event B. B is described by a property that three of the possible outcomes have. If the die gives a 1, a 3, or a 5, we say that the event B occurred.

Exercise 1.3. When rolling a fair die many times, what fraction of the rolls (approximately) will result in an odd number?

[Solution]

Remark 1.8 (Comparing experiments). Rolling a fair die is physically different from the experiment of selecting an object from a box containing six possible choices, as described in Section 1.5. However, in both cases there are six basic events, everything can be described in terms of those events, and the probabilities of the basic events are equal to 1/6 in both cases. Thus one can translate any problem dealing with one of these experiments into a similar problem dealing with the other, and the corresponding numerical answers must agree.

This observation applies to the fair case. Unfair dice certainly exist, perhaps due to variations in the density of the material. On the other hand, there isn't an obvious way to modify the experiment in Section 1.5, in order to have different probabilities for the six basic events.

Exercise 1.4 (Lottery tickets). This book does not advocate buying lottery tickets. But we can think about them without making a purchase. Suppose that a company offers n lottery tickets for sale, where n may be a large integer. Exactly one of these tickets is the winning ticket, and the purchaser will receive a large sum of money. The remaining tickets are worthless, and of course we don't know which ticket is the winner. You have purchased one ticket. Let W be the event that your ticket turns out to be the winner.

(i) Let $\mathbf{P}(W)$ be the probability of W. Find $\mathbf{P}(W)$.

Note that the experiment of Section 1.5, using Box 1, essentially solves this problem for $n = 6$.

Common sense probability likely gives you the answer as well.

(ii) A certain wealthy gambler buys k lottery tickets, where k may be any number less than or equal to n Let G be the event that the gambler wins the lottery with one of purchased tickets. Find $\mathbf{P}(G)$.

[Solution]

Remark 1.9. Let W and G be the events described in the lottery of Exercise 1.4. Suppose that n is equal to 10^6. Is $\mathbf{P}(W)$ a physically meaningful

probability value? Think about deciding whether the price of the ticket is reasonable. $\mathbf{P}(W)$ is certainly relevant to that decision.

We found $\mathbf{P}(W)$ theoretically, using Probability Fact 1.1. Suppose that you wish to use the frequency interpretation to test the validity of the value calculated for $\mathbf{P}(W)$. In principle this can be done. However, the whole lottery is part of the experiment, and a ridiculously large number of repetitions of the lottery would be required to accurately measure the frequency with which W occurs.

On the other hand, when k is comparable in size to n, the value of $\mathbf{P}(G)$ could be tested experimentally with fewer repetitions. This is indirectly a test of $\mathbf{P}(W)$.

Example 1.10 (Sampling from a population). One can think about making a random selection from a population as an experiment. Pollsters do this all the time, of course.

It's easier to think about a population of jelly beans than about a population of people, so suppose you have a large bowl containing many jelly beans, some yellow and some red. In this experiment we assume that there are n jelly beans altogether, k yellow ones and $n - k$ red ones. In the experiment, you randomly select exactly one bean, and record its color.

Specifying the experiment includes specifying the *actual number of beans* of each color that are in the bowl.

We prepare for the experiment by stirring the jelly beans vigorously, so that the beans in the bowl are thoroughly mixed. That is not the experiment, just part of the setup.

Let C be the event that the selected bean is yellow. We would like to know $\mathbf{P}(C)$, that is the probability that the selected bean is yellow.

Calculations in the setting of this experiment will be similar to calculations for the lottery described in Exercise 1.4. The event that your own ticket is the winner corresponds to the event that a particular jelly bean is selected. The set of tickets bought by the wealthy gambler in part (ii) of Exercise 1.4 would correspond to a subset of the beans in the bowl, for example, to the yellow jelly beans in the bowl.

1.10 Probability distributions

One usually wants to know the probabilities for the possible outcomes of an experiment, and perhaps for some of the possible events. Here's some standard terminology.

Definition 1.11 (Probability distributions). A rule which assigns probabilities for some family of related events is called a *probability distribution* for the events. The probability which the rule prescribes for an event A is usually denoted by $\mathbf{P}(A)$.

A simple example of a probability distribution is a rule which gives the probability of each possible outcome of an experiment. We might find such a distribution experimentally, as in the next section.

The use of the word "distribution" in Definition 1.11 may reflect the fact that the probability values for all the outcomes must add up to one. In that respect, assigning probabilities to various possible events is a bit like splitting up a unit quantity of material and distributing it to various locations.

The phrase "family of related events" in Definition 1.11 is not precise. It might refer to all events, or to some limited collection of events which are of interest at the moment. We will see examples of distributions in specific settings later.

1.11 Collecting statistics

It's more fun to talk about frequencies than to actually perform experiments and measure them. But perhaps we should take a moment to look at some examples.

Statistical data

We will refer to experimental data which is systematically recorded and tabulated as *statistical data* (and see [8] for a discussion of correct grammatical usage of the word "data"!).

General features of such data are referred to as *statistical properties*. If our data is the result of a sequence of repeated experiments, one statistical

property is the frequency of a particular event. Of course one can calculate many other statistical properties in this setting, such as the frequency of obtaining the same outcome twice in a row, or the degree of variation in the data, etc. But at present we will just focus on the frequency.

Collecting data to learn probabilities

In the case of the experiment of rolling a die, a probability distribution gives the probability of each possible result. For a *fair* die, the probabilities for the values $1, 2, 3, 4, 5, 6$ are $1/6, 1/6, 1/6, 1/6, 1/6, 1/6$, respectively, but of course a die may be unfair.

Suppose we have a die, but we don't know the probability distribution associated with this die. In this situation, one might roll the die repeatedly and use the frequency interpretation to get an idea about the distribution.

Imagine rolling the die 20 times, recording the results. (To save time, we can use a computer to *simulate* rolling the die. This means that a computer program produces numbers that have similar statistical properties to the results of performing the actual repeated experiments. It's not obvious that this can be made to work, but it does work, pretty well.)

For a particular sequence of 20 trials, the outcomes happen to be

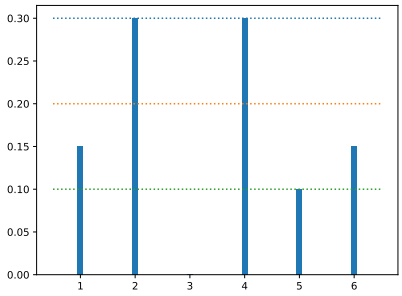
$$2, 5, 5, 5, 6, 2, 2, 4, 2, 1, 1, 1, 2, 4, 3, 2, 4, 4, 2, 6$$

You can check that the counts for the outcomes $1, 2, 3, 4, 5, 6$, are $3, 7, 1, 4, 3, 2$. Thus the frequencies for outcomes $1, 2, 3, 4, 5, 6$, are $0.15, 0.35, 0.05, 0.2, 0.15, 0.1$, respectively. See Figure 1.2a.

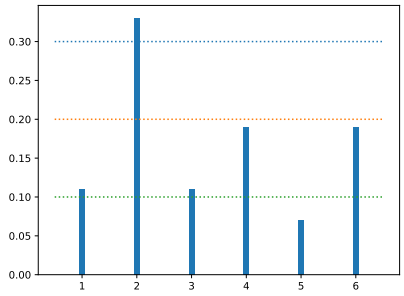
These numbers are not probabilities, of course. They are just numbers that tell us something about the recorded outcomes for a particular experiment. But if we think about making additional rolls of the same die, we can hope that these numbers give us some idea of the probability of each possible outcome.

That hope is based on the frequency interpretation of probability, which says that the probability of obtaining a particular value on one roll of the die should be similar to the observed frequency for that value, when we have a *long* sequence of repeated trials.

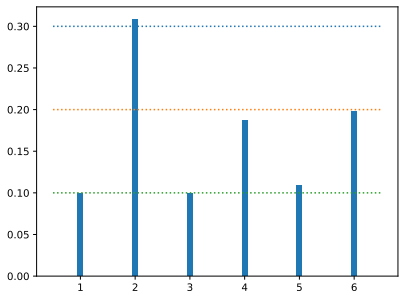
However, a sequence of 20 trials does **not** seem long, especially when there are six possible outcomes. So it seems rash to draw a conclusion about probabilities based on these frequencies.



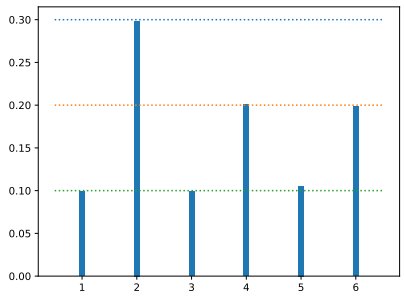
(a) frequencies: 20 rolls



(b) frequencies: 100 rolls



(c) frequencies:1000 rolls



(d) frequencies:10000 rolls

Longer sequences of trials

Let's try to get a more accurate estimate for the probability of each possible result. If we roll the die 100 times, recording the results, the frequencies for $1, 2, 3, 4, 5, 6$ are $0.11, 0.33, 0.11, 0.19, 0.07, 0.19$, respectively. See Figure 1.2b.

Even 100 repetitions is not very many. So let's do more repetitions.

If we roll the die 1000 times, recording the results, the frequencies for $1, 2, 3, 4, 5, 6$ are $0.099, 0.308, 0.099, 0.187, 0.109, 0.198$. See Figure 1.2c.

This is fairly consistent with the results for 100 trials, but of course is likely to be more reliable.

Let p_i be the probability of obtaining the value i when rolling this particular die. If we have to start playing a gambling game using this particular die, as a practical choice we might as well assume that

$$p_1 = 0.099, p_2 = 0.308, p_3 = 0.099, p_4 = 0.187, p_5 = 0.109, p_6 = 0.198.$$

If you happen to know that this example was made up by a person who

likes simple numbers, then you may suspect that the actual probabilities for outcomes $1, 2, 3, 4, 5, 6$ are $.1, .3, .1, .2, .1, .2$, respectively. However, in real world situations we should not expect such convenient values for the probabilities.

Let's do a sequence of 10000 trials, to check for consistency. This time we find that the frequencies for $1, 2, 3, 4, 5, 6$ are $0.0989, 0.2984, 0.09870, 0.20070, 0.1053, 0.198$, respectively. (See Figure 1.2d.)

Now we feel reasonably confident that we have a good approximation for the probability distribution for this die.

Remark 1.12 (Messy data!). By now you will have noticed that when randomness is involved, recorded observations seem rather messy. If we display all the data in a plot, we are unlikely to obtain obtain a nice neat picture. This is in contrast to, for example, the beautiful curves we get when plotting solutions of differential equations. We can deal with randomness but we cannot eliminate it.

With this in mind, it is striking that elegant patterns of behavior do emerge in data associated with *large* random systems. The Central Limit Theorem of probability shows this for a long series of coin tosses ([10] and Chapter 18). It is also one of the key insights of statistical physics.

The data for die rolls was obtained by simulation using a computer. We won't take time to discuss how a computer actually carries out such simulations. The next exercise asks you to consider a different kind of simulation.

Exercise 1.5 (Simple simulations). Suppose you are thinking about some experiment with three possible outcomes, each of which is supposed to have probability $1/3$. For convenience, let's give the three outcomes labels: a, b, c.

The physical apparatus for this experiment is complicated and expensive, so you won't actually perform the experiment today. But you would like to play with some statistical data corresponding to these probabilities. You can try to simulate this experiment using different equipment. That is, instead of actually doing the experiment, you will do some other experiment (perhaps something that is easier to perform repeatedly), which will produce the same values, with the same statistical properties as the real experiment.

What matters is that your simulation is supposed to produce one of the labels a, b, c, with equal probability for each label. You may not have the equipment you need, though. Here are some cases to consider.

(i) Suppose you have in your possession a fair die. How can you perform the simulation?

(ii) Suppose you don't have the fair die, but you have a fair coin and an unfair coin, and the unfair coin is known to produce a head with probability 1/3. How can you perform the simulation?

[Solution]

1.12 Brownian motion

Our examples have been rather simple, although the principles they illustrate also apply to very complex situations. Randomness seems to exist everywhere, and is almost unavoidable.

When discussing coin-tossing in Section 1.4, we suggested that most people seem to have some kind of "shakiness" in their arms and hands, which causes the result of the coin toss to be unpredictable. One might try to express this in a more general way by saying that the small motions of their arms are unpredictable, and this unpredictability then leads to unpredictable results for coin tosses. But then one can ask, "why are the small arm motions unpredictable?". This type of questioning can be continued. It seems to lead us consider more and more detailed pictures of physical processes, at smaller and smaller scales. Randomness and unpredictability apparently exist at all known levels of description.

This book has no ultimate explanation for randomness. However, to illustrate randomness on a small scale, and how its effects can spread, let's briefly consider a famous example: *Brownian motion*.

It's 1827, and biologist Robert Brown is peering into his microscope ([4], [6]). He sees little particles moving around in fluid, in a very irregular manner. The original particles come from pollen grains, but as he continues his observations he finds that all little particles in fluid seem to move in a rather similar way. They constantly change direction and do not seem to get "tired". He finds that even water that has been trapped inside rocks can contain moving particles, and they must have kept moving during millions of years. Apparently the statistical properties of this particle motion do not change.

What Brown saw is related to heat. Nowadays we interpret heat as *disorderly motion* of atoms and molecules. So let's think about disorder. You can create disorder, for example by dropping something, so that the energy of its fall is transformed into heat when it hits the ground. You can move disorder around, for example when a hot object is placed in contact with a cold object. But it is very difficult to make a large disorderly collection become more orderly.

The universe seems to be full of disorder, especially at small scales. The particles that Brown observed were large compared to molecules. That's why he could see them. But physicists think that the motion of Brown's particles is caused by collisions with molecules of the fluid which contains the particles.

These "invisible" molecules in the fluid are moving in a disorderly way. We can't predict the details of the movement of the molecules, and we think of their movement as random behavior.

It's interesting to consider how the fluid molecules interact with a particle that Brown observes. Any such particle will receive many impacts per second on all sides, from the tiny molecules. At normal temperatures the particle is going to be hit a lot.

The effect of the collisions on the particle is roughly the same in all directions, because of the disorderly motion of the molecules. However, the number of impacts on each side naturally *fluctuates*, so that briefly one side of the particle receives more collisions than the other.

We shouldn't be surprised that there are fluctuations. Fluctuations are part of random behavior. If you think of tossing a fair coin many times, there will inevitably be periods when more heads than tails occur, just by chance. It all evens out in the long run, of course.

But random fluctuations are what cause the particle movement that Brown observed. When more molecules hit a particle on one side than the other, it will move. Since the resulting particle motion is large enough to be observable in a microscope, those tiny invisible molecules must have a lot of energy.

By our standards molecules move rather violently! If the molecules of your body somehow became orderly, and all moved in a single direction, your body would hurtle away at a speed of hundreds of meters per second.

Brownian motion provides us with a vivid picture of disorder. It also gives us an example of how random behavior on a small scale is pervasive, and can lead to random behavior on a larger scale.

1.13 Solutions for Chapter 1

Solution (Exercise 1.1). Yep!

You knew, that didn't you? The success frequency using the new version of Box 1 will even worse than before (1/10 rather than 1/6).

Solution (Exercise 1.2). Either a head or a tail must be obtained. Hence events H, T cover all possibilities. These events are mutually exclusive, since the coin cannot come up both heads and tails!

By equation 1.4, with $D_1 = H$ and $D_2 = T$, $\mathbf{P}(T) + \mathbf{P}(H) = 1$.

This is equation (1.5).

Solution (Exercise 1.3). Let D_i be the event that outcome i occurs, for $i = 1, \ldots, 6$. By equation (1.4),

$$\mathbf{P}(D_1) + \ldots + \mathbf{P}(D_6) = 1.$$

For a fair die, $\mathbf{P}(D_1) = \mathbf{P}(D_2) = \ldots = \mathbf{P}(D_6)$, and so we have

$$6\mathbf{P}(D_1) = 1.$$

Thus $\mathbf{P}(D_1) = 1/6$, and so $\mathbf{P}(D_i) = 1/6$ for each $i = 1, \ldots, 6$. (Yup, we used the same argument in Section 1.8.)

Let B be the event that an odd number is obtained. Clearly

$$B = D_1 \cup D_3 \cup D_5.$$

By equation (1.3),

$$\mathbf{P}(B) = \mathbf{P}(D_1) + \mathbf{P}(D_3) + \mathbf{P}(D_5) = \frac{3}{6} = \frac{1}{2}.$$

Using the Frequency Interpretation, we expect that in a large number of rolls, approximately 1/2 of the rolls will result in an odd number.

Solution (Exercise 1.4). Suppose that each ticket has an identification number.

Let D_j be the event that ticket j is the winning ticket.

The events D_j, $j = 1, \ldots, n$ are clearly mutually exclusive and cover all possibilities.

As far as we know, no ticket is favored, and we will calculate probabilities based on that. Since no ticket is favored, $\mathbf{P}(D_i)$ is the same for every j.

By equation 1.4, $\mathbf{P}(D_1) + \ldots + \mathbf{P}(D_n) = 1$.
Hence

$$\mathbf{P}(D_j) = \frac{1}{n}$$

for every j.

(i) Suppose you purchased ticket t. W is the event that ticket t is the winning ticket. Thus $W = D_t$, and so $\mathbf{P}(W) = 1/n$.

(ii) We can always number the tickets so that tickets $1, \ldots, k$ are the ones that the wealthy gambler purchases. This is just to make it easier to write down the argument. Then

$$G = D_1 \cup \ldots \cup D_k,$$

and so by equation (1.3) we know that

$$\mathbf{P}(G) = \mathbf{P}(D_1) + \ldots + \mathbf{P}(D_k) = \frac{k}{n}.$$

Solution (Exercise 1.5).

(i) For each roll of the fair die, report the result as label a if the die gave a 1 or a 2, report label b if the die gave a 3 or a 4, and report label c if the die gave a 5 or a 6.

The die will give 1 on approximately 1/6 of the tosses and the die will give a 2 on approximately 1/6 of the tosses. Hence label a will be reported approximately 1/3 of the time, which is what is desired. Similarly labels b and c will each be reported 1/3 of the time.

(ii) Toss the unfair coin. If the coin gives a head, report label a. Otherwise, continue the simulation by tossing the fair coin. If the fair coin gives a head, report label b. If the fair coin gives a tail, report label c.

Clearly label a will be reported on approximately 1/3 of the times you perform the simulation. You will report label b during approximately 1/2 of the times that you *don't* report a. Since 1/2 of 2/3 is 1/3, this is what is desired. Similarly label c will be reported approximately 1/3 of the times.

Chapter 2

Assumptions for probability, and their consequences

In this chapter we lay out the general structure of mathematical probability. General statements are necessarily abstract, but the abstractions of probability theory are fairly pleasant.

2.1 Abstract outcomes

Often we want to know whether or not the result of a given experiment has a certain property that we are interested in. The occurrence of this property is what we call an "event". The complete result of the experiment is called the "outcome". There may be many possible outcomes for a given experiment, some of which have the property we are interested in. Calculating the probability of an event typically requires us to consider all possible outcomes. With that in mind, let's think about representing outcomes in a mathematical model.

In a calculation, we necessarily restrict our attention to abstract representations of outcomes. These mathematical representations of physical outcomes will also be called "outcomes", or perhaps "abstract outcomes" if we want to emphasize that these are objects of thought.

Each abstract outcome is a mathematical object, from which all inessential properties have been ruthlessly stripped. Thus if a botanist is experimenting in breeding roses, a beautiful new plant in the real world might be represented abstractly by a single letter which indicates its color. In general,

the representation must include whatever properties of the outcome that we are interested in, but need not have more details.

If we base our calculations on the possible outcomes, then the events that we are interested in must be represented in terms of the outcomes. When a physical event is defined by a certain property, we will represent the event as the *set* of outcomes which have this property.

Is that an adequate way to represent an event? As a set? The next example tests this approach.

Example 2.1 (Brown hair as a set of outcomes). Consider an experiment in which one person is randomly selected from a population. The person selected is the physical outcome of the experiment!

Since the outcome is a person, the outcome has a lot of properties. One of the properties of the outcome is hair color. Let A be the event that the selected person has brown hair.

Notice we have defined A physically in terms of a property of the outcome. Now suppose we wish to represent this event in an abstract model.

We can give each person in the population an identification code. An abstract outcome would be the ID of the randomly selected person. The abstract version of A would be the set of all IDs of people that have brown hair.

The question is whether this representation of A is sufficient for our needs.

Suppose that no one told you what property defines A, but instead showed you the entire set of people who have that property, would you be able to guess what the property was?

The entire set of people with the property defining A consists of exactly those members of the population who have brown hair. If you became aware of that fact, you might *guess* that hair color was the property that defines A. On the other hand, it is conceivable that some other property might occur in exactly the same set of people. So we must admit that knowing the set of outcomes does not really tell you what physical property is under consideration.

However, since you know the abstract representation of A as a set of outcomes, then, if a particular outcome occurs as a result of the experiment, you can tell *whether or not event A occurred*: just check whether the outcome is in the set which represents A. And that sort of information should be enough for a probability calculation.

We should keep in mind that mathematical terminology steals words from ordinary language. If a mathematical term is well chosen, then its meaning in ordinary language will suggest its mathematical meaning, but one cannot simply rely on ordinary language to guess the exact mathematical definition. The word "event" in ordinary language usually suggests that something interesting has happened. In mathematical probability theory an event is just a set of outcomes.

The following is some standard terminology for situations where we want to systematically represent all possible outcomes for an experiment.

Definition 2.2 (Sample space models). The set of all possible abstract outcomes is called the *sample space*, often denoted by the uppercase Greek letter Ω ("Omega"). The abstract outcomes are the "points" making up the sample space, and they are traditionally called "sample points". A sample point is often denoted by the lowercase Greek letter ω ("omega").

The sample space is said to be a "model" for an experiment when its sample points can be interpreted as the possible outcomes of that experiment.

Certain subsets of the sample space will be referred to as "events", although of course they are mathematical objects rather than physical events. When we use the sample space as a model for an experiment, these subsets provide mathematical representations for actual physical events.

Any one-point set $\{\omega\}$ is an event in the model. It represents the event that the result of the experiment is the outcome represented by ω.

Since ω represents a possible result of an experiment, it would not be unreasonable to also say that ω itself is an event. However, since we are representing general events as *sets* of sample points, probably it's less confusing to stick to that, and use $\{\omega\}$ rather than ω when we are talking about events.

It should be emphasized that a sample space is a mental concept. It represents something about the real world, but only indirectly. Even a very large sample space has no weight!

When Ω is a sample space which represents an experiment, for any property that you can express in terms of the outcome of the experiment, there

is a corresponding set of sample points in Ω which represents that property. Conversely, when the sample space consists of a finite number of outcomes, every subset of the sample space can be interpreted as representing some physical event, although not necessarily an interesting one. When we study infinite sample spaces later, there will be subsets of the sample space that are not given a physical interpretation, and such sets will not be called events.

We will find later that the general properties of probability are often sufficient to solve a problem efficiently without committing our thoughts to any explicit choice of a sample space. On the other hand, if we cannot think of any sample space at all to represent the outcomes of a proposed experiment, it might be a good idea to investigate whether the experiment makes sense.

Here are some standard examples of experiments and corresponding sample spaces.

Example 2.3 (Tossing a coin once). The points of the sample space Ω should represent exactly two physical outcomes, the occurrence of a head, and the occurrence of a tail. So we can take $\Omega = \{1, 0\}$, a set with only two points. Here the point 1 represents the outcome in which a head is obtained, and the point 0 represents obtaining a tail. There are four events in the sample space Ω: $\{1, 0\}, \{1\}, \{0\}, \emptyset$, where $\emptyset$ denotes the empty set, i.e. the set with no members. The event $\{1, 0\}$ in the sample space describes a physical event which always happens. The empty set $\emptyset$ contains no sample points. There is no outcome which is a member of set, so this event never happens. But we will still call it an event.

Like all sample spaces, the set Ω is a thought in our heads, not something in the real world. We could use letters rather than numbers to represent sample points, so that "h" would mean a head was obtained and "t" would mean a tail. Then we would have $\Omega = \{\text{“h”}, \text{“t”}\}$. What matters is the *interpretation*. The interpretation associates one sample point with the result in which the coin toss gives a head, and the other sample point with the result in which the coin toss gives a tail.

For brevity, sometimes we'll refer to getting a head as "success", and getting a tail as "failure". Of course, the name doesn't matter, and we could switch, and call getting a tail "success", if we felt like it.

Example 2.4 (Rolling a die once). Much as in the case of a coin toss, we can take $\Omega = \{1, 2, 3, 4, 5, 6\}$, so that the outcome is simply the number obtained on the roll of the die.

There are 64 possible subsets of this sample space, and each subset is an event in the sample space which represents a physical event. For instance the event $\{2, 4, 6\}$ in the sample space represents the physical event that an even number was obtained.

Incidentally, we claimed that there are 64 possible events in the sample space for one roll of a die. Did that number make sense?

In general, it useful to know the following fact.

Lemma 2.5 (Number of subsets). Any set of size k has exactly 2^k subsets. (The empty set is one of the 2^k sets.)

Proof. We can build a subset by making a decision for each object in the set: "include" or "don't include". Thus we build a subset by making k decisions, each of which has two choices. This gives $\underbrace{2 \times \ldots \times 2}_{k \text{ factors}}$ ways to build the subset.

□

We can apply Lemma 2.5 to the sample space for rolling a die. In that case $k = 6$, and $2^k = 64$. Each of the 64 subsets is the mathematical representation of a possible event.

After considering tossing a coin once, we might consider tossing it n times, where n can be $1, 2, \ldots$.

Example 2.6 (Tossing a coin a million times). An outcome in the sample space for the experiment of tossing a coin one million times must record the result of each toss! Our choice for a sample point is a sequence $(x_1, \ldots, x_{1000000})$, where each x_i is either 1 or 0, and x_i tells what happened on the ith toss. We could use a similar sample space for tossing a coin n times, for any n.

Tossing a coin one million times would not be practical for an individual, but it would be perfectly feasible in an industrial setting. Notice however

that Ω contains $2^{1000000}$ sample points. (We did say that a sample space is a mental concept, rather than a real object, didn't we?) Every subset is an event in the sample space, and so there are $2^{2^{1000000}}$ possible abstract events! Each of these abstract events has a physical interpretation, though very very few of the details of such events are significant.

It may seem absurd to consider such a large sample space. Nevertheless, since we are able to reason precisely in an abstract setting, we are able to reliably establish useful facts.

Exercise 2.1. Consider the experiment of tossing a coin 400 times, and recording the result of all 400 tosses. This is not a very complicated experiment, and could easily be carried out by hand by one person.

You can use a sample space for this experiment similar to that in Example 2.6. Let N be the number of sample points in this sample space.

According to google, the number 10^{82} is likely an upper bound for the number of atoms in the observable universe. How does the number 10^{82} compare with N?

[Solution]

Example 2.7 (Drawing a card). A standard deck of playing cards consists of 52 distinct cards. There are four types of cards. The types are called "suits", and every card belongs to exactly one suit. Each suit has 13 cards, and the names of the suits are "spades", "hearts", "diamonds" and "clubs".

If the deck is *shuffled* a few times, cards become arranged in a fairly random order. Drawing the top card from the deck is equivalent to selecting one member of a population of 52 (with no member of the population favored). What would be a reasonable model for this sampling experiment?

We could certainly number the cards, in an arbitrary manner. A number in $\{1, \ldots, 52\}$ is then an abstract representation for a card, and we could build our model using these abstract "cards". Let's agree to call each number in $\{1, \ldots, 52\}$ an abstract outcome of the experiment of drawing a card.

Suppose that we are interested in the physical event A that a "heart" card is drawn. Since our abstract model contains an abstract outcome (a number label) representing each possible outcome, we can represent the event A as

the set of abstract outcomes that represent "heart" cards. Thus A contains 13 sample points.

We have already discussed experiments involving sampling from a bowl of jelly beans (Example 1.10), or from the population of a country. The reader will have no difficulty constructing appropriate sample spaces for these experiments, when needed.

2.2 Distributions and set-functions

Mathematical probability theory tells us how to reliably calculate new probabilities from given probabilities. Mathematics doesn't tell us how to get the probabilities that we start with. To quote the physicist E.T.Jaynes, " No matter how profound your mathematics is, if you hope to come out eventually with a probability distribution, then at some point you have to put in a probability distribution" ([5]). The probabilities we start with must somehow come from the physical description of an experiment.

Probability Assumption 2.1 (Existence of a distribution). When we work with a model, and represent events as subsets of a sample space, it is assumed that there is a mathematical probability $\mathbf{P}(A)$ for each event, although we may not know the value of every probability.

This probability $\mathbf{P}(A)$ is of course a function of A. Since the domain of $\mathbf{P}$ is made up of sets, one often speaks of $\mathbf{P}$ as a probability set-function.

Definition 2.8 (Probability models and probability terminology). A sample space, together with a given probability set-function, will be called a *probability model*.

Any rule which specifies probabilities can be called a distribution (Definition 1.11). So a probability set-function can also be called a probability distribution, and we frequently use that terminology.

Of course we often start analyzing a problem by thinking directly about probabilities for physical events connected with a particular experiment.

There need not be any sample space chosen at that stage, so the probability values $\mathbf{P}(A)$ are associated with the actual physical events A, or rather with our mental conceptions of them. In this case we would not think of $\mathbf{P}$ as a set-function, but one can still refer to the family of values $\mathbf{P}(A)$ as a distribution.

Let's pause for a moment to compare what we are doing here with our discussions in Chapter 1. In that chapter we talked about probability *facts*, namely the frequency interpretation and additivity. But in Probability Assumption 2.1, we are apparently starting to make *assumptions*. What happened? Did we lose the courage of our convictions?

Here's what's going on. In Chapter 1 we were talking about physical probability, for real-world situations. Now we are talking about abstract models, things which we can reason about mathematically. Our abstract models are indeed relevant to the real world, but only if we build in the correct mathematical assumptions. It is those assumptions that we are talking about here.

Remark 2.9 (Interpreting a model). A "model" in mathematics may represent an experiment, but Definition 2.8 doesn't say much by itself about the physical situation that a probability model represents. The connecting link between a probability model and the real world is the *interpretation* of the model, and the interpretation is not part of the mathematical definition. But we usually need to have at least a rough interpretation in mind to work successfully with a model.

In a *valid* interpretation of a probability model, the value of the probability for the abstract event A should be approximately equal to the probability of the physical event represented by A.

Making sure that a model is valid is ultimately a physical problem rather than a mathematical one, although mathematics may help us to test the validity of a model. When we discuss the applications of a mathematical probability model in this book, we will confidently assume that our model is a valid one. In the real world such confidence can be misplaced.

In this book we will study the general mathematical properties that probability models have, and then apply those properties when we use a probability model to represent an experiment. Some simple examples are given in

Sections 2.4 and 2.6. In later chapters we will deal with more complicated models. The same rules apply in all situations.

2.3 Events defined in terms of other events

Events in a mathematical model are represented by sets, and so relationships are often expressed using set language. Consequently, readers will need to know the basic terminology for set operations. This material is likely familiar, but Section 2.8 reviews all the concepts and notations which are needed. It's a good idea to look through that section, since notations and terminology for set operations can vary slightly.

Here are some set concepts and notations which are often used.

For sets $A_1, \ldots, A_k$, the *union* of $A_1, \ldots, A_k$ is the set consisting of every element which is a member of *at least one* of the sets $A_1, \ldots, A_k$. This set is denoted by $A_1 \cup \ldots \cup A_n$. When $A_1, \ldots, A_n$ are events, the union $A_1 \cup \ldots \cup A_n$ represents the event that at least one of the events $A_1, \ldots, A_n$ occurred.

For sets $A_1, \ldots, A_k$, the *intersection* of $A_1, \ldots, A_k$ is the set consisting of every element which is a member of *all* of the sets $A_1, \ldots, A_k$. This set is denoted by $A_1 \cap \ldots \cap A_n$. When $A_1, \ldots, A_n$ are events, the intersection $A_1 \cap \ldots \cap A_n$ represents the event that every one of the events $A_1, \ldots, A_n$ occurred.

We often consider situations in which some events $A_1, \ldots, A_k$ are *mutually exclusive*, meaning that *at most one* of these events can occur. In that case no sample point can be a member of more than one of the sets $A_1, \ldots, A_n$, and we say that these sets are *disjoint*.

The *empty* set is denoted by $\emptyset$. Note that the definition of disjointness implies that for any set A, the sets A and $\emptyset$ are disjoint.

For any sets A, B, the *set difference* $B - A$ is the set of all elements which are members of B but not A. And in situations where all sets are subsets of some fixed set U, it is convenient to write $U - A$ as A^c. The set A^c is referred to as the *complement* of A. If A is defined by some property, notice that A^c is the set of all elements in U which do *not* have this property.

We often denote of elements in a finite set S by $|S|$. If a set if not finite we say it is infinite, and say that $|S| = \infty$.

See Section 2.8 for more discussion of sets.

Probability Assumption 2.2 (Set operations and sample space events). If $A_1, \ldots, A_k$ are events in a sample space, then so are $A_1 \cup \ldots \cup A_k$ and

$A_1 \cap \ldots \cap A_k$. If A and B are events in the sample space, then so are $A - B$ and A^c.

To justify this assumption, recall that events in the sample space correspond to meaningful statements about the physical result of an experiment.

If we think that given statements $\alpha_1, \ldots, \alpha_k$ are meaningful, then surely we must also think that the statement "at least one of the statements $\alpha_1, \ldots, \alpha_k$ holds" is meaningful, and "all of the statements $\alpha_1 \ldots \alpha_k$ hold" is meaningful.

It is also meaningful to say "α_1 is true and α_2 is not true".

Translating such observations into set language gives us Probability Assumption 2.2.

Remember Probability Fact 1.2 , which dealt with adding probabilities of mutually exclusive events. Suppose now that we are given events $D_1, \ldots, D_k$ which happen to be disjoint subsets of a sample space. Then there is no outcome ω which is a member of more than one of these sets. Whatever properties these events describe must therefore be *mutually exclusive*. Thus we can rephrase Probability Fact 1.2 using set notation as follows.

Probability Assumption 2.3 (Additivity of probability). Let $D_1, \ldots, D_k$ be *disjoint* events in some probability model. Then

$$\mathbf{P}(D_1 \cup \ldots \cup D_k) = \mathbf{P}(D_1) + \ldots + \mathbf{P}(D_k). \tag{2.1}$$

Also, probabilities in the model are such that

$$\mathbf{P}(\Omega) = 1. \tag{2.2}$$

If we think think of a probability simply as a number that measures degree of belief, we could scale all our probability values up or down by a factor, without changing their usefulness. Since Ω represents an event that always happens, equation (2.2) tells us that we are using a belief scale for which certainty is 1. Of course this scale fits the statement of the frequency interpretation, so it is the natural scale for probability.

Remark 2.10. If $D_1, \ldots, D_k$ are disjoint events in some probability model, and $D_1 \cup \ldots \cup D_k = \Omega$, Probability Assumption 2.3 implies that

$$\mathbf{P}(D_1) + \ldots + \mathbf{P}(D_k) = 1. \tag{2.3}$$

Thus Probability Assumption 2.3 includes the abstract version of equation (1.4).

Remark 2.11 (Set notations without sets!). If A and B represent physical events, we may still use set notation to describe combinations of these events, even when we are not representing A and B as sets. For example, the event that A occurs *and* B occurs will still be expressed as $A \cap B$.

This convention can be justified in two ways. First, it is a convenient brief notation. Second, for any experiment, one *could* define some sample space model to represent the experiment. In that case the event that both A and B occur would indeed be represented by the intersection of two sets in the sample space.

Many examples in probability theory involve experiments which only have a finite number of possible outcomes. Each possible outcome is represented by a sample point ω in the sample space. As we noted in Definition 2.2, the event $\{\omega\}$ is a particularly simple event, since it is a one-point set. We can make some formulas a bit neater by introducing the following special notation for the probability of a one-point set.

Definition 2.12 (Probability mass functions). For any probability model, we can optionally write $\mathbf{P}(\{\omega\})$ as $\boldsymbol{p}(\omega)$, for brevity.

The function $\boldsymbol{p}$ is referred to as as a *probability mass function*, or more briefly as a probability function.

The word "mass" in the name "probability mass function" is intended to suggest a lump of probability attached to each sample point. Theorem 2.13 states that the probability of an event can be pictured as the sum of the masses of all the sample points in the event.

We include a proof of the next theorem, Theorem 2.13, to emphasize that this rule follows from the assumptions that we have already made: additivity, and the fact that one-point sets are events.

Theorem 2.13 (Finite events). Let Ω and $\mathbf{P}$ be a probability model. If A is a finite set of sample points, then A is an event, and

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}(\{\omega\}) = \sum_{\omega \in A} \boldsymbol{p}(\omega). \tag{2.4}$$

Proof. By the definition of union,

$$A = \bigcup_{\omega \in A} \{\omega\}. \tag{2.5}$$

The sets $\{\omega\}$ in equation (2.5) are obviously disjoint. Applying the additivity of probability to equation (2.5) then gives equation (2.4).

□

In the proof just given, if equation (2.5) does not seem clear, please check a concrete example. For example, show from the definition of union that

$$\{1, 2\} = \{1\} \cup \{2\}.$$

Equations (2.2) and (2.4) of course tell us that when Ω is a finite set,

$$\sum_{\omega \in \Omega} \boldsymbol{p}(\omega) = \mathbf{P}(\Omega) = 1. \tag{2.6}$$

When setting up a probability model with a finite sample space, if we can decide on the value of $\mathbf{P}(\{\omega\})$ for each sample point ω, then (by equation (2.4)) all other probabilities $\mathbf{P}(A)$ are determined. So a simple probability model is usually defined by listing the probabilities of the outcomes.

2.4 Some basic examples

Example 2.14 (Probabilities for a single coin toss). There are only two possible outcomes. As in Example 2.3, we can choose to represent these outcomes by 0 and 1. The outcome is 1 if a head is obtained, and the outcome is 0 if a tail is obtained. Thus the sample space Ω is given by $\Omega = \{0, 1\}$.

By equation (2.3), $\mathbf{P}(\{1\}) + \mathbf{P}(\{0\}) = 1$. Using the notation of Definition 2.12, this says that $\boldsymbol{p}(1) + \boldsymbol{p}(0) = 1$.

If the probability of a head is p and the probability of a tail is q, then $p + q = 1$. For a *fair* coin, $p = q = 1/2$.

Example 2.15 (Probabilities for a single roll of a die). We take $\Omega = \{1, 2, 3, 4, 5, 6\}$, with the same interpretations as Example 2.4.

If the die is fair, then the probability of each possible outcome is the same, so $\mathbf{P}(\{1\}) = \mathbf{P}(\{2\}) = \mathbf{P}(\{3\}) = \mathbf{P}(\{4\}) = \mathbf{P}(\{5\}) = \mathbf{P}(\{5\})$. Using our probability mass function notation, this says that $\boldsymbol{p}(1) = \boldsymbol{p}(2) = \boldsymbol{p}(3) = \boldsymbol{p}(4) = \boldsymbol{p}(5) = \boldsymbol{p}(6)$.

By equation (2.3), $\boldsymbol{p}(1) + \boldsymbol{p}(2) + \boldsymbol{p}(3) + \boldsymbol{p}(4) + \boldsymbol{p}(5) + \boldsymbol{p}(6) = 1$.

Thus in the fair case $\boldsymbol{p}(i) = 1/6$ for each i.

Exercise 2.2. Suppose that $\Omega = \{1, 2, 3, 4, 5, 6\}$, and assume that $\mathbf{P}(\{\omega\}) = 1/6$ for each $\omega \in \Omega$. Let $A = \{2, 4, 6\}$, so that A represents the physical event that an even number is obtained. Show from the definitions that $\mathbf{P}(A) = 1/2$.

[Solution]

2.5 Symmetry in probability

In games, we generally try to use a fair coin.

The true test of fairness is to toss the coin a large number of times, and see if we obtain approximately the same fraction of heads and tails. If we can't do that, we can at least examine the coin carefully, to see if there is anything about the physical properties of the coin which would favor heads

or tails. If the physical properties of the coin seem similar when viewed from either side, we would say that the coin is *symmetric* with respect to heads and tails. Since there is nothing that would lead us to assign a higher probability to one side over the other, it seems reasonable to assign equal probability to each of the two possible outcomes.

Symmetry in probability calculations has been used for a long time, and in the old days it was sometimes described as "the Principle of Indifference", or "the Principle of Insufficient Reason". This principle says we should assign equal probabilities to possible outcomes if we have no positive reason to do otherwise. We already used a somewhat similar approach in Section 1.8.

The use of symmetry is dangerous if it is based on ignorance. For example, suppose you decide to gamble with someone who is tossing a coin, and you know very little about the person and the coin. As a believer in the Principle of Insufficient Reason, you may feel you have no choice but to assign a probability of 1/2 to the occurrence of a head. If your new friend obtains five tails in a row you may regret this probability assignment.

More generally, even if you make a careful examination of the setting of an experiment, you may overlook some factor. Then the setting of the experiment may be less symmetrical than you think.

Of course, in a real-life situation, you need not stick to your original assumptions, when new information starts to come in. Chapter 4 (on conditional probability) deals with rules for updating probability assessments, when you obtain additional information.

2.6 More examples

Example 2.16 (Probabilities for tossing a fair coin twice). Consider the experiment of tossing a coin twice. As in Example 2.3, we think of 1 as representing a head, and zero as representing a tail. The result of each toss is represented that way, and there are two tosses, so we take every sample point to be an ordered pair of numbers, each of which is either one or zero. The first number represents the result of the first toss, and the second number represents the result of the second toss.

There are two choices for the first number, and two for the second number, so there are four sample points, and $\Omega = \{(1,1),(1,0),(0,1),(0,0)\}$. Our interpretation is that $(1,1)$ represents obtaining two heads, $(1,0)$ represents

getting a head followed by a tail, $(0, 1)$ represents getting a tail and then a head, and $(0, 0)$ represents getting two tails.

We need to find $\boldsymbol{p}(\,(1, 1)\,)$, $\boldsymbol{p}(\,(1, 0)\,)$, $\boldsymbol{p}(\,(0, 1)\,)$, $\boldsymbol{p}(\,(0, 0)\,)$.

It will be easy to find these probability values, once we introduce the general concept of *independence* (Section 5.1). If you have used independence in any previous study of probability, you must be impatient to use it here! But for the moment we'll just consider the fair case, and calculate probabilities based on an extra assumption: that all outcomes should be equally likely.

By equation (2.6), the four outcome probabilities should add to one, and so

$$\boldsymbol{p}(\,(1, 1)\,) = \boldsymbol{p}(\,(1, 0)\,) = \boldsymbol{p}(\,(0, 1)\,) = \boldsymbol{p}(\,(0, 0)\,) = \frac{1}{4}.$$

Exercise 2.3. In the two-toss experiment of Example 2.16, when the coin is fair, use the four-point sample space Ω to calculate the probability that the same result is obtained on both coin tosses.

[Solution]

Exercise 2.4. In the two-toss experiment of Example 2.16, when the coin is fair, use the four-point sample space Ω to calculate the probability that the *first toss* produces a head.

You know the answer already, but we are checking here that the sample space for two tosses is consistent with the sample space for one toss.

Exercise 5.3 will show that the same result holds when we model tossing a general coin, one which is not necessarily fair.

[Solution]

Exercise 2.4 is an example of what we do when getting familiar with a new tool. We check that it works properly!

Exercise 2.5 (First toss of a million). In Exercise 2.4 you considered finding the probability of success on one toss of a fair coin, when using the model for two tossses. To no one's suprise, the model for two tosses agrees with the model for one toss.

How about using the model for tossing a fair coin a million times, as in Example 2.6? That has to work too, doesn't it? But you will check that now.

You are only allowed to work with the big sample space. Any event you consider must be a subset of that space, which has $2^{1000000}$ points.

Just as in the case of tossing a fair coin twice (Example 2.16), we will assume that all sample points have the same probability. And that probability is ....

Let A be the event that the very first toss of the coin results in a head. Using the big sample space, find $\mathbf{P}(A)$.

(And yes, we will rerun this problem in Exercise 7.9 for the case of a coin which might be unfair. That works too.)

[Solution]

Example 2.17 (Probabilities for two rolls of a fair die). Just as we can toss a coin twice, we can roll a die twice, or roll two different dice at the same time. The sample space is larger, but the principle is the same. $\Omega = \{(i,j) : \ i = 1, \ldots, 6, \ j = 1, \ldots 6\}$. There are 36 sample points.

Assume that the die is fair. We would like to know the probability distribution on this sample space. We are willing to assume that all sample points of Ω have the same probability p.

By equation (2.6), $36p = 1$. Hence $\boldsymbol{p}(\,(i,j)\,) = 1/36$ for all i, j

In problems involving experiments with two steps, it is often helpful to list the sample points in a table. Let the row indices refer to the first element in a pair and column indices refer to the second element in a pair. Then we list Ω as:

	1	**2**	**3**	**4**	**5**	**6**
1	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
2	$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$
3	$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$
4	$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$	$(4,5)$	$(4,6)$
5	$(5,1)$	$(5,2)$	$(5,3)$	$(5,4)$	$(5,5)$	$(5,6)$
6	$(6,1)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$	$(6,6)$

(2.7)

We will revisit this experiment after introducing the concept of independence (Chapter 5).

One is often interested in the sum of the scores on the two dice. Let A_k be the event that the sum of the numbers obtained on the two rolls is equal to k.

The largest possible sum is 12. So we see that A_k is empty for $k > 12$.

The smallest possible sum is 2. So A_1 is empty.

To find $\mathbf{P}(A_k)$, we need to count the number of outcomes in (x_1, x_2) in A_k, for each k with $2 \leq k \leq 12$. Each outcome has probability $1/36$, and we add these probabilities, as usual.

$$
\begin{aligned}
A_2 &= \{(1,1)\}\,,\ \mathbf{P}\,(A_2) = \frac{1}{36},\\
A_3 &= \{(1,2),(2,1)\}\,,\ \mathbf{P}\,(A_3) = \frac{2}{36},\\
A_4 &= \{(1,3),(2,2),(3,1)\}\,,\ \mathbf{P}\,(A_4) = \frac{3}{36},\\
A_5 &= \{(1,4),(2,3),(3,2),(4,1)\}\,,\ \mathbf{P}\,(A_5) = \frac{4}{36},\\
A_6 &= \{(1,5),(2,4),(3,3),(4,2),(4,1)\}\,,\ \mathbf{P}\,(A_6) = \frac{5}{36},\\
A_7 &= \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}\,,\ \mathbf{P}\,(A_7) = \frac{6}{36},\\
A_8 &= \{(2,6),(3,5),(4,4),(5,3),(6,2)\}\,,\ \mathbf{P}\,(A_8) = \frac{5}{36},\\
A_9 &= \{(3,6),(4,5),(5,4),(6,3)\}\,,\ \mathbf{P}\,(A_9) = \frac{4}{36},\\
A_{10} &= \{(4,6),(5,5),(6,4)\}\,,\ \mathbf{P}\,(A_{10}) = \frac{3}{36},\\
A_{11} &= \{(5,6),(6,5)\}\,,\ \mathbf{P}\,(A_{11}) = \frac{2}{36},\\
A_{12} &= \{(6,6)\}\,,\ \mathbf{P}\,(A_{12}) = \frac{1}{36}
\end{aligned}
\tag{2.8}
$$

Exercise 2.6. In the experiment of Example 2.17, let A be the event that the first roll produces the number 5.

Find $\mathbf{P}(A)$, using the sample space Ω of Example 2.17.

You are checking that the two-roll model is consistent with the one-roll model. And yes, yes, again this is obvious physically. We are just testing for bugs in our mathematical machinery.

[Solution]

Exercise 2.7. Consider the experiment of rolling a fair die twice.

Find the probability that the first roll produces an even number *and* the second roll produces a number larger than four.

As in Example 2.17, let Ω consists of the pairs (x_1, x_2), where $x_i = 1, \ldots, 6$ and $x_2 = 1, \ldots, 6$.

We will return to this problem in Exercise 5.5.

[Solution]

Exercise 2.8. Again consider the experiment of rolling a fair die twice. Find the probability that the sum of the numbers obtained on the two rolls is less than or equal to 5.

[Solution]

Exercise 2.9. When rolling a fair die twice, let C be the event that the sum of the numbers obtained on the two rolls is an even number.

Find $\mathbf{P}(C)$.

Let D be the event that that sum of the numbers obtained on the two rolls is larger than 6.

Find $\mathbf{P}(D)$ and $\mathbf{P}(C \cap D)$.

[Solution]

Example 2.18 (Probability of drawing a card from a deck). This is the experiment defined in Example 2.7. We said that drawing the top card from a deck is equivalent to selecting a member of a population of 52, with no member of the population favored. Thus each card has same probability to be drawn, and we know these probabilities sum to one. Hence each card has the probability 1/52 to be drawn.

Sometimes people think about *dealing* cards, which means removing cards repeatedly, starting from the top of the deck. The deck is *shuffled* before dealing, to arrange the cards of the deck in random order. Thus the third card dealt from the deck is a random sample from the deck, just as the top card is a random sample. And so the probability that any particular card will be the third card dealt is exactly the same as the probability that it is the first card dealt, 1/52 in both cases.

We can picture the process more concretely if we think of randomly laying out all 52 cards face down on a table, forming a long row. Instead of drawing the top card from the deck, one might think of turning over the first card in the row. The third card drawn is the third card that we turn over, and so on. The probability that a particular card is in the first position is clearly the same as the probability that it is in the third position.

Exercise 2.10. Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ be a sample space with associated probability mass function p. Suppose that $\boldsymbol{p}(\omega_2) = 2\boldsymbol{p}(\omega_1)$, $\boldsymbol{p}(\omega_3) = 3\boldsymbol{p}(\omega_2)$, $\boldsymbol{p}(\omega_4) = 4\boldsymbol{p}(\omega_3)$, $\boldsymbol{p}(\omega_5) = 5\boldsymbol{p}(\omega_4)$. Find $\boldsymbol{p}(\omega_3)$.

[Solution]

Exercise 2.11. A certain combination lock will only open when the correct code is entered. The code consists of 4 digits in order. The allowable digits are $0, \ldots, 9$. A stranger who does not know the correct code attempts to open the lock by entering 4 arbitrarily chosen digits. Find the probability that the lock opens. Express your reasoning in terms of an appropriate sample space and a probability mass function. If it seems appropriate with your model, you may assume that all sample points are equally probable.

[Solution]

Exercise 2.12. An experiment consist of tossing a certain coin six times, and counting the number of heads which are obtained. If we regard the outcome of the experiment to be the number of heads which are obtained, then an appropriate sample space for this experiment is $\Omega = \{0, 1, 2, 3, 4, 5, 6\}$. This

sample space is only adequate for *listing* the outcomes. It is definitely not an adequate sample space for *computing* probabilities of outcomes.

Suppose that the coin which is used in this experiment is unfair, and actually the probability of a head on each toss is 1/3. We will show later that the correct probability mass function $\boldsymbol{p}$ for Ω is given by

$$\boldsymbol{p}(j) = \binom{6}{j} \left(\frac{1}{3}\right)^j \left(\frac{2}{3}\right)^{6-j},$$

where $\binom{6}{j}$ is the *binomial coefficient* given by

$$\binom{6}{j} = \frac{6!}{j!(6-j)!}.$$

(We will not use this particular sample space Ω when we *derive* this formula. This sample space is too simple to represent what is going on in the experiment, which involves a number of steps.)

As a small test of whether our formula for $\boldsymbol{p}$ is correct, use the binomial theorem (if you happen to know it) to verify that the values of $\boldsymbol{p}$ sum to one. If you haven't met the binomial theorem before, omit this problem. And do not worry, the binomial theorem is derived in Section 8.2.

[Solution]

Exercise 2.13 (The number wheel experiment). At a booth in a fairground, we find a large wheel marked with the numbers from 0 to 100. By spinning the wheel, and seeing where it stops, a random number is chosen. This will be considered as the outcome of an experiment.

(a) Provide a suitable sample space for this experiment. Assume that each outcome has equal probability, and find the probability mass function.

(b) Answer the following questions.

(i) What is the probability that the number is 3?

(ii) What is the probability that the number is even?

(iii) Let A be the event that the number is smaller than 20, and let B be the event that the number is larger than 60. Find $\mathbf{P}(A \cup B)$.

(iv) What is the probability that the number is less than 50 and is divisible by 3? Remember that zero is divisible by any number.

[Solution]

Exercise 2.14 (Probability of a complement). The following obvious consequence of additivity is often surprisingly useful. Let A be an event for some experiment. Prove that

$$\mathbf{P}\left(A^c\right) = 1 - \mathbf{P}(A). \tag{2.9}$$

[Solution]

Exercise 2.15. In Exercise 2.11, find the probability that the lock does not open.

[Solution]

Exercise 2.16. Using the probability model in Exercise 2.12, find the probability that at least one head is obtained in the six tosses.

[Solution]

Exercise 2.17. Let A, B be any events. Show that A and $B - A$ are disjoint, and

$$B = (A \cap B) \cup (B - A), \tag{2.10}$$

and so by additivity,

$$\mathbf{P}(B) = \mathbf{P}(A \cap B) + \mathbf{P}(B - A). \tag{2.11}$$

Thus

$$\mathbf{P}(B - A) = \mathbf{P}(B) - \mathbf{P}(A \cap B). \tag{2.12}$$

See Figure 2.1.

[Solution]

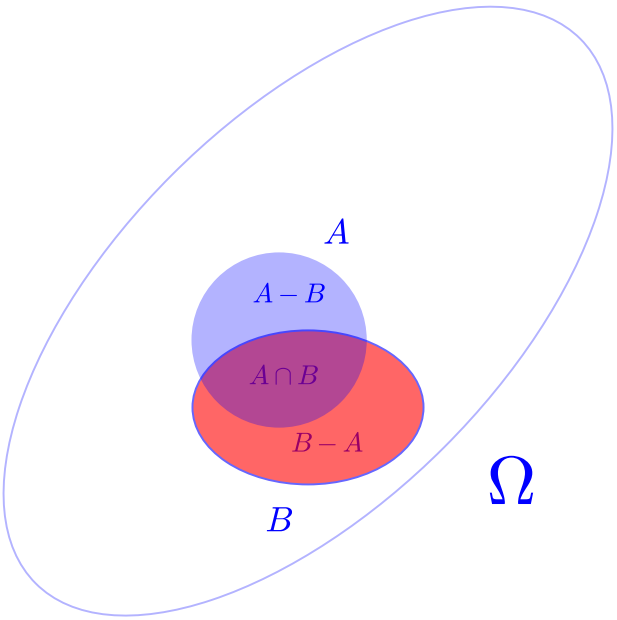


Figure 2.1: Exercise 2.17: $B = (A \cap B) \cup (B - A)$. $B - A$ is red, $A - B$ is blue, and $A \cap B$ is purple. $A = (A \cap B) \cup (A - B)$.

Example 2.19 (Choosing a positive integer). Suppose someone says to you: "Think of a number, any number." Probably they mean that you should choose a positive integer, and they don't want you to favor any particular number. Strictly speaking, this is impossible! To check that, consider the following argument.

Let p_k is the probability that you choose k. Assume that p_k is the same for all k. Let c be the value of p_k.

By additivity, the probability that you choose a number less than or equal to n is exactly

$$p_1 + \ldots + p_n.$$

Any probability is less than or equal to one, so you must have

$$p_1 + \ldots + p_n \leq 1.$$

If it is really true that $p_k = c$ for all k, then

$$nc \leq 1, \text{ i.e. } c \leq \frac{1}{n}.$$

This can only be true for all n if $c = 0$. But then the probability of choosing a number less than or equal to n is zero, for every n. So the chance that you choose a number less than a million is zero, and the chance that you choose a number less than a trillion trillion is also zero, and so on. So it seems you will stand silently. No one will want to play this game with you!

In real life, if someone asks you to think of a number, you will probably not have a precise recipe in mind, but you likely have a finite range of possible numbers in mind, and try to choose one of them without being too predictable.

Here is one more fact that is often useful.

Lemma 2.20 (Monotonicity). If A_1 and A_2 are events,

$$A_1 \subset A_2 \implies \mathbf{P}(A_1) \leq \mathbf{P}(A_2). \tag{2.13}$$

Here we use $\implies$ to mean "implies".

In words, we can say that probability is *monotone increasing* as a function of events, i.e. bigger sets give bigger probabilities. No surprise here, and that's good!

Proof. From the definitions, $A_2 = A_1 \cup (A_2 - A_1)$, and the sets $A_1, A_2 - A_1$ are disjoint (see Figure 2.2).

Hence $\mathbf{P}(A_2) = \mathbf{P}(A_1) + \mathbf{P}(A_2 - A_1)$. □

Notice the technique in this proof. We broke sets up into disjoint pieces, and then used additivity. This is a general trick. It is used, for example, in the proof of equation 2.14.

The next exercise tells us that if an event has probability one, then we might as well think of that event as being the whole same space, since it includes everything that has a chance of happening.

Exercise 2.18 (Probability one includes essentially everything). Suppose that $\mathbf{P}(A) = 1$.

For any event B, prove the following statements.

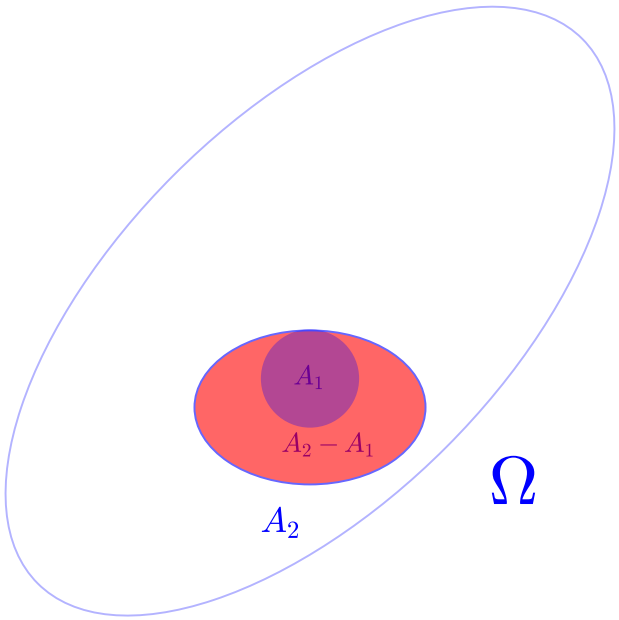


Figure 2.2: Lemma 2.20: $A_2 = A_1 \cup (A_2 - A_1)$. A_1 is purple, $A_2 - A_1$ is red.

(i) $\mathbf{P}(B - A) = 0$.

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

Lemma 2.20 is useful for writing down the argument. Figure 2.1 shows the general relation between the events, but explain why under the assumptions of this exercise, you have $\mathbf{P}(B - A) = 0$.

[Solution]

Definition 2.21 (Uniform distribution on a finite set). When a probability model uses a finite sample space and assigns the same probability to every sample point, we will refer to this assignment of probabilities as a uniform distribution on the finite set Ω.

Theorem 2.22 (Fair sampling). Let S be a set of n objects, and let T be a subset of S containing j objects. Suppose an object is chosen randomly from S, in such a way that all objects in S are treated the same way by the selection process. Then the probability that the chosen object is a member of T is j/n.

Proof. The simplest choice for a sample space is $\Omega = S$.

Then the set T is also the abstract representation of the event that the selected object lies in T.

We want to show that $\mathbf{P}(T) = j/n$.

We are told that there is symmetry in the selection process: all objects are treated in exactly the same way. Hence $\mathbf{P}(\{\omega\})$ is the same for all $\omega \in \Omega$. Let's call this number p.

Since $\mathbf{P}(\Omega) = 1$, we know by additivity that

$$\sum_{\omega \in \Omega} \boldsymbol{p}(\omega) = 1.$$

Hence $np = 1$, so $p = 1/n$.

Using additivity again,

$$\mathbf{P}(T) = \sum_{\omega \in T} \boldsymbol{p}(\omega) = jp = \frac{j}{n}.$$

□

It should be emphasized that Theorem 2.22 is not a surprising fact. If there are n ways that something can happen, and if j of those ways are "good", and all ways seem equally likely, we would naturally think that the likelihood of a good result depends on how big j is, compared to n. So j/n is the value we expect for the probability of a good result.

With that in mind, the proof of Theorem 2.22 can be thought of as yet another test of the theory of probability. The general theory gives the answer we expect.

Exercise 2.19. Theorem 2.22 deals with the situation of Example 1.10. So let's review jelly bean selection.

Consider picking a jelly bean randomly from a bowl. Suppose that there are 75 yellow beans, 53 red beans, 27 purple beans, and 18 green beans in the bowl. Find the probability that the selected jelly bean is red.

[Solution]

Exercise 2.20.

(a) A box contains 25 white marbles and 13 blue marbles. Our experiment consists of randomly selecting one marble. We assume that each marble has the same probability of being selected. What is the probability that the selected marble is blue?

(b) Now we prepare a new experiment, which we will call experiment 2. We replace every blue marble in the box by 10 blue marbles, and we replace every white marble in the box by 10 white marbles. The actual procedure for experiment 2 is the same as before: randomly select one marble, in such a way that every marble has the same chance of being selected. What is the probability that the selected marble is blue?

[Solution]

Example 2.23 (Choosing two beans). Return to the setting of Exercise 2.19. A new experiment in this setting consists of randomly selecting two jelly beans. If the chooser is planning to eat the jelly beans, it seems clear that the precise manner in which the beans are extracted from the bowl does not matter. The outcome here should be the *set* of two beans that is selected.

Let A be the event that a red bean and a green bean are selected. We would like to find $\mathbf{P}(A)$.

No jelly bean is favored in the choosing, so any set of two beans has the same chance of being selected. This is the crucial fact, since it allows us to use Theorem 2.22.

Using the chosen set as the sample point, Theorem 2.22 tells us that

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

where as usual we denote the number of elements in a set S by $|S|$.

However, we still need to calculate $|A|$ and $|\Omega|$. The necessary formula is given later, in equation 8.3. But we don't have to wait until we get to that equation. The idea that is used in deriving equation 8.3 can already be used right here: we will think about selecting the jelly beans one at a time.

This may seem unnecessarily complicated, since when we eat the two jelly beans we don't care which one was chosen first. However, it seems to be a good way to calculate the probability that a particular set of two jelly beans is chosen.

Notice that by thinking about getting the jelly beans one at a time we have modified our experiment. Now it is a *two-step* experiment. We must define a new sample space Ω. Now a sample point ω is not a set of two jelly beans from the bowl, it is an *ordered pair* (b_1, b_2), where b_1 represents the first jelly bean chosen, and b_2 represents the second.

A key point: We definitely want to eat *two* jelly beans, so we only allow sample points with $b_2 \neq b_1$. That is, after the first bean is selected, it is removed from the bowl, and is no longer available for the second selection.

No jelly bean is favored, so again all sample points are equally likely. Using Theorem 2.22 in this new sample space, we have

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}.$$

What is the event A using this sample space? A is the set of all pairs (b_1, b_2), where either b_1 is red and b_2 is green, or else b_1 is green and b_2 is red.

There are 53×18 ways of choosing a red bean and then a green. There are 18×53 ways of choosing a green bean and then a red. Thus $|A| = 2 \times (53 \times 18)$.

We find $|\Omega|$ in much the same way. The total number of beans is $75 + 53 + 27 + 18 = 173$. Hence there are 173 ways to choose the first jelly bean. Having chosen the first bean, there are then 172 ways to choose the second jelly bean.

Notice that the choice of first bean determines the available choices for the second bean, but the *number* of choices for the second bean is always the same, and does not depend on what the first bean was.

Combining our facts,

$$\mathbf{P}(A) = \frac{2 \times 53 \times 18}{173 \times 172}.$$

Exercise 2.21 (Choosing two red beans). In the setting of Example 2.23, let R be the event that both chosen beans are red. Find $\mathbf{P}(R)$.
[Solution]

2.7 Beyond additivity

It is useful to say something about probabilities for unions which are *not* disjoint!

Theorem 2.24 (Inclusion-Exclusion formula). Let A and B be any events, and let $\mathbf{P}$ be a probability set-function. Then

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B). \tag{2.14}$$

The reason for the name of this formula will be evident from the proof.

Proof. See Figure 2.1 for the general relation between the events $A, B, A \cap B, B - A$. $A \cup B$ consists of all the colored regions in the figures. $A - B$ is blue, $A \cap B$ is purple, and $B - A$ is red.

If an outcome is in $A \cup B$, and it is not in both events, then either the outcome is in A but not in B, or else the outcome is in B but not in A. It follows that $A \cup B$ is the disjoint union of $A \cap B$, $A - B$, or $B - A$. By additivity,

$$\mathbf{P}(A \cup B) = \mathbf{P}(A - B) + \mathbf{P}(B - A) + \mathbf{P}(A \cap B). \tag{2.15}$$

Similar arguments show even more easily that $\mathbf{P}(A) = \mathbf{P}(A - B) + \mathbf{P}(A \cap B)$ and $\mathbf{P}(B) = \mathbf{P}(B - A) + \mathbf{P}(A \cap B)$. This is also clear from Figure 2.1. Adding these two equations gives

$$\mathbf{P}(A) + \mathbf{P}(B) = \mathbf{P}(A - B) + 2\mathbf{P}(A \cap B) + \mathbf{P}(B - A)$$

Comparing this equation to equation (2.15) gives equation (2.14).

□

Relating the proof of Theorem 2.24 to the name of the formula, note that $A \cap B$ is the set of outcomes which are *included* in both A and B, while $A - B$ is the set of outcomes we obtain from A when we *exclude* the outcomes in B.

In the proof of Theorem 2.24 we used the technique of breaking up non-disjoint sets into disjoint pieces. This is often a useful procedure. We have seen it already in the proof of Lemma 2.20.

In the statement of Theorem 2.24, consider the special case of finite sets for which all the outcomes have the same probability. Remember that in this situation we find the probability of any event by simply counting the number of outcomes in the event, and then multiplying by the probability of a single outcome. We can then say that the term $-\mathbf{P}(A \cap B)$ in (2.14) compensates for "double-counting" outcomes, since any outcome in $A \cap B$ contributes both to $\mathbf{P}(A)$ and to $\mathbf{P}(B)$.

When events are not disjoint, we don't have additivity, but we still have an *inequality*, as the next theorem shows. Please work Exercise 2.22 after reading the next theorem.

Theorem 2.25 (Subadditivity property). Let $A_1, \ldots, A_k$ be any events for some probability model. Then

$$\mathbf{P}\left(A_1 \cup \ldots \cup A_k\right) \leq \sum_{j=1}^{k} \mathbf{P}\left(A_j\right). \tag{2.16}$$

Proof. Consider the case $k = 2$. Let $A = A_1$, $B = A_2$. Equation (2.16) follows at once from (2.14).

This proves the theorem for $k = 2$.

The statement is obviously true for $k = 1$. (Right?)

A proof by induction for the case $k > 2$ is left to the reader, in Exercise 2.23.

□

The title for Theorem 2.25 uses the word "subadditivity". Since one of the meanings of "sub" is "below", subadditivity seems like a suitable name for property expressed in equation (2.16). This inequality says that the probability of a union of events is never greater than the sum of the separate probabilities.

Exercise 2.22 (Obtaining an estimate). In Exercise 2.1 on sample spaces, we considered the experiment of tossing a fair coin 400 times.

(a) We accept that for each sample point ω, $\mathbf{P}(\{\omega\})$ is exactly the same. What is this probability value?

(b) Consider the event A of ever obtaining 20 heads in succession during these 400 tosses. By definition, A occurs if there is any index k such that you get a head on toss k, toss $k+1$, toss $k+2$, ..., toss $k+19$.

Does this event feel likely or unlikely?

(c) Use subadditivity to find an estimate for the probability of A, and decide whether A is likely or unlikely.

[Solution]

Exercise 2.23 (The Old Induction Trick). Prove the case $k > 2$ of Theorem 2.25. (This sort of argument, passing from $k = 2$ to general k, is useful in many situations. If you haven't seen it before it is worth working through.)

[Solution]

2.8 A review of set operations

Since we represent physical events by sets of abstract outcomes in a sample space, set operations will play a basic role.

This section contains definitions and notations for all the standard set operations. Readers can quickly skim through it, and then refer back again as needed. Notations and terminology for sets can differ slightly, so even an experienced reader might benefit from a quick survey. You might want to recall something that J.R.R. Tolkien said about hobbits: "they liked to have books filled with things that they already knew, set out fair and square with no contradictions". We are entering hobbit-mode now.

Here we go. The members of a set can be called "elements" of the set, or "points" of the set, although of course such points need not have any geometrical meaning. Sometimes we may list the contents of a set as a

sequence. The order in which the contents are listed is irrelevant, since sets are not ordered. The words "set" and "collection" have the same meaning throughout this book. Use of the word "collection" makes it possible to avoid too many repetitions of the word "set". For example if we happen to be dealing with a set of sets, we would tend to use the phrase "collection of sets" rather than "set of sets".

Definition 2.26 (Unions of sets). Let $A_1, \ldots, A_k$ be any sets. The union of $A_1, \ldots, A_k$ is the set consisting of every element which is in at least one of the sets $A_1, \ldots, A_k$. We can write the union of two sets A_1, A_2 in symbols as $A_1 \cup A_2$, and the union of $A_1, \ldots, A_k$ as $A_1 \cup \ldots \cup A_k$.

It is easy to check from the definition that $A \cup B = B \cup A$, or in other words that union is a *commutative* operation. It is also easy to check that $A \cup (B \cup C) = A \cup B \cup C = (A \cup B) \cup C$, so that union is an *associative* operation.

Definition 2.27 (Intersections of sets). Let $A_1, \ldots, A_k$ be any sets. The intersection of $A_1, \ldots, A_k$ is the set consisting of those elements which are in every one of $A_1, \ldots, A_k$. We can write the intersection of two sets A_1, A_2 in symbols as $A_1 \cap A_2$, and the intersection of $A_1, \ldots, A_k$ as $A_1 \cap \ldots \cap A_k$.

Like union, intersection is a commutative and associative operation, as can easily be checked.

Usually a set that we deal with is defined by some property, i.e. a sample space event is the set of all sample points which have a certain property. For sets we have the option of using **property language** as an alternative to set language. Union corresponds to "or" and intersection corresponds to "and". That is, if set A is the collection of objects that satisfy property α, and set B is the collection of objects that satisfy property β, then $A \cup B$ is the collection of all objects for which "α **or** β " is true, and $A \cap B$ is the collection of all objects for which "α **and** β " is true. Writing "$A \cup B$" seems a little shorter than writing "α **or** β", but is not necessarily clearer.

Remark 2.28 (The inclusive sense of the word "or"). It should be emphasized that when we say that $A \cup B$ is the collection of all objects for

which "α **or** β " is true, we are using the word "or" in the *inclusive sense*, which includes the possibility that both statements might be true.

The inclusive sense is one of the two correct uses of the word "or" in English. For example, if I say, "I dream of being rich or famous", this does not mean that I would be heartbroken if I were both, so I am using the inclusive sense.

On the other hand, suppose you are ordering supper at your favorite diner, and your order comes with a free dessert. When the waiter says: "You can have jello or rice pudding", this is very likely an example of using "or" in the *exclusive sense*, meaning that exactly one of two possibilities is true.

In mathematics, if we mean "or" in the exclusive sense, we will say so explicitly, unless it is obvious.

Definition 2.29 (Set difference and complement). For any sets A and B, $A - B$ denotes the set difference, simply meaning the set of elements which are in A but not in B. The set difference $A - B$ is sometimes written as $A \setminus B$, but we won't use that notation.

If you think that your reader knows the "universe" U of elements that you are currently interested in, then for any set A contained in U, the set $U - A$ can be written more briefly as A^c. The set A^c is called the *complement* of A. In probability theory the set U is often the sample space Ω.

Just as union corresponds to "**or**" in property language, and intersection corresponds to "**and**" in property language, set difference and complement correspond to "**not**" in property language.

If a subset A of the sample space represents the occurrence of a certain physical event E, then A^c represents the event that E does **not** occur. Notice that

"Two events are equal if and only if their complements are equal."

Complements are sometimes more convenient than set differences.

Exercise 2.24 (De Morgan's Laws). Please verify the following facts:

$$(A^c)^c = A, \tag{2.17}$$

$$(A \cup B)^c = A^c \cap B^c, \tag{2.18}$$

$$(A \cap B)^c = A^c \cup B^c. \tag{2.19}$$

In property language, equation (2.17) expresses the meaning of a "double negative". Equations (2.18) and (2.19) are known as De Morgan's laws. Using equation (2.17) one can deduce either of De Morgan's laws from the other.

[Solution]

Definition 2.30 (Venn diagrams). Visual images seem to aid our thinking at times, and books often represent sets and their relationships with pictures, called Venn diagrams. Venn diagrams for sets are not pictures of actual sets, but they are schematic representations which show certain properties.

Most readers will have seen such diagrams, often outside mathematics. Figure 2.1 is a good example.

In general, readers are encouraged to follow any urge to draw pictures when thinking about any problems or concepts!

Definition 2.31 (Set membership). We can express *membership* in a set by "$\in$", Thus $x \in A$ means that x is a member of A.

Using $\in$ takes less space than using the word "in", so we'll tend to use $\in$ in formulas later.

Definition 2.32 (Set comparison). We write $A \subset B$ to mean that every member of A is also a member of B. In this case we say that A is a subset, or that A is *included* in B.

If A is a subset of B, but A is not equal to B, we say in words that A is a *proper* subset of B. We do not have a separate notation for proper inclusion. (The inclusion relation is sometimes written $A \subseteq B$, in which case $A \subset B$ might denote proper inclusion, but we won't use that convention.)

The word "contains" is used in two ways for sets. If $x \in A$ we say that A contains x, but occasionally if $A \subset B$ one also says B contains A. The context usually makes the meaning clear.

Definition 2.33 (Disjoint sets and the empty set). Any sets A, B are **disjoint** if there is no element which is in both sets. Thus A and B are disjoint if $A \cap B = \emptyset$, where $\emptyset$ denotes the *empty set*.

Sets $A_1, \ldots, A_k$ are disjoint if there is no point which is a member of more than one of the sets $A_1, \ldots, A_k$.

In property language, disjoint sets correspond to **mutually exclusive** properties. If $A_1, \ldots, A_k$ are disjoint events in a sample space, then a sample point can be a member of *at most one* of $A_1, \ldots, A_k$. That is, at most one of the corresponding physical events can occur.

Exercise 2.25. When manipulating sets we often use simple observations such as

$$\begin{aligned} A \subset B \implies A \cap B = A, \\ A \cap (B - A) = \emptyset. \end{aligned} \tag{2.20}$$

Here we use $\implies$ to mean "implies".
Please prove the facts in equation (2.20).
[Solution]

Number of elements in a set A set can be finite or infinite. The number of elements in a finite set S will be denoted by $|S|$. If S is an infinite set we will write $|S| = \infty$.

Exercise 2.26 (Intersection distributes over union). Prove that

$$B \cap (A_1 \cup \ldots \cup A_k) = (B \cap A_1) \cup \ldots \cup (B \cap A_k)\,. \tag{2.21}$$

[Solution]

Equation (2.21) can be expressed by saying that "and distributes over or".

You are asked to show in the next exercise that "or distributes over and". This second fact is not hard either, but it is worth checking, especially since we only have one distributive law in the case of numbers!

Exercise 2.27 (Union distributes over intersection). Write an equation in terms of set operations expressing the fact that union distributes over intersection. Then prove this equation. Give two proofs. The first proof should only depend on the basic definitions of union and intersection. The second proof should use equation (2.21) and De Morgan's Laws.

[Solution]

In practical situations, using "and" and "or", we easily recognize the truth of the rules in Exercises 2.26 and 2.27, even though we may not think of them abstractly.

George Boole seems to have been the first person to observe (in 1854) that such general algebraic properties can be formulated for logical statements involving "and", "or", and "not".

2.9 Solutions for Chapter 2

Solution (Exercise 2.1). (a) We can take the sample space to be the set of all sequences $(x_1, \ldots, x_{400})$, where each x_i can be either H or T. Since there are two choices for each x_i, the sample space contains 2^{400} points.

(b) Writing 2^{400} as $(2^4)^{100} = 16^{100}$, we see that the number of points in the sample space is much larger than 10^{100}, and hence it is much larger than the number of atoms in the observable universe.

Of course the number of *abstract events* for the sample space is 2^N, which is a far bigger number than N.

Solution (Exercise 2.2). Since $A = \{2, 4, 6\}$,

$$A = \{2\} \cup \{4\} \cup \{6\},$$

so by the additivity of probability we have

$$P(A) = \mathbf{P}(\{2\}) + \mathbf{P}(\{4\}) + \mathbf{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

If you think that we essentially repeated the proof of Theorem 2.13, you are correct.

Solution (Exercise 2.3). Let A be the event that the same result is obtained on both coin tosses. Then

$$A = \{(1,1),(0,0)\} = \{(1,1)\} \cup \{(0,0)\}\,.$$

By the additivity of probability,

$$\mathbf{P}(A) = \mathbf{P}(\,\{(1,1)\}\,) + \mathbf{P}(\,\{(0,0)\}\,) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Solution (Exercise 2.4). Let $\Omega = \{(1,1),(1,0),(0,1),(0,0)\}$.

Remember that the sample point $(1,0)$ represents the result that the first toss gives success (a head) and the second toss does not, and so on.

Let H_1 denote the event that the first toss produces a head. Then $H_1 = \{(1,1),(1,0)\}$, so

$$\mathbf{P}\,(H_1) = \mathbf{P}(\{(1,1)\}) + \mathbf{P}(\{(1,0)\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

in agreement with the probability found using the sample space for one coin toss.

Solution (Exercise 2.5). To save writing, let $N = 1000000$.

Let $(x_1,\ldots,x_N)$ be a sample point in A.

Then $x_1 = 1$, and there are two choices for each of the remaining x_i, for $i = 2,\ldots,N$. Thus $|A| = 2^{N-1}$.

Since the coin is fair, each of the 2^N sample points is equally likely. Thus $\mathbf{P}(\{\omega\}) = 2^{-N}$ for each sample point ω, and so (by additivity)

$$\mathbf{P}(A) = 2^{N-1}\ 2^{-N} = 2^{-1} = \frac{1}{2},$$

as we knew.

Solution (Exercise 2.6). Using the sample space of Example 2.17,

$$A = \{(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\}\,.$$

Hence

$$\mathbf{P}(A) = \mathbf{P}(\{(5,1)\})+\mathbf{P}(\{(5,2)\})+\mathbf{P}(\{(5,3)\})+\mathbf{P}(\{(5,4)\})+\mathbf{P}(\{(5,5)\})+\mathbf{P}(\{(5,6)\}).$$

Hence $\mathbf{P}(A) = 6/36 = 1/6$, consistent with the model for rolling a single die.

Solution (Exercise 2.7). Let A the event that the first roll gives an even number. Let B be the event that the second roll gives a number larger than four. Our sample space consists of all pairs (x_1, x_2), where each x_i can be $1, 2, 3, 4, 5$ or 6.

Each outcome (x_1, x_2) has probability $1/36$.

To obtain an outcome in $A \cap B$, it is easy to see that there are 3 ways to choose x_1 and 2 ways to choose x_2. Thus there are $3 \times 2 = 6$ outcomes in $A \cap B$, so $\mathbf{P}(A \cap B) = 6(1/36) = 1/6$.

Solution (Exercise 2.8). Using the sample space of Example 2.17, let A be the event that the sum of the numbers on the two dice is at most equal to 5. Consider an outcome (x_1, x_2) in A.

Since x_2 is greater than zero, x_1 cannot be larger than 4. When $x_1 = 1$, x_2 can be any of $1, 2, 3, 4$. When $x_1 = 2$, x_2 can be any of $1, 2, 3$. When $x_1 = 3$, x_2 can be 1 or 2. When $x_1 = 4$, x_2 must be 1. Thus the number of outcomes in A is equal to $4+3+2+1 = 10$. Hence $\mathbf{P}(A) = 10(1/36) = 5/18$.

Solution (Exercise 2.9). The sets $A_2, \ldots, A_{12}$ are *disjoint*.

We will use equation (2.8).

Since $C = A_2 \cup A_4 \cup A_6 \cup A_8 \cup A_{10} \cup A_{12}$,

$$\begin{aligned}\mathbf{P}(C) &= \mathbf{P}(A_2) + \mathbf{P}(A_4) + \mathbf{P}(A_6) + \mathbf{P}(A_8) + \mathbf{P}(A_{10}) + \mathbf{P}(A_{12}) \\ &= \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} = \frac{18}{36} = \frac{1}{2}. \quad (2.22)\end{aligned}$$

Since $D = A_7 \cup A_8 \cup A_9 \cup A_{10} \cup A_{11} \cup A_{12}$,

$$\mathbf{P}(D) = \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{7}{12}.$$

Since $C \cap D = A_8 \cup A_{10} \cup A_{12}$.

$$\mathbf{P}(C \cap D) = \mathbf{P}(A_8) + \mathbf{P}(A_{10}) + \mathbf{P}(A_{12}) = \frac{5}{36} + \frac{3}{36} + \frac{1}{36} = \frac{1}{4}.$$

Solution (Exercise 2.10). Since $\mathbf{P}(\Omega) = 1$ we have $\boldsymbol{p}(1) + \boldsymbol{p}(2) + \boldsymbol{p}(3) + \boldsymbol{p}(4) + \boldsymbol{p}(5) = 1$.

We find easily that $p(\omega_n) = n! p(\omega_1)$, for $n = 1, 2, 3, 4, 5$.

Hence $(1 + 2 + 6 + 24 + 120)\, p(\omega_1) = 1$, and so $p(\omega_1) = 1/153$. This gives $p(\omega_3) = 3!/153 = 6/153$.

Solution (Exercise 2.11). The sample space Ω can be taken to be the set of all sequences (k_1, k_2, k_3, k_4), where each k_i is in $\{0, \ldots, 9\}$. Since there are 10 choices for each k_i, the number of sample points is 10^4. We have no reason to think that the owner of the lock prefers a particular code, so we consider that each sample point ω has the same probability $\boldsymbol{p}(\omega)$ of being the correct code. These probabilities add to one, so $\boldsymbol{p}(\omega) = 1/10000$ for all ω. Hence for any given sample point, such as the sequence which the stranger enters, the probability that this particular sequence is the correct code is 1/10000.

Solution (Exercise 2.12). By the binomial theorem,

$$(a+b)^6 = \sum_{j=0}^{6} \binom{6}{j} a^j b^{6-j}.$$

If we take $a = 1/3$ and $b = 2/3$, the right side of this equation is $p(0) + \ldots + p(6)$. The left side is clearly equal to one.

Solution (Exercise 2.13).

(a) $\Omega = \{0, 1, \ldots, 100\}$. Since the probability values are equal and sum to 1, $p(\omega) = 1/101$ for every ω.

(b)

(i) $\mathbf{P}(\{3\}) = p(3) = 1/101$.

(ii) There are 51 even numbers in the sequence $0, 1, \ldots, 100$. Summing up $p(\omega)$ for these ω, the probability is 51/101.

(iii)

Since A contains 20 numbers, $\mathbf{P}(A) = 20/101$. Since B contains 40 numbers, $\mathbf{P}(B) = 40/101$. A and B are disjoint, so $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

(iv) Numbers divisible by 3 are of the form $3 \times k$. Numbers of this form which are less than 50 are the numbers $3 \times 0, 3 \times 1, 3 \times 2, \ldots, 3 \times 16$. Hence there are 17 numbers in the event described, and the event has probability 17/101.

Solution (Exercise 2.14). $\Omega = A \cup A^c$, and this is a disjoint union. By additivity, $\mathbf{P}(A) + \mathbf{P}(A^c) = \mathbf{P}(\Omega) = 1$.

Solution (Exercise 2.15). From the solution to Exercise 2.11, the probability that the lock opens is $1/10000$. Hence the probability that the lock does not open is $1 - 1/10000 = 9999/10000$,

Solution (Exercise 2.16). By the statement of Exercise 2.12, the probability that no head is obtained is $\binom{6}{0}(\frac{2}{3})^6 = 64/729$. Hence the probability that at least one head is obtained is $1 - 64/729 = 665/729$.

Solution (Exercise 2.17). By definition, if $\omega \in B - A$ then $\omega \notin A$, so A and $B - A$ are disjoint.

If $\omega \in A \cap B$ then $\omega \in B$, by definition. If $\omega \in B - A$ then $\omega \in B$, by definition. Hence if $\omega \in (A \cap B) \cup (B - A)$ then necessarily $\omega \in B$.

On the other hand, if $\omega \in B$ then either $\omega \in A$ or $\omega \notin A$. In the first case $\omega \in A \cap B$, and in the second case $\omega \in B - A$. Thus in all cases $\omega \in (A \cap B) \cup (B - A)$.

We have shown that B and $(A \cap B) \cup (B - A)$ are the same set. This proves equation (2.10).

Since $A \cap B \subset A$, and since we know that A and $B - A$ are disjoint, we know that $A \cap B$ and $B - A$ are disjoint.

Hence by additivity we obtain equation (2.11).

Solution (Exercise 2.18). (i) For any events A, B, $B - A \subset A^c$, so monotonicity tells us that $\mathbf{P}(B - A) \leq \mathbf{P}(A^c)$. When $\mathbf{P}(A) = 1$, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A) = 0$, so $\mathbf{P}(B - A) = 0$.

(ii) For any events A, B, $B = (B \cap A) \cup (B - A)$, and this is a disjoint union, so $\mathbf{P}(B) = \mathbf{P}(B \cap A) + \mathbf{P}(B - A)$.

If $\mathbf{P}(A) = 1$ then $\mathbf{P}(B - A) = 0$ by part (i).

Solution (Exercise 2.19). There are 173 beans in the bowl. By Theorem 2.22 the probability of picking a red bean is $53/173$.

Solution (Exercise 2.20).

(a) There are 38 sample points, all of equal probability $1/38$. Let A be the event. A contains 13 sample points, so $\mathbf{P}(A) = 13/38$, by Theorem 2.22.

(b) There are 380 sample points, all of equal probability $1/380$. Let B be the event. A contains 130 sample points, so $\mathbf{P}(A) = 130/380 = 13/38$, by Theorem 2.22.

Solution (Exercise 2.21). Imagine choosing one bean at a time.

As in Example 2.23, Ω is the set of all pairs (b_1, b_2), where b_1 represents the first bean selected and b_2 presents the second bean selected, with $b_2 \neq b_1$.

There are 173 beans. There are 173 ways to choose the first bean, and, having selected the first bean, there are 172 ways to choose the second bean. Thus

$$|\Omega| = 173 \times 172,$$

and

$$\mathbf{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{173 \times 172}$$

for every ω.

When choosing two red beans, there are 53 ways to choose the first bean and, having chosen the first bean, there are 52 ways to choose the second bean. Thus

$$|R| = 53 \times 52,$$

and

$$\mathbf{P}(R) = \frac{53 \times 52}{173 \times 172}.$$

Solution (Exercise 2.22).

(a) There are 2^{400} sample points ω, each one of the same probability p. Hence $2^{400}p = 1$, so $p = 2^{-400}$.

(c) Let A_j be the event that a head is obtained during tosses $j, j+1, \ldots, j+19$. This event is defined for $j = 1, \ldots, 381$.

Since the result of the *other* tosses is not specified, each set A_j contains 2^{380} sample points, and each sample point has probability 2^{-400}. Thus

$$\mathbf{P}(A_j) = 2^{380}2^{-400} = 2^{-20}.$$

Notice that we get the same probability for A_j, if we think of tosses $j, j+1, \ldots, j+19$ as a small experiment by itself.

By the definition of A, $A = A_1 \cup \ldots \cup A_{381}$. By subadditivity,

$$\mathbf{P}(A) \leq \mathbf{P}(A_1) + \ldots + \mathbf{P}(A_{381}) = \frac{381}{2^{20}} \approx 0.00036335.$$

This is a small value.

Solution (Exercise 2.23). Assume it is known that for any events A_1, A_2,

$$\mathbf{P}\left(A_1 \cup A_2\right) \leq \mathbf{P}(A_1) + \mathbf{P}(A_2). \tag{2.23}$$

We will prove by induction that equation (2.16) holds for all $k \geq 2$.

For some integer $k \geq 2$, suppose it is known that for any events $A_1, \ldots, A_k$,

$$\mathbf{P}\left(A_1 \cup \ldots \cup A_k\right) \leq \sum_{j=1}^{k} \mathbf{P}\left(A_j\right). \tag{2.24}$$

Let $A_1, \ldots, A_{k+1}$ be given. Define

$$A = A_1 \cup \cdots \cup A_k.$$

By the meaning of union,

$$A_1 \cup \ldots \cup A_k \cup A_{k+1} = A \cup A_{k+1}.$$

By equation (2.23),

$$\mathbf{P}(A_1 \cup \ldots \cup A_{k+1}) \leq \mathbf{P}(A) + \mathbf{P}(A_{k+1}).$$

Since we have assumed the truth of equation (2.24), we know that

$$\mathbf{P}(A) \leq \sum_{j=1}^{k} \mathbf{P}\left(A_j\right).$$

Combining the last two equations,

$$\mathbf{P}(A_1 \cup \ldots \cup A_{k+1}) \leq \sum_{j=1}^{k} \mathbf{P}\left(A_j\right) + \mathbf{P}(A_{k+1}) = \sum_{j=1}^{k+1} \mathbf{P}\left(A_j\right).$$

Thus equation (2.16) holds with k replaced by $k+1$.

By induction, equation (2.16) holds for all $k \geq 2$.

Solution (Exercise 2.24). To verify $(A^c)^c = A$, note that by definition A^c is the set of elements which are in the "universe" but not in A.

By definition, $(A^c)^c$ is the set of elements which are in the universe but not in A^c. Thus $(A^c)^c$ is the set of elements x in the universe such that the

statement "x is not in A" is false, i.e. the statement "x is in A" is true. This shows that equation (2.17) holds.

To verify that $(A \cup B)^c = A^c \cap B^c$, note that by definition $(A \cup B)^c$ is the set of elements x in the universe for which the statement "x is in $A \cup B$" is false. That is $(A \cup B)^c$ is the set of elements x in the universe for which the statement "x is in A or x is in B" is false.

Equivalently, $(A \cup B)^c$ is the set of elements x in the universe such that both of the two statements "x is in A" or "x is in B" are false. Thus $(A \cup B)^c$ is the set of elements x in the universe such that $x \in A^c$ and $x \in B^c$. This shows that equation (2.18) holds.

To verify that $(A \cap B)^c = A^c \cup B^c$, note that by definition $(A \cap B)^c$ is the set of elements x in the universe for which the statement "x is in $A \cap B$" is false. That is $(A \cap B)^c$ is the set of elements x in the universe for which the statement "x is in A and x is in B" is false.

Thus $(A \cap B)^c$ is the set of elements x in the universe such that at least one of the statements "$x \in A^c$", " $x \in B^c$" is true.

Thus $(A \cap B)^c$ is the set of elements x in the universe such that $x \in A^c \cup B^c$. This shows that equation (2.19) holds.

One way to deduce equation (2.19) from (2.18) using (2.17), is to start by letting $H = A^c$ and $K = B^c$.

Since equation (2.18) holds for any sets A, B, it also holds with A replaced by H and B replaced by K.

This gives:

$$(H \cup K)^c = (H)^c \cap (K)^c = (A^c)^c \cap (B^c)^c .$$

By equation (2.17),

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

Now take the complement on both sides of this equation:

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

By equation (2.17),

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

This is equation (2.19).

Solution (Exercise 2.25). Yep, these follow from the definitions.

To prove that $A \subset B \implies A \cap B = A$, assume that $A \subset B$.

Then if $x \in A$ then $x \in B$. Hence $x \in A \cap B$.

On the other hand, if $x \in A \cap B$ then by definition $x \in A$ and $x \in B$, so in particular $x \in A$.

We have shown that $A = A \cap B$.

This proves the first equality.

To prove that $A \cap (B - A) = \emptyset$, consider $x \in A$. If $x \in B - A$ then by definition x is not a member of A, which is false. Thus $x \notin B - A$. Since A and $B - A$ have no members in common, $A \cap (B - A) = \emptyset$.

Solution (Exercise 2.26). Let x be a point in $B \cap (A_1 \cup \ldots \cup A_k)$. By definition, $x \in B$ and there is some index j such that $x \in A_j$. Then $x \in B \cap A_j$. Hence by definition $x \in (B \cap A_1) \cup \ldots \cup (B \cap A_k)$.

Let y be a point in $(B \cap A_1) \cup \ldots \cup (B \cap A_k)$. By definition, there is some index j such that $y \in B \cap A_j$. Then $y \in B$ and $y \in A_j$. Hence $y \in B$ and $y \in A_1 \cup \ldots \cup A_k$, so by definition $y \in B \cap (A_1 \cup \ldots \cup A_k)$.

We have proved that $B \cap (A_1 \cup \ldots \cup A_k)$ and $(B \cap A_1) \cup \ldots \cup (B \cap A_k)$ contain exactly the same points, so they are the same set.

Solution (Exercise 2.27). We must show that:

$$B \cup (A_1 \cap \ldots \cap A_k) = (B \cup A_1) \cap \ldots \cap (B \cup A_k)\,. \tag{2.25}$$

First proof: Let x be a point in $B \cup (A_1 \cap \ldots \cap A_k)$. By definition, this means that at least one of the following statements holds:

(i) $x \in B$.

(ii) $x \in A_1 \cap \ldots \cap A_k$.

If statement (i) is true then $x \in B \cup A_j$ for every j, and so by definition $x \in (B \cup A_1) \cap \ldots \cap (B \cup A_k)$.

If statement (ii) is true, then $x \in A_j$ for every j. Hence again we have $x \in B \cup A_j$ for every j, so again $x \in (B \cup A_1) \cap \ldots \cap (B \cup A_k)$.

Thus in all cases, $x \in (B \cup A_1) \cap \ldots \cap (B \cup A_k)$.

Let y be a point in $(B \cup A_1) \cap \ldots \cap (B \cup A_k)$.

By definition, for every index $j = 1, \ldots, k$, $y \in B \cup A_j$.

If $y \in B$, then statement (i) holds with x replaced by y.

If $y \notin B$, then for every index j, $y \in A_j$ must hold, since otherwise $y \in B \cup A_j$ would be false. Hence in this case $y \in A_1 \cap \ldots \cap A_k$, so statement (ii) holds with x replaced by y.

We have shown that either statement (i) or statement (ii) holds, so $y \in B \cup (A_1 \cap \ldots \cap A_k)$.

We have proved that $B \cup (A_1 \cap \ldots \cap A_k)$ and $(B \cup A_1) \cap \ldots \cap (B \cup A_k)$ contain exactly the same points, so they are the same set.

This proves equation (2.25).

Second proof: Let C be any set, and let $D_1, \ldots, D_k$ be any sets.

By equation (2.21),

$$C \cap (D_1 \cup \ldots \cup D_k) = (C \cap D_1) \cup \ldots \cup (C \cap D_k) .$$

Then

$$(C \cap (D_1 \cup \ldots \cup D_k))^c = ((C \cap D_1) \cup \ldots \cup (C \cap D_k))^c .$$

Using equations (2.19) and (2.18),

$$C^c \cup (D_1 \cup \ldots \cup D_k)^c = (C \cap D_1)^c \cap \ldots \cap (C \cap D_k)^c .$$

Using equation (2.19),

$$C^c \cup (D_1 \cup \ldots \cup D_k)^c = (C^c \cup D_1^c) \cap \ldots \cap (C^c \cup D_k^c) . \tag{2.26}$$

Let $C = B^c$, and let $D_j = A_j^c$. By equation (2.17), $C^c = B$ and $(D_j)^c = A_j$. Thus equation (2.26) gives equation (2.25).

Incidentally, using equation (2.17) here is correct, but we could express things in another way: since C can be any set, C^c can be any set, and since D_j can be any set, D_j^c can be any set. And so, in equation (2.26) we can replace C^c by any B and D_j^c by any A_j.

Chapter 3

Models with continuous sample spaces

Probability models come in many forms, and the theory of probability applies to all of them. Having a wide range of examples deepens our understanding of general properties. In this chapter we discuss models that have continuous sample spaces. These will used in applications later.

Our discussion of the general principles of probability resumes in Chapter 4. Impatient readers can read through the present chapter quickly, and return later as needed. Exercises 3.1, 3.2, 3.3, 3.4, 3.5 will be useful in seeing the main ideas here.

3.1 Choosing a point in a continuous interval

In this section we introduce a new class of sample space models. These models are more abstract than the simple models described in Section 2.1, but the general properties of probability remain true.

The particular model we discuss here has a sample space which is made up of an infinite number of points. And not just that: the sample space forms a *continuous interval*, meaning an interval with no gaps.

Consider the physical experiment of choosing a location at random on a yardstick. Since a yardstick is three feet long, one might represent the yardstick as the interval $[0, 3]$ of the real line. We can then think of the experiment more abstractly as choosing a point in the interval $[0, 3]$. The outcome is the point chosen, and the sample space Ω is simply the interval

$[0, 3]$ itself.

A point in Ω is a real number, so evidently we have chosen to represent a physical point on the yardstick as a real number. However, specifying a real number means specifying an infinite number of decimal digits! For an actual physical location, this is very wasteful, since an infinitely precise description of position has no experimental meaning. Thus for a sample point ω in $[0, 3]$ it seems that we can only use the first few digits from the decimal expansion of ω, and the number ω is a misleadingly precise description of a physical location.

But this rather vague interpretation of ω seems acceptable for practical purposes. We know that a mathematical interval is an extreme idealization, and no one could expect that $[0, 3]$ would match up perfectly with a real physical yardstick.

Still though, since we acknowledge the imprecision in the interpretation, a reader may suspect that we are cluttering up the sample space with a lot of irrelevant details. Can't we find a simpler model?

As an *alternative* sample space, if we are satisfied by specifying positions with an accuracy of six decimal places, we could agree to conceptually divide the yardstick into subintervals of length 10^{-6}, and just let the sample space Ω be a set consisting of integers that label these sections. That sample space would be less complicated mathematically, and we could adequately describe any location by simply stating which subinterval contains it.

However, notice that using the interval $[0, 3]$ as the sample space preserves much more of the geometrical setting for the experiment. And we will find that meaningful calculations of probabilities are actually clearer and more elegant if we use real numbers as sample points. So we will stick with using an interval of the real line as the sample space for this experiment, and for similar situations.

Does that choice seem strange? It actually should not come as a great surprise that using a continuous interval of real numbers can make life easier, when modeling the physical world. Readers have likely already experienced the benefits of using the real line in calculus, to help solve problems about physical objects and physical processes.

Now let's think about how to assign probabilities to events, when the sample space is a continuous interval of the real line.

3.2 Probabilities of subsets of an interval

Since we are going to be talking quite a bit about intervals, let's state a definition, just to make sure we all agree on what we are talking about.

Definition 3.1 (Intervals). An interval of the real line is defined as a subset of the form $[a, b]$, or $[a, b)$, or $(a, b]$, or (a, b). A one-point set $\{b\} = [b, b]$ counts as an interval too.

As a warmup to defining probabilities, let's look at what doesn't work. What doesn't work is equation (3.1) in Theorem 2.13. The problem is that Theorem 2.13 dealt with the probability of an event A which is composed of a *finite* number of points. Until now, that has been the only situation that we have had to deal with, and it's a nice situation, because the additivity of probability basically tells us everything we need.

Theorem 2.13 says that when there are only a finite number of sample points in A, then

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}(\{\omega\}) \tag{3.1}$$

If you have a probability model with a finite sample space, every event is a finite set of points, so we can define the probability of all possible events by figuring out the appropriate value of $\mathbf{P}(\{\omega\})$ for each sample point ω. Equation (3.1) then gives you the probability of any event A. Very convenient.

But now we are in a new world. When the sample space is an interval of the real line, the sample space certainly contains an infinite number of points. Furthermore, for any single sample point ω, the one-point set $\{\omega\}$ seems rather useless all by itself when modeling the choice of a random location, since the idea of specifying a physical position with infinite precision is a fantasy. And if A is any set which contains only a finite number of points, the same argument suggests that A is not going to help in modeling real events either. So we cannot avoid dealing with events which are infinite sets of points.

After thinking about it, it seems that when choosing a random point in an interval, the most useful events will be subintervals. If $A = [u, v]$, then A is the event that the chosen point lies somewhere in the interval $[u, v]$. This event seems physically meaningful, at least if the length of $[u, v]$ is not too small to measure.

So now we have a specific question to think about. Given a subset Ω of the real line, and modeling the random selection of a point from Ω, how can we define a probability distribution for intervals which are contained in Ω?

3.3 The uniform probability distribution on an interval

We are considering randomly choosing a point from a subset Ω of the real line. We take the sample space to be Ω, so that any point x in Ω represents the outcome that x is the chosen point. Events are sets of outcomes, so events are subsets of the Ω.

Consider a *special case*. Let's add the assumption that the random point will be chosen from Ω, in such a way that no point of Ω is favored. For simplicity, we'll also assume that Ω is an interval of positive length, or is made up of a finite number of such intervals.

How should we define probabilities in this case?

Physically, it seems clear that a long interval is more likely to contain that chosen point than is a short one. Building on that insight, it seems reasonable to make a specific mathematical assumption:

> **The probability that a point lies in a subinterval A of Ω should be proportional to the length of A.**

This means that there is some constant c such that for any subinterval A of Ω,

$$\mathbf{P}(A) = c\,\mathbf{length}(A). \tag{3.2}$$

Definition 3.2 (Uniform distributions on subsets of the real line). Let Ω be a subset of the real line which is an interval, or the union of a finite number of intervals.

Let $\mathbf{P}$ be a probability distribution on Ω such that for some constant c, equation (3.2) holds for every subinterval A of Ω.

Then we say that the probability distribution $\mathbf{P}$ is the *uniform distribution* on Ω.

When using Definition 3.2, how do we find the constant c in equation (3.2)?

For simplicity, let's assume that Ω is an interval.

We also must have $\mathbf{P}(\Omega) = 1$, so equation (3.4) tells us that

$$c\,\mathbf{length}(\Omega) = 1,$$

i.e.

$$c = \frac{1}{\mathbf{length}(\Omega)}. \tag{3.3}$$

We conclude that when A is a subinterval of Ω,

$$\mathbf{P}(A) = \frac{\mathbf{length}(A)}{\mathbf{length}(\Omega)}. \tag{3.4}$$

Exercise 3.1. When Definition 3.2 holds, and Ω is not a single interval, but is the union of several disjoint intervals $B_1, \ldots, B_k$, what is the correct formula for $\mathbf{P}(A)$, instead of equation (3.4)?

[Solution]

In an experiment, if you want to describe properties of positions, using a ruler or some other measuring device, you will likely describe one or more subintervals which are ranges specified by your measurements. Thus a typical event when choosing a random point seems likely to be a finite union of intervals. If A denotes such an event, then there are disjoint intervals $I_1, \ldots, I_m$ such that

$$A = I_1 \cup \ldots \cup I_m. \tag{3.5}$$

See Figure 3.1. Mathematically, other events are certainly possible, but we don't need to consider these at the moment.

Remark 3.3 (One-point events never happen here!). In the situation of Definition 3.2, let ω be a point of Ω, and let $A = \{\omega\}$.

Since $A = [\omega, \omega]$, by formula (3.4) we have

$$\mathbf{P}(A) = \frac{\mathbf{length}([\omega, \omega])}{\mathbf{length}(\Omega)} = 0.$$

Figure 3.1: $A = I_1 \cup I_2 \cup I_3 \cup I_4$.

This event A has probability zero, so, loosely speaking, it never happens! Is that right? Let c denote a particular point. Suppose that you randomly choose a point from the interval every second of every day for a trillion years. What's the probability that the particular point c will be one of the points that is chosen during that period? The probability of obtaining c on any choice is zero, as we just showed. And adding up lots of zeros still gives zero. So the probability of ever getting the point c is indeed equal to zero.

On the other hand, every time you perform the experiment, *some* location is chosen. So *some* one-point event *always* happens!

This sounds a bit like a paradox. To see that there is really no problem, let's look at an analogous story about length.

Every one-point set has zero length, right? So the unit interval is made up of sets which have zero length. But the good old unit interval has length equal to one.

Is that a paradox? Well, the length of the unit interval is definitely not found as an "infinite sum" by adding up the lengths of the points that compose it. So the length story seems ok.

We should keep in mind that the real line is an abstraction, and points of the real line are not physical objects. We can often think about models as if points of the real line are physical objects, but it's not so.

Incidentally, when one-point events have zero probability, the probability of an interval does not depend on whether of not we include the endpoints. That is, when choosing a random point,

$$\mathbf{P}([a,b]) = \mathbf{P}([a,b)) = \mathbf{P}((a,b]) = \mathbf{P}((a,b)). \tag{3.6}$$

Why is that true? By additivity,

$$\begin{aligned}
\mathbf{P}([a,b]) &= \mathbf{P}([a,a]) + \mathbf{P}((a,b)) + \mathbf{P}([b,b]) \\
\mathbf{P}([a,b)) &= \mathbf{P}([a,a]) + \mathbf{P}((a,b)) \\
\mathbf{P}((a,b]) &= \mathbf{P}((a,b)) + \mathbf{P}([b,b]))
\end{aligned}$$

And we are assuming that $\mathbf{P}([a,a]) = \mathbf{P}([b,b]) = 0$.

Exercise 3.2. A certain factory makes special telephone cables. Occasionally defects occur in the pieces of cable which are produced. The defects are rare, and seem equally likely to occur at any point in a cable.

There are four communication centers, A, B, C, D. They are connected using cables of the sort just described. One cable runs from A to B, another from B to C, and a final cable runs from C to D. Cable AB is 3 miles long, cable BC is 4 miles long, and cable CD is 2 miles long.

After the cables are installed, the staff discovers that a signal is unable to pass from A to D via the three cables. However, a signal passes successfully from B to C.

Assuming that there is only one defect in the three cables, the defect must lie in either the cable from A to B or in the cable from C to D. Find the probability that the cable from C to D is the one with the defect.

[Solution]

Exercise 3.3. A certain street is 600 feet long. Sam lost his lucky penny somewhere along this street. He knows he lost it there, but has no idea in what part of the street it has fallen.

His friends Alice, Bob and Clancy decide to search for for Sam's coin. Alice searches the first 300 feet, Bob searches the next 200 feet, and Clancy searches the final 100 feet. The searchers are careful, so they will not miss the coin.

(i) Let A be the event that the coin is located in the interval that Alice is searching. Find $\mathbf{P}(A)$.

(ii) Suppose that we learn the following additional information. After five minutes, the coin has not yet been found. Alice has already searched two-thirds of her section. Bob has searched half of his section, and Clancy has searched three-quarters of his section.

Let A be the event that Alice eventually finds the coin. Based on all the information we *now* have, find the probability of A.

This part of the problem is an example of a conditional probability calculation, and we have not yet covered conditional probability. However, you can solve this problem by building a new model, with a new sample space.

[Solution]

3.4 Probability densities on intervals

Let Ω be an interval of the real line, or perhaps a finite union of intervals.

We will think of Ω as part of a model for the experiment of choosing a random point. Of course we have to have a probability distribution defined too. So far we have talked about uniform distributions. But that might not match the physical conditions of the experiment. It might be that the random point is more likely to be chosen from one region rather than another.

A probability distribution which is not uniform can be represented by using a *probability density function* which is larger in some regions and smaller in others.

As the name suggests, a probability density which is defined on a portion of a line tells us the "probability per unit length".

Definition 3.4 (Probability densities). A probability density f is a function such that

(i) f is nonnegative, and

(ii) the integral of f over Ω is equal to one.

If $\mathbf{P}$ is a probability distribution such that

$$\mathbf{P}([a,b]) = \int_a^b f(x)\,dx, \tag{3.7}$$

for every interval $[a,b]$ which is contained in Ω, then we say that f is a probability density for $\mathbf{P}$.

In calculus, $\int_a^b f(x)\,dx$ is usually referred to as "the integral of f over the set $[a,b]$". Equation (3.7) says that the probability of $[a,b]$ is given by the integral of the probability density over the set $[a,b]$.

Conditions (i) and (ii) in Definition 3.4 are needed because probabilities are nonnegative, and because we must have $\mathbf{P}(\Omega) = 1$.

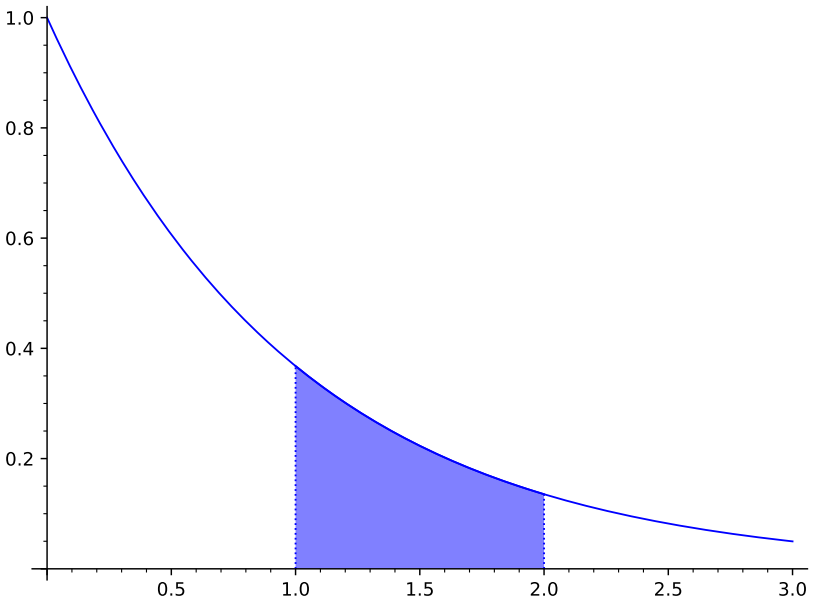


Figure 3.2: Exercise 3.4: the probability of choosing from a set is the integral of the density over the set.

Exercise 3.4. Let $\Omega = [0,3]$. Let f be a probability density on Ω which is a multiple of e^{-x}. A point is chosen at random from $[0,3]$, with a probability

distribution given by the density f. Find the probability that the point is chosen from the interval $[1, 2]$. See Figure 3.2.

[Solution]

Exercise 3.5 (Constant densities give uniform distributions). Let Ω be a subset of the real line which is the union of a finite number of disjoint intervals. Let f be a constant function on Ω such that the integral of f over Ω is equal to one.

Let $\mathbf{P}$ be the distribution with density f. Prove that $\mathbf{P}$ is the uniform distribution on Ω, in the sense of Definition 3.2.

[Solution]

Remark 3.5 (With densities, one-point events never happen). We noted in Remark 3.3 that in the case of a uniform probability distribution on an interval, one-point events always have probability zero. The same is true for any distribution that is given by a probability density. To see that, consider equation (3.7) with $b = a$.

Since $[a, a] = \{a\}$, in this case equation (3.7) says that

$$\mathbf{P}(\{a\}) = \int_a^a f(x)\,dx,$$

and the integral over an interval of zero length is zero.

Thus equation (3.6) holds, just as it did in the case of a uniform density, and we don't need to be fussy about endpoints:

$$\mathbf{P}([a, b]) = \mathbf{P}([a, b)) = \mathbf{P}((a, b]) = \mathbf{P}((a, b)). \tag{3.8}$$

A general form of Definition 3.4 will be given in Definition 15.5, The general definition applies to a wide range of sample spaces, not just the real line.

3.5 Cleaning up integral notations

Formulas are easier to understand if we simplify the notation.

For example, in equation (3.7), unless we want to show the formula for f explicitly, there is really no need to write $f(x)\,dx$ in the traditional calculus manner. It is clearer to just write

$$\mathbf{P}([a,b]) = \int_a^b f. \tag{3.9}$$

Even this notation can be improved. The best notation for the integral of f over a general set A is:

$$\textbf{integral of f over } A = \int_A f. \tag{3.10}$$

Thus equation (3.9) becomes

$$\mathbf{P}([a,b]) = \int_{[a,b]} f. \tag{3.11}$$

And actually, if equation (3.11) holds for all intervals $[a,b]$, then for *any* event A it is true that

$$\mathbf{P}(A) = \int_A f. \tag{3.12}$$

This form is convenient, but what does it mean, when A is not an interval?

Presumably we know what the event A means, or we wouldn't be talking about it. And so we know what $\mathbf{P}(A)$ means. But what about $\int_A f$?

The concept of integrating a function over a set actually makes sense for lots of sets, not just sets which are intervals.

To see this physically, think about a wire whose *mass density* might vary along the wire. The mass of any part of the wire is found by integrating the mass density function over that part of the wire. This makes sense even if the part of the wire that you are interested in consists of many separate pieces. Just find the mass of each piece (by integrating the mass density) and then add up the masses.

In calculus this is how we can find the integral of a function f over a set A, if A is the union of two disjoint intervals $[a,b]$ and $[c,d]$:

$$\int_A f = \int_a^b f + \int_c^d f. \tag{3.13}$$

For example: if we are doing a calculus problem where we have to integrate a function which has a different formula on different parts of an interval A, we often calculate the integral over A as the sum of the integrals over the separate parts.

For general sets we can do the same thing. We can find the integral over a set by integrating over the pieces of the set, and then adding up the results. We will call this the additive property for integration over a set.

And notice that the additivity of integration over sets is exactly what we need if equation (3.12) is used to define a probability distribution. After all, if $A = D_1 \cup D_2$, where D_1, D_2 are disjoint events, the additivity of probability says we must have $\mathbf{P}(A) = \mathbf{P}(D_1) + \mathbf{P}(D_2)$. That probability equation can only true if:

$$\int_A f = \int_{D_1} f + \int_{D_2} f. \tag{3.14}$$

And equation (3.14) expresses additivity for integrating over a set.

So we know how densities define probabilities, and we know how integration over a set works. That's all we need to understand probability densities. But let's nail this down by giving a nice general definition of the process of integrating over a set.

Our definition should capture the idea that the integral of f over a set A is the integral using the values of f on the set A, and nothing else. Using that idea, it is pleasantly simple to give a general definition of $\int_A f$, as follows.

Definition 3.6 (Integration over a set). Let f be a function and let A be a set.

Define a new function g as follows:

$$g(x) = \begin{cases} f(x) & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

Then by definition

$$\int_A f = \int g. \tag{3.16}$$

We might say that g in equation (3.15) is formed by discarding the values of f on A^c.

Incidentally, when it comes to actually computing an integral over a set, which notation you use won't make much difference. But the modern $\int_A$ notation seems clearer for thinking.

Remark 3.7 (Intervals characterize densities). Let Ω be an interval of the real line, or perhaps a finite union of intervals.

Suppose you are studying a distribution $\mathbf{P}$ on Ω, and you come up with a function f such that $\mathbf{P}([s,t]) = \int_{[s,t]} f$ for every subinterval $[s,t]$ of the sample space. Then by definition f is a valid density for $\mathbf{P}$. Does that mean that for *any* event you can go ahead and calculate $\mathbf{P}(A)$ using $\mathbf{P}(A) = \int_A f$? One would hope so, and happily that is actually true! Very convenient. We won't write down a formal proof, but it illustrates a general principle: knowing that an equation is true for all intervals is good evidence that it holds for all sets.

We'll return to this subject in Remark 9.12.

More examples of densities on the real line are given in Section 3.7.

3.6 Choosing a point in $\mathbb{R}^2$: throwing darts

Consider the experiment of throwing darts at a target called the dart board. Assume that the thrower is rather inaccurate, so the point where the dart hits is random. (If the thrown dart misses the target completely, we will ignore that throw, and consider that the experiment did not occur.)

The outcome of the experiment is the point of impact, i.e. the location at which the dart hits the board. We can represent this outcome as a point on an idealized copy of the dart board, which we take to be a region called T in $\mathbb{R}^2$, where $\mathbb{R}^2$ is the set of all coordinates (x_1, x_2) in the plane, i.e. $\mathbb{R}^2$ is the set of all pairs of of real numbers.

The dart board region T is our sample space Ω in this model. An event is then simply a sub-region of the dart board region. See Figure 3.3 for a picture of Ω and an event A.

If the thrower under consideration is very inaccurate, for simplicity we might assume that every part of the target has the same chance of being hit. In that case it seems natural to assume that the probability of hitting a particular region on the target is proportional to the *area* of the region.

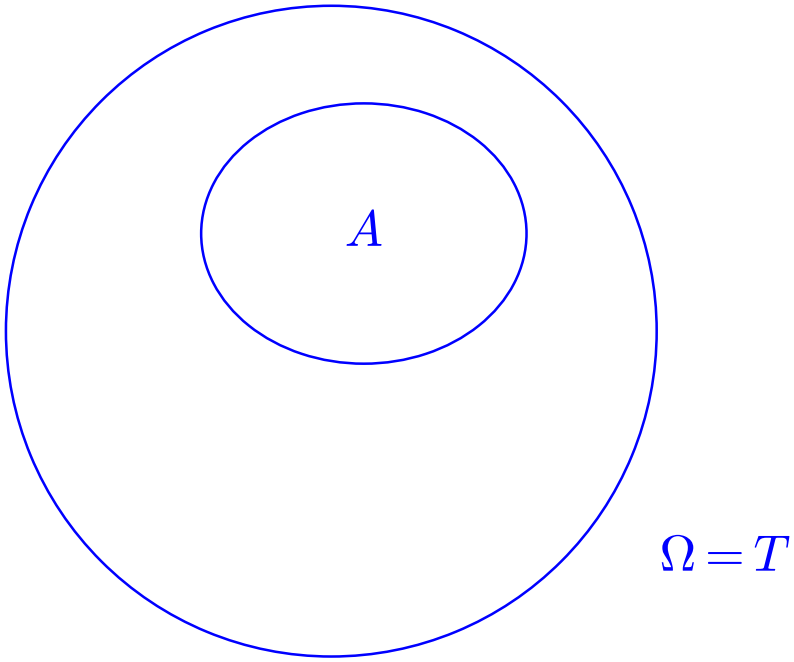


Figure 3.3: An event on the dart board

Since the probability of hitting the whole target must be 1 (remember that we disregard any throw that hits elsewhere), this means that the probability of hitting some region A of the dart board is given by:

$$\mathbf{P}(A) = \frac{\mathbf{area}(A)}{\mathbf{area}(\Omega)}. \tag{3.17}$$

Like the formula for subsets of the real line which was given in equation (3.4), equation (3.17) is a continuous analog of Theorem 2.22.

Definition 3.8 (Uniform distribution on a region in the plane). If a probability set-function $\mathbf{P}$ is defined on subsets of a region T of the plane, and is such that probability is proportional to area, it will be said to be a *uniform probability distribution* on T.

Example 3.9 (Probability of missing the central region). Someone is throwing darts at a target represented by a disc of radius 5, centered at the origin of $\mathbb{R}^2$.

The point of impact (x, y) is random, with a uniform distribution on the target.

Let A be the set of points (x, y) in the target such that $\sqrt{x^2 + y^2} > 2$.

Since $\sqrt{x^2 + y^2} = r$, the distance of the point (x, y) from the origin, A represents the physical event that the dart lands more than two units of distance from the center. See Figure 3.4.

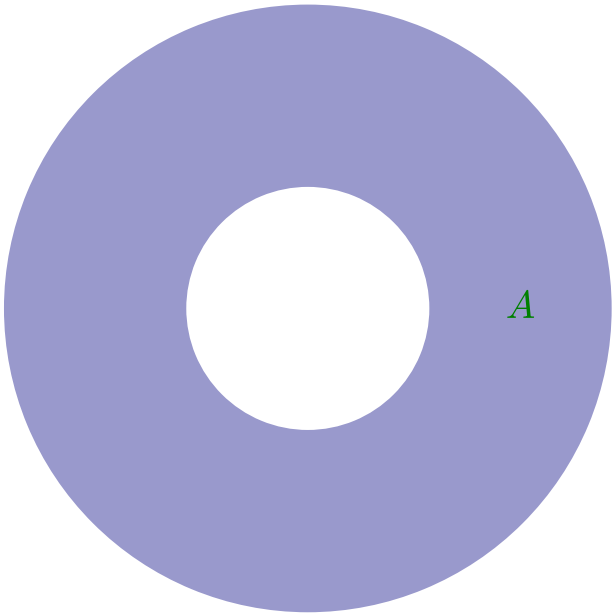


Figure 3.4: A is the event that the dart misses the center region

Let us find $\mathbf{P}(A)$.

Since the probability distribution is uniform,

$$\mathbf{P}(A) = \frac{\mathbf{area}(A)}{\mathbf{area}(T)},$$

where $T = \Omega$ is the target region, and of course $\mathbf{area}(A) = \mathbf{area}(T) - \mathbf{area}(A^c) = 25\pi - 4\pi = 21\pi$. Thus

$$\mathbf{P}(A) = \frac{21}{25}.$$

The next two exercises are mostly a test to see if you can still calculate areas. It's ok to skip them, as long as the statements of the questions make sense to you.

Exercise 3.6. Let Ω be the rectangle consisting of all points x, y such that $0 \leq x \leq 2$ and $0 \leq y \leq 5$. Let $\mathbf{P}$ be the uniform probability distribution on Ω, so that this sample space and distribution form a model for choosing a point at random from Ω.

(i) Let A be the event that the chosen point (x, y) is such that $x < y$. Find $\mathbf{P}(A)$.

(ii) Let B be the event that $y < 4 - 2x$. Find $\mathbf{P}(A \cap B)$.

[Solution]

Exercise 3.7. Someone is throwing darts at a target represented by the unit disc. The point of impact (x, y) is random, with a uniform distribution on the target.

Let A be an event defined in terms of the height of the point of impact: A is represented by the set of points (x, y) in the target such that $-\frac{1}{\sqrt{2}} \leq y \leq \frac{1}{\sqrt{2}}$.

Find $\mathbf{P}(A)$.

[Solution]

Just as in the case of an interval, we can easily define probability densities for regions in the plane. Two-dimensional integrals take more calculation than one-dimensional integrals, but the basic idea is the same. Such densities, and general densities, are considered in Section 15.3.

3.7 More examples of densities

Readers likely don't have an urgent need for more examples at the moment. But it may be enlightening to glance over the examples here, and return later, when using densities in later chapters.

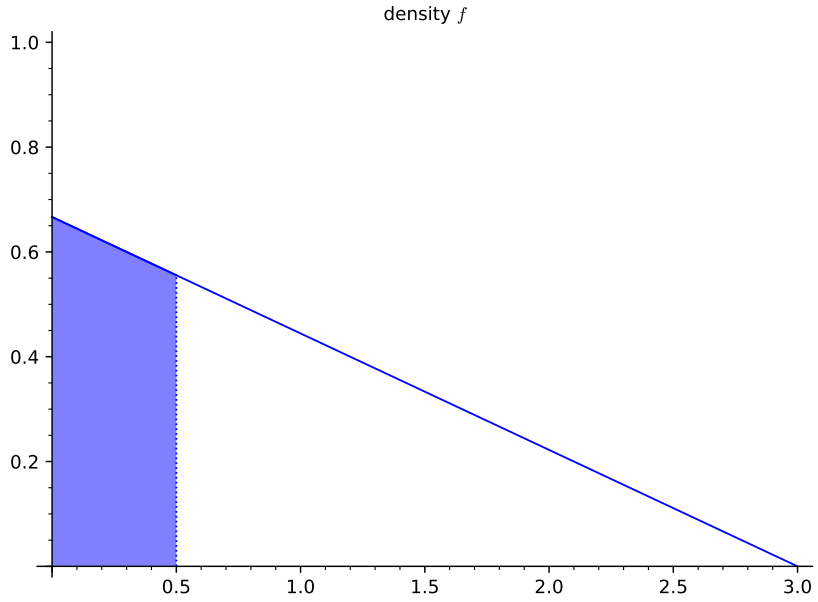


Figure 3.5: Exercise 3.8

Exercise 3.8. For the experiment of choosing a number from the interval $[0, 3]$, suppose that points near 0 are more likely to be chosen, specifically that the probability set-function is given by a density f of the form $f(x) = c(3-x)$, where c is some constant.

(i) Find c.

(ii) Calculate the probability that selected number is less than $1/2$.

See Figure 3.5.

[Solution]

Exercise 3.9. Consider a probability model with sample space Ω equal to $[0, 4]$ and probability density $f(t) = \frac{1}{8}t$.

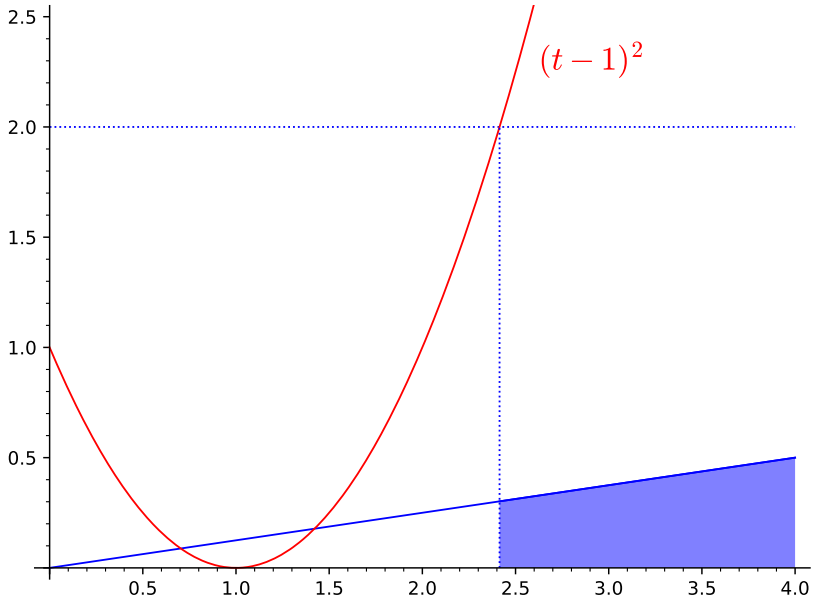


Figure 3.6: Exercise 3.9

(i) Check that f is a probability density.

(ii) Let $\mathbf{P}$ be the probability set-function with density f. Suppose that a random number t is selected. Let A be the event that $(t-1)^2 > 2$. Find $\mathbf{P}(A)$.

See Figure 3.6.
[Solution]

Exercise 3.10. Consider the probability model in Exercise 3.9. Using the $\mathbf{P}$ with density f, let t be the randomly selected point. Let A be the event that $2t - 2 \leq (t-1)^2$. Find $\mathbf{P}(A)$

See Figure 3.7.

(You finally get to use equation (3.12) in a situation where A is not an interval!)

[Solution]

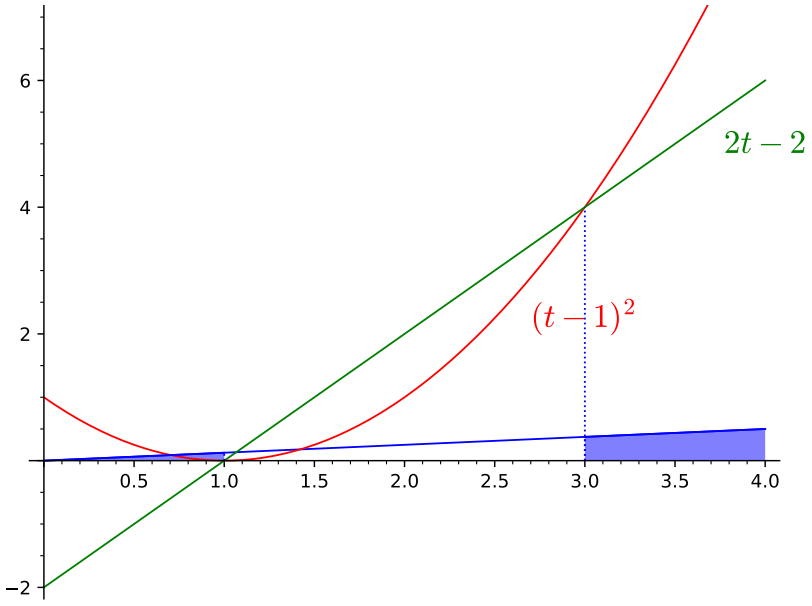


Figure 3.7: Exercise 3.10

Remark 3.10 (To what extent is the probability density unique?). The purpose of a density f is to define a probability set-function $\mathbf{P}$.

If one just changes f at few points, it makes no difference in any integral involving f, and so it makes no difference in the definition of $\mathbf{P}$. So such a modified function works, i.e. is a correct probability density for the probability set-function $\mathbf{P}$. It is just as valid as the original function f, even though it may have a more cumbersome definition.

We might think about the density f as a kind of probability machine. One turns the crank on this machine (i.e. integrates f) to get a probability. That is the sole purpose of f, its raison d'être. Any other function h such that $\int_A h = \int_A f$ for all events A also deserves the honor of being called a probability density for $\mathbf{P}$.

Example 3.11 (A uniform density on an infinite interval?). Much as in Example 2.19, consider a constant probability density f on an *infinite* interval, like $[0, \infty)$ for example. Does such a density make sense?

Let k be the constant value of f. k is a nonnegative number since f is

a density. Since $\int_0^\infty f = \int_\Omega f = \mathbf{P}(\Omega) = 1$, k cannot be zero. On the other hand, since

$$1 \geq \mathbf{P}([0,n]) = \int_{[0,n]} f = k\ n,$$

for every n, we are forced to conclude that k must be zero. This contradiction shows that a constant probability density on an infinite interval does not exist.

A *nonconstant* probability density on an infinite interval is certainly possible, as the next exercise illustrates.

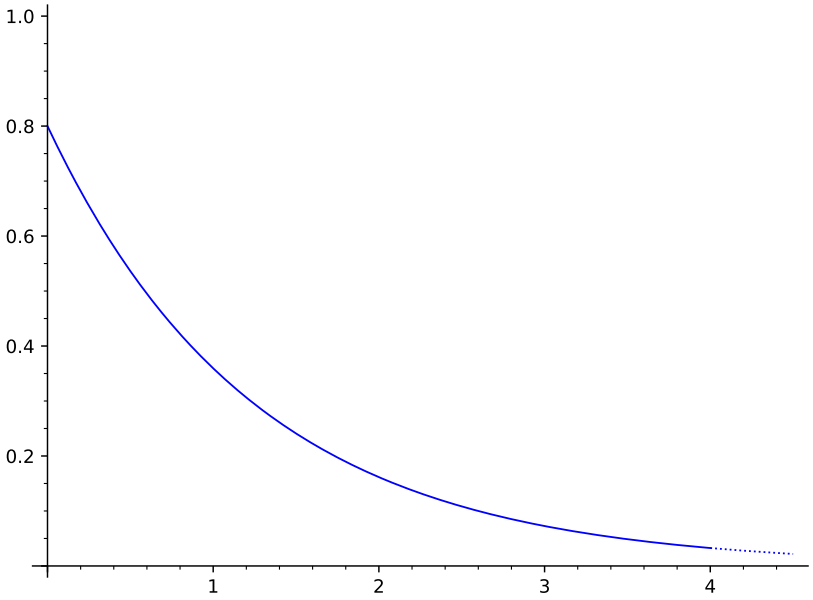


Figure 3.8: $f(x) = ce^{-.8\,x}$

Exercise 3.11 (The exponential density). Let λ be a positive constant. Let $f(x) = ce^{-\lambda x}$ for $x \geq 0$, $f(x) = 0$ otherwise. Assuming f is a probability density on $\mathbb{R}$, find c.

See Figure 3.8.

[Solution]

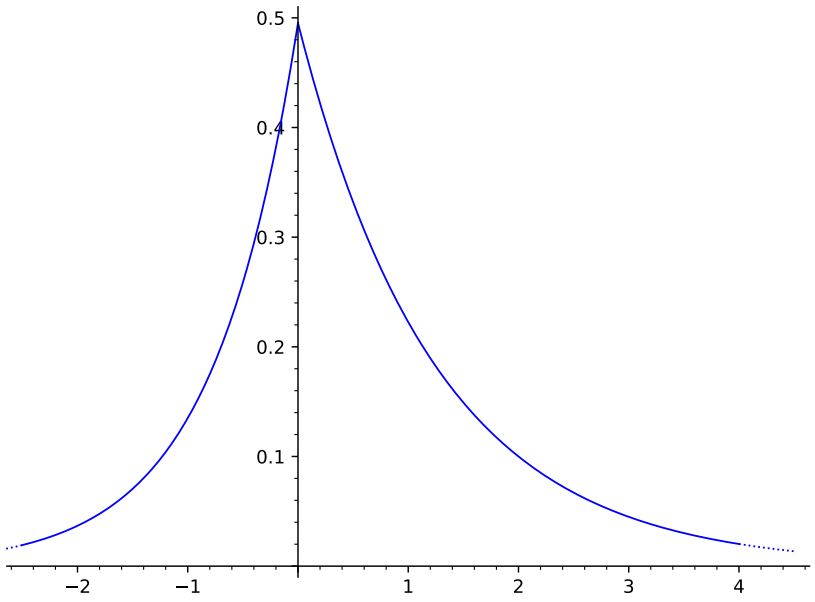


Figure 3.9: $\alpha = .8$, $\beta = 1.3$

Exercise 3.12. Let α and β be positive constants. Let $f(x) = ce^{-\alpha x}$ for $x > 0$, $f(x) = ce^{\beta x}$ for $x < 0$. Assuming f is a probability density on $\mathbb{R}$, find c.

See Figure 3.9.

[Solution]

3.8 Solutions for Chapter 3

Solution (Exercise 3.1). Since $\Omega = B_1 \cup \ldots \cup B_k$, and the sets $B_1, \ldots, B_k$ are disjoint, we know by additivity that

$$\mathbf{P}(\Omega) = \mathbf{P}(B_1) + \ldots + \mathbf{P}(B_k).$$

But $\mathbf{P}(\Omega) = 1$, and by equation (3.2) we know that $\mathbf{P}(B_i) = c\,\mathbf{length}(B_i)$.

Thus

$$1 = \mathbf{length}(B_1) + \ldots + \mathbf{length}(B_k),$$

so

$$c = \frac{1}{\mathbf{length}(B_1) + \ldots + \mathbf{length}(B_k)}.$$

Replacing c in equation (3.3) by this value gives the correct form of equation (3.4):

$$\mathbf{P}(A) = \frac{\mathbf{length}(A)}{\mathbf{length}(B_1) + \ldots + \mathbf{length}(B_k)}. \tag{3.18}$$

Remark 3.12. You could extend the definition of length to include sets which are not intervals. So in the situation of this problem, when Ω is the union of disjoint intervals $B_1, \ldots, B_k$, we could agree to define

$$\mathbf{length}(\Omega) = \mathbf{length}(B_1) + \ldots + \mathbf{length}(B_k).$$

Then equation (3.4) would still be valid.

Solution (Exercise 3.2). Choose numbers a, b, c, d such that the length of $[a, b]$ is equal to the length of the cable from A to B, and the length of $[c, d]$ is equal to the length of the cable from C to D. Choose the numbers so that the intervals $[a, b]$ and $[c, d]$ are also disjoint.

Let $\Omega = [a, b] \cup [c, d]$.

We can think of a point in one of the intervals $[a, b], [c, d]$ as a position coordinate which describes the possible location of the cable defect.

Let $\mathbf{P}$ be the uniform probability distribution on Ω.

Let H represent the event that the defect lies in the cable from C to D. Then $H = [c, d]$.

By equation (3.18),

$$\mathbf{P}(H) = \frac{\mathbf{length}(H)}{\mathbf{length}([a, b]) + \mathbf{length}([c, d])} = \frac{d - c}{(b - a) + (d - c)} = \frac{2}{3 + 2} = \frac{2}{5}.$$

Solution (Exercise 3.3). Let Ω be the union of disjoint intervals $U, V.W$, where $\mathbf{length}(U) = 300$, $\mathbf{length}(V) = 200$, and $\mathbf{length}(W) = 100$.

We can think of a point in one of the intervals U, V, W as a position coordinate which describes the possible location of the lost penny.

Let $\mathbf{P}$ be the uniform distribution on Ω.

(i) The abstract event representing A is U.

$$\mathbf{P}(U) = \frac{\mathbf{length}(U)}{\mathbf{length}(\Omega)} = \frac{300}{600} = \frac{1}{2}.$$

(ii) Let $\bar{U}, \bar{V}, \bar{W}$ be the unsearched parts of U, V, W, respectively. We will assume that these unsearched parts are intervals. (If the unsearched parts were made up of many pieces, the same method would work, it would just take longer to write down.)

Then $\bar{U}, \bar{V}, \bar{W}$ are intervals with $\mathbf{length}(\bar{U}) = (1/3)\mathbf{length}(U) = 100$, $\mathbf{length}(\bar{V}) = (1/2)\mathbf{length}(U) = 100$, $\mathbf{length}(\bar{W}) = (1/4)\mathbf{length}(W) = 25$.

Now let the sample space be $\bar{\Omega} = \bar{U} \cup \bar{V} \cup \bar{W}$. This represents the unsearched road.

A point in $\bar{\Omega}$ represents the possible position of the missing coin, in the unsearched road.

The original description of the problem gives no reason to treat any section of the whole road differently from any other section. The decision about where to search for the coin does not seem connected in any way to the actual location of the coin. So it seems that there is still no reason to treat any section of the unsearched road differently from any other section.

So the appropriate distribution for the location of the missing coin is the uniform distribution on $\bar{\Omega}$. Let $\bar{\mathbf{P}}$ be the uniform distribution on $\bar{\Omega}$.

In this model, the event that Alice eventually finds the coin is $\bar{U}$. Using equation (3.18),

$$\bar{\mathbf{P}}(\bar{U}) = \frac{\mathbf{length}(\bar{U})}{\mathbf{length}(\bar{U}) + \mathbf{length}(\bar{V}) + \mathbf{length}(\bar{W})} = \frac{100}{100 + 100 + 25} = \frac{4}{9}.$$

Solution (Exercise 3.4). The requested probability is $\mathbf{P}([1, 2])$, where the density for $\mathbf{P}$ is given by ce^{-x} on $[0, 3]$, for some constant c.

Since $\mathbf{P}(\Omega) = 1$,

$$\int_0^3 ce^{-x}\, dx = 1,$$

so

$$1 = -ce^{-x}\Big|_0^3 = c(1 - e^{-3}).$$

Thus

$$c = \frac{1}{1 - e^{-3}}.$$

Thus

$$\mathbf{P}([1,2]) = \int_1^2 ce^{-x}\,dx = -ce^{-x}\Big|_1^2 = c(e^{-1} - e^{--2}) = \frac{e^{-1} - e^{-2}}{1 - e^{-3}}.$$

Solution (Exercise 3.5). We are told that f is a constant function. Let c be the vaue of f.

Let A be any subinterval of Ω. Then

$$\int_A f = \int_A c.$$

From calculus we know that integrating c over an interval A gives $c\,\mathbf{length}(A)$.

We are given that f is a probability density for $\mathbf{P}$. Thus $\mathbf{P}(A) = \int_A f$.

We have shown that $\mathbf{P}(A) = c\,\mathbf{length}(A)$ for every subinterval A,

By Definition 3.2, $\mathbf{P}$ is the uniform probability distribution on Ω.

Solution (Exercise 3.6).

(i) The set $A^c = \{(x,y) :\ y \le x\} \cap \Omega$ is a triangle with base 2 and altitude 2. Hence its area is $(1/2)4 = 2$. The set Ω is a 2×5 rectangle, so its area is 10. Thus $\mathbf{P}(A^c) = 2/10 = 1/5$, and so $\mathbf{P}(A) = 4/5$.

Alternatively, note that

$$\mathbf{area}(A) = \int_0^2 \int_x^5 1\ dy\,dx = \int_0^2 (5-x)\,dx = \frac{-(5-x)^2}{2}\Big|_0^2 = -\frac{9}{2} + \frac{25}{2} = 8.$$

Thus $\mathbf{P}(A) = \mathbf{area}(A)/\mathbf{area}(\Omega) = 8/10 = 4/5$.

(ii) The line $y = 4 - 2x$ crosses the line $y = x$ at the point $(4/3, 4/3)$.

Let's find the area of $(A \cap B)^c$. The integral of a function of the form $mx + b$ over an interval is equal to the length of the interval times the value of the function at the midpoint. Thus

$$\mathbf{area}(\,(A\cap B)^c\,) = \int_0^{4/3} (4-2x)\ dx + \int_{4/3}^2 x\,dx = \frac{4}{3}\left(4 - 2\left(\frac{2}{3}\right)\right) + \frac{2}{3}\left(\frac{5}{3}\right) = \frac{14}{3}.$$

Hence

$$\mathbf{P}(A \cap B) = \frac{\mathbf{area}(A \cap B)}{\mathbf{area}(\Omega)} = \frac{\frac{16}{3}}{10} = \frac{16}{30} = \frac{8}{15}.$$

Solution (Exercise 3.7). Let $M_+ = \left\{(x,y) : y > \frac{1}{\sqrt{2}}\right\} \cap \Omega$, and let $M_- = \left\{(x,y) : y < -\frac{1}{\sqrt{2}}\right\} \cap \Omega$.

Then $A = \Omega - (M_+ \cup M_-)$, so $\mathbf{area}(A) = \pi - \mathbf{area}(M_+) - \mathbf{area}(M_-) = \pi - 2\ \mathbf{area}(M_+)$.

Let S be the square with vertices $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$, $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$, $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$.

Then $\Omega - S$ is the union of four regions, M_+ and three other regions with the same area as M_+. Hence

$$\pi - 4 = 4\ \mathbf{area}(M_+),$$

so

$$\mathbf{area}(M_+) = \frac{\pi - 4}{4}.$$

Thus

$$\mathbf{area}(A) = \pi - 2\ \mathbf{area}(M_+) = \pi - \frac{\pi - 4}{2} = \frac{\pi}{2} + 2.$$

Hence

$$\mathbf{P}(A) = \frac{\mathbf{area}(A)}{\mathbf{area}(\Omega)} = \frac{\frac{\pi}{2} + 2}{\pi} = \frac{1}{2} + \frac{2}{\pi}.$$

Solution (Exercise 3.8).

(i)

$$1 = \mathbf{P}(\Omega) = \int_0^3 c(3-x)\,dx = -c\frac{(3-x)^2}{2}\bigg|_0^3 = \frac{9}{2}.$$

Thus $c = 2/9$.

(ii) The probability is

$$\int_0^{1/2} cf(x)\,dx = \frac{2}{9}\int_0^{\frac{1}{2}} (3-x)\,dx = -\frac{1}{9}(3-x)^2\bigg|_0^{1/2} = \frac{9 - \left(\frac{5}{2}\right)^2}{9} = \frac{11}{36}.$$

Solution (Exercise 3.9).

(i) f is clearly nonnegative on Ω, so we only need to check that $\int_\Omega f = 1$.

$$\int_\Omega f = \int_0^5 \frac{1}{8}t\,dt = \frac{1}{16}t^2\bigg|_0^4 = \frac{4^2}{16} = 1.$$

(ii) We need to write A more explicitly.

So we must rewrite the inequality $(t-1)^2 > 2$ in an appropriate form.

If $t \geq 1$, then $t - 1 \geq 0$, and $(t-1)^2 > 2$ is equivalent to $t - 1 > \sqrt{2}$, i.e. $t > 1 + \sqrt{2}$.

If $t < 1$, then $t - 1 < 0$, so $1 - t > 0$, and $(t-1)^2 > 2$ is equivalent to $1 - t > \sqrt{2}$, i.e. $t < 1 - \sqrt{2}$.

Combining these statements with the fact that $0 \leq t \leq 4$, we see that $A = (1 + \sqrt{2}, 4]$.

Thus

$$\mathbf{P}(A) = \int_{1+\sqrt{2}}^{4} \frac{1}{8} t \, dt = \frac{t^2}{16}\Big|_{1+\sqrt{2}}^{4} = \frac{1}{16}\left(16 - \left(1+\sqrt{2}\right)^2\right)$$
$$= \frac{1}{16}\left(16 - \left(1 + 2\sqrt{2} + 2\right)\right) = \frac{1}{16}\left(13 - 2\sqrt{2}\right).$$

Solution (Exercise 3.10). We could consider cases, as in the previous problem, but perhaps it's faster to rewrite the inequality which defines A in a different way first. A is the set of t such that $(t-1)^2 - 2t + 2 \geq 0$.

This inequality says $t^2 - 4t + 3 \geq 0$, i.e. $(t-1)(t-3) \geq 0$.

The polynomial $(t-1)(t-3)$ is zero at $t = 1$ and $t = 3$, positive for $t < 1$, positive for $t > 3$, and negative otherwise.

Hence $A = [0, 1] \cup [3, 4]$, and so

$$\mathbf{P}(A) = \int_0^1 \frac{1}{8} t \, dt + \int_3^4 \frac{1}{8} t \, dt = \frac{1}{16}\left(t^2\Big|_0^1 + t^2\Big|_3^4\right) = \frac{1}{16}(1 - 0 + 16 - 9) = \frac{8}{6} = \frac{1}{2}.$$

Solution (Exercise 3.11). Since

$$1 = \int_{-\infty}^{\infty} f = c\int_{-\infty}^{0} 0 \, dx + c\int_0^{\infty} e^{-\lambda x} \, dx = c\frac{1}{-\lambda}\Big|_0^{\infty} e^{-\lambda x} = \frac{c}{\lambda},$$

we must have $c = \lambda$.

Solution (Exercise 3.12). Since

$$1 = \int_\Omega f = c\int_0^{\infty} e^{-\alpha x} \, dx + c\int_{-\infty}^{0} e^{-\beta x} \, dx = c\frac{1}{-\alpha}\Big|_0^{\infty} e^{-\alpha x} + c\frac{1}{\beta}\Big|_{-\infty}^{0} e^{\beta x} = \frac{c}{\alpha} + \frac{c}{\beta},$$

we must have

$$c = \frac{1}{\frac{1}{\alpha} + \frac{1}{\beta}} = \frac{\alpha\beta}{\alpha + \beta}.$$

Note that the answer here agrees with the answer to Exercise 3.11 if we set $\lambda = \alpha$ and let $\beta \to \infty$. Why should that be the case?

Chapter 4

Conditional probability

4.1 Conditional probability defined

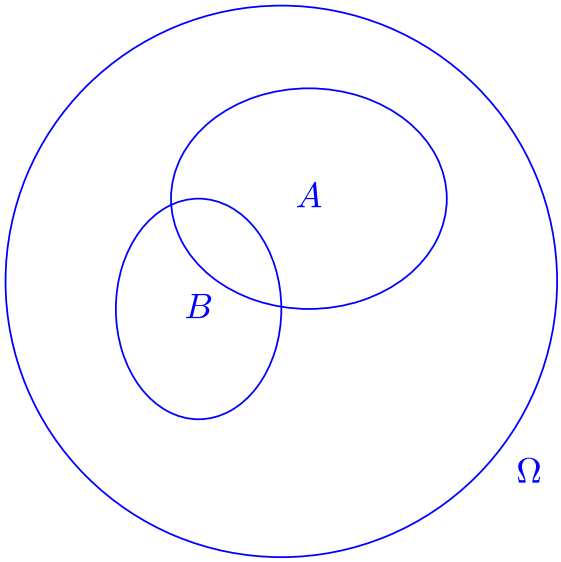


Figure 4.1: Events on the dart board

Consider any experimental situation, and any possible events A, B for this experiment. Suppose that, based on your knowledge of the experimental setup, you know the value of probabilities such as $\mathbf{P}(B)$ and $\mathbf{P}(A)$.

Now suppose the experiment has been performed. Although you do not yet know the result, someone tells you that the event B did occur. This extra knowledge, combined with what you already knew, gives you a new experimental situation, and a new probability for the event A. We call this new probability "the **conditional probability** that A occurred **given** that B occurred". This probability value is written as $\mathbf{P}(A \mid B)$.

How do you find $\mathbf{P}(A \mid B)$?

As a simple example, we can think about the experiment of throwing darts at a dart board, described in section 3.6. The throw takes place, but we are not looking. We ask someone a specific question: "Did the dart land in region B?" (See Figure 4.1.). Imagine that the answer is "yes", so we have one additional piece of information about the experiment. We do not have any other additional information.

The question is: what probability should we now assign to the event that the dart landed in region A?

It should be emphasized that conditional probabilities are not different from any other probabilities. Every probability is conditional on *some* information! Mathematicians use the word "conditional" here merely to emphasize the way in which your knowledge has *changed* from what you started with.

We are currently thinking about physical probabilities for an actual experiment. There is simple formula for the conditional probability .

Fact 4.1 (The conditional probability formula). Let A and B be physical events for some experiment. If $\mathbf{P}(B) \neq 0$,

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}, \tag{4.1}$$

or equivalently

$$\mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A \mid B). \tag{4.2}$$

If you construct a probability model for the experiment, you will define theoretical probabilities for the abstract events in your model. The probabilities $\mathbf{P}(A)$ in a valid model will be the appropriate probabilities based on your initial knowledge. In the case of a probability model the conditional probability formula holds by definition, but the definition must follow the physical rule.

Definition 4.2 (The conditional probability formula). Let A and B be events in a model for some experiment. If $\mathbf{P}(B) \neq 0$, the $\mathbf{P}(A \mid B)$ is defined by equation (4.1).

Both equation (4.1) and equation (4.2) are useful forms of the conditional probability formula. We might call equation (4.2) the "mutiplied-through" form of the conditional probability formula.

Much of our real-world knowledge about people or things can be regarded as approximate forms of conditional probability assessments! For example, if we are getting ready to deal with a person's possible reactions to some situation that might occur, we may be using a thought process analogous to (4.2). B would represent the event that the situation occurs, and A would represent the event that the person reacts in a particular way.

Example 4.3 (Computing conditional dart probabilities). Let's go back to the experiment of throwing darts (section 3.6). We said in section 3.6 that if the thrower is very inaccurate, one might use a model in which the probability of hitting any region of the target is proportional to the *area* of that region. (We also agreed that if the thrower misses the target entirely then we will ignore the throw.)

In the model for this experiment, we will let Ω be the region in the plane representing the target. For any region A inside the target, in our model the set A is used to represent the event that the dart lands in A. We are assuming that

$$\mathbf{P}(A) = \frac{\mathbf{area}(A)}{\mathbf{area}(\Omega)}. \tag{4.3}$$

Now consider the target shown in Figure 4.1. Then $\mathbf{P}(A \mid B)$ is the probability that a dart which hits B also hits A. Of course this can only happen if the dart lands in $A \cap B$.

If you tell me that the dart hit B, and nothing else, then I don't know in what part of B the dart landed. So to understand $\mathbf{P}(A \mid B)$ it seems appropriate to think of a new experiment, in which the target is B. In this new experiment we ignore any throw which does not hit B. By the same reasoning which made equation (4.3) seem valid, we now assume that

$$\text{probability to hit } A \cap B \ = \frac{\mathbf{area}(A \cap B)}{\mathbf{area}(B)}.$$

So we suspect that $\mathbf{P}(A \mid B)$ is simply given by

$$\mathbf{P}(A \mid B) = \frac{\mathbf{area}(A \cap B)}{\mathbf{area}(B)}. \tag{4.4}$$

Is equation (4.4) consistent with the general conditional probability formula given in equation (4.1)?

Well, using equation (4.3),

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

So yes, (4.4) is exactly what equation (4.1) tells us.

4.2 Why the conditional probability formula holds

Like additivity, the conditional probability formula is a fundamental rule in probability. In this section we will spend some time justifying this formula.

Thinking about information To show that equation (4.1) is correct physically, think first about events S_1 and S_2 which are *subsets* of B. Physically, this means that S_1 and S_2 are special cases of event B.

Suppose that $\mathbf{P}(S_1) = \mathbf{P}(S_2)$. This means that, based on everything you know initially, these two events are equally likely.

If someone tells you now that B occurred, does that extra piece of information give you any reason to believe that one of the events S_1, S_2 is now more likely than the other? It is hard to see how that could be the case. The extra information does not treat either event differently. You may know ways in which events S_1 and S_2 differ from each other physically, but you already knew that when you initially decided that S_1 and S_2 had the same probability.

Now let's think more generally, about any events S_1 and S_2 which are subsets of B. If we learn that B occurred, that extra information does not seem to treat either event differently. So it is plausible that the *relative* sizes

of $\mathbf{P}(S_1 \mid B)$ and $\mathbf{P}(S_2 \mid B)$ should be the same as the relative sizes of $\mathbf{P}(S_1)$ and $\mathbf{P}(S_2)$. In other words, it is plausible that for some constant c,

$$\mathbf{P}(S \mid B) = c\,\mathbf{P}(S) \tag{4.5}$$

for every event S which is a subset of B.

Applying equation (4.5) with $S = B$, we see that $\mathbf{P}(B \mid B) = c\mathbf{P}(B)$, and of course $\mathbf{P}(B \mid B) = 1$! Thus $c = 1/\mathbf{P}(B)$, and so we have

$$\mathbf{P}(S \mid B) = \frac{\mathbf{P}(S)}{\mathbf{P}(B)} \tag{4.6}$$

for every event S which is a *subset* of B.

That's the story when the event S is a subset of B. What about the general case, when we are interested in an event A which need not be a subset of B?

For any event A, we can see that A is the union of the two parts, the part $S = A \cap B$ that is contained in B and the part $A - B$ that is outside B: that is,

$$A = S \cup (A - B).$$

The rules of probability apply to conditional probability, so conditional probability is additive:

$$\mathbf{P}(A \mid B) = \mathbf{P}(S \mid B) + \mathbf{P}(A - B \mid B), \tag{4.7}$$

Since $S = A \cap B$ is a subset of B, we know by equation (4.6) that

$$\mathbf{P}(S \mid B) = \frac{\mathbf{P}(S)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}. \tag{4.8}$$

Also, since B and $A - B$ are disjoint, they are mutually exclusive. If B happens then $A - B$ certainly does not happen. That is,

$$\mathbf{P}(A - B \mid B) = 0.$$

Then equation (4.7) says that $\mathbf{P}(A \mid B) = \mathbf{P}(S \mid B)$. Hence equation (4.8) tells us that equation (4.1) holds, which is the general conditional probability formula.

Thinking about frequencies We've given one justification for the conditional probability formula. Now let's give another, this time using frequencies.

The frequency interpretation applies to conditional probabilities, since it applies to all probabilities. So we can use the frequency interpretation (Probability Fact 1.1) to derive a formula for $\mathbf{P}(A \mid B)$.

The frequency interpretation says that, when the experiment is performed, $\mathbf{P}(B)$ is roughly equal to the fraction of the time that we see B occur. The frequency interpretation for probabilities given B tells us that if we only look at times when B occurs, the fraction of the time that we see A occur is $\mathbf{P}(A \mid B)$.

If we want to know the fraction of the time that we see both A and B occur, we can find this number as a "fraction of a fraction": take the fraction of the time that B occurs, and multiply that fraction by the fraction of **those** times when we see A occur. Replacing the fractions by the corresponding probabilities, this says that

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

This is the multiplied-through form of the conditional probability formula, equation (4.2).

This completes the derivation of the conditional probability formula from the frequency interpretation of probability. The next example just writes out the same "fraction of a fraction" computation more explicitly.

Example 4.4 (Writing out the fractions). $\mathbf{P}(A \mid B)$ is the frequency with which A occurs in the situation in which the original experiment was performed **and** B occurred. The frequency with which A occurs in *this* experimental situation is the right approximation to $\mathbf{P}(A \mid B)$.

To get this frequency, repeat the original experiment many times, say N times, but only record results for those times when the physical event B occurs. The fraction of *those* recorded times for which the physical event A occurs will give us a good approximation to $\mathbf{P}(A \mid B)$.

We are assuming that N is large. Suppose that during the N repetitions of the experiment, the physical event B occurred M times.

By the frequency interpretation for the unconditional probability, we know it is likely that

$$\frac{M}{N} \approx \mathbf{P}(B). \tag{4.9}$$

We are only interested here in the case that $\mathbf{P}(B) > 0$, so that B can happen. When $\mathbf{P}(B) > 0$, equation (4.9) tells us that M will be large when N is large.

Suppose that during the M times that B occurred, the physical event A occurred L times.

By the frequency interpretation for the conditional probability, it is likely that

$$\frac{L}{M} \approx \mathbf{P}(A \mid B). \tag{4.10}$$

Thus

$$\mathbf{P}(A \mid B) \approx \frac{\frac{L}{N}}{\frac{M}{N}}. \tag{4.11}$$

Let's look at the fraction L/N. L counts the times when B occurred *and* A occurred. Thus, by the frequency interpretation for the unconditional probability, we know it is likely that

$$\frac{L}{N} \approx \mathbf{P}(A \cap B). \tag{4.12}$$

Apply equations (4.9) and (4.12) to equation (4.11). This allows us to conclude:

$$\mathbf{P}(A \mid B) \approx \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}. \tag{4.13}$$

Equation (4.13) is based on approximations that become more and more accurate as the number of trials increases. Thus the approximation in equation (4.13) tells us that $\mathbf{P}(A \mid B)$ and $\frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$ must be equal, and so equation (4.1) holds.

4.3 Using the conditional probability formula

Exercise 4.1. In the experiment described in Exercise 2.19, you are choosing a jelly bean from a bowl containing 75 yellow beans, 53 red beans, 27 purple beans, and 18 green beans.

Let A be the event that the selected bean is yellow, red or green. Let B be the event that the selected bean is red, purple or green.

Find $\mathbf{P}(A \mid B)$, using the conditional probability formula.

[Solution]

Exercise 4.2. In the setting of Exercise 4.1, when the experiment consists of choosing *two* jelly beans in succession, let B be the event that the first bean chosen is red.

(i) Find $\mathbf{P}(B)$.

(ii) From the description of the experiment, R is the event that both beans chosen are red. Thus $\mathbf{P}(R \mid B)$ is the probability of choosing a a red bean in the second selection, from the *remaining* beans, after the first selection resulted in a red bean. Use this observation to find $\mathbf{P}(R \mid B)$ by applying Theorem 2.22 to an appropriate population, without using the conditional probability formula.

(iii) Fnd $\mathbf{P}(R)$ using the multiplied-through form of the conditional probability formula, equation (4.2).

Check that your answer agrees with the probability found in Exercise 2.21.

[Solution]

Remark 4.5 (Conditional probabilities are often simpler). The calculation in part (ii) of Exercise 4.2 illustrates a general fact: a conditional probability often holds in an experimental situation which is simpler than the original model. While calculating such a conditional probability, we temporarily live in the simpler model, and forget everything else. Then, as in Exercise 4.2, we can return to the original model, and use the conditional probability we have found to calculate something there.

Exercise 4.3. In the experiment of Exercise 4.2, suppose you learn that the second jelly bean chosen was purple. What is the probability that the first jelly bean chosen was also purple?

[Solution]

Exercise 4.4. Let A, B be events with $\mathbf{P}(B) \neq 0$. Show that

$$\mathbf{P}(A \mid B) = \mathbf{P}(A \cap B \mid B). \tag{4.14}$$

[Solution]

Just for fun, the next exercise takes the "fraction of a fraction" idea to a new level.

Exercise 4.5 (Telescoping conditional probabilities). Simplify

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

It is assumed that $\mathbf{P}(A)$, $\mathbf{P}(A \cap B)$, and $\mathbf{P}(A \cap B \cap C)$ are nonzero.

[Solution]

The mathematical properties of conditional probability mirror the way we think. The next exercises illustrate this, and provide some practice in manipulating conditional formulas.

Exercise 4.6 (Conditioning on an additional event). Suppose that B, C are events for some probability model. Suppose that $\mathbf{P}(B) \neq 0$. For any event D, define $\mathbf{Q}(D) = \mathbf{P}(D \mid B)$. This is just to simplify notation.

When using $\mathbf{Q}$ as your distribution, the fact that B occurred is "built into" your probability model.

Assume that $\mathbf{P}(B \cap C) \neq 0$.

Show that $\mathbf{Q}(C) \neq 0$, and that for any event A we have

$$\mathbf{Q}(A \mid C) = \mathbf{P}(A \mid B \cap C). \tag{4.15}$$

[Solution]

The event $B \cap C$ is the event that B occurred *and* C occurred. Thus Equation (4.15) is exactly what we expect from the idea that a conditional probability uses additional information, since in calculating $\mathbf{Q}(A \mid C)$ we are adding still more information to the extra information that we already used to calculate $\mathbf{Q}(A)$.

Exercise 4.7 (Conditioning on stronger information). Suppose that B, C are events for some probability model. Suppose that $\mathbf{P}(B) \neq 0$. For any event D, define $\mathbf{Q}(D) = \mathbf{P}(D \mid B)$.

Let C be an event with $\mathbf{P}(C) > 0$, such that C is a subset of B.

Show that $\mathbf{Q}(C) \neq 0$, and that for any event A we have

$$\mathbf{Q}(A \mid C) = \mathbf{P}(A \mid C). \tag{4.16}$$

[Solution]

Exercise 4.8. Let A, B be events in some experiment.

Let C be an event such that $\mathbf{P}(B \cap C) \neq 0$. Show that

$$\mathbf{P}(A \cap C \mid B \cap C) = \mathbf{P}(A \mid B \cap C). \tag{4.17}$$

[Solution]

4.4 Total probability

The name of the next theorem seems a little pretentious, since the statement is a simple consequence of additivity and the conditional probability formula. However, breaking a problem up into cases is a fundamental technique, and is frequently used.

Theorem 4.6 (The Law of Total Probability). Let $D_1, \ldots, D_k$ be disjoint events with union D, and let M be an event. Then

$$\mathbf{P}(M \cap D) = \sum_{i=1}^{k} \mathbf{P}(D_i)\, \mathbf{P}(M \mid D_i). \tag{4.18}$$

In this equation, it appears that we must assume that $\mathbf{P}(D_i) > 0$, so that $\mathbf{P}(M \mid D_i)$ will be defined. However, we can use the equation in all situations, with the following convention: if $\mathbf{P}(D_i) = 0$ then we simply interpret the whole term $\mathbf{P}(D_i)\, \mathbf{P}(M \mid D_i)$ as zero.

Proof. We checked in Exercise 2.26 that intersection distributes over union, so we know that

$$M \cap D = \bigcup_{i=1}^{k} M \cap D_i. \tag{4.19}$$

(Actually, one likely doesn't even think about "the distributive property" when writing down equation (4.19). Instead one can just think about cases, i.e. think about the possible ways that $M \cap D$ can happen: the possible ways are $M \cap D_1, \ldots, M \cap D_k$.)

Since $D_1, \ldots, D_k$ are disjoint, additivity gives

$$\mathbf{P}(M \cap D) = \sum_{i=1}^{k} \mathbf{P}\left(M \cap D_i\right). \tag{4.20}$$

Consider a term $\mathbf{P}(M \cap D_i)$ on the right side of equation (4.20). If $\mathbf{P}\left(D_i\right) \neq 0$ then

$$\mathbf{P}(M \cap D_i) = \mathbf{P}(D_i)\mathbf{P}(M \mid D_i) \tag{4.21}$$

by the conditional probability formula given in equation (4.2).

If $\mathbf{P}(D_i) = 0$, then, since $M \cap D_i \subset D_i$ we also have $\mathbf{P}(M \cap D_i) = 0$. So equation (4.21) holds with $\mathbf{P}(D_i)\mathbf{P}(M \mid D_i)$ replaced by zero.

Substituting for $\mathbf{P}(M \cap D_i)$ throughout equation (4.20) gives equation (4.18). □

When applying this theorem, we typically look for cases D_i where we know how to find $\mathbf{P}\left(M \mid D_i\right)$. In this way we can break up problems into simpler parts.

Often the event M is such that $M \subset D_1 \cup \ldots \cup D_k$. Then $M \cap D = M$, so equation (4.18) becomes

$$\mathbf{P}(M) = \sum_{i=1}^{k} \mathbf{P}\left(D_i\right) \mathbf{P}\left(M \mid D_i\right), \tag{4.22}$$

The following simple corollary to the Law of Total Probability is sometimes convenient.

Corollary 4.7. For some experiment, let $D_1, \ldots, D_k$ and M be events. Let D be the event that at least one of $D_1, \ldots, D_k$ occurs.

Suppose that at most one of the events $D_1, \ldots, D_k$ can occur, and that $\mathbf{P}(M \mid D_i) = p$ for $i = 1, \ldots, k$, where p is some number.

Then $\mathbf{P}(M \mid D) = p$.

Exercise 4.9. Prove Corollary 4.7.

[Solution]

Does Corollary 4.7 seem physically obvious? (Think of a hall with many doors, and suppose that for *every* door i, a hungry tiger waits behind that door with probability p. Given that you must pass out through one of the doors, is it hard to calculate your chance of survival?)

Example 4.8 (Sampling without replacement). Consider the setting described in Exercises 4.1 and 4.2, where the bowl contains 75 yellow beans, 53 red beans, 27 purple beans, and 18 green beans.

As in Exercise 4.2, think about an experiment with two steps. In the first step we stir the bowl, and then select one jelly bean randomly, with no jelly bean in the bowl favored. We note the color of the chosen jelly bean, but do *not* replace the jelly bean in the bowl.

In the second step, we stir the bowl again, and then select a second jelly bean, again with no jelly bean favored.

Let A_1 be the event that the bean selected in step 1 is yellow or red. Let B_2 be the event that the bean selected in step 2 is yellow or green.

Goal: As an exercise, let's find $\mathbf{P}(A_1 \cap B_2)$.

Because the bowl is stirred, we are confident that the only way the first step can affect the second step is by altering the numbers of jelly beans of each color in the bowl.

We are going to use the conditional probability formula, conditioning on A_1. By the multiplied-through form of the conditional probability formula,

$$\mathbf{P}(A_1 \cap B_2) = \mathbf{P}(A_1)\,\mathbf{P}(B_2 \mid A_1).$$

Let Y_1 be the event that the first bean selected is yellow, and let R_1 be the event that the first bean selected is red. There are two possible cases for A_1, namely Y_1 and R_1. That is, $A_1 = Y_1 \cup R_1$.

So we think about applying equation (4.18) (the law of total probability), with $D = A_1$, $D_1 = Y_1$ and $D_2 = R_1$.

Thus, by the law of total probability,

$$\mathbf{P}(A_1 \cap B_2) = \mathbf{P}(Y_1)\,\mathbf{P}(B_2 \mid Y_1) + \mathbf{P}(R_1)\,\mathbf{P}(B_2 \mid R_1).$$

Remember, the initial numbers of jelly beans in the bowl are as follows: 75 yellow beans, 53 red beans, 27 purple beans, and 18 green beans.

By Theorem 2.22, $\mathbf{P}(Y_1) = 75/173$ and $\mathbf{P}(R_1) = 53/173$.

We can also use Theorem 2.22 to find $\mathbf{P}(B_2 \mid Y_1)$ and $\mathbf{P}(B_2 \mid R_1)$. The point here is that step 2 of the experiment is a "self-contained" sampling experiment, that is, a sampling experiment that can be considered by itself.

Given Y_1: Given Y_1, we know that the bowl contains 74 yellow beans and 18 green beans, and 172 beans altogether. Thus there are 92 beans in the bowl that are yellow or green. Thus by Theorem 2.22, $\mathbf{P}(B_2 \mid Y_1) = 92/172$.

Given R_1: Similarly, given R_1, we know that the bowl contains 75 yellow beans and 172 beans altogether, and there are 93 beans in the bowl that are yellow or green. Thus by Theorem 2.22, $\mathbf{P}(B_2 \mid R_1) = 93/172$.

The rest is arithmetic.

$$\mathbf{P}(A_1 \cap B_2) = \frac{75}{173}\frac{92}{172} + \frac{53}{173}\frac{93}{172}.$$

Exercise 4.10. Solve part (ii) of Exercise 3.3 again, applying the Law of Total Probability (Theorem 4.6).

In the setting of Exercise 3.3, let N be the event that the lost coin has not been found after searching two-thirds of Alice's section, half of Bob's section, and three-quarters of Clancy's section.

The event A is defined as the event that the lost coin is located somewhere in Alice's section, and you are asked to find $\mathbf{P}(A \mid N)$.

[Solution]

Exercise 4.11. There are two boxes on the table. Box 1 contains 10 red balls and 30 green balls. Box 2 contains 50 red balls and 10 green balls. Our experiment takes place in two steps.

(1.) First, toss an unfair coin. The probability of a head is 2/3 for this unfair coin.

(2.) If the result of the coin toss is a head, choose one ball at random from Box 1. Otherwise, choose one ball at random from Box 2. All the balls in Box 1 have the same chance to be selected. All the balls in Box 2 have the same chance to be selected.

Let A be the event that green ball is selected.

(i) Find $\mathbf{P}(A)$, using the following sample space argument.

Take Ω to be the set of all pairs (i, b), where i is the number of the box that is chosen, and b identifies the ball that is chosen from Box i. Let C_1 be the event that Box 1 is chosen, and let C_2 be the event that Box 2 is chosen. You may assume from the physical description that every outcome in C_1 has the same probability, and every outcome in C_2 has the same probability. The physical description also tells us the values of $\mathbf{P}(C_1)$ and $\mathbf{P}(C_2)$. Do not use the conditional probability formula or the law of total probability.

(ii) Find $\mathbf{P}(A)$ again, using the law of total probability.

[Solution]

We will return to the next exercise in Example 9.15. There we will see how to use random variable concepts to obtain more information.

Exercise 4.12 (Choosing from overlapping intervals). A fair coin is tossed. Suppose that if the result of the coin toss is a head, a point is chosen at random from $[0, 3]$, with no point favored. If the result of the coin toss is a tail, a point is chosen at random from $[2, 4]$, with no point favored. (Uniform distributions on continuous intervals are discussed in Section 3.3.)

Let J be an interval of the real line, and let A be the event that the chosen point is in J. Using the Law of Total Probability, find $\mathbf{P}(A)$ in each of the following four cases: (i) $J \subset [0, 2)$, (ii) $J \subset [2, 3)$, (iii) $J \subset [3, 4]$, (iv) J disjoint from $[0, 4]$.

You are not required to define a sample space for this experiment. In your solution you can simply work with the laws of probability, without specifying a sample space.

Notice that if you want to have a sample space that represents *everything* that happens in the two steps of the experiment, it will be a bit more complicated than usual.

[Solution]

4.5 The theorem of Bayes

Theorem 4.9 (Bayes). Let A and B be events for some probability model, such that $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$. Then

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(B)}. \tag{4.23}$$

Theorem 4.9 is an immediate consequence of the conditional probability formula, applied twice. First write $\mathbf{P}(A \cap B)$ as $\mathbf{P}(A)\mathbf{P}(B \mid A)$ using equation (4.2). Then find $\mathbf{P}(A \mid B)$ using equation (4.1).

The formula in equation (4.23) was found by Thomas Bayes, and published in a posthumous volume of his work in 1763. Although this formula bears his name, the same formula was independently found by Laplace. Despite its simplicity, the formula is frequently used, often repetitively as experimental results are accumulated.

If we think of A as describing a "cause", and B as describing an "effect" due to this cause, we might think of the formula of Bayes as showing how to calculate the probability of a possible cause when a certain effect is observed.

The quantity $\mathbf{P}(B)$ in the denominator of equation (4.23) can often be calculated using the Law of Total Probability.

The number $\mathbf{P}(A)$ is sometimes called a "prior" probability, meaning the probability of A *before* an experiment takes place, while $\mathbf{P}(A \mid B)$ is the "posterior" probability of A, meaning that it is the probability of A *after* the event B is observed in the experiment.

Notice that we need to have some idea of the value of $\mathbf{P}(A)$ to use Bayes.

Let's write out a form of equation (4.23) using the Law of Total Probability. Let an event M be a subset of the union of disjoint events $D_1, \ldots, D_k$.

Suppose that we observe M, and we wonder which of the events D_j occurred. By equation (4.22),

$$\mathbf{P}(D_j \mid M) = \frac{\mathbf{P}(D_j \cap M)}{\mathbf{P}(M)} = \frac{\mathbf{P}(D_j)\,\mathbf{P}(M \mid D_j)}{\sum_{i=1}^{k} \mathbf{P}(D_i)\,\mathbf{P}(M \mid D_i)}. \tag{4.24}$$

In ordinary life we frequently use reasoning similar to the Theorem of Bayes. Consider the following.

Example 4.10 (An everyday mystery). Grandma has just finished baking one of her delectable cherry pies. She places it in an open window to cool. Shortly thereafter, she observes that the pie is missing. There are several people who may have passed her window during the interval. Only one, however, has an extreme fondness for pie. Rolling pin in hand, Grandma knows where to focus the next stage of her investigation.

Exercise 4.13 (Putting some numbers on Example 4.10). In the situation of Example 4.10, the pie was only in the window for a short period of time. Suppose that there are only three people who could have passed by Grandma's window during this time period: Alice, Brandon, and Clyde. Let A be the event that Alice passed by, and let B and C the corresponding events for Brandon and Clyde. Grandma thinks it is very unlikely that two people passed her window during this period, so she considers these events to be disjoint.

Grandma originally had no reason to think any of the three people is more likely than the others to pass by. She sets $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = \delta$, where δ is some positive number. Here $\mathbf{P}$ represents the probability that Grandma would have assigned to an event, *before* she discovers that her pie is missing.

Let T be the event that the pie in the window is taken. Alice and Brandon are highly reliable and have never shown any tendency to eat excessively large quantities of pie. Clyde, on the other hand, has a bad track record. Based on past events, Grandma sets $\mathbf{P}(T \mid A) = .01$, $\mathbf{P}(T \mid B) = .01$, and $\mathbf{P}(T \mid C) = .5$.

Calculate $\mathbf{P}(C \mid T)$.

[Solution]

Grandma can deal with her problem without using numbers, of course. In other situations the precision of mathematical calculation may be needed, as the next exercise illustrates.

Exercise 4.14 (A positive result in a test for disease). A rare but serious disease is present in approximately .01% of the people in a large population, i.e. a fraction 1/10000 of the population have the disease.

There is a test for this disease. A positive result for this test is an indication of disease.

The test is good but not perfect. When a healthy person is tested, the probability of a false positive is .01, i.e. one percent.

For simplicity, assume that the test never misses an actual case of the disease. That is, assume that the probability of a false negative is zero.

Suppose that someone is randomly selected from the population and tested. The result of the test is positive. Find the probability that the person has the disease.

[Solution]

Exercise 4.15. In the experiment of Exercise 4.11, suppose you learn that a red ball was selected. Find the probability that the toss of the coin for this experiment produced a head.

[Solution]

Exercise 4.16. Return to the experiment of Exercise 4.12.

Let B be the event that a point in the interval $(2, 3)$ is obtained. Find $\mathbf{P}(H \mid B)$.

[Solution]

Exercise 4.17 (Bayes and the chosen coin). (i) To practice using the theorem of Bayes, let's model a situation in which one of two coins is randomly chosen and then tossed. As usual when tossing a coin, we'll think of getting a head as "success" for the toss. The coins are named coin 1 and coin 2. Coin 1 has success probability 2/5 and coin 2 has

success probability 4/7. Suppose that each of the two coins has the same probability to be chosen.

After the coin was chosen and tossed, you find that the result was a tail. You don't know which coin was tossed. Find the probability that coin 2 is the coin that was tossed.

Before calculating this probability, decide whether you think the probability is greater than 1/2 or less than 1/2.

(ii) Suppose now that in addition to coin 1 and coin 2 we also have coin 3. Like coin 2, this coin also has success probability 4/7. A new experiment is carried out, in which one of these three coins is selected with equal probability, and the selected coin is tossed. The result is a tail. Find the probability that the selected coin had success probability equal to 4/7.

[Solution]

The next exercise gives you a chance to practice with algebra and inequalities. Even if you don't do the problem, think about equation (4.25) and see if it agrees with your own feelings about physical probabilities.

Exercise 4.18. Consider two coins, coin a and coin b. Coin a has success probability p_a, and coin b has success probability p_b, where $p_a > p_b$. That is, coin a is luckier than coin b.

Suppose now that one of these two coins is randomly selected. Let A be the event that coin a is selected, and let B be the event that coin b is selected.

Assume that $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$, so either coin could be chosen. If we don't choose coin a then we must choose coin b. So of course $\mathbf{P}(A) + \mathbf{P}(B) = 1$.

After the selection, suppose that the selected coin is tossed. Let H be the event that this toss gives success. Show that $\mathbf{P}(H) > 0$ and

$$\mathbf{P}(A \mid H) > \mathbf{P}(A). \tag{4.25}$$

Thus obtaining a success with the coin makes you more confident that it is the lucky coin.

[Solution]

4.6 Tree diagrams

Pictures are helpful in any field of mathematics. In probability problems it can be helpful to represent events using a "tree of possibilities", a.k.a. a tree diagram. Tree diagrams do not introduce any new concepts, but they can assist us in seeing what is going on, when a computation involves several events.

Drawing a tree diagram seems to be an art rather than a science, since the goal is to display ideas visually within a limited space. We can only make a few general remarks here, and then give some examples.

When using a tree diagram, we only need to draw the part of a tree that represents events which we are interested in.

There is no general rule about whether a tree diagram will be useful. When drawing your own diagram, just starting a tree may be enough to suggest how to actually approach the problem, and you can switch to using equations.

There is some standard terminology for describing tree diagrams. Every tree has a *root*. The *branches* spread out from the root. The trees in tree diagrams may be drawn upside down, with the root at the top, or lying on one side! We'll draw our diagrams here with the root on the left side.

Every branch has two ends. The ends of the branches are often called "nodes". The root is a node. A branch *begins* at one node and *ends* at another, and you have to remember which is which (the starting node is the one which is closer to the root).

In a probability tree diagram, the nodes represent events. The root of the tree represents Ω. The ending node of any branch represents an event which is a subset of the event represented by the starting node. So a branch represents *inclusion*. And the end of one branch can be the starting node for another branch. So tree diagrams can potentially be large.

If a branch starts with an event A and ends with an event B, then we often label the connecting branch with the conditional probability $\mathbf{P}(B \mid A)$. Each node along a chain of branches is always a subset of the earlier nodes in the chain. A key fact: the probability of a node which lies at the end of a chain of branches is equal to the product of the conditional probabilities

along the chain! (This follows from repeated use of the multiplied-through form of the conditional probability formula.)

Figure 4.2 shows a small tree diagram for Exercise 4.2. Notice that only the relevant parts of the tree are shown. Since this is such a simple situation, a tree diagram is not needed, but the picture does show how trees work. Note the informal labelling.

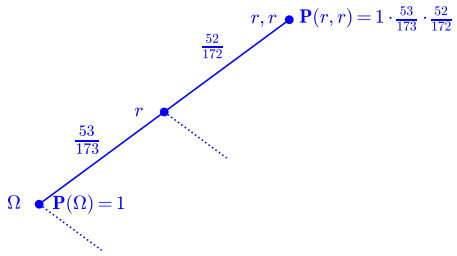


Figure 4.2: Obtaining two red jelly beans, one at a time (Exercise 4.2).

Let's change Exercise 4.2. Instead of calculating the probability of two red jelly beans, let's find the probability of winding up with a red and a green. Figure 4.3 shows a tree diagram for that calculation. To get the final answer for this problem, note that you add the probabilities of two events: getting a red and then a green, and getting a green and then a red. So you add the probabilities associated with two paths on the tree, and the final answer is

$$2 \cdot \frac{53 \cdot 18}{173 \cdot 172}.$$

Notice that in a tree diagram, every fork creates nodes that represent disjoint events. Thus nodes which are not on the same chain of branches necessarily represent disjoint events. That's why you add probabilities for

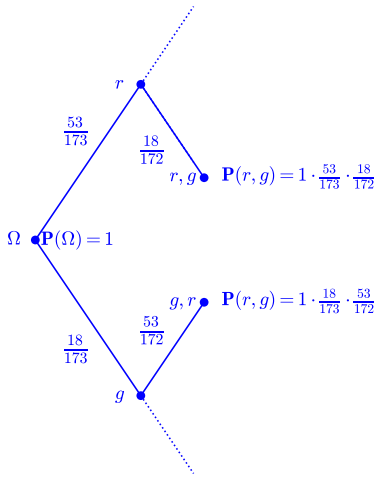


Figure 4.3: Obtaining one red and one green jelly bean, one at a time, in either order.

these nodes. Thus interpreting a tree diagram uses the Law of Total Probability, in an informal manner.

When looking at tree diagrams in probability books, you will notice various ways of labelling nodes in practical situations. Informality is the order of the day, and making ideas clear is the priority.

If a branch starts at a node called A and ends at a node representing $A \cap M$, you may see the ending node labelled with "M", rather than $A \cap M$. In other words, we often label a node using only the *additional* properties which distinguish it from the preceding node. However, to make sense of the diagram you should think of the ending node as representing the event $A \cap M$.

Example 4.11. Just for fun, let's make a tree diagram which is a bit bigger than the one shown in Figure 4.3. Think of randomly selecting jelly beans, one at a time, from a bowl containing two red jelly beans, one yellow jelly bean, and one green jelly bean. We want the red ones, so we will stop as soon as we obtain both red beans!

Let A_n be the event that it takes exactly n tries to get both red ones. Here n might be 2, 3, or 4. Suppose we would like to find $\mathbf{P}(A_n)$.

Figure 4.4 is a tree diagram for this problem, where getting a red bean is represented by an upward branch, getting a yellow bean is represented by a horizontal branch, and getting a green bean is represented by a downward branch. To find $\mathbf{P}(A_n)$, add up the probabilities for the paths with length n. This gives $\mathbf{P}(A_2) = 1/6$, $\mathbf{P}(A_3) = 1/3$ and $\mathbf{P}(A_4) = 1/2$.

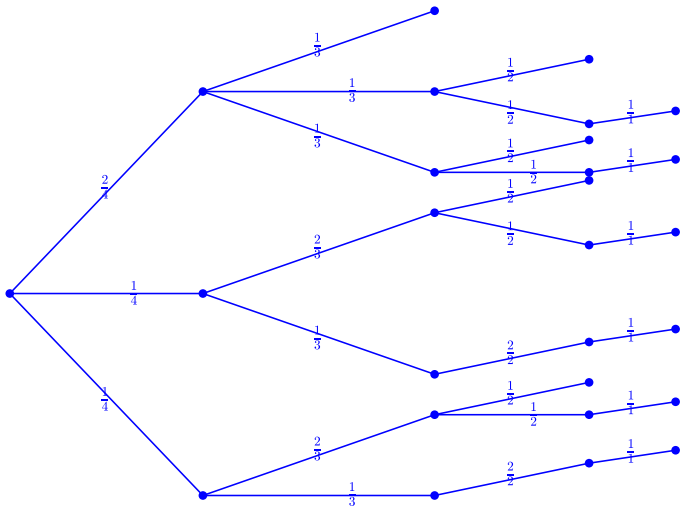


Figure 4.4: Sampling until two red jelly beans are obtained, starting with 2 red, 1 yellow, and 1 green. Upward indicates a red bean, horizontal indicates a yellow bean, and downward indicates a green bean. There is one path of length two, four paths of length three, and six paths of length four.

A nice example of a tree diagram is given in [12], for the Monty Hall problem (see Section 6.2 for information about Monty Hall).

4.7 Solutions for Chapter 4

Solution (Exercise 4.1). Since $|B| = 53 + 27 + 18 = 98$ and $|\Omega| = 75 + 53 + 27 + 18 = 173$, $\mathbf{P}(B) = 98/173$ by Theorem 2.22.

$A \cap B$ is the event that the selected bean is red or green. Hence $\mathbf{P}(A \cap B) = 71/173$ by Theorem 2.22.

By the conditional probability formula, $\mathbf{P}(A \,|\, B) = (71/173)/(98/173) = 71/98$.

Solution (Exercise 4.2). There are 75 yellow beans, 53 red beans, 27 purple beans, and 18 green beans in the bowl.

(i) There are 173 beans in the bowl, and 53 of them are red. By Theorem 2.22,

$$\mathbf{P}(B) = \frac{53}{173}.$$

(ii) After the selection of a red bean, the bowl contains 75 yellow beans, 52 red beans, 27 purple beans, and 18 green beans. This is the setting for the second selection, given that B occurred. So we are solving part (i) again, but in a new setting. By Theorem 2.22,

$$\mathbf{P}(R \,|\, B) = \frac{52}{172}.$$

(iii) By the mutiplied-through version of the conditional probability formula, equation (4.2),

$$\mathbf{P}(R) = \mathbf{P}(B)\mathbf{P}(R \,|\, B) = \frac{53}{173}\,\frac{52}{172}.$$

Notice that this agrees with the probability found in Exercise 2.21, which we solved without using conditional probabilities.

Solution (Exercise 4.3). Let P_1 be the event that the first jelly bean selected was purple, and let P_2 be the event that the second jelly bean selected was purple. We would like to find $\mathbf{P}(P_1 \,|\, P_2)$.

We can use a sample space consisting of all pairs of jelly beans (j_1, j_2), where $j_1 \neq j_2$. Since there are $75 + 53 + 27 + 18 = 173$ jelly bean altogether, the sample space contains $173 \cdot 172$ sample points. The physical description of the experiment tells us that all sample points are equally likely.

Let p denote the probability of a sample point. Then

$$p = \tfrac{1}{173 \cdot 172}.$$

The number of purple jelly beans is 27, and the number of non-purple jelly beans is $173 - 27 = 146$.

Since there are 27 purple jelly beans, $P_1 \cap P_2$ contains $27 \cdot 26$ sample points.

We have

$$\mathbf{P}(P_1 \cap P_2) = (27 \cdot 26)p.$$

Using the same sample space, a sample point in P_2 consists of all sample points (j_1, j_2), such that j_2 is purple. There are 27 choices for j_2, and for each possible j_2 there are 172 choices for j_1. Thus P_2 contains $172 \cdot 27$ sample points, and so $\mathbf{P}(P_2) = (172 \cdot 27)p$.

Hence

$$\mathbf{P}(P_1 \mid P_2) = \frac{\mathbf{P}(P_1 \cap P_2)}{\mathbf{P}(P_2)} = \frac{(27 \cdot 26)p}{(172 \times 27)p} = \frac{27 \cdot 26}{172 \times 27} = \frac{13}{86}. \tag{4.26}$$

Thinking backwards As an alternative method, we might ignore the physical times at which the steps occur, and just think about the possible pairs of jelly beans (j_1, j_2) that are obtained. One could think about building a pair by (mentally) selecting j_2 first, and then selecting j_1. There are 173 choices for j_2, and then, having chosen j_2, there are 172 choices for j_1.

Thus to find $\mathbf{P}(P_1 \mid P_2)$, think that a purple jelly bean has already been chosen for j_2. $\mathbf{P}(P_1 \mid P_2)$ is the probability that the choice of j_1 now gives a purple jelly bean. Thus

$$\mathbf{P}(P_1 \mid P_2) = \frac{26}{172} = \frac{13}{86}, \tag{4.27}$$

as before.

Notice that in equation (4.26) we apply the conditional probability formula to obtain $\mathbf{P}(P_1 \mid P_2)$, while in equation (4.27) we think of a physical situation in which P_2 has occurred, and then perform a calculation to find $\mathbf{P}(P_1)$ in that situation.

Solution (Exercise 4.4). By the conditional probability formula (equation (4.1)),

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

The first and last equalities hold by the conditional probability formula. The middle equality holds because $B \cap B = B$.

Solution (Exercise 4.5).

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

Cancelling does the rest.

Solution (Exercise 4.6). By definition,

$$\mathbf{Q}(C) = \frac{\mathbf{P}(C \cap B)}{\mathbf{P}(B)}.$$

Since we assume that $\mathbf{P}(C \cap B) > 0$, $\mathbf{Q}(C) > 0$ also.

By definition,

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

Solution (Exercise 4.7). Since $C \cap B = C$, everything follows from Exercise 4.6.

Solution (Exercise 4.8).

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

The first and last equalities hold by the conditional probability formula. The middle equality holds because $C \cap C = C$.

Solution (Exercise 4.9). By total probability,

$$\mathbf{P}(M \cap D) = \sum_{i=1}^{k} \mathbf{P}\left(D_i\right) \mathbf{P}\left(M \mid D_i\right) = \sum_{i=1}^{k} \mathbf{P}\left(D_i\right) p = p\mathbf{P}(D).$$

Dividing by $\mathbf{P}(D)$ gives the result.

Solution (Exercise 4.10). As in the solution for part (i) of Exercise 3.3, $\mathbf{P}(A) = 300/600 = 1/2$.

Suppose that the situation described in part (ii) of Exercise 3.3 holds. That is, Alice has searched two-thirds of her section, Bob has searched half of his section, Clancy has searched three-quarters of his section. We use probabilities based on this information.

Let N be the event that the coin has not yet been found.

Let A be the event that the coin is located in Alice's interval, let B be the event that the coin is located in Bob's interval, and let C be the event that the coin is located in Clancy's interval. The events A, B, C are disjoint, and $A \cup B \cup C = \Omega$. Part (ii) of Exercise 3.3 asks us to find $\mathbf{P}(A \mid N)$.

By equation (4.22),

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

Let I be Alice's interval and let J the part of Alice's interval which has not yet been searched. Then $\mathbf{P}(N \mid A)$ is simply the probability that the coin is located in J. Thus

$$\mathbf{P}(N \mid A) = \frac{\mathbf{length}(J)}{\mathbf{length}(I)} = \frac{1}{3}.$$

The values of $\mathbf{P}(N \mid B)$ and $\mathbf{P}(N \mid C)$ are found similarly.

This gives

$$\mathbf{P}(N) = \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{4}\right) = \frac{3}{8}.$$

We need to find $\mathbf{P}(A \mid N)$. By the Conditional Probability Formula,

$$\mathbf{P}(A \mid N) = \frac{\mathbf{P}(A \cap N)}{\mathbf{P}(N)} = \frac{\mathbf{P}(A)\mathbf{P}(N \mid A)}{\mathbf{P}(N)} = \frac{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)}{\frac{3}{8}} = \frac{4}{9}.$$

Solution (Exercise 4.11).

(i) The probability of a head is given to be 2/3.

Hence $\mathbf{P}(C_1) = 2/3$. There are 40 outcomes in C_1, each of equal probability. Hence every outcome in C_1 has probability $(2/3)(1/40)$.

$\mathbf{P}(C_2) = 1/3$. There are 60 outcomes in C_2, each of equal probability. Hence every outcome in C_2 has probability $(1/3)(1/60)$.

There are 30 outcomes in $A \cap C_1$, each with probability 1/60. Hence $\mathbf{P}(A \cap C_1) = 1/2$.

There are 10 outcomes in $A \cap C_2$, each of probability 1/180. Hence $\mathbf{P}(A \cap C_2) = 1/18$.

Thus $\mathbf{P}(A) = 1/2 + 1/18$.

(ii)

$$\mathbf{P}(A) = \mathbf{P}(C_1)\,\mathbf{P}(A \,|\, C_1) + \mathbf{P}(C_2)\,\mathbf{P}(A \,|\, C_2) = \frac{2}{3}\frac{30}{40} + \frac{1}{3}\frac{10}{60} = \frac{1}{2} + \frac{1}{18}.$$

Solution (Exercise 4.12). Let H be the event that the coin toss gives a head. Let T be the event that a tail is obtained. Since the coin is fair, $\mathbf{P}(H) = \mathbf{P}(T) = 1/2$.

Let J be subinterval of $[0, 4]$. Let A be the event that the chosen point is in A.

By the Law of Total Probability,

$$\mathbf{P}(A) = \mathbf{P}(H)\mathbf{P}(A \,|\, H) + \mathbf{P}(T)\mathbf{P}(A \,|\, T). \tag{4.28}$$

Case (i): $J \subset [0, 2)$ When H occurs, the point is chosen from $[0, 3)$ with uniform probability on $[0, 3)$. By equation (3.4),

$$\mathbf{P}(A \,|\, H) = \frac{\mathbf{length}(J)}{\mathbf{length}([0, 3))} = \frac{1}{3}\,\mathbf{length}(J). \tag{4.29}$$

When T occurs, the point is chosen from $[2, 4]$, so $\mathbf{P}(A \,|\, T) = 0$.

Substituting in equation (4.28),

$$\mathbf{P}(A) = \frac{1}{6}\,\mathbf{length}(J \cap [0, 3)) \tag{4.30}$$

Case (ii): $J \subset [2,3)$ When H occurs, the point is chosen from $[0,3)$ with uniform probability on $[0,3)$. By equation (3.4),

$$\mathbf{P}(A \mid H) = \frac{\mathbf{length}(J)}{\mathbf{length}([0,3))} = \frac{1}{3}\,\mathbf{length}(J). \tag{4.31}$$

When T occurs, the point is chosen from $[2,4]$ with uniform probability. By equation (3.4),

$$\mathbf{P}(A \mid T) = \frac{\mathbf{length}(J)}{\mathbf{length}([2,4])} = \frac{1}{2}\,\mathbf{length}(J). \tag{4.32}$$

Substituting in equation (4.28),

$$\mathbf{P}(A) = \frac{1}{2}\,\frac{1}{3}\,\mathbf{length}(J) + \frac{1}{2}\;\;\frac{1}{2}\,\mathbf{length}(J) = \frac{5}{12}\,\mathbf{length}(J). \tag{4.33}$$

Case (iii): $J \subset [3,4]$ When H occurs, the point is chosen from $[0,3)$, so $\mathbf{P}(A \mid H) = 0$.

When T occurs, the point is chosen from $[2,4]$ with uniform probability. By equation (3.4),

$$\mathbf{P}(A \mid T) = \frac{\mathbf{length}(J)}{\mathbf{length}([2,4])} = \frac{1}{2}\,\mathbf{length}(J). \tag{4.34}$$

Substituting in equation (4.28),

$$\mathbf{P}(A) = \frac{1}{4}\,\mathbf{length}(J). \tag{4.35}$$

Case (iv): J **disjoint from** $[0,4]$ In all cases, the point is chosen from $[0,4]$, so $\mathbf{P}(A) = 0$.

Solution (Exercise 4.13). Grandma considers that the events A, B, C are mutually exclusive. Let D be their union.

The physical statement of the problem tells us that Alice, Brandon and Clyde are the only people that could have taken the pie. Hence $T \subset D$, so $T \cap D = T$. By the Law of Total Probability (Theorem 4.6),

$$\begin{aligned}\mathbf{P}(T) = \mathbf{P}(A)\mathbf{P}(T \mid A) + \mathbf{P}(B)\mathbf{P}(T \mid B) + \mathbf{P}(C)\mathbf{P}(T \mid C)\\ = \delta(.01) + \delta(.01) + \delta(.5).\end{aligned} \tag{4.36}$$

By definition,

$$\mathbf{P}(C \mid T) = \frac{\mathbf{P}(T \cap C)}{\mathbf{P}(T)} = \frac{\mathbf{P}(C)\mathbf{P}(T \mid C)}{\mathbf{P}(T)} = \frac{\delta(.5)}{\delta(.52)} = \frac{50}{52}.$$

This number is close to one, and directs Grandma's attention to Clyde.

Note that in our solution we could have appealed to equation (4.24), but instead we simply repeated the derivation of that equation. This is often natural.

Solution (Exercise 4.14). Let A be the event that the person who is tested actually has the disease. Let B be the event that the test is positive.

Using equation (4.24) or its derivation,

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}\left(B \mid A^c\right)}. \tag{4.37}$$

We are given that $\mathbf{P}(A) = .0001$. Hence $\mathbf{P}(A^c) = .9999$.

We are given that $\mathbf{P}(B \mid A^c) = .01$ and $\mathbf{P}(B \mid A) = 1$.

Thus

$$\mathbf{P}(A \mid B) = \frac{.0001(1)}{.0001(1) + .9999(.01)} = \frac{1}{1 + 99.99} = \frac{1}{100.99}.$$

Remark 4.12 (The worst case for a positive test recipient). By equation (4.37),

$$\mathbf{P}(A \mid B) = \frac{1}{1 + \left(\frac{\mathbf{P}(A^c)}{\mathbf{P}(A)}\right)\left(\frac{\mathbf{P}(B \mid A^c)}{\mathbf{P}(B \mid A)}\right)}. \tag{4.38}$$

Notice that for any given values of $\mathbf{P}(A)$, $\mathbf{P}(A^c)$ and $\mathbf{P}(B \mid A^c)$, the quantity $\frac{\mathbf{P}(B \mid A^c)}{\mathbf{P}(B \mid A)}$ decreases when $\mathbf{P}(B \mid A)$ increases.

Thus the denominator of the fraction in equation (4.38) decreases when $\mathbf{P}(B \mid A)$ increases.

We conclude that $\mathbf{P}(A \mid B)$ increases when $\mathbf{P}(B \mid A)$ increases.

So taking $\mathbf{P}(B \mid A) = 1$ gives the largest possible value for $\mathbf{P}(A \mid B)$ when the other numbers are known. Thus in Exercise 4.14 we have calculated $\mathbf{P}(A \mid B)$ in the worst case.

Solution (Exercise 4.15). Let R be the event that a red ball was selected, and let H be the event that the coin which was tossed produced a head.

We wish to find $\mathbf{P}(H \mid R)$.

$$\mathbf{P}(H \mid R) = \frac{\mathbf{P}(H \cap R)}{\mathbf{P}(R)} = \frac{\mathbf{P}(H)\mathbf{P}(R \mid H)}{\mathbf{P}(H)\mathbf{P}(R \mid H) + \mathbf{P}\left(H^c\right)\mathbf{P}\left(R \mid H^c\right)} = \frac{\frac{1}{3}\frac{10}{30}}{\frac{1}{3}\frac{10}{30} + \frac{2}{3}\frac{50}{60}} = \frac{1}{6}.$$

Solution (Exercise 4.16). Using Bayes,

$$\mathbf{P}(H \mid B) = \frac{\mathbf{P}(H \cap B)}{\mathbf{P}(B)}.$$

Thus

$$\mathbf{P}(H \mid B) = \frac{\mathbf{P}(H)\mathbf{P}(B \mid H)}{\mathbf{P}(B)}.$$

By the solution to Exercise 4.12,

$$\mathbf{P}(H \mid B) = \frac{\frac{1}{2}\frac{1}{3}\,\mathbf{length}((2,3))}{\frac{5}{12}\,\mathbf{length}((2,3))} = \frac{2}{5}.$$

Solution (Exercise 4.17).

(i) The coins had equal chances of being chosen, and the result (a tail) is more likely if coin 1 was tossed. This makes it more likely that coin 1 was used, so we should expect that the probability that coin 2 was used is less than 1/2.

Let A be the event that coin 2 was used, and let B be the event that the result was a tail. We wish to calculate $\mathbf{P}(A \mid B)$.

Using Bayes,

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}(B \mid A^c)}$$
$$= \frac{\frac{1}{2}\left(\frac{3}{7}\right)}{\frac{1}{2}\left(\frac{3}{7}\right) + \frac{1}{2}\left(\frac{3}{5}\right)} = \frac{5}{12} < \frac{1}{2}.$$

(ii) Now we let A be the event that the selected coin had probability equal to 4/7, i.e. the event that coin 2 or coin 3 was used. Similarly to part (i),

we then have

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}(B \mid A^c)}$$
$$= \frac{\frac{2}{3}\left(\frac{3}{7}\right)}{\frac{2}{3}\left(\frac{3}{7}\right) + \frac{1}{3}\left(\frac{3}{5}\right)} = \frac{10}{17} > \frac{1}{2}.$$

Solution (Exercise 4.18).

$$\mathbf{P}(A \mid H) = \frac{\mathbf{P}(A \cap H)}{\mathbf{P}(H)} = \frac{\mathbf{P}(A)\mathbf{P}(H \mid A)}{\mathbf{P}(A)\mathbf{P}(H \mid A) + \mathbf{P}(B)\mathbf{P}(H \mid B)}.$$

That is,

$$\mathbf{P}(A \mid H) = \frac{\mathbf{P}(A)p_a}{\mathbf{P}(A)p_a + \mathbf{P}(B)p_b}. \tag{4.39}$$

Take a look at the final expression in equation (4.39). If we replace p_b by p_a, then the denominator gets bigger. So the fraction gets smaller. This tells us that

$$\mathbf{P}(A \mid H) > \frac{\mathbf{P}(A)p_a}{\mathbf{P}(A)p_a + \mathbf{P}(B)p_a} = \frac{\mathbf{P}(A)}{\mathbf{P}(A) + \mathbf{P}(B)} = \mathbf{P}(A).$$

Chapter 5

Independence and its consequences

5.1 Independence defined

Consider an experiment. Let A be an event which describes one property of the result, and let B be an event which describes another property of the result.

Definition 5.1 (Physical independence). We will say that physical events A and B are *independent* if knowledge that A occurred does nothing to change your opinion about $\mathbf{P}(B)$, and vice versa.

More precisely, if $\mathbf{P}(A) \neq 0$ we have

$$\mathbf{P}(B \mid A) = \mathbf{P}(B). \tag{5.1}$$

and if $\mathbf{P}(B) \neq 0$ we have

$$\mathbf{P}(A \mid B) = \mathbf{P}(A). \tag{5.2}$$

Equation (5.1) says that

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

and so

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B). \tag{5.3}$$

In the same way, Equation (5.2) also implies equation (5.3).

Notice that equation (5.3) also holds when $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 0$, and this equation is symmetric in A and B.

Exercise 5.1. Please check that equation (5.3) holds when $\mathbf{P}(A) = 0$.

[Solution]

Furthermore, equation (5.3) implies equation (5.1) when $\mathbf{P}(A) \neq 0$, and equation (5.3) implies equation (5.2) when $\mathbf{P}(B) \neq 0$.

So equation (5.3) is a single equation that completely expresses the physical independence of events A, B. This equation is naturally used as the definition of independence in a mathematical model.

Definition 5.2 (Mathematical Independence). Let A, B be events in a mathematical model. Whenever (5.3) holds, we say that the events A and B are independent. Equivalently, we say that the *pair* A, B is independent.

Sometimes it is convenient to express independence more colloquially, by saying that A is independent of B. Of course this also implies that B is independent of A.

Equation (5.3) is the *whole definition* of mathematical independence. Sometimes one refers to mathematical independence as "statistical independence". This reflects the fact that mathematical independence will hold in a model whenever the experimental statistics fit equation (5.3). We can assert that events are independent without identifying an underlying physical cause to explain why they are independent.

Incidentally, when we say "A and B are independent events", it may sound as if there is a property called "independence" that each event can have separately. That is not the case, and one should keep in mind that independence expresses a relationship, and is a property of two events considered *together*.

When (5.3) holds for two events, we also say that the probability of the events is *multiplicative*, meaning that the probability of the intersection is equal to the product of the separate probabilities.

Remark 5.3. Let A, B be events with $\mathbf{P}(B) \neq 0$. Our discussion shows that following statements are equivalent.

(i) A, B are independent.

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

We can use whichever formulation is convenient.

Example 5.4 (Tossing a coin twice). Consider the experiment of tossing a coin twice. Let H_1 be the event that the result of the first toss is a head, and let H_2 be the event that the result of the second toss is a head.

Let $p = \mathbf{P}(H_1)$. Based on our experience with coins, we expect that also $\mathbf{P}(H_2) = p$. What about the probability of obtaining two heads in succession, i.e. $\mathbf{P}(H_1 \cap H_2)$?

In ordinary experience, neither the coin nor the tosser is significantly altered by the result of the first coin toss. So we expect that when $\mathbf{P}(H_1) > 0$, $\mathbf{P}(H_2 \mid H_1) = \mathbf{P}(H_2)$. And indeed experience shows us that the probability of a head on the second toss is unaffected by the result of the first toss. Thus

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

By Definition 5.2, H_1, H_2 are independent.

Equation (5.3) is easy to verify directly when $\mathbf{P}(H_1) = 0$! Thus in all cases, H_1, H_2 are independent. Thus

$$\mathbf{P}(H_1 \cap H_2) = \mathbf{P}(H_1)\mathbf{P}(H_2) = p^2.$$

The same argument works for any combination of heads or tails on the two tosses. Thus, with $q = 1 - p$ we also have

$$\mathbf{P}(H_1 \cap H_2^c) = \mathbf{P}(H_1)\mathbf{P}(H_2^c) = pq,$$

$$\mathbf{P}(H_1^c \cap H_2) = \mathbf{P}(H_1^c)\mathbf{P}(H_2) = qp,$$

and

$$\mathbf{P}(H_1^c \cap H_2^c) = \mathbf{P}(H_1^c)\mathbf{P}(H_2^c) = q^2.$$

Exercise 5.2 (Sample space for two tossses). We can choose a particular sample space Ω to model tossing a coin twice. For example, let 1 denote a head and let 0 denote a tail, and take $\Omega = \{(1,1),(1,0),(0,1),(0,0)\}$.

The interpretation is that $(1,0)$ represents the physical outcome that a head is obtained on the first toss and a tail is obtained on the second toss. The other sample points are interpreted similarly.

(i) Define $H_1 = \{(1,1),(1,0)\}$ and define $H_2 = \{(1,1),(0,1)\}$. (Please check that H_1 and H_2 represent the physical events H_1, H_2 in Example 5.4.)

Find $H_1 \cap H_2$ as a set of sample points.

(ii) Let $q = 1 - p$. Using the coin tossing interpretation, show that the correct definition for $\mathbf{P}$ on this sample space is the following.

$$\begin{aligned} \mathbf{P}(\{(1,1)\} &= p^2, \\ \mathbf{P}(\{(1,0)\} &= pq, \\ \mathbf{P}(\{(0,1)\} &= qp, \\ \mathbf{P}(\{(0,0)\} &= q^2. \end{aligned} \tag{5.4}$$

(iii) Verify that the values in part (ii) give $\mathbf{P}(\{(1,1)\}) + \mathbf{P}(\{(1,0)\}) + \mathbf{P}(\{(0,1)\}) + \mathbf{P}(\{(0,0)\}) = 1$.

[Solution]

In previous comments we have mentioned that an abstract sample point need only represent the properties of an outcome that we currently wish to analyze. Example 2.14 provided a model that represents tossing a coin once. Now in Exercise 5.2 we have considered a model for two tosses. The model for two tosses also gives a representation for a single toss, since each of the two tosses is a single toss by itself. Are these representations consistent? The next exercise addresses that question.

Exercise 5.3. (This extends Exercise 2.4 from the case of a fair coin to the case of a general coin.)

In the model for Exercise 5.2, let A be the event that the first of two tosses results in a head. Check that $\mathbf{P}(A) = p$, *using the sample space for two tosses*. That means you can use equation (5.4) but nothing else.

Do the same for the event B that the second of two tosses results in a tail.

[Solution]

Exercise 5.4 (Independence for two tosses). In Exercise 5.2 we derived equation (5.4) by assuming independence for the results of the tosses.

For the events A and B of Exercise 5.3, show mathematically that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. You must base your answer entirely on equation (5.4).

[Solution]

Exercise 5.5. Return to the situation of Exercise 2.7. This deals with the experiment of rolling a fair die twice. You are asked to find the probability that the first roll produces an even number *and* the second roll produces a number larger than four.

Use independence to obtain the answer. You may use the fact that any event which only involves the first roll is independent of any event which only involves the second roll.

(And of course, if your answer now does not agree with the value found in Exercise 2.7, something is wrong.)

[Solution]

Exercise 5.6. When rolling a fair die twice, let A be the event that the sum of the numbers obtained on the two rolls is an even number.

Find $\mathbf{P}(A)$. You have solved this problem in Exercise 2.9. This time use independence to save work.

[Solution]

Exercise 5.7 (Simple cases of independence). Please check the following two easy special cases of independence.

For any events A, B,

$$\mathbf{P}(A) = 0 \text{ or } \mathbf{P}(A) = 1 \implies A \text{ and } B \text{ are independent.} \tag{5.5}$$

[Solution]

Exercise 5.8 (Independent and disjoint?). Suppose that events A, B are disjoint. Under what conditions will A, B also be independent?

[Solution]

We will see many examples of independent events in the rest of this book. Independence simplifies probability calculations immensely, *if it holds*. But an unjustified assumption of independence can lead to disaster (see for example [11]).

5.2 Independence for sampling with replacement

As in Example 4.8 and Example 2.23, we consider a two-step experiment. We select two jelly beans, one at a time. Each selection is random, and is such that no jelly bean in the bowl is favored.

However, unlike Example 4.8 and Example 2.23, after we have noted the color of the first jelly bean that is selected, we *replace* it in the bowl before proceeding to make the second selection.

Let A_1 be the event that the bean selected in step 1 is yellow or red. Let B_2 be the event that the bean selected in step 2 is yellow or green. We would like to find $\mathbf{P}(A_1 \cap B_2)$.

Assume that in the bowl, before any selections, there are y yellow beans, r red beans, and g green beans.

Consider each step as an experiment in itself. As usual, by Theorem 2.22 we have

$$\mathbf{P}(A_1) = \frac{y+r}{y+r+g}, \quad \mathbf{P}(B_2) = \frac{y+g}{y+r+g}. \tag{5.6}$$

Because we stir the bowl before each selection, experience tells us that the results of step 1 and step 2 are physically independent, and so we confidently assume that A_1 and B_2 are mathematically independent. Hence we can find $\mathbf{P}(A_1 \cap B_2)$ at once:

$$\mathbf{P}(A_1 \cap B_2) = \mathbf{P}(A_1)\,\mathbf{P}(B_2). \tag{5.7}$$

Notice that we did not need to actually specify a sample space for the two-step experiment. Instead we simply followed the rules of probability, using independence.

Remark 5.5 (A bit more about the "why" of independence). In an experimental situation, we would expect statistical independence for events A and B if there is no connection between the processes involved in A and the processes involved in B. But that's not the only case. Even when there is a connection, we frequently still expect independence.

Consider a person tossing a coin twice. There is a very direct physical connection between the two tosses, since the same person is doing the tossing. Nevertheless, experience shows that it doesn't matter, at least as far as the statistics of the two tosses is concerned.

The jelly bean experiment might be a little easier to analyze. Let's think about that.

When choosing a jelly bean, we prepare for the experiment by stirring the bowl of jelly beans vigorously, so that the beans in the bowl are thoroughly mixed. When we select a jelly bean twice, we have a two-step experiment, and we will stir the bowl of jelly beans before each step of the experiment (although the second stirring may not be necessary). We feel that that the two selections have statistically independent results. Why?

To fix our ideas, let's imagine that in the experiment we always select the top bean in the center of the bowl. Call that the "pickup location".

For simplicity, let's also assume that the bowl only contains yellow and red jelly beans, and that there are exactly the same number of yellow and red jelly beans.

Given these assumptions, we are confident that the chance of a yellow bean being selected on the first choice is $1/2$. And we actually don't believe this depends on the state of the bowl *before* the beans are stirred. Suppose, for example, that all the red beans were initially in one part of the bowl, and all the yellow beans were in a different region. We still think that a vigorous stirring is just as likely to move a red bean into the pickup location as a yellow bean. So knowing where the beans are before the stirring doesn't seem to help at all in predicting the color of the bean that is chosen after the stirring.

That statement applies to the both choices. Knowing the result of the first choice simply gives us a bit of information about the state of the jelly beans before the second stirring. So stirring the bowl between the two choices

should make the results of the two choices statistically independent.

To explore this idea further, let us now change the experimental procedure. Suppose that we didn't stir the bowl between the two choices, but we gave it a good stir at the beginning, before any beans were selected. If we chose the top bean at the pickup location when we made the the first selection, then we would select the bean just below it in the second selection. Could this spoil independence? Apparently not. All our experience suggests that a reasonably thorough stirring of the bowl will make the colors of adjacent beans statistically independent, or close to it.

It would be nice to give a mathematical model showing precisely why independence holds. But that seems to be an unsolved hard problem, even though we believe the conclusion. (Notice that such a model would have represent the positions of all the jelly beans and their shapes, and we would have to somehow show that they typically move in a disorderly manner.) Our probability model for choosing jelly beans doesn't concern itself with the details of stirring. We do not try to give a *mathematical* explanation of why red or green beans are equally likely, and why the two choices are independent. Our judgement about the probability of selecting a bean is just built into the model, based on our general practical experience.

5.3 Independence applies to complements

Lemma 5.6 (Independence and complements). Let A and B be any events in a probability model. Each of the following statements is mathematically equivalent to any of the others.

(i) A, B are independent.

(ii) A, B^c are independent.

(iii) A^c, B are independent.

(iv) A^c, B^c are independent.

The phrase "mathematically equivalent" for two statements means that if either one of the statements is true then the other is true also. The equivalence stated in Lemma 5.6 seems totally reasonable if we think about information. If you know whether or not A occurred, then you know whether or not A^c occurred, and so on!

Using the definition of independence, Lemma 5.6 can be restated as follows.

Lemma 5.7 (Equations for independence and complements). Let A and B be any events in a probability model. Each of the following four equations is mathematically equivalent to any of the others.

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B), \tag{5.8}$$

$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c), \tag{5.9}$$

$$\mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B), \tag{5.10}$$

$$\mathbf{P}(A^c \cap B^c) = \mathbf{P}(A^c)\mathbf{P}(B^c). \tag{5.11}$$

The sets $A \cap B$, $A \cap B^c$, $A^c \cap B$ and $A^c \cap B^c$ are represented in Figure 5.1.

A proof for Lemma 5.7 is requested in Exercise 5.9. To give a mathematical proof we will have to think about the precise definition of mathematical independence, not just the physical meaning. As usual, you are encouraged to work at the proofs, but the physical meaning is the most important thing.

Exercise 5.9. Prove Lemma 5.7.

For efficiency, we might start by proving the following.

Substitution Fact For any events D_1, D_2, suppose that:

$$\mathbf{P}\left(D_1 \cap D_2\right) = \mathbf{P}\left(D_1\right)\mathbf{P}\left(D_2\right) \tag{5.12}$$

holds. Then

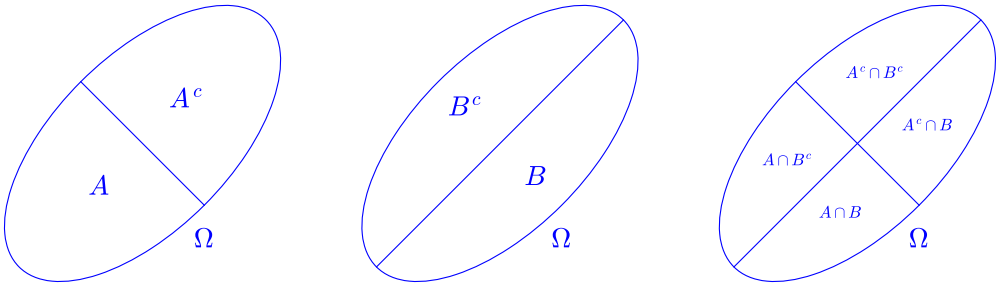
$$\mathbf{P}\left(D_1^c \cap D_2\right) = \mathbf{P}\left(D_1^c\right)\mathbf{P}\left(D_2\right). \tag{5.13}$$

In other words, replacing D_1 by D_1^c throughout the first equation gives another true statement.

Of course, since order doesn't affect the intersection operation, and order doesn't affect multiplication either, the Substitution Fact also implies that a true statement is also obtained from equation (5.12) if D_2 is replaced by D_2^c.

Once the Substitution Fact is proved, you can apply it to proving the lemma.

[Solution]



(a) A divides the space (b) B divides the space (c) The pieces

Figure 5.1: The pieces of Ω generated by A and B

Exercise 5.10 (A test for independence). Let A, B be events with $\mathbf{P}(A) > 0$ and $\mathbf{P}(A^c) > 0$.

(i) Suppose that

$$\mathbf{P}(B \mid A) = \mathbf{P}(B \mid A^c) \tag{5.14}$$

Show that A, B are independent.

(ii) Suppose that equation (5.14) does *not* hold. Show that A, B are *not* independent.

[Solution]

5.4 Using independence to simplify calculations

We often use independence to justify *ignoring* events. For example, suppose that in some large and complicated experiment we think about three events, A, B, and C. Physically, suppose we believe that A and B depend on certain properties of the experimental setup which are unrelated to the occurrence or non-occurrence of C.

As a practical matter, when calculating something about A and B, for example $\mathbf{P}(B \mid A)$, we can completely ignore C, even if we know whether or not C occurred. We do this automatically in our problems. Calculations of probabilities would be hopelessly complex if we could not make simplifications of this sort!

That's the physical picture. It would be interesting to consider mathematical ways to express the fact that C can be ignored, but we won't take time for that, except in the next exercise.

Exercise 5.11. Assume that

$$\mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C), \tag{5.15}$$

$$\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C), \tag{5.16}$$

$$\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A \cap B)\mathbf{P}(C). \tag{5.17}$$

Show that then

$$\mathbf{P}(B \mid A \cap C) = \mathbf{P}(B \mid A) \tag{5.18}$$

Equation (5.18) is an example of ignoring C, when the three independence statements (5.15), (5.16) and (5.17) all hold. One might be guess the first two independence statements should be sufficient, but they ain't. Something like condition (5.17) is needed too.

[Solution]

5.5 Extending independence to unions

This section gives us a chance to play a little more with the general definition of independence.

The following lemma is not surprising, but it states a useful fact.

Lemma 5.8 (Independence using cases). Let $A_1, \ldots, A_k$ be disjoint events in some probability model. Let B be an event such that A_i, B are independent for each $i = 1, \ldots, A_k$. Then $A_1 \cup \ldots \cup A_k$ and B are also independent.

Physically, Lemma 5.8 seems obvious. After all, for each i, being given information about the occurrence or non-occurrence of A_i has no affect on our opinion about B. If someone tells us the exciting news that at least one of the events A_i occurred, that is *less* information than telling us about a particular A_i.

A mathematical proof of Lemma 5.8 is a good exercise.

Exercise 5.12. Prove Lemma 5.8.
[Solution]

5.6 Solutions for Chapter 5

Solution (Exercise 5.1). Since $A \cap B \subset A$, we have $\mathbf{P}(A \cap B) \leq \mathbf{P}(A) = 0$. Thus $\mathbf{P}(A) = 0$ implies that $\mathbf{P}(A \cap B) = 0$.

Thus when $\mathbf{P}(A) = 0$, equation (5.3) is equivalent to the assertion that $0 = 0$.

Solution (Exercise 5.2).

(i) $(1, 1)$ is the only point in both H_1 and H_2, so $H_1 \cap H_2 = \{(1, 1)\}$.

(ii) We already showed that $H_1 \cap H_2 = \{(1, 1)\}$.
Similarly $H_1 \cap H_2^c = \{(1, 0)\}$, $H_1^c \cap H_2 = \{(0, 1)\}$, and $H_1^c \cap H_2^c = \{(0, 0)\}$.
Comparing this with the facts in Example 5.4 gives equation (5.4).

(iii)

$$\begin{aligned}\mathbf{P}(\{(1,1)\}) + \mathbf{P}(\{(1,0)\}) + \mathbf{P}(\{(0,1)\}) + \mathbf{P}(\{(0,0)\}) &= p^2 + pq + qp + q^2 \\ &= p(p+q) + q(p+q) = (p+q)^2 = 1.\end{aligned}$$

Solution (Exercise 5.3). $A = \{(1,1),(1,0)\}$, so $\mathbf{P}(A) = \mathbf{P}(\{(1,1)\}) + \mathbf{P}(\{(1,0)\}) = p^2 + pq = p(p+q) = p$.

$B = \{(1,0),(0,0)\}$, so $\mathbf{P}(B) = \mathbf{P}(\{(1,0)\}) + \mathbf{P}(\{(0,0)\}) = pq + q^2 = (p+q)q = q$.

Solution (Exercise 5.4). From the solution of Exercise 5.3, $\mathbf{P}(A) = p$ and $\mathbf{P}(B) = q$.

Also $A \cap B = \{(1,0)\}$, so $\mathbf{P}(A \cap B) = pq$ by equation (5.4).

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

Solution (Exercise 5.5). Let A the event that the first roll gives an even number. Let B be the event that the second roll gives a number larger than four.

We can look at the first roll as a separate experiment. The sample space has 6 outcomes of equal probability, so $\mathbf{P}(A) = 3(1/6) = 1/2$.

Similarly we can look at the second roll as a separate experiment, so $\mathbf{P}(B) = 2(1/6) = 1/3$.

By independence, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) = (1/3)(1/2) = 1/6$.

This approach is more efficient than the method used to solve Exercise 2.7. Physically we are sure that the two methods are both valid.

Solution (Exercise 5.6). Let B_1 be the event that the first roll gives an even number, and let C_1 be the event that the first roll gives an odd number.

Let B_2 be the event that the second roll gives an even number, and let C_2 be the event that the second roll gives an odd number.

Clearly $\mathbf{P}(B_1) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$. Similarly $\mathbf{P}(C_1) = \frac{1}{2}$, $\mathbf{P}(B_2) = \frac{1}{2}$ and $\mathbf{P}(C_2) = \frac{1}{2}$.

The sum of an even number and an odd number is odd. Even plus even is even, and odd plus odd is even.

Thus

$$A = (B_1 \cap B_2) \cup (C_1 \cap C_2)\,.$$

Using additivity and independence,

$$\mathbf{P}(A) = \frac{1}{2}\,\frac{1}{2} + \frac{1}{2}\,\frac{1}{2} = \frac{1}{2}.$$

Solution (Exercise 5.7). Suppose that $\mathbf{P}(A) = 0$. Then for any B, $\mathbf{P}(A \cap B) \leq \mathbf{P}(A) = 0$, so $\mathbf{P}(A \cap B) = 0 = \mathbf{P}(A)\mathbf{P}(B)$. Thus by definition A, B are independent.

Now suppose that $\mathbf{P}(A) = 1$. By Exercise 2.18, $\mathbf{P}(A \cap B) = \mathbf{P}(B) = \mathbf{P}(B)\mathbf{P}(A)$, so by definition A, B are independent.

Solution (Exercise 5.8). Suppose that A, B are disjoint. Then $\mathbf{P}(A \cap B) = 0$.

If A, B are also independent then $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. Since $0 = \mathbf{P}(A)\mathbf{P}(B)$, at least one of the events A, B has zero probability.

If $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 0$, then A, B are independent by Exercise 5.7.

This shows that when A, B are disjoint, then A, B are independent if and only $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 0$.

Solution (Exercise 5.9). First let us prove the Substitution Fact.

Suppose that equation (5.12) holds. We need to show that equation (5.13) holds.

That is, suppose that $\mathbf{P}(D_1 \cap D_2) = \mathbf{P}(D_1)\mathbf{P}(D_2)$. We need to show that if we replace D_1 by D_1^c in this equation we get another true equation.

To prove this, note that $D_1^c \cap D_2$ is exactly the part of D_2 which is not in D_1. Thus

$$\mathbf{P}(D_1^c \cap D_2) = \mathbf{P}(D_2) - \mathbf{P}(D_1 \cap D_2).$$

(For a justification, see equation (2.12) and Figure 2.1.)

Since $\mathbf{P}(D_1 \cap D_2) = \mathbf{P}(D_1)\mathbf{P}(D_2)$,

$$\begin{aligned}\mathbf{P}(D_1^c \cap D_2) &= \mathbf{P}(D_2) - \mathbf{P}(D_1)\mathbf{P}(D_2) \\ &= (1 - \mathbf{P}(D_1))\mathbf{P}(D_2) = \mathbf{P}(D_1^c)\mathbf{P}(D_2),\end{aligned}$$

as claimed.

This proves the stated Substitution Fact.

To prove the lemma, we consider some applications of the Substitution Fact.

Suppose that equation (5.8) is true. Replacing A by A^c gives equation (5.10), so this equation must also be true.

On the other hand, suppose that equation (5.10) is true. Replacing A^c by $(A^c)^c = A$ gives equation 5.8, so equation 5.8 must be true also.

We have shown that the truth of either one of equations (5.8) and (5.10) implies the truth of the other.

Switching between B and B^c shows that equations(5.8) and (5.9) are equivalent.

Switching between A and A^c shows that equations (5.9) and (5.11) are equivalent.

Thus we can change any one of the equations into any of the other equations, using one or two substitution operations, and these substitution operations preserve truth.

Solution (Exercise 5.10).

(i) Suppose that $\mathbf{P}(B \mid A) = \mathbf{P}(B \mid A^c)$.
By the Law of Total Probability (Theorem 4.6),

$$\mathbf{P}(B) = \mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}(B \mid A^c). \tag{5.19}$$

Hence

$$\mathbf{P}(B) = \mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}(B \mid A). \tag{5.20}$$

Since $\mathbf{P}(A) + \mathbf{P}(A^c) = 1$,

$$\mathbf{P}(B) = \mathbf{P}(B \mid A). \tag{5.21}$$

By Remark 5.3, A, B are independent.

(ii) We must show that if A, B are independent then equation (5.14) must hold! (Does it make sense that this formulation is equivalent to what is asked in part (ii) of the question? It certainly will if you think about it. We could get fancy here and talk about the "contrapositive" form of a statement, but we don't need to.)

Since A, B are independent, Remark 5.3 tells us that

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

Also, by Lemma 5.6, if A, B are independent then also A^c, B are independent. So we can replace A by A^c in the equation just obtained. This gives

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

so $\mathbf{P}(B \mid A^c) = \mathbf{P}(B \mid A)$.

Solution (Exercise 5.11).

Proof. By the conditional probability formula,

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

proving equation (5.18).

□

Solution (Exercise 5.12).

$$\mathbf{P}\left(B \cap (A_1 \cup \ldots \cup A_k)\right) = \mathbf{P}\left((B \cap A_1) \cup \ldots \cup (B \cap A_k)\right).$$

By assumption, $A_1, \ldots, A_k$ are disjoint, so $B \cap A_1, \ldots, B \cap A_k$ are disjoint. Hence

$$\begin{aligned}\mathbf{P}\left(B \cap (A_1 \cup \ldots \cup A_k)\right) &= \mathbf{P}\left(B \cap A_1\right) + \ldots + \mathbf{P}\left(B \cap A_k\right) \\ &= \mathbf{P}(B)\mathbf{P}\left(A_1\right) + \ldots + \mathbf{P}(B)\mathbf{P}\left(A_k\right) \\ &= \mathbf{P}(B)\left(\mathbf{P}\left(A_1\right) + \ldots + \mathbf{P}\left(A_k\right)\right) = \mathbf{P}(B)\mathbf{P}\left(A_1 \cup \ldots \cup A_k\right).\end{aligned}$$

By definition, this shows that B and $A_1 \cup \ldots \cup A_k$ are independent.

Chapter 6

Tricky little problems

Sometimes very simple problems are enlightening, for example if they illustrate the need to be careful when setting up an abstract model of a real-world setting. In this short chapter we'll work through two well-known examples.

6.1 One or two successes

The happy Sam problem A local charity has a booth at the fair. This booth offers donors the opportunity to play a game. In this game, your chance of winning is p, and if you win, you receive a small prize.

Let us think about a donor named Sam, who visits the fair one afternoon. He plays the game exactly twice during the afternoon. Sam is easy to please, so he is happy if he wins at least one prize. If he does not win any prize, he is unhappy.

Let A be the event that Sam wins both times he plays the game. We assume that the results of the two games are independent, so the probability of A is p^2. This is what we will call the unconditional probability of A in this setting.

His friends meet Sam some time after he leaves the fair. They know that Sam played two games at the fair, but they do not know the results of the two games. However, they observe that Sam is happy. Thus his friends know that Sam won at least one game at the fair. Based on their information about Sam, what probability should his friends assign to A?

We can use a model for Sam's games which is similar to the model for two coin tossses (Example 5.4). Let W_1 be the event that Sam won the first time

he played the game, and let L_1 be the event that he lost. Then $\mathbf{P}(W_1) = p$ and $\mathbf{P}(L_1) = q$, where $q = 1 - p$. Define W_2, L_2 similarly.

Let B be the event that Sam won at least one game at the fair. There are three ways that could have happened: Sam could have won both games, Sam could have won the first and lost the second, or Sam could have lost the first and won the second. That is:

$$B = (W_1 \cap W_2) \cup (W_1 \cap L_2) \cup (L_1 \cap W_2)\,.$$

After they meet Sam, his friends know that the event B has occcurred.

Sam's friends want to know $\mathbf{P}(A \mid B)$, where $A = W_1 \cap W_2$.

Using additivity and independence,

$$\mathbf{P}(B) = \mathbf{P}(W_1)\mathbf{P}(W_2) + \mathbf{P}(W_1)\mathbf{P}(L_2) + \mathbf{P}(L_1)\mathbf{P}(W_2) = p^2 + pq + qp.$$

Notice that $A \subset B$, so $A \cap B = A$. Since $\mathbf{P}(B) = p^2 + 2pq$, the conditional probability formula (4.1) tells us that

$$\mathbf{P}(A \mid B) = \frac{p^2}{p^2 + 2pq} = \frac{p}{p + 2q} = \frac{p}{1 + q}.$$

When $p = 1/2$, this says that the probability that Sam won both games is only $1/3$.

For comparison, consider a different problem about Sam's games.

The Sam's witness problem The afternoon of the fair, you are strolling through the fair, and you happen to pass by the charity booth at a moment when Sam is playing one of his two games. You observe that Sam wins.

You don't see Sam again that day but you are told that he played the game twice.

Like Sam's friends, you know that the event B has occurred. However, you also have some additional information. What probability should you assign to A?

Exercise 6.1. Solve the Sam's witness problem. [Solution]

The difference between these two problems about Sam may be evident, but in many problems such differences may be obscured by the wording.

Here is an example of a loosely phrased version of the happy Sam problem: "Sam played two games at the fair, and he won at least one game. Find the probability that he won the other game."

A common variant of this problem: "A couple has two children. Given that one of the children is a girl, find the probability that the other child is a girl."

Remark 6.1 (Using a sample space for independence). In our analysis of the happy Sam problem, we didn't bother to define a sample space, but just worked with events. Just as in Exercise 5.2, we could have defined a sample space. For example, we could let $\Omega = \{(1,1),(1,0),(0,1),(0,0)\}$. Here $(1,1)$ represents the outcome that Same won both games, $(1,0)$ represents the outcome that Sam won the first game but lost the second game, and so on. Would that be better? For example, is it easy to see that $\mathbf{P}(\{(0,1)\}) = qp$? Since we are familiar with the sample space for two coin tosses, it likely is easy. But working with events like W_i and L_i lets us use the the definition of independence directly, and that seems better for our thinking.

6.2 The Monty Hall problem!

This is well-known, but worth reviewing. A good history of this problem is given in [9]. Apparently some mathematicians refused to believe the correct answer. The embarrassing details are given in section 1.10 of [9].

At any rate, here is the problem. It is inspired by a game which was sometimes played on the television show *Let's Make a Deal*, hosted by Monty Hall. The idealized version of the game which is described here may not match what actually used to happen, so our Monty Hall is not quite the real Monty Hall.

We assume that in this game, three doors are visible to the contestant, and the contestant is asked to choose one of these doors. The contestant will be awarded whatever prize is concealed behind the selected door. There is a valuable prize, perhaps a sports car, behind one door, and something very disappointing is behind each of the other two doors.

Of course we assume that the prize can lie behind any door, with no door favored. The contestant has no idea which door has the prize.

So far, so good. Now comes the twist. We assume that after the contestant has chosen a door, but before revealing whether the contestant's guess was correct, Monty Hall often opens one of the two doors which were *not* selected by the contestant, always revealing one of the disappointing non-prizes when he does so. Then, Monty Hall offers the anxious contestant an opportunity to switch his or her choice to the *other unopened* door.

The basic question here is whether the contestant would benefit by switching.

We begin by focusing our attention on the original choice by the contestant. Let C be the event that the contestant's choice is correct. Since no door was favored in setting up the game, $\mathbf{P}(C) = 1/3$.

Here $\mathbf{P}$ refers to probabilities based on the information we have before Monty Hall opens a door.

But notice that Monty Hall does not physically move the valuable prize.

So if the contestant's choice is correct at the moment of the choosing, the contestant's choice is correct for ever. If the choice is wrong, it stays wrong. We also know there is only one alternative left after Monty Hall has opened a door. Thus, if the contestant's original choice was wrong, the contestant should switch to the other unopened door.

The contestant will choose correctly approximately 1/3 of the time, and incorrectly 2/3 of the time. Hence the policy of "always switching" pays off 2/3 of the time, while "never switching" pays off 1/3 of the time. So switch!

That answers the basic question. But there seems to be something about the Monty Hall problem that makes people doubt the answer. It is not completely clear why. Monty Hall's actions do complicate the problem. But sloppy wording of the problem can also cause trouble, if Monty Hall's procedure is not explained precisely.

Consider the following variation of the Monty Hall problem.

Example 6.2 (The defective door problem). Suppose Monty Hall is on vacation. In the absence of a skilled host, the manager of the game show decides that they can only provide a simplified version of the game. The contestant will choose a door, and will then be given whatever prize lies behind the door.

However, fate is about to intervene.

After the contestant has chosen a door, one of the other doors suddenly swings open. The door must have been defective in some way, although no one knew about this until now. Perhaps a vibration in the floor, or a gust of

wind, has now made the door open.

We think that the location of the prize does not effect the condition of any of the doors, and the location of the prize cannot cause any door to open or not open. However, it happens that the door which opened is not the one concealing the valuable prize. So the contestant's door and one other door remain unopened, and we know that one of them hides the valuable prize.

The manager of the game show notices that this accident has presented the audience with exactly the same situation that would have been the result of Monty Hall's usual antics. To live up to the expectations of the audience, the manager offers the contestant the opportunity of switching his or her choice to the other unopened door.

We ask the same question as before. Does the contestant benefit by switching?

Exercise 6.2. Please solve the defective door problem.

[Solution]

Example 6.2 seems more natural than the Monty Hall problem, and sometimes people may solve the wrong problem.

Exercise 6.3 (Mega-Monty). In order to convince people that switching is the right policy for the standard Monty Hall problem, the following variation is sometimes presented. An argument which is claimed to work for the standard Monty Hall problem can be "stress-tested" on this version of the problem.

Suppose that for a special edition of the game show, a long hallway is used, with 100 doors. The prize is behind one of the doors, and the miserable contestant must choose one door. After the choice has been made, in a surprising act of generosity Monty Hall opens 98 of the remaining doors, none of which have prizes, and then offers the contestant a chance to switch his or her choice to the other unopened door.

Again we ask, does the contestant benefit by switching?

[Solution]

Exercise 6.4 (Monty with a tell).

Part 1 Returning to the Monty Hall game with the usual three doors, imagine you have been invited to appear as a contestant. The game will be hosted by Monty's sister, Ivy Hall, who sometimes replaces Monty.

You prepare by carefully watching recordings of all previous shows for which Ivy was the host. By the end of each show the audience knows where the prize was located for that show, so that information is in the recording.

The three doors for this contest are arranged in a line going from left to right. You excitedly notice the following behavior pattern. Whenever the door originally chosen by the contestant is the door with the valuable prize, so that Ivy Hall can *choose* which of the two remaining doors to open, Ivy always opens the remaining door on the *left*.

Eventually you appear on the game show, and select your door. And then Ivy Hall opens . . . the remaining door on the left!

At this point Ivy Hall offers you the usual opportunity to switch your choice. As you stand there, weighing your chances, Ivy notices your indecision, and makes an unusual extra offer. She will pay you an additional $100, win or lose, if you do *not* switch.

What should you do?

Part 2 Suppose the same situation arises as in Part 1, except that in this case you observe that Ivy Hall has opened the remaining door on the *right*. Everything else is the same.

What should you do?

[Solution]

6.3 Solutions for Chapter 6

Solution (Exercise 6.1). The witness doesn't just know that Sam won one game, the witness can also specify which game it was, namely the game that was played while the witness walked by.

The *other* game Sam played is well-defined, and the result of that game is of course independent of anything that happens in the game that the witness saw.

Sam won the game that the witness saw, so the probability that Sam won *both* games is just the probability that Sam won the other game, and this is p.

.

Solution (Exercise 6.2). As usual, the choice made by the contestant is not influenced by the location of the valuable prize.

In this problem the contestant's choice is also unrelated to the location of the defective door. This is in contrast to the situation when Monty Hall opens a door, since Monty never opens the door chosen by the contestant.

Method 1 The idea of the solution is easy to state: we will use symmetry.

After noting that the choice of a door by the contestant has no effect on anything else, we can ignore the contestant, and simply look at the doors. After the defective door has opened, we see two remaining unopened doors. One of these doors has the valuable prize. Nothing in the description of the problem (except the contestant's choice) treats either of these doors differently. So any probability statement that we derive, concerning the location of the valuable prize, must also treat both these doors in the same way. Thus the valuable prize is equally likely to reside behind either door.

There is no reason to switch.

Method 2

If you include the contestant's choice of a door in the discussion, you might say something like the following.

From the contestant's viewpoint, the opening of the door is a random event, independent of everything else. The chance of any particular door opening is the same, a small probability.

Let C be the event that the contestant's original choice of a door was correct, meaning that it is the one with the valuable prize.

As usual $\mathbf{P}(C) = 1/3$, so $\mathbf{P}(C^c) = 2/3 = 2\mathbf{P}(C)$.

Let M be the event which describes the new situation after the door opened. When deciding whether or not to switch, the contestant should be interested in $\mathbf{P}(C \mid M)$.

The key idea in this approach: we are dealing with the situation in which the door that opened was neither the door with the prize nor the door picked by the contestant. Common sense probability tells us that this is twice as likely to happen if the contestant chose correctly, since then the door with

the valuable prize and the door picked by the contestant are one and the same door, and there are *two* possible choices for the defective door. (We could justify this probability statement more formally, but right now we are focused on the physical meaning.)

Using our notation, $\mathbf{P}(M \mid C) = 2\mathbf{P}(M \mid C^c)$.

The multiplied-through version of the conditional probability formula (equation (4.2)) tells us that

$$\mathbf{P}(M \cap C) = \mathbf{P}(C)\mathbf{P}(M \mid C) \text{ and } \mathbf{P}(M \cap C^c) = \mathbf{P}(C^c)\mathbf{P}(M \mid C^c).$$

The fact that $\mathbf{P}(M \mid C) = 2\mathbf{P}(M \mid C^c)$ exactly compensates for the fact that $\mathbf{P}(C^c) = 2\mathbf{P}(C)$, and we have

$$\mathbf{P}(M \cap C) = \mathbf{P}(M \cap C^c).$$

In frequency language, this equation says that contestants will find themselves in situation M just as often when the chosen door is correct as when it is incorrect.

And

$$\mathbf{P}(C \mid M) = \frac{\mathbf{P}(M \cap C)}{\mathbf{P}(M)} = \frac{\mathbf{P}(M \cap C^c)}{\mathbf{P}(M)} = \mathbf{P}(C^c \mid M).$$

Thus the chance of getting the valuable prize in this situation is the same, whether or not you switch, and there is no reason to switch.

Solution (Exercise 6.3). Now the probability that the original guess was correct is 1/100. This happens approximately 1/100 of the time. And so switching brings success approximately 99 times out of 100.

Switch!

Solution (Exercise 6.4).

Part 1 Let L be the event that Ivy Hall opens the remaining door on the left.

Let C be the event that the contestant picked the correct door. From your study of past Ivy Hall shows, you know that $\mathbf{P}(L \mid C) = 1$.

If C^c occurs, then the prize is behind one of the two doors which the contestant did not pick. Knowing that C^c occurred does not give us information which favors either of those two doors.

If C^c occurs, Ivy Hall must of course open the remaining door which does not have the prize, wherever it is, left or right. Thus $\mathbf{P}(L \mid C^c) = 1/2$.

By the Law of Total Probability (Theorem 4.6),

$$\mathbf{P}(L) = \mathbf{P}(C)\mathbf{P}(L \mid C) + \mathbf{P}(C^c)\mathbf{P}(L \mid C^c) = \frac{1}{3}1 + \frac{2}{3}\frac{1}{2} = \frac{2}{3}.$$

Thus

$$\mathbf{P}(C \mid L) = \frac{\mathbf{P}(C \cap L)}{\mathbf{P}(L)} = \frac{\mathbf{P}(C)\mathbf{P}(L \mid C)}{\mathbf{P}(L)} = \frac{\frac{1}{3}1}{\frac{2}{3}} = \frac{1}{2}.$$

Since switching does not improve your chances of winning the valuable prize, stick with your choice and take the \$100.

Part 2 If the door you chose had been wrong, Ivy would have chosen the remaining door on the left. She didn't do that.

Don't switch!

Chapter 7

Independent sequences

The most important facts in this chapter are Definitions 7.1 and 7.7, and the formulas in Section 7.2.

We've discussed independence for two events. But often we want to consider more than two events, perhaps even a long sequence of events.

7.1 Sequences of experiments

Consider a big experiment which consists of a sequence of repetitions of a smaller experiment. For example think about tossing a coin many times, or rolling a die many times, or treating many patients with a particular drug. We've considered repeated experiments in the past, but here we are thinking of the whole sequence of smaller experiments as making up one big experiment. We can call each repetition of the smaller experiment a "trial".

In this situation we typically assume that the smaller experiments do not influence each other in any significant way, so that properties of different trials are described by independent events. Then we say that we have an independent sequence of trials.

We would like to study some mathematical formulas connected with independent trials. Before doing that, we should state a more precise definition of independence in this situation.

Definition 7.1 (Independent trials). We will say that a sequence of n experiments is independent if, for each $k < n$, information about the results for trials $1, \ldots, k$ does not change our opinion about the probability of any properties of the result of trial $k + 1$.

For each i, let D_i be a physical event that is defined completely by the outcome of the i-th experiment. If the sequence of n experiments is independent then we will say that the sequence $D_1, \ldots, D_n$ is an independent sequence of physical events.

Suppose that, based on our experience, we think that a certain sequence of experiments is independent in the sense of Definition 7.1.

Let D_i be an event that is defined completely by the outcome of the i-th experiment. What can we say about these events?

It should be emphasized that Definition 7.1 is concerned with real physical experiments, not abstract sample spaces. Consider the event $D_1 \cap \ldots \cap D_k$. Here (as in Remark 2.11) $D_1 \cap \ldots \cap D_k$ is just a convenient way of expressing the event that every one of the physical events $D_1, \ldots, D_k$ occurs. The occurrence or non-occurrence of $D_1 \cap \ldots \cap D_k$ is certainly determined by the results of trials $1, \ldots, k$. So, by Definition 7.1, information about $D_1 \cap \ldots \cap D_k$ does not change our opinion about the probability of D_{k+1}. If $\mathbf{P}(D_1 \cap \ldots \cap D_k) \neq 0$, this says that

$$\mathbf{P}\left(D_{k+1} \,|\, D_1 \cap \ldots \cap D_k\right) = \mathbf{P}\left(D_{k+1}\right). \tag{7.1}$$

And as in Definition 5.1 we conclude that

$$D_1 \cap \ldots \cap D_k \text{ and } D_{k+1} \text{ are independent events.} \tag{7.2}$$

Please do the next exercise!

Exercise 7.1. Let $D_1, \ldots, D_n$ be an independent sequence of events. Show that

$$\mathbf{P}\left(D_1 \cap \ldots \cap D_n\right) = \mathbf{P}\left(D_1\right) \ldots \mathbf{P}\left(D_n\right). \tag{7.3}$$

[Solution]

Definition 7.1 says what we mean by independence for a sequence of physical events. It is not a definition of mathematical independence for an abstract model, although of course consequences such as equation (7.3) must hold in any valid model for an independent sequence of experiments. We'll think later about making a precise mathematical definition of an independent sequence.

For now, let's calculate some consequences of Definition 7.1.

7.2 Outcome probabilities when tossing a coin n times

In the present section we perform a key computation.

Consider n tosses of a coin. The coin need not be fair. The probability of a head is some number p and the probability of a tail is $q = 1 - p$.

We mentioned in Example 2.3 that when the result of a toss is a head, we sometimes say that result is *success*. Using this sort of language can be briefer, and it also makes it a bit easier to adapt our results about coin-tossing to other situations which are similar.

We'll usually record the success or failure of a toss using a number, either 0 or 1. Success is represented by 1 and failure is represented by 0.

The record of successes and failures for a whole sequence of n repeated tosses is then a sequence $(x_1, \ldots, x_n)$, where for each j, x_j is either zero or one. We can call this sequence the "success record".

Clearly there are 2^n possible success records.

Should we use $(x_1, \ldots, x_n)$ as the sample point that records the whole result of the experiment when the coin is tossed n times? Then our sample space will simply be the set of all possible sequences of this sort.

We certainly can use that sample space, but the probability argument may clearer if we simply just talk about events, without committing to a particular sample space representation.

For each $j = 1, \ldots, n$, let W_j be the event that toss j produced success. Let $D_j^1 = W_j$ and let $D_j^0 = W_j^c$.

For any sequence $(x_1, \ldots, x_n)$ of zeros and ones, $D_1^{x_1} \cap \ldots \cap D_n^{x_n}$ is the event that:

> (toss 1 produced x_1) **and** (toss 2 produced x_2) **and** ... **and** (toss n produced x_n).

Thus

$$D_1^{x_1} \cap \ldots \cap D_n^{x_n} \text{ is the event that the success record is } (x_1, \ldots, x_n). \quad (7.4)$$

For coin tosses, the tosses are independent trials. Thus events defined in terms of different tosses are physically independent. Thus by equation (7.3), for any success record $(x_1, \ldots, x_n)$ we know that

$$\mathbf{P}\left(D_1^{x_1} \cap \ldots \cap D_n^{x_n}\right) = \mathbf{P}\left(D_1^{x_1}\right) \cdots \mathbf{P}\left(D_n^{x_n}\right). \quad (7.5)$$

Equation (7.5) is the key probability fact we need.

We know that $D_n^{x_j} = p$ if $x_j = 1$ and $D_n^{x_j} = q$ if $x_j = 0$. Thus

$$\mathbf{P}\left(D_1^{x_1} \cap \ldots \cap D_n^{x_n}\right) = p^k q^{n-k}, \tag{7.6}$$

where k is the number of indices j such that $x_j = 1$.

We have proved:

Lemma 7.2 (Coin toss outcome probabilities). For any sequence $(x_1, \ldots, x_n)$ of zeros and ones, the probability of obtaining exactly that success record is $p^k q^{n-k}$, where k is the number of successes in the sequence $(x_1, \ldots, x_n)$.

Lemma 7.2 lets us calculate a very useful probability, in the next lemma.

Lemma 7.3 (Probability of obtaining k successes). Let G_k be the probability that n tosses produce exactly k successes. Then

$$\mathbf{P}(G_k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}. \tag{7.7}$$

Here 0! is interpreted as 1 in case that $k = 0$ or $k = n$.

Proof. Suppose a particular sequence of trials produces exactly k successes. What would the success record look like?

It is a sequence $(x_1, \ldots, x_n)$ made up of zeros and ones. Since there are k successes, exactly k ones must appear in the sequence. As noted in equation (7.4), $D_1^{x_1} \cap \ldots \cap D_n^{x_n}$ is the event that the success record is $(x_1, \ldots, x_n)$. We don't care in what order the successes occur, so for general n and k there are many such success records. G_k is the union all the corresponding events.

That is, G_k is the union of events $D_1^{x_1} \cap \ldots \cap D_n^{x_n}$, over all sequences $(x_1, \ldots, x_n)$ which have exactly k successes. These events $D_1^{x_1} \cap \ldots \cap D_n^{x_n}$ are clearly disjoint, so $\mathbf{P}(G_k)$ must be the sum of the probabilities of the various events $D_1^{x_1} \cap \ldots \cap D_n^{x_n}$, where $(x_1, \ldots, x_n)$ which has exactly k successes.

Let m be the number of distinct sequences $(x_1, \ldots, x_n)$ which contain exactly k ones.

Lemma 7.2 and the additivity of probability tells us that

$$\mathbf{P}(G_k) = m p^k q^{n-k}. \tag{7.8}$$

The number m depends on n and k.

Equation (7.8) will give us equation (7.7), once we show that

$$m = \frac{n!}{k!(n-k)!}. \tag{7.9}$$

How do we show that equation (7.9) holds? This is a *counting problem.*

We could explain right now how to count the number of sequences containing the k ones. But it seems more efficient to do that in Lemma 8.2, as part of a general discussion. So we will leave equation (7.9) as an I.O.U. for the moment. This obligation will be fully discharged in Lemma 8.2.

□

Exercise 7.2. Equation (7.9) tells us that $m = n$ when $k = 1$. Verify in this special case that $m = n$ is the correct value.

[Solution]

Exercise 7.3. Consider the experiment of tossing a fair coin 5 times. Let A be the event that the first three tosses produce at most 1 head in total. Let B be the event that the last two tosses produce exactly 1 head in total. Find $\mathbf{P}(A \cap B)$.

[Solution]

Exercise 7.4 (Overlapping sequence segments). Consider the experiment of tossing a fair coin 8 times. Let A be the event that the first six tosses produce exactly 4 heads. Let B be the event that the last five tosses produce exactly 3 heads. Find $\mathbf{P}(A \cap B)$.

Hint: Let $C = A \cap B$. A reasonable approach to finding $\mathbf{P}(C)$ is to break up the problem into cases. Let M_j be the event that the fourth, fifth and sixth tosses produce exactly j heads. Then $\mathbf{P}(C) = \mathbf{P}(C \cap M_0) + \mathbf{P}(C \cap M_1) + \ldots$.

[Solution]

Exercise 7.5. Consider n tosses of a fair coin. As in Example 2.6, you can use a sample space Ω whose points are sequences of length n, made up of zeros and ones.

Let A be any event on your sample space Ω. Prove that $\mathbf{P}(A)$ can be written as a fraction whose denominator is a power of 2.

Exercise 1.5 considered methods of simulating an experiment with three equally likely outcomes. In such an experiment, each possible outcome must have probability 1/3. The present exercise shows that tossing a fair coin, even many times, can never perfectly simulate such an experiment.

[Solution]

Exercise 7.6 (Counting sequences). Consider tossing a coin 30 times.

Let D_i^1 denote the event that toss i produces a head and let D_i^0 denote the event that toss i produces a tail.

Using the sequence sample space Ω of Exercise 7.5, with $n = 30$, please answer the following questions.

(i) How many sample points are there in D_5^1?

(ii) How many sample points are there in $D_5^1 \cap D_7^0$?

(iii) List all the sample points in $D_1^1 \cap D_1^0 \cap D_3^1 \cap D_4^0 \cap \ldots \cap D_{29}^1 \cap D_{30}^0$?

[Solution]

7.3 Bernoulli trials terminology

Like coin-tossing, many experimental situations involve repeated independent experiments, each of which either results in an event called "success" or an event called "failure". The next definition provides a convenient name for such experiments.

Definition 7.4 (Bernoulli trials). Let $W_1, \ldots, W_n$ be independent events, each of which has the same probability p. We will say that the sequence $W_1, \ldots, W_n$ form a sequence of *Bernoulli trials*. We will often refer to the occurrence of W_i as *success* on trial i, and the occurrence of W_i^c as *failure* on trial i. The probability p will be called the probability of success.

We also speak of the experiments and models associated with the events $W_1, \ldots, W_n$ as Bernoulli trials.

Tossing a coin n times, when the probability of a head is p, gives an example of Bernoulli trials, provided we interpret a head as "success" on any toss. The event W_i in this situation is simply the event that a head is obtained on toss i.

Any mathematical statement about a Bernoulli trial sequence can be translated into a mathematical statement about a coin-tossing sequence with the same success probability. Thus we are free to use either Bernoulli trial language or coin-tossing language to describe the relevant concepts.

Translating Lemma 7.3 into the language of Bernoulli trials gives the following.

Theorem 7.5 (Probability of k successes). Let $W_1, \ldots, W_n$ be a sequence of Bernoulli trials with success probability p.

Let $\mathbf{P}$ be the appropriate probability for the model. Let G_k be the event that exactly k successes occur. Then

$$\mathbf{P}(G_k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}, \tag{7.10}$$

where 0! is interpreted as 1 in case that $k = 0$ or $k = n$.

One often writes $\frac{n!}{k!(n-k)!}$ as $\binom{n}{k}$, so equation (7.10) can also be written as

$$\mathbf{P}(G_k) = \binom{n}{k} p^k q^{n-k}, \tag{7.11}$$

where

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

The expression $\binom{n}{k}$ is called a "binomial coefficient", since it appears in the statement of the Binomial Theorem (equation (8.6)).

Definition 7.6 (The binomial distribution). The distribution given by equation (7.10) is called the *binomial distribution* with parameter p.

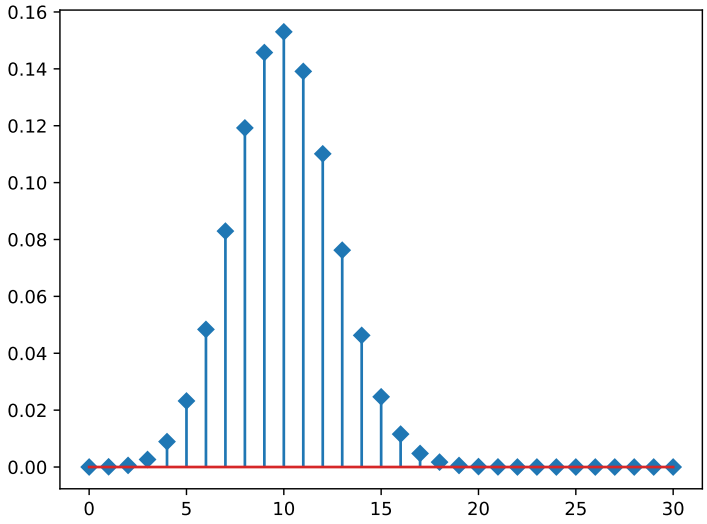


Figure 7.1: $\mathbf{P}(k \text{ heads})$ in 30 tosses, success prob 1/3.

Figure 7.1 shows a plot of $\mathbf{P}(G_k)$ versus k for 30 trials, with success probability 1/3. Notice that the graph appears to be centered around $k = 10$, although it is not symmetric around that point. Since the success probability is 1/3, the average number of successes over many repetitions of a sequence of 30 trials is also equal to 10, since $(1/3) * 30 = 10$.

Also notice that those probabilities in Figure 7.1 get awfully small when you move a moderate distance away from 10.

7.4 Mathematical independence for a sequence

Building on our experience with coin-tossing, let's give a general mathematical definition. A precise definition of this sort is interesting, but it is not really necessary for our work in this book. Our instincts about independence will tell us what to do in most calculations. So readers can skip this section if desired, or just skim it quickly.

Definition 7.7 (Independence for n abstract events). Let $W_1, \ldots, W_n$ be a sequence of events in some probability model. Suppose that for every

choice of events $D_1, \ldots, D_n$, where each D_i is either W_i or W_i^c, the following equation holds.

$$\mathbf{P}(D_1 \cap \ldots \cap D_n) = \mathbf{P}(D_1) \cdots \mathbf{P}(D_n). \tag{7.12}$$

Then we say that the sequence $W_1, \ldots, W_n$ is an independent sequence of events in the probability model.

Be careful to note that independence is a property of the whole sequence, not of each event W_i by itself! Nevertheless, for brevity we do often express independence by saying that "the events are independent", rather than saying that the events form an independent sequence.

Notice that equation (7.12) gives us 2^n equations, when we substitute for $D_1, \ldots, D_n$ in all possible ways. That's a lot of equations!

Why should we think that equation (7.12) is a reasonable definition?

Well, suppose your model deals with a sequence of independent physical trials, and the abstract event W_i represents a physical event defined entirely in terms of the result of trial i. Since D_i is either W_i or W_i^c, we know that D_i also represents a physical event defined entirely in terms of the result of trial i, and so by Exercise 7.1 we must believe that equation (7.12), holds for any choice of $D_1, \ldots, D_n$.

But is that enough? Perhaps physically independent events have more properties, which are not captured by those 2^n equations given in equation (7.12). Should we worry about that? A reassuring answer is given by the fact that the 2^n events $D_1 \cap \ldots \cap D_n$ cover all the possible cases of what can happen in n tosses. In other words, anything that can be said about the outcomes can be expressed in terms of set operations on events of the form $D_1 \cap \ldots \cap D_n$.

So it is plausible that Definition 7.7 is a sound definition. But is it beautiful? It is expressed using a lot of equations. On the other hand, all 2^n equations follow the same pattern. So it's not too bad.

There is one exceptional case: for $n = 2$, Lemma 5.6 shows that the single equation $\mathbf{P}(W_1 \cap W_2) = \mathbf{P}(W_1)\mathbf{P}(W_2)$ implies all four of the equations obtained by substituting in equation (7.12).

This shows that the sequence W_1, W_2 is independent in the sense of Definition 7.7 if and only if W_1, W_2 are independent in the sense of Definition 5.2. That is good, since it avoids ambiguity when we use the word "independent".

It's too bad things are more complicated when n is greater than 2.

But don't worry! You will usually find that your physical understanding of independent events will let you *guess* the correct equations for any practical problem. We used this approach in Method 1 for the solution of Exercise 7.3. And of course that's how our whole discussion of independent sequences got started, leading us to equation (7.1). Physical reasoning lets us go directly to the calculations we need for independent sequences, although it is not sufficient for a general proof.

Incidentally, when we cover independent random variables in Chapter 12, you will see a neater way to describe mathematical independence.

Remark 7.8 (Order does not matter for independence of sequences). Note that if $D_1, \ldots, D_n$ is an independent sequence as defined in Definition 7.7, then any *reordering* of the sequence is also independent. This is true because the intersection of sets does not depend on the order in which they are listed, and the product of their probabilities is also the same regardless of the order of the factors.

Your physical understanding of independence will make you confident that the next exercise is correct. But working out the solution is a good way to get a feeling for the mathematical definition.

Exercise 7.7. Let A, B, C be three sets which are mathematically independent in the sense of Definition 7.12. Based only on the *mathematical definition*, prove the following.

(i) Show that A, B are independent. (Suggestion: consider $A \cap B \cap C$ and $A \cap B \cap C^c$.)

(ii) Show that $A \cap B$ and C are independent.

[Solution]

Remark 7.9 (The length-one case). When $n = 1$, we should interpret $D_1 \cap \ldots \cap D_n$ simply as D_1. It follows that any length one sequence satisfies Definition 7.7. Thus every sequence of length one is a (rather boring) independent sequence.

Notice we are not saying that D_1 is independent of itself. In our present terminology that would be a statement about the length-two sequence D_1, D_1.

Whether or not you work Exercise 7.8, please be aware of the danger it points out.

Exercise 7.8 (Pairwise independence is not enough). Here's an important observation that comes up once in a while. For three events A, B, C, suppose that you know all possible *pairwise* independence statements hold, i.e.

- A, B is an independent pair, and
- B, C is an independent pair, and
- A, C is an independent pair.

You still *cannot* be sure that A, B, C is an independent sequence.

Here's an example. Consider tossing a fair coin twice. Let A be the event that the first toss produces a head, and let B be the event that that second toss produces a head. Let C be the event that the results of the two tosses *agree*, that is, $C = (A \cap B) \cup (A^c \cap B^c)$.

The statement of the experiment tells us that A, B are independent.

Show:

(i) that also B, C are independent and A, C are independent, but

(ii)
$$\mathbf{P}(A \cap B \cap C) \neq \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

Thus A, B, C is not an independent sequence.

(See Figure 5.1. The event C is the union of two of the four pieces shown in part (c) of the figure. Notice that each of the four pieces in part (c) has probability 1/4, so calculations should not be hard.)

[Solution]

7.5 Thinking about consistency again

In the case of two tosses of a coin, it was mentioned earlier that our physical experience makes us confident that probability models for two tosses are consistent with models for one toss. And we checked that in Exercise 2.4.

Let's look now at more general sequences of tosses. As in Example 2.6, suppose you are studying tossing a coin $1,000,000$ times, and using the sample space consisting of sequences $(x_1, \ldots, x_{1000000})$, where each x_i is either 1 or 0.

In this section we will check a couple of things.

First, let's check that the probabilities of the outcomes add up to one.

The following notation is handy.

Let $\theta(1) = p$ and let $\theta(0) = 1-p$. Notice, by the way, that $\theta(1)+\theta(0) = 1$.

Using the θ notation, equation (7.6) can be written neatly as

$$\mathbf{P}\left(D_1^{x_1} \cap \ldots \cap D_n^{x_n}\right) = \theta(x_1)\ldots\theta(x_n). \tag{7.13}$$

If we want to show that the probabilities of the outcomes add up to one, we must show that

$$\sum_{x_1,\ldots,x_n} \theta(x_1)\ldots\theta(x_n) = 1,$$

where the sum in this equation is over all possible values for $x_1, \ldots, x_n$, and each x_i can be 1 or 0.

We have to do something with that big sum on the left side of the equation.

Using the distributive law as much as possible, we see that

$$\underbrace{\big(\theta(1)+\theta(0)\big)\ldots\big(\theta(1)+\theta(0)\big)}_{n \text{ factors}} = \sum_{x_1,\ldots,x_n} \theta(x_1)\ldots\theta(x_n).$$

Since $\theta(1)+\theta(0) = 1$,

$$\sum_{x_1,\ldots,x_n} \theta(x_1)\ldots\theta(x_n) = \underbrace{1\times\ldots\times 1}_{n \text{ factors}} = 1,$$

so the probabilities of the outcomes do indeed add up to one!

Here's another exercise in checking.

Exercise 7.9 (Consistency). You've tackled this problem already in the case of a fair coin, in Exercise 2.5. Now consider the general case, when the coin is not necessarily fair, and has success probability p.

Fnd $\mathbf{P}(D_1^1)$ using the million-toss sample space. Remember, no peeking at the one-toss space!

[Solution]

7.6 Solutions for Chapter 5

Solution (Exercise 7.1). When $n = 2$, equation (7.3) is simply the statement of equation (7.2) with $k = 1$.

In general, using equation (7.2) repeatedly, we have

$$\begin{aligned}
\mathbf{P}\left(D_1 \cap \ldots \cap D_n\right) &= \mathbf{P}\left(D_1 \cap \ldots \cap D_{n-1}\right) \mathbf{P}\left(D_n\right), \\
&= \mathbf{P}\left(D_1 \cap \ldots \cap D_{n-2}\right) \mathbf{P}\left(D_{n-1}\right) \mathbf{P}\left(D_n\right), \\
&= \mathbf{P}\left(D_1 \cap \ldots \cap D_{n-3}\right) \mathbf{P}\left(D_{n-2}\right) \mathbf{P}\left(D_{n-1}\right) \mathbf{P}\left(D_n\right), \\
&\vdots \\
&= \mathbf{P}(D_1) \ldots \mathbf{P}(D_n).
\end{aligned}$$

A more formal solution would phrase this as an induction argument.

Solution (Exercise 7.2). Let $(x_1, \ldots, x_n)$ be a sequence of ones and zeros, for which exactly one of the numbers x_i is equal to one.

There are n choices for the index i with $x_i = 1$. Hence there are exactly n sequences of this type.

Solution (Exercise 7.3). Method 1 We consider the first three tosses as a separate experiment to find $\mathbf{P}(A)$. The sample space consists of sequences of length 3. $A = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1)\}$. Thus $\mathbf{P}(A) = 4(1/8) = 1/2$.

We consider the last two tosses as a separate experiment to find $\mathbf{P}(B)$. $B = \{(1,0), (0,1)\}$, so $\mathbf{P}(B) = 2(1/4) = 1/2$.

Using our physical understanding of independence, A and B should be independent, since they depend on separate tosses. Thus $\mathbf{P}(A \cap B) = (1/2)(1/2) = 1/4$.

Method 2 The sample space for the five tosses consists of 32 sequences of zeros and ones. All have the same probability, so $\mathbf{P}(\{\omega\}) = 1/32$ for every sample point ω.

A consists of all sequences of the form

$$(0,0,0,x_4,x_5) \text{ or } (1,0,0,x_4,x_5) \text{ or } (0,1,0,x_4,x_5) \text{ or } (0,0,1,x_4,x_5),$$

where x_4, x_5 can be zero or one. Thus A contains $4 \times 2 \times 2$ points, so $\mathbf{P}(A) = 16/32 = 1/2$.

B consists of all sequences of the form

$$(x_1,x_2,x_3,1,0) \text{ or } (x_1,x_2,x_3,0,1),$$

where x_1, x_2, x_3 can be zero or one. Thus B contains $2 \times 2 \times 2 \times 2$ points, so $\mathbf{P}(B) = 16/32 = 1/2$.

Consider a sample point $(x_1,x_2,x_3,x_4,x_5) \in A \cap B$. There are two possible cases:

- $x_i = 0$ for all $i = 1,2,3$ and $x_i = 1$ for exactly one of the indices $i = 4,5$. There are $1 \times 2 = 2$ ways to choose x_1,x_2,x_3,x_4,x_5, so there are two sample points for this case.

- $x_i = 1$ for exactly one of the indices $i = 1,2,3$, and $x_i = 1$ for exactly one of the indices $i = 4,5$, There are $3 \times 2 = 6$ ways to choose x_1,x_2,x_3,x_4,x_5, so there are six sample points for this case.

Since $A \cap B$ contains 8 sample points, $\mathbf{P}(A \cap B) = 8/32 = 1/4$.

Of course, Method 1 is more efficient, and conceptually clearer.

Solution (Exercise 7.4). Let L_i be the event that the first three tosses produce exactly i heads.

Let M_j be the event that the fourth, fifth and sixth tosses produce exactly j heads.

Let R_k be the event that the last two tosses produce exactly k heads.

Then

$$\mathbf{P}(L_i) = \binom{3}{i}\frac{1}{3^i},\ \mathbf{P}(M_j) = \binom{3}{j}\frac{1}{2^j},\ \text{and } \mathbf{P}(R_k) = \binom{2}{k}\frac{1}{2^k}.$$

Also, for any i,j,k, L_i, M_j, R_k is an independent sequence of events.

Clearly M_j is empty for $j > 3$. Thus $\mathbf{P}(C) = \mathbf{P}(C \cap M_0) + \mathbf{P}(C \cap M_1) + \mathbf{P}(C \cap M_2) + \mathbf{P}(C \cap M_3)$.

A is the event that the first six tosses produce 4 heads. Thus $A \cap M_0$ is empty, while $A \cap M_1 = L_3$, $A \cap M_2 = L_2$, and $A \cap M_3 = L_1$.

B is the event that the last five tosses produce 3 heads. Thus $B \cap M_0$ is empty, while $B \cap M_1 = R_2$, $B \cap M_2 = R_1$, $B \cap M_3 = R_0$.

Thus $C \cap M_0$ is empty, $C \cap M_1 = L_3 \cap M_1 \cap R_2$, $C \cap M_2 = L_2 \cap M_2 \cap R_1$, $C \cap M_3 = L_1 \cap M_3 \cap R_0$.

It follows that

$$\mathbf{P}(C) = \mathbf{P}(L_3)\mathbf{P}(M_1)\mathbf{P}(R_2) + \mathbf{P}(L_2)\mathbf{P}(M_2)\mathbf{P}(R_1) + \mathbf{P}(L_1)\mathbf{P}(M_3)\mathbf{P}(R_0).$$

Substituting,

$$\begin{aligned}\mathbf{P}(C) &= \binom{3}{3}\binom{3}{1}\binom{2}{2}\frac{1}{2^8} + \binom{3}{2}\binom{3}{2}\binom{2}{1}\frac{1}{2^8} + \binom{3}{1}\binom{3}{3}\binom{2}{0}\frac{1}{2^8} \\ &= \frac{1}{2^8}\left(1 \cdot 3 \cdot 1 + 3 \cdot 3 \cdot 2 + 3 \cdot 1 \cdot 1\right) \\ &= \frac{24}{2^8}.\end{aligned}$$

Solution (Exercise 7.5). Each outcome has probability $1/2^k$. By Theorem 2.13, if an event A contains ℓ outcomes then $\mathbf{P} = \ell/2^k$.

Solution (Exercise 7.6). (i) Let $(x_1, \ldots, x_{30})$ be a sample point in D_5^1.

Then $x_5 = 1$. For each index $i \neq 5$, there are two possible values for x_i. Hence there are 2^{29} choices for $(x_1, \ldots, x_{30})$, so $|D_5^1| = 2^{29}$.

(ii) Let $(x_1, \ldots, x_{30})$ be a sample point in $D_5^1 \cap D_7^0$.

Then $x_5 = 1$ and $x_7 = 0$. For each index distinct from 5 and 7, there are two possible values for x_i. Hence there are 2^{28} choices for $(x_1, \ldots, x_{30})$, so $|D_5^1 \cap D_7^0| = 2^{29}$.

(iii) Let $(x_1, \ldots, x_{30})$ be a sample point in $D_1^1 \cap D_1^0 \cap D_3^1 \cap D_4^0 \cap \ldots \cap D_{29}^1 \cap D_{30}^0$.

Then $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, $x_4 = 0$, etc. Thus $x_i = 1$ if i is odd and $x_i = 0$ if i is even. This is the only sample point in the event.

Solution (Exercise 7.7).

(i) From the definition of mathematical independence, $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$, and also $\mathbf{P}(A \cap B \cap C^c) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C^c)$.

Since $\mathbf{P}(C) + \mathbf{P}(C^c) = 1$ we have $\mathbf{P}(A \cap B \cap C) + \mathbf{P}(A \cap B \cap C^c) = \mathbf{P}(A)\mathbf{P}(B)$.

Also $(A \cap B \cap C) \cup (A \cap B \cap C^c) = A \cap B$, and the sets in this union are disjoint because C and C^c are disjoint. Hence by additivity we have $\mathbf{P}(A \cap B) = \mathbf{P}(A \cap B \cap C) + \mathbf{P}(A \cap B \cap C^c)$.

Thus we have shown that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

(ii) By definition, $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$. By part (i), $\mathbf{P}(A)\mathbf{P}(B) = \mathbf{P}(A \cap B)$. Hence $\mathbf{P}((A \cap B) \cap C) = \mathbf{P}(A \cap B)\mathbf{P}(C)$, as claimed.

Solution (Exercise 7.8). From the statement of the problem $\mathbf{P}(A) = 1/2$, $\mathbf{P}(B) = 1/2$, and A, B are independent.

Since A, B are independent, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) = 1/4$.

Using Lemma 5.6, $\mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c) = 1/4$, $\mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B) = 1/4$, and $\mathbf{P}(A^c \cap B^c) = \mathbf{P}(A^c)\mathbf{P}(B^c) = 1/4$.

Since $C = (A \cap B) \cup (A^c \cap B^c)$, $\mathbf{P}(C) = 1/4 + 1/4 = 1/2$.

(i) Then $\mathbf{P}(A \cap C) = \mathbf{P}(A \cap B) = 1/4$, $\mathbf{P}(B \cap C) = \mathbf{P}(A \cap B) = 1/4$, so A, C and B, C are independent.

(ii) However, $A \cap B \cap C = A \cap B$, so $\mathbf{P}(A \cap B \cap C) = 1/4$, while of course $\mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C) = 1/8$.

Solution (Exercise 7.9). Let $n = 1,000,000$.

When the coin has success probability p, and $p \neq 1/2$, sample points are not equally likely. So we need to use independence.

The event D_1^1 consists of points of the form $(1, x_2, \ldots, x_n)$.

That is,

$$D_1^1 = \{(1, x_2, \ldots, x_n) : \text{ where } x_i = 0 \text{ or } 1 \text{ for } i = 2, \ldots, n\}. \tag{7.14}$$

Thus

$$\mathbf{P}(D_1^1) = \sum_{x_2,\ldots,x_n} \theta(1)\theta(x_2)\ldots\theta(x_n) = p \sum_{x_2,\ldots,x_n} \theta(x_2)\ldots\theta(x_n),$$

where the sum is over all possible values of $x_2, \ldots, x_n$, and each x_i can be either 0 or 1, for $i = 2, \ldots, n$.

By using the distributive law as much as possible, we see that

$$\underbrace{(\theta(1) + \theta(0))\ldots(\theta(1) + \theta(0))}_{n-1 \text{ factors}} = \sum_{x_2,\ldots,x_n} \theta(x_2)\ldots\theta(x_n).$$

Hence

$$\sum_{x_2,\ldots,x_n} \theta(x_2)\ldots\theta(x_n) = \underbrace{1\times\ldots\times 1}_{n-1 \text{ factors}} = 1,$$

and so $\mathbf{P}(D_1^1) = p$, as it must.

Chapter 8

Counting

8.1 Counting ordered and unordered choices

Basic combinatoric methods, i.e. counting "permutations and combinations", are essential in analyzing many probability problems.

8.1.1 Ordered choices

Imagine that you are choosing a small administrative board to serve a club with n members. Three positions must be filled: president, vice-president and treasurer. No person can fill more than one position. The members of the board must be members of the club. How many possible boards are there?

Think of choosing the president first. This can be done in n ways. Next choose the vice president. Since one member of the group is already assigned to a position, you have $n-1$ choices for the vice-president. Finally, choose the treasurer from the remaining $n-2$ people. The total number of ways to do this is then $n(n-1)(n-2)$.

Is it obvious that we should count the total number of ways by multiplying the number of choices at each step? Sometimes people picture a "tree of possibilities" to see this. (Branches represent choices. There are n branches coming from the root, $n-1$ sub-branches coming from the tip of each of those branches, and then $n-2$ sub-sub-branches coming from the tip of each sub-branch.)

We can also compare this calculation with another one.

Suppose that a company owns 5 apartments buildings, and each building has 3 floors, and each floor has 7 apartments. If you are asked to choose one of the company's apartments to live in, you have a total of $5 \times 3 \times 7$ choices. Notice that in this situation, if you make a different choice at step one, you have a completely different set of choices for step two, and so on. Perhaps that makes it slightly easier to picture what is going on. In the case of choosing a board, if you make a different choice at step one, the set of possible choices for step two is only altered slightly. However, because you have already made a different choice at step one, there is no danger of counting the same board twice.

Of course, in any calculation like this, the fact that we can simply multiply the number of choices at each step depends on the fact that the *number* of possible choices at each step does not depend on what choices were made in previous steps.

The number of sequences of k distinct elements chosen from a set of n elements is often denoted by P_k^n. In case it is needed in a formula, we interpret P_0^n as 1, which means that we think there is only one way to choose zero elements.

The argument just given for choosing a board tells us that

$$P_k^n = n(n-1)\dots(n-k+1). \tag{8.1}$$

Formulas sometimes become neater if we use factorials. Equation (8.1) can be written as:

$$P_k^n = \frac{n!}{(n-k)!}. \tag{8.2}$$

For $k = n$ we use the standard convention that $0! = 1$ in this formula.

As a matter of terminology, a sequence of k distinct elements chosen from a set S is sometimes called a *permutation* of length k chosen from S.

8.1.2 Unordered choices

Now imagine that you are choosing a "clean-up" committee consisting of 3 members, from the club with n members. There are no special roles for the members of the this committee. They are simply supposed to work together to clean up after the next club meeting. You can still choose the members one at a time, but choosing the same three people in a different order just gives you the same committee.

If we think about choosing the committee members one by one, and count the number of ways to do that, we know that there are $n(n-1)(n-2)$ ways. But counting these ordered choices means counting the same committee multiple times. How many times?

A given clean-up committee is a set of 3 people. Equation (8.1), with $k = n = 3$, tells us there are 3! ways to perform the ordered choices which give us the same committee.

So the actual number of distinct possible clean-up committees is $n(n-1)(n-2)/3!$.

Generalizing the situation just described, let C_k^n denote the number of *subsets* of size k from a set of n members. A brief way to refer to C_k^n is "n choose k"". Sometimes people refer to a chosen subset as a "combination". Then C_k^n is "the number of combinations of n things taken k at a time".

Lemma 8.1. C_k^n is given by

$$C_k^n = \frac{P_k^n}{k!} = \frac{n(n-1)\dots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}, \tag{8.3}$$

where $\binom{n}{k}$ is the *binomial coefficient*, defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{8.4}$$

By the definition, $C_0^n = 1$. This is consistent with equation (8.1) with the standard convention that $0! = 1$.

Proof. We could imitate the argument just given when $k = 3$. But perhaps it's neater to rearrange the argument, as follows.

Consider choosing a sequence of k distinct elements from a set S containing n elements. When we thought about this choice, we chose the members in order, one at a time. But we can also carry out the choice in two stages.

In stage 1, select an unordered subset A of size k. By the definition of C_k^n, that can be done in C_k^n ways. We don't know the actual numerical value for C_k^n yet, but by definition C_k^n is the number of ways to choose A.

In stage 2, arrange the elements of A in order. By equation (8.1), with $n = k$, this can be done in $k!$ ways.

Clearly P_k^n is found by multiplying the number of ways to perform stage 1 times the number of ways to perform stage 2. Hence $P_k^n = C_k^n k!$, proving equation (8.3).

□

We define $\binom{n}{k} = 0$ if $k < 0$ or $k > n$. This makes equation (8.3) true for those values of k.

The reason $\binom{n}{k}$ is called the binomial coefficient will be clear from equation (8.6) below (the binomial theorem).

Lemma 8.2 (Counting successes using zeros and ones). Let S be the set of all sequences $(x_1, \ldots, x_n)$, where each x_i is zero or one.

Let A_k be the subset of S consisting of all sequences $(x_1, \ldots, x_n)$ such that $x_i = 1$ for exactly k indices i. Then

$$|A_k| = \binom{n}{k}. \tag{8.5}$$

Proof. We can specify any sequence $(x_1, \ldots, x_k)$ by simply specifying the *set* of indices i for which $x_i = 1$. Hence the number of sequences $(x_1, \ldots, x_k)$ which have k successes is exactly equal to the number of ways to choose a subset of size k from a set of size n. That is, $|A_k| = \binom{n}{k}$.

□

Lemma 8.2 is what we need to finish deriving equation (7.9). That equation gives the formula for the Binomial Distribution (Theorem 7.5).

8.2 The binomial theorem

We have used the binomial theorem from time to time in examples. Let's give a general statement and proof of this theorem now, for comparison with the proof that was just given for Theorem 7.5.

Consider expanding $(a + b)^n$. The usual first step is to write

$$(a+b)^n = \underbrace{(a+b)(a+b)\ldots(a+b)}_{n \text{ times}}.$$

The next step is to apply the distributive law energetically, resulting in 2^n terms. Notice that each of your 2^n terms is a product of n factors. The factors are a's and b's, where we choose either a or b from each of the n factors in the original expression for $(a+b)^n$.

To record a term, we could simply note the *set* of factors $(a+b)$ from which we chose a. That completely specifies the term.

For example, one of the terms in the expansion of

$$(a+b)(a+b)(a+b)(a+b)(a+b)$$

is *ababb*. We can record that term by saying that we chose a from the first and third factors and chose b from the other factors.

The final step is called "collecting like terms". Suppose you would like to combine all terms which are equal to $a^k b^{n-k}$. How many such terms are there?

The number of such terms is exactly the same as the number of ways in which you can select k factors from the n factors in the product $(a+b)^n$. Hence there are $\binom{n}{k}$ terms which are equal (after rearranging the order) to $a^k b^{n-k}$.

This proves the binomial theorem:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{8.6}$$

8.3 Two recursion formulas

To practice using counting arguments, we'll prove two recursive formulas for the binomial coefficients.

Here's the first one.

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}. \tag{8.7}$$

To prove (8.7), take any set of n elements, and choose one particular element for a special role.

When choosing a subset consisting of k elements, there are two possibilities. Either your subset contains the special element, or it does not.

If your subset does not contain the special element, then it is chosen from the other $n-1$ non-special elements. That can be done in $\binom{n-1}{k}$ ways.

If your set does contain the special element, then your subset is characterized by the $k-1$ non-special elements it contains. Those elements can be chosen in $\binom{n-1}{k-1}$ ways.

Combining the two cases proves (8.7).

Exercise 8.1. Check equation (8.7) using algebra.

[Solution]

Remark 8.3 (Pascal's triangle). This is a pictorial device based on equation (8.7). It is used to quickly find small binomial coefficients. One often writes coefficients in rows, with $\binom{n}{k}$, $k = 0, 1, \ldots, n$ in the n-th row. Elements to left or right of the binomial coefficients are taken to be zero, and the rows are *staggered*, meaning that each element in row n is placed in between the two nearest elements above it in row $n-1$. Equation (8.7) tells us that each element in row n is the sum of the two nearest elements in the preceding row. Thus:

$$\begin{array}{ccccccc} & & & 0 & 1 & 0 & \\ & & 0 & 1 & 1 & 0 & \\ & 0 & 1 & 2 & 1 & 0 & \\ 0 & 1 & 3 & 3 & 1 & 0 & \end{array}$$

and so on.

Here's another recursion formula. For any $n \geq 1$ and any $k \geq 1$,

$$\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}. \tag{8.8}$$

This formula has an easy algebraic proof.

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

Just for fun, let's make up a counting proof too. Think about choosing a special clean-up committee made up of k members from a club with n members. The members of the committee all must clean, but one member of the committee is chosen to be the boss of the committee. That member of the committee has a special role: the boss is responsible for making sure that the job is done well.

We can choose the special clean-up committee in two stages:

First stage: Choose a set of k members. That can be done in $\binom{n}{k}$ ways.

Second stage: Select one member from the k people in the chosen set, and make that person the boss. This can be done in k ways.

Hence the total number of possible special committees is

$$k\binom{n}{k}.$$

Alternatively, we can choose the special clean-up committee in a different way.

Alternate first stage: Select one person from n people in the club. The selected person will be the boss.

Alternate second stage: Select the remaining members of the clean-up committee from the remaining $n-1$ members of the club. This can be done in $\binom{n-1}{k-1}$ ways.

Hence total number of possible special committees is

$$n\binom{n-1}{k-1}.$$

Equating the two different expressions for the number of possible special committees gives equation (8.8).

The general theory of counting is known as combinatorics. It would be enjoyable to explore this interesting area, but we need to restrain ourselves and return to probability.

8.4 Random sets

8.4.1 Choosing a subset

Let S be a set containing a total of N elements. Consider the experiment of *randomly* choosing a subset consisting of n elements, in such a way that no element is favored. It seems most natural to represent a sample point by the actual subset of elements that are chosen. Let Ω be the collection of subsets of size n.

Let x be a particular point in the set S. The probability that x will be one of the n random elements chosen is n/N, as was shown in Theorem 2.22. Here we consider probabilities of choosing several particular points.

Exercise 8.2.

(i) For the experiment in this section, find the number of sample points in Ω.

(ii) Suppose you have a special interest in two of the elements in S, called x and y. Let A be the event that both x and y are in the selected set. Assume that $n > 1$. Find $\mathbf{P}(A)$.

(iii) Generalize your result in part (ii) to the situation where you are interested in a particular set T of elements, with $|T| = K$. Let A_T be the event that all the elements in T are in the selected set. Assume that $n \geq K$. Find $\mathbf{P}(A_T)$.

[Solution]

Example 8.4. In Section 4.6 we considered the situation of Exercise 4.2 and found two probabilities. Let's find the same probabilities using our counting tools.

In Exercise 4.2 we are choosing a set of two jelly beans from a bowl which contains 75 yellow beans, 53 red beans, 27 purple beans, and 18 green beans.

Thus there are a total of 173 jelly beans in the bowl.

Let R be the event that a set of two red jelly beans is obtained, and let M be the event that a set containing one red and one green jelly bean is obtained. We wish to find $\mathbf{P}(R)$ and $\mathbf{P}(M)$.

Let Ω be the collection of all two-point subsets of the set of jelly beans in the bowl. We assume that all sample points are equally likely. Clearly

$$|\Omega| = \binom{173}{2} = \frac{173 \cdot 172}{2},$$

so for each $\omega \in \Omega$ we have

$$\mathbf{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{2}{173 \cdot 172}.$$

The event R is the collection of all subsets made up of two red jelly beans. Thus

$$|R| = \binom{53}{2} = \frac{53 \cdot 52}{2}.$$

Hence

$$\mathbf{P}(R) = \frac{53 \cdot 52}{2} \frac{2}{173 \cdot 172} = \frac{53 \cdot 52}{173 \cdot 172}.$$

The event M is the collection of all subsets made up of one red jelly bean and one green jelly beans. There are 53 ways to choose the red bean and 18 ways to choose the green. Thus

$$|M| = 53 \cdot 18.$$

Hence

$$\mathbf{P}(M) = 53 \cdot 18 \frac{2}{173 \cdot 172} = 2 \frac{53 \cdot 18}{173 \cdot 172}.$$

Exercise 8.3. Consider the situation of Exercise 8.2, part (ii). In addition to x and y, suppose you are also interested in a third point z. Let B be the event that y and z are both in the selected set. Find $\mathbf{P}(B \mid A)$.

[Solution]

When choosing sets, one has to be careful in labelling the sizes correctly, and counting. It's not hard, just takes care. Let's do a little practicing with that, next.

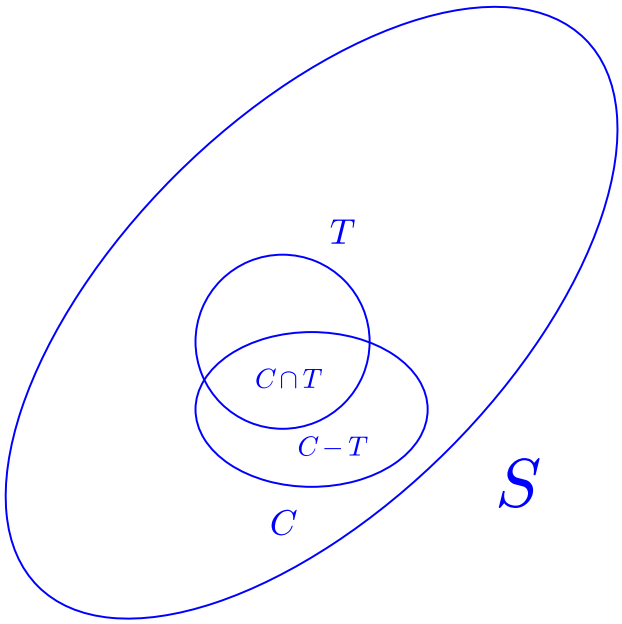


Figure 8.1: Lemma 8.5: $|S| = N$, $|T| = K$, $|C| = n$, $|C \cap T| = i$, $|C - (C \cap T)| = n - i$

Exercise 8.4. A bowl contains N marbles. K of the marbles are red, the rest are green. A subset consisting of n marbles is selected. No marble is favored. Let R_i be the event that there are exactly i red marbles in the selected set.

From the description of the problem,

$$0 \leq K \leq N, \ 0 \leq n \leq N. \tag{8.9}$$

(i) Suppose that all the following inequalities hold:

$$0 \leq i \leq K, \tag{8.10}$$

$$i \leq n, \tag{8.11}$$

$$K - i \leq N - n. \tag{8.12}$$

Find $\mathbf{P}(R_i)$.

(ii) Show that equations (8.10), (8.11), and (8.12) must hold in for any possible outcome. Thus R_i must be empty, i.e. $R_i = \emptyset$, when i does not satisfy all the inequalities in those equations.

(iii) For what value of i is your answer to part (i) already given by Exercise 8.2?

[Solution]

Abstractly, Exercise 8.4 deals with a set S of size N, having a specified subset T of size K. In this situation another subset C, having size n, is chosen.

For the moment, don't think about how C is chosen. Let's just consider a simple question: if all we know is that C is a subset of S with size n, what are the possible values for the size of $C \cap T$?

We'll repeat the arguments for Exercise 8.4, starting by writing down inequalities. Clearly $K \leq n$ and $n \leq n$.

Let i denote the size of $C \cap T$. Then we must have $i \geq 0$, $i \leq K$, $i \leq n$. Must i satisfy any other inequalities?

Well, note that $C = (C \cap T) \cup (C - (C \cap T))$. So $|C - (C \cap T)| = n - i$, and the elements of $C - (C \cap T)$ are in $S - T$. So we also have $n - i \leq N - K$, or equivalently $n + K \leq N + i$, which is equivalent to $K - i \leq N - n$.

Notice that in the solution for Exercise 8.4 we argue slightly differently to obtain the same inequality, as follows. $T = (C \cap T) \cup (T - (C \cap T))$. So $|T - (C \cap T)| = K - i$, and the elements of $T - (C \cap T)$ are in $S - C$. So we also have $K - i \leq N - n$.

Incidentally, we might rewrite equations (8.10), (8.11), and (8.12) more symmetrically:

$$\begin{aligned} 0 &\leq i, \\ i &\leq K, \\ i &\leq n, \\ K + n &\leq N + i. \end{aligned} \tag{8.13}$$

These inequalities in equation (8.13) are symmetric in K and n. They had to be, because we have not used any information here which treats T and C differently.

We have established that the size i of $C \cap T$ must satisfy equation (8.13). As shown in Exercise 8.4, no more conditions are needed. The following lemma asserts all this for the record. The remarks already given can easily be turned into a more formal proof.

Lemma 8.5 (Possible intersections of two subsets). Let S be a set with $|S| = N$, and let T be a subset with $|T| = K$. Let n be an integer with $0 \leq n \leq n$.

For any integer i satisfying equation (8.13), there exists a subset C with $|C| = n$, such that $|C \cap T| = i$.

Conversely, if C is a subset of S with size n, then $i = |C \cap T|$ satisfies equation (8.13). See Figure 8.1.

Definition 8.6 (The hypergeometric distribution). In Exercise 8.4 it was shown that

$$\mathbf{P}(R_i) = \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{8.14}$$

when i satisfies the inequalities in equation (8.13) (or equivalently when i satisfies equations (8.10), (8.11), and (8.12)).

For any distribution, if equation (8.14) holds when i satisfies the inequalities in equation (8.13), with $\mathbf{P}(R_i) = 0$ otherwise, we say that the distribution is the *hypergeometric distribution*, with parameters N, K, n.

Since notations in other books will differ, for applications remember that:

- N is the total size of the population from which a random sample of size n is selected,
- K is the size of a set of special elements in the population, and
- i is the number of special elements that are in the sample.

8.4.2 Choosing a sequence

Let S be a set containing n elements. Consider the experiment of *randomly* choosing a sequence consisting of n elements, in such a way that no element

is favored during the successive choices. It seems most natural to represent a sample point as the actual sequence of elements that are chosen. Let Ω be the collection of subsequences of size n.

Exercise 8.5.

(i) Find the number of sample points in Ω.

(ii) Suppose you have a special interest in two distinct elements in S, called x and y. Let A be the event that x is the third point chosen and y is the seventh point chosen. (Assume that $n > 6$.) Find $\mathbf{P}(A)$. [Solution]

Exercise 8.6. In the setting of Exercise 8.5, suppose that, in addition to x and y, you are also interested in a point z, which is different from x and y. Let B be the event that y is the seventh point chosen and z is fifth point chosen. Find $\mathbf{P}(B \mid A)$. (Assume $n > 6$.)

[Solution]

Exercise 8.7. Recall Exercise 8.4. In that experiment, a subset consisting of n marbles is chosen randomly from a bowl of n marbles, and R_i is the event that exactly i of the chosen marbles are red. The total number of red marbles in the bowl is K.

The final goal is to find $\mathbf{P}(R_i)$, but one may decide to choose the marbles in the subset one at a time, afterwards ignoring the order in which the marbles are chosen. Let's try that approach.

The description of the experiment shows that the values of $\mathbf{P}(R_i)$ follow the *hypergeometric distribution*, with parameters N, K, n (Definition 8.14. So the approach we are trying now must eventually eventually produce the formula for this distribution which was already given in equation (8.14).

To compare our model with Bernoulli trials (Section 7.3), let a sample point be $\omega = (\omega_1, \ldots, \omega_n)$, where ω_ℓ is the marble chosen at step ℓ. The number of red marbles chosen is then the number of indices ℓ such that ω_ℓ is red.

R_i is the set of all outcomes $(\omega_1, \ldots, \omega_n)$ such that exactly i of the marbles ω_ℓ are red.

Much as when we studied coin tossing, let W_ℓ be the event that the ℓ-th marble chosen is red, so that W_ℓ^c is the event that the ℓ-th marble chosen is not red. Then for any $\omega = (\omega_1, \ldots, \omega_n)$,

$$\{\omega\} = D_1 \cap \ldots \cap D_n, \tag{8.15}$$

where $D_\ell = W_\ell$ if ω_ℓ is red and $D_\ell = W_\ell^c$ if ω_ℓ is not red.

Thus R_i is the union of all events of the form $D_1 \cap \ldots \cap D_n$, where for each t, either $D_t = W_t$ or $D_t = W_t^c$, and where the $D_t = W_t$ for exactly i of the times t.

The next step in this approach would be to calculate $\mathbf{P}(D_1 \cap \ldots \cap D_n)$. But at this point there is an obstacle, since we don't have independence.

Explain why the events $D_1, \ldots, D_n$ are not independent.

[Solution]

The next exercise shows that the sequence model considered in Exercise 8.7 eventually leads to equation (8.14), as it should. There are more steps this way, but the steps are not hard.

Exercise 8.8. In the setting of Exercise 8.7, consider the events $D_1 \cap \ldots \cap D_n$ which make up R_i. Show that every event $D_1 \cap \ldots \cap D_n$ has the same probability, and find this probability.

A good way to think about this is to use the sample space of Exercise 8.7.

Thus a sample point is $\omega = (\omega_1, \ldots, \omega_n)$, where ω_ℓ is the marble chosen at step ℓ. The number of red marbles chosen is then the number of indices ℓ such that ω_ℓ is red.

Since no marble is favored, all sequences $(\omega_1, \ldots, \omega_n)$ have the same probability.

[Solution]

Example 8.7 (Plotting a hypergeometric distribution). In the setting of Exercise 8.7, suppose we have a bowl with 120 marbles, 40 of which are red and the rest green. If you randomly select a single marble from this bowl, the probability of a red marble is 1/3 (by Theorem 2.22, say).

In this setting, consider the experiment of Exercise 8.4 with the total number of marbles equal to 120, the total number of red marbles equal to 40, and 30 marbles are chosen from the bowl.

In our present notation $N = 120$, $K = 40$ and $n = 30$.

We are interested in $\mathbf{P}(R_i)$, where R_i is the event that i red marbles are obtained.

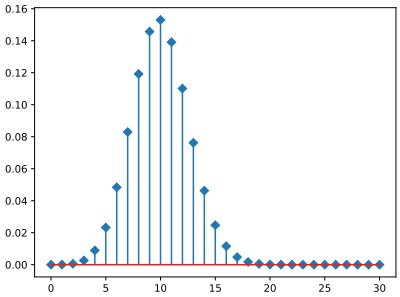
This experiment differs from n Bernoulli trials, since the successive colors obtained in the n selections are not independent, and the values of $\mathbf{P}(R_n)$ follow the hypergeometric distribution with parameters N, K, n. We will use the formula for this distribution derived earlier in Exercise 8.4.

Figure 8.2b shows the graph of $\mathbf{P}(R_i)$ versus i for one experiment ($N = 120$, $K = 40$, $n = 30$). We see that this graph is similar to the coin-tossing graph in Figure 8.2a, but the graphs are not identical.

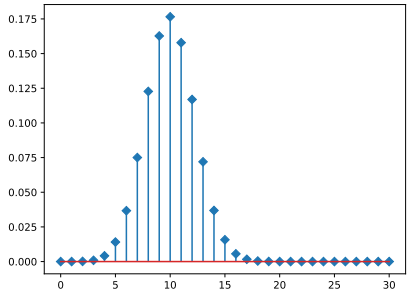
For comparison, Figure 8.2c shows the graph of $\mathbf{P}(R_i)$ versus i when $N = 12000$, $K = 4000$, $n = 30$. We see that this graph is almost identical to the graph in Figure 8.2a. Why is that??

Exercise 8.9. Suggest an answer to the question posed at the very end of Example 8.7.

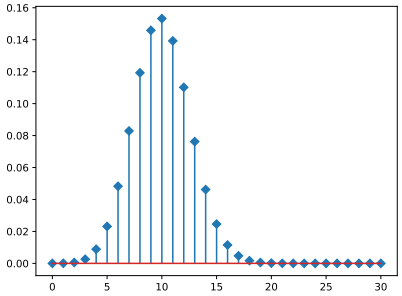
[Solution]



(a) Figure 7.1 again. $\mathbf{P}(k \text{ heads})$, 30 tosses, success prob 1/3.



(b) Probability that a randomly selected subset of size 30 from a set of 120 contains i red marbles, when 40 out of the 120 are red.



(c) Probability that a randomly selected subset of size 30 from a set of 12000 contains i red marbles, when 4000 out of the 12000 are red.

8.5 Solutions for Chapter 8

Solution (Exercise 8.1).

Proof.

$$\binom{n-1}{k} + \binom{n-1}{k-1} = \frac{(n-1)!}{k!(n-k-1)!} + \frac{(n-1)!}{(k-1)!(n-k)!}.$$

Bringing the summands to a common denominator gives

$$\frac{(n-k)(n-1)! + k(n-1)!}{k!(n-k)!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

□

Solution (Exercise 8.2).

(i) By definition, a sample point is a subset of S having size n. Hence there are exactly $\binom{N}{n}$ sample points.

(ii) Since no element is favored in the selection, every sample point must have the same probability. Hence each sample point has probability $1/\binom{N}{n}$. By additivity,

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}(\{\omega\}) = |A| \, \frac{1}{\binom{N}{n}}.$$

A sample point in A is a subset of S containing x and y and $n-2$ additional elements from S. $|A|$ is equal to the number of ways to choose $n-2$ elements from $S - \{x, y\}$. Hence

$$\mathbf{P}(A) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}}.$$

(iii) Similar arguments show that

$$\mathbf{P}\left(A_T\right) = \frac{\binom{N-K}{n-K}}{\binom{N}{n}}. \tag{8.16}$$

Solution (Exercise 8.3). Using, say, Exercise 8.2, part (iii),

$$\mathbf{P}(A) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}}, \ \mathbf{P}(A \cap B) = \frac{\binom{N-3}{n-3}}{\binom{N}{n}}.$$

Hence

$$\mathbf{P}(B \mid A) = \frac{\frac{\binom{N-3}{n-3}}{\binom{N}{n}}}{\frac{\binom{N-2}{n-2}}{\binom{N}{n}}} = \frac{\binom{N-3}{n-3}}{\binom{N-2}{n-2}} = \frac{\frac{(N-3)!}{(n-3)!\,((N-3)-(n-3))!}}{\frac{(N-2)!}{(n-2)!(N-2)-(n-2))!}}$$

$$= \frac{\frac{(N-3)!}{(n-3)!(N-n)!}}{\frac{(N-2)!}{(n-2)!(N-n)!}} = \frac{(N-3)!}{(N-2)!}\frac{(n-2)!}{(n-3)!} = \frac{n-2}{N-2}.$$

We have followed the general pattern of the conditional probability formula here. However, we would arrive at the same value for $\mathbf{P}(B \mid A)$ if we considered the selection of x and y as part of the setting of the experiment, so that the experiment consists of choosing the *rest* of the sample. Now the sample space Ω' is the set of all subsets of $S - \{x, y\}$, and we want the probability that when choosing $n-2$ points from Ω', the point z is in the chosen set. By Theorem 2.22, this probability is $(n-2)/(N-2)$.

In this way we don't need use of the conditional probability formula. The second method is a common approach to conditional probability problems.

Solution (Exercise 8.4).

(i) A sample point ω is a subset of size n. Suppose that $\omega \in R_i$.

Since $i \leq K$ and $i \leq n$, there are $\binom{K}{i}$ choices for the red marbles in ω.

Since $n - i \leq N - K$, there are $\binom{N-K}{n-i}$ choices for the non-red marbles in ω.

Hence there are $\binom{K}{i}\binom{N-K}{n-i}$ choices for ω in R_i, that is, $|R_i| = \binom{K}{i}\binom{N-K}{n-i}$. As usual, $|\Omega| = \binom{N}{n}$. Hence

$$\mathbf{P}(R_i) = \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}. \tag{8.17}$$

(ii) Equation (8.10) says that every chosen red marble is a red marble.

Equation (8.11) says that every chosen red marble is a chosen marble.

Equation (8.12) says that every remaining red marble is a remaining marble.

So all three of these equations must hold.

(iii) R_K is simply the event that *all* the red marbles are chosen. If we think of the red marbles as the elements of interest in the set, then part (iii) of Exercise 8.2. tells us that $\mathbf{P}(R_K) = \frac{\binom{N-K}{n-K}}{\binom{N}{n}}$.

Solution (Exercise 8.5).

(i) A sample point is a sequence of distinct elements, having length n, so there are exactly P_n^N sample points, where P_n^N is given by equation (8.1),

$$P_n^N = N(N-1)\dots(N-n+1) = \frac{N!}{(N-n)!}.$$

(ii) Since no element of S is favored in selecting the sequence, all sample points have the same probability. Thus $\mathbf{P}(\{\omega\}) = 1/P_n^N$ for all ω.

For any sequence ω in the event A, we are given the positions of x and y in the sequence. Thus the ω is determined once the other elements in the sequence are determined. The other $n-2$ elements in the sequence form a sequence consisting of distinct elements from $S-\{x,y\}$. Since $|S-\{x,y\}| = N-2$, there are P_{n-2}^{N-2} choices for the other elements in the sequence. This shows that $|A| = P_{n-2}^{N-2}$.

$$\begin{aligned}\mathbf{P}(A) = \sum_{\omega\in A}\mathbf{P}(\{\omega\}) = |A|\,\frac{1}{P_n^N} &= \frac{P_{n-2}^{N-2}}{P_n^N} \\ &= \frac{(N-2)\dots(\,(N-2)-(n-2)+1)}{N\dots(N-n+1)} = \frac{1}{N(N-1)}. \quad (8.18)\end{aligned}$$

The positions of x and y in the sequence were given, but we see that the probability is the same for any fixed choices of the positions.

Solution (Exercise 8.6).

$$\begin{aligned}\mathbf{P}(B\,|\,A) = \frac{\mathbf{P}(A\cap B)}{\mathbf{P}(A)} = \frac{\frac{P_{n-3}^{N-3}}{P_n^N}}{\frac{P_{n-2}^{N-2}}{P_n^N}} &= \frac{P_{n-3}^{N-3}}{P_{n-2}^{N-2}} \\ &= \frac{(N-3)\dots(N-3-(n-3)+1)}{(N-2)\dots((N-2)-(n-2)+1} = \frac{N-3}{N-2}.\end{aligned}$$

As in the solution for Exercise 8.3, we won't need the conditional probability formula if we take the setting of our experiment to include the fact that x is the third point chosen and y is the seventh point chosen.

Solution (Exercise 8.7). To see what is going on, take $n = 2$.

Suppose $D_1 = W_1$ and $D_2 = W_2$. Then $\mathbf{P}(D_1) = K/N$ and $\mathbf{P}(D_2) = K/N$.

However, we can find $\mathbf{P}(D_2 \,|\, D_1)$ by thinking of the second choice as a self-contained experiment, with $N-1$ marbles $K-1$ red marbles. Hence

$$\mathbf{P}(D_2 \,|\, D_1) = \frac{K-1}{N-1} \neq \mathbf{P}(D_2).$$

Thus independence does not hold.

Solution (Exercise 8.8). We'll use the sample space of Exercise 8.7.

Thus a sample point is $\omega = (\omega_1, \ldots, \omega_n)$, where ω_ℓ is the marble chosen at step ℓ. The number of red marbles chosen is then the number of indices ℓ such that ω_ℓ is red.

By equation (8.2),

$$|\Omega| = \frac{N!}{(N-n)!}.$$

Since all marbles are treated the same way, all sample points have the same probability. For any $(\omega_1, \ldots, \omega_n)$,

$$\mathbf{P}(\{(\omega_1, \ldots, \omega_n)\}) = \frac{1}{|\Omega|} = \frac{(N-n)!}{N!}. \tag{8.19}$$

Let K be total number of red marbles in the bowl.

Fix i, $1 \leq i \leq n$ and $i \leq K$. Let $D_1 \cap \ldots \cap D_n$ be such that $D_1 \cap \ldots \cap D_n \subset R_i$.

Consider $(\omega_1, \ldots, \omega_n) \in D_1 \cap \ldots \cap D_n$. Then ω_ℓ is red for exactly i of the indices ℓ. Let V be the set of indices ℓ such that ω_ℓ is red. Let $W = \{1, \ldots, n\} - V$ be the other indices.

There are i indices in V. So, using equation (8.2), the number of ways to choose x_ℓ for $\ell \in V$ is

$$\frac{K!}{(K-i)!}.$$

That is the number of ways to fill the red indices. The number of non-red indices is $n - i$ And the number of non-red marbles is $N - K$. Hence the number of ways to fill the non-red indices is: is

$$\frac{(N-K)!}{((N-K)-(n-i))!}.$$

Hence

$$|D_1 \cap \ldots \cap D_n| = \frac{K!}{(K-i)!} \; \frac{(N-K)!}{((N-K)-(n-i))!}.$$

Using equation (8.19),

$$\mathbf{P}(D_1 \cap \ldots \cap D_n) = \frac{K!}{(K-i)!} \; \frac{(N-K)!}{((N-K)-(n-i))!} \frac{(N-n)!}{N!}.$$

Thus

$$\mathbf{P}(D_1 \cap \ldots \cap D_n) = \frac{K!(N-K)!}{N!} \frac{(N-n)!}{(K-i)!\,((N-K)-(n-i))!}. \tag{8.20}$$

Equation (8.20) shows in particular that $\mathbf{P}(D_1 \cap \ldots \cap D_n)$ has the same value for any set $D_1 \cap \ldots \cap D_n$ contained in R_i.

To check the probability value given by equation (8.20), note that a particular event $D_1 \cap \ldots \cap D_n$ is determined once the set V of red indices is chosen. Hence there are $\binom{n}{i}$ events $D_1 \cap \ldots \cap D_n$ contained in R_i, and so

$$\mathbf{P}(R_i) = \binom{n}{i} \frac{K!(N-K)!}{N!} \frac{(N-n)!}{(K-i)!\,((N-K)-(n-i))!},$$

i.e.

$$\mathbf{P}(R_i) = \frac{n!}{i!(n-i)!} \frac{K!(N-K)!}{N!} \frac{(N-n)!}{(K-i)!\,((N-K)-(n-i))!}. \tag{8.21}$$

That is,

$$\mathbf{P}(R_i) = \frac{K!}{i!(K-i)!} \; \frac{(N-K)!}{(n-i)!\,((N-K)-(n-i))!} \; \frac{n!(N-n)!}{N!}.$$

This agrees with equation (8.14), which says that

$$\mathbf{P}(R_i) = \frac{\binom{K}{i}\binom{(N-K)}{n-i}}{\binom{N}{n}}.$$

Solution (Exercise 8.9). When the total number of marbles is large, and the total number of red marbles is large, choosing one marble has little effect, meaning that the chance of a red marble on a second choice is almost the

same as it was on the first choice, regardless of whether or not the first marble was red. Thus the outcomes (red or non-red) are almost independent, and choosing a 30 marbles is almost like 30 independent trials in coin-tossing.

Incidentally, that graph in Figure 8.2b looks a bit narrower than the graph in Figure 8.2a, doesn't it? Could there be a reason for that?

A reason is given Remark 16.26.

Chapter 9

Random variables

In this chapter we introduce a new concept, random variables. The benefits of using this concept will become evident in the chapters that follow. The current chapter has important definitions and examples, but not much in the way of applications. Readers should try to enjoy the peaceful contemplation of well-chosen concepts. This investment will pay off later.

9.1 Random variables defined

Definition 9.1 (Random variables). Physically, a random variable for an experiment is a quantity whose value depends on the outcome of the experiment. In our discussions the quantity will usually be represented by a number, but it might be represented by a vector, a set, a symbol, or some other property.

Mathematically, a random variable is a function whose domain is the sample space Ω of a model, and whose values can be of any kind. A real-valued random variable is a function from a sample space Ω to the real numbers.

To save words, we often simply use the phrase "random variable" to mean a real-valued random variable, since that is the most common case for us. Since we are following that convention, when we are dealing with a random variable whose values are *not* numbers, we'll try to say what the values are, or at least add an adjective to make that clear. For example we might speak of a "vector-valued random vector", or a "random vector", to indicate that our random variable has vector values.

By convention, random variables are normally denoted by uppercase letters, with X, Y, Z being the most common choices.

Definition 9.1 is not quite complete, since it omits a mathematical technicality. This technicality has no practical significance for our work in this book, and can be safely ignored, but a brief discussion is given below in Section 9.7.

Notation for properties and sets of values

Suppose that X is a real-valued random variable, and that for some reason we want to consider the probability that the value of X is greater than five and less than eight.

The usual notation for this probability is $\mathbf{P}(5 < X < 8)$. That expresses the probability using "property language".

We can also write the same probability as $\mathbf{P}(\, X \in (5, 8)\,)$. That expresses the probability using "set language".

The same kind of notation is used in general. If S is the set of all possible values that have the property that we are interested in, we can write $\mathbf{P}(X \in S)$ to denote the probability that the value of X lies in the set S.

We could use a more formal mathematical notation for $\mathbf{P}(X \in S)$. If X is a function on a sample space Ω, we could define an event A by

$$A = \{\omega : \ X(\omega) \in S\}\,. \tag{9.1}$$

Then $\mathbf{P}(A)$ would be $\mathbf{P}(X \in S)$.

But usually it is more convenient to use the briefer notations which are common in probability. So we just write $\mathbf{P}(X \in S)$ instead of defining A. The notations used in probability theory are the following:

$$\begin{aligned} \{X \in S\} &= \{\omega : \ X(\omega) \in S\}\,, \\ \mathbf{P}(X \in S) &= \mathbf{P}(\{X \in S\}) = \mathbf{P}(\{\omega : \ X(\omega) \in S\}). \end{aligned} \tag{9.2}$$

Readers will find that this type of notation is quite clear and easy to read.

Notation for random variables Since a random variable is a function, why not use a typical function name, such as "f", instead of an uppercase letter? Perhaps an uppercase letter is used to remind the reader that the domain of a random variable is the set of possible outcomes for an experiment.

This can be very different from the domain of a function in calculus, which is usually an interval of the real line.

Calling a random variable "X" offers another benefit. If we want to refer to a value of the random variable X, we can denote the value by the lowercase letter "x". This reminds the reader of the source of the value.

In this chapter we will mainly consider a random variable whose set of possible values is finite, i.e. a random variable with *finite range*. But most of the concepts make sense for general mathematical random variables.

Example 9.2 (The result of one coin toss). For a single toss of a coin, let $X = 1$ if the result is a head, and let $X = 0$ if the result is a tail. In other words, X is equal to the number of heads obtained by this toss.

If the probability of a head is p, then the probability that $X = 1$ is equal to p, and the probability that $X = 0$ is equal to $1 - p$.

X is a very boring random variable! However, we will soon see that more interesting random variables can be built using random variables like X.

To represent X mathematically, if the sample space Ω is equal to the two-point set $\{1, 0\}$, as in Example 2.14, then $X(\omega) = \omega$, but, as usual in applications, there is no need to use any particular sample space. Given a physical random variable X, the mathematical random variable representing X is valid if it has the correct values and produces those values with the correct probabilities.

Remark 9.3 (An example of an alternate sample space for one coin toss). To emphasize the fact that the sample space is not unique, here's a extreme example of an alternate sample space. We could take Ω equal to the whole unit interval $[0, 1]$, and use the uniform probability $\mathbf{P}$ on $[0, 1]$. In this case, define a random variable $\tilde{X}$ by $\tilde{X}(\omega) = 1$ if $0 \leq \omega < p$, and $\tilde{X}(\omega) = 0$ if $p \leq \omega \leq 1$. (Think about randomly choosing a point in the unit interval (with a uniform probability distribution) and shouting "success!" or "Pay me!" if the chosen point lies in the interval $[0, p)$.)

Notice that with this definition we have arranged matters so that the possible values of $\tilde{X}$ are 1 and 0, the probability that $\tilde{X} = 1$ is equal to p, and the probability that $\tilde{X} = 0$ is equal to $1 - p$. This exactly matches the physically observable behavior of the random variable X which was defined on the two-point set $\{0, 1\}$.

Remember, what matters are the values, and the probabilities of those values. That's what is "real" about the mathematical random variable.

One might certainly say that using $\Omega = [0, 1]$ is wasteful, since we don't need such a big sample space, but the sample space is not incorrect. It might even be appropriate in a complicated experiment, if there are additional properties that must be represented using the same sample space.

Exercise 9.1 (Notation check). Let $\tilde{X}$ be the random variable defined in Remark 9.3. Let $S = \{0\}$. Find $\left\{\tilde{X} \in S\right\}$.

[Solution

Example 9.4 (The result of one roll of a die). For a single roll of a die, let X be the number that shows on the die when it comes to rest. Then the possible values for X are $1, 2, 3, 4, 5, 6$.

If the die is fair, then the probability that $X = i$ is equal to $1/6$ for all i. In general, the probability that $X = i$ will be some probability p_i, where $p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1$.

If the sample space Ω is equal to the six-point set $\{1, 2, 3, 4, 5, 6\}$, then $X(\omega) = \omega$, but, as Remark 9.3 illustrates, we could can always use some other sample space.

The language of random variables gives us a new way to describe some events, but we still find probabilities using the same rules. The next exercise illustrates this.

Exercise 9.2. Let X be the random variable defined in Example 9.4.

- Let $A = \{X > 4\}$, and let B be the event that X is an even number. Find A and B, as subsets of $\{1, 2, 3, 4, 5, 6\}$.
- Find the probability that X is an even number.

[Solution]

Example 9.5 (Number of successes in Bernoulli trials (coin tosses)). Let $A_1, \ldots, A_n$ be Bernoulli trials (see Section 7.3) with success probability p. That is, $A_1, \ldots, A_n$ are independent and $\mathbf{P}(A_i) = p$ for each i.

As our most typical example, the experiment consist of tossing a coin n times, and A_i might be the event that toss i gives a head.

Let S_n be the total number of successes. (This notation is *not* related to our earlier use of S to denote some set.) By definition, S_n is the number of indices i such that A_i occurs. For the experiment of tossing a coin n times, S_n is the number of heads which are obtained.

The possible values for S_n are $0, 1, \ldots n$, so S_n has a fairly simple range.

The event $\{S_n = k\}$ is exactly the event G_k described in Theorem 7.5. Thus equation (7.10) states that

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{9.3}$$

9.2 The probability of obtaining a value in a set

Random variables are implicitly present in any probability model. Using the terminology of random variables explicitly is often convenient, even when we are performing the same old calculations.

Readers might want to look one more time at the calculation in the second solution of Exercise 9.2. The calculation is trivial, of course, but it feels liberating to simply write

$$\mathbf{P}(X \text{ is even}) = \mathbf{P}(X = 2) + \mathbf{P}(X = 4) + \mathbf{P}(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2},$$

without giving a thought to the sample space.

It's useful to state a general version of the same argument.

Let X be any random variable, and let S be any set such that S only contains a finite number of points in the range of X. Let $x_1, \ldots, x_k$ be the numbers in the range of X that are members of S, listed in any order, without repetitions. If the value of X is a member of S, then the value of X must be

equal to one of the numbers $x_1, \ldots, x_k$. Thus

$$\{X \in S\} = \{X = x_1\} \cup \ldots \cup \{X = x_k\}. \tag{9.4}$$

Since the values $x_1, \ldots, x_k$ are distinct, the sets $\{X = x_1\}, \ldots, \{X = x_k\}$ are disjoint. By the additivity of probability we have

$$\mathbf{P}(X \in S) = \mathbf{P}(X = x_1) + \ldots + \mathbf{P}(X = x_k). \tag{9.5}$$

We typically use facts like equation (9.5) without comment. Equations of this sort help us to think about events in terms of what they mean, rather than as subsets of the abstract sample space.

Remark 9.6 (Adding some values which are outside the range). In equations (9.4) and (9.5), suppose we increased the list $x_1, \ldots, x_k$ by including some additional numbers which are *not* in the range of X. If a is a number which is not in the range of X, then of course $\{X = a\} = \emptyset$, the empty set. So equations (9.4) and (9.5) will continue to hold.

The next exercise extends equation (9.5) to general sets. All readers should note the statements, and think about them enough to see that they are true.

Exercise 9.3 (Cases for a random variable). Let X be any random variable.

(i) For any sets $S_1, \ldots, S_k$, which need not be subsets of the range of X, if $S = S_1 \cup \ldots \cup S_k$, show that

$$\{X \in S\} = \{X \in S_1\} \cup \ldots \cup \{X \in S_k\}. \tag{9.6}$$

Some of the sets $\{X \in S_i\}$ may be empty, but that's fine.

(ii) Suppose now that the sets $S_1, \ldots, S_k$ are disjoint. Show that the sets $\{X \in S_1\}, \ldots, \{X \in S_k\}$ are disjoint. Let $S = S_1 \cup \ldots \cup S_k$. Show that

$$\mathbf{P}(X \in S) = \mathbf{P}(X \in S_1) + \ldots + \mathbf{P}(X \in S_k). \tag{9.7}$$

(By taking $S_j = \{x_j\}$, we see that equation (9.7) includes equation (9.5) as a special case.)

[Solution]

9.3 Estimating probability sums

For the random variable S_n in Example 9.5, equations (9.3) and (9.5) can be used to find the probability of any event defined in terms of the value of S_n. However, when n is large it may take some work to extract the information we need.

For example, toss a fair coin a million times. Let $n = 1000000$, so S_n is the number of heads obtained. We are rarely interested in the tiny probability that S_n is exactly equal to $499,500$. But we might be interested in, say, the probability that at most 49.95% of the tosses resulted in a head, i.e. $\mathbf{P}(S_n \leq 499,500)$. How do we find this probability, or at least estimate this probability in some way? We know from equation (9.3) and additivity that the probability is given by

$$\mathbf{P}(S_n \leq 499,500) = \sum_{j=0}^{499,500} \binom{1,000,000}{j} \left(\frac{1}{2}\right)^{1,000,000}. \tag{9.8}$$

True, but the size of this number does not exactly leap out at us. Not only are there many terms, but a typical term in this sum is the product of a very large number times a very small number.

The Central Limit Theorem ([10], Chapter 18) is a powerful method for estimating probabilities like $\mathbf{P}(S_n \leq 499,500)$. Incidentally, the Central Limit Theorem tells us that $\mathbf{P}(S_n \leq 499,500)$ is approximately equal to .159 (Exercise 18.7.).

9.4 Random variable distributions

Recall that we introduced the general idea of a probability distribution in Definition 1.11. Any rule which assigns probabilities for a family of events can be called a probability distribution. The next definition is a very important example of this terminology.

Definition 9.7 (The distribution of a general random variable). For any real-valued random variable X associated with any probability model, the probability distribution of X is the rule that specifies $\mathbf{P}(X \in S)$ for subsets S of $\mathbb{R}$.

If X and Y are random variables with the same distribution, we often write $X \sim Y$ to express that fact.

Notice that since a distribution specifies probabilities, the probability distribution of a random variable is something that can be measured experimentally, or at least tested. If X is a random variable associated with a repeatable experiment, and if someone asserts that $\mathbf{P}(X \geq 5)$ is equal to .3, then we could in principle perform many repetitions of the experiment, and measure the average number of times that $X \geq 5$ occurs, to see if this frequency is close to .3. So distributions are "real".

On the other hand, a sample space is an abstract concept in our minds, which is useful but can never be directly measured. If two people are separately creating probability models for the same experiment, they may come up with very different sample spaces. But they must agree about the distributions of any physically meaningful random variables.

At present we are mainly dealing with finite-range random variables. Equation (9.5) shows that it is easy to find the distribution of a finite-range random variable once we know the probability of each point in the range. Sometimes it's convenient to use the probability mass function notation (introduced in Definition 2.12).

Definition 9.8 (The probability mass function for the distribution of a random variable). Let X be a real-valued random variable for a probability model. The probability mass function for the distribution of X is the function $\boldsymbol{q}$ on $\mathbb{R}$ defined by $\boldsymbol{q}(x) = \mathbf{P}(X = x)$.

Clearly $\boldsymbol{q}(x) = 0$ for any x which is not in the range of X, so $\boldsymbol{q}$ is determined by its values as a function on the range of X.

Let X be a finite-range random variable whose distribution has probability mass function $\boldsymbol{q}$. Let S be a subset of $\mathbb{R}$, and let $x_1, \ldots, x_k$ be any list of distinct numbers which includes all the numbers in the range of X that are members of S. We can rewrite Equation (9.5) using $\boldsymbol{q}$:

$$\mathbf{P}(X \in S) = \boldsymbol{q}(x_1) + \ldots \boldsymbol{q}(x_k). \tag{9.9}$$

We see from equation (9.9) that the probability mass function of a finite-range random variable determines its distribution. So the probability mass function itself is sometimes referred to as the distribution of the random variable.

Example 9.9 (A random variable with a binomial distribution). For the random variable S_n defined in Example 9.5, using equation (9.3) we have

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{9.10}$$

The distribution of S_n is exactly the binomial distribution defined in Definition 7.6. We say that S_n has a binomial distribution, and may also refer to S_n as a binomial random variable.

Example 9.10 (A random variable with a hypergeometric distribution). Consider a set of N objects, K of which are in a certain target class. Let a set of n objects be randomly selected from the N objects (sampling without replacement). We assume that the inequalities in equation (8.9) hold, i.e. $K \leq N$ and $n \leq N$.

Let $L_{N,K,n}$ be the number of target objects in the selected set. $\mathbf{P}(L_{N,K,n} = i)$ is the value $\mathbf{P}(R_i)$ studied in Exercise 8.4 and Definition 8.6.

Thus the distribution of $L_{N,K,n}$ is the hypergeometric distribution, with parameters N, K, n, which was defined in Definition 8.6.

By definition, the range of $L_{N,K,n}$ is the set of all i such that equation (8.13) holds.

By equation (8.17),

$$\mathbf{P}(L_{N,K,n} = i) = \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{9.11}$$

for i such that equation (8.13) holds. Otherwise $\mathbf{P}(L_{N,K,n} = i) = 0$.

Can we graph a random variable? Our main experience with functions has likely been in the setting of calculus, and in calculus we certainly can

understand a function better by plotting its graph. It can be difficult to graph a random variable, though, since the domain of the random variable might be very different from the real line. Example 9.5 illustrates this, since for this example the domain is the sample space, and that might be the set of all sequences of zeros and ones that have length n. There seems to be no convenient way to portray the set of such sequences visually, at least when n is greater than 2 or 3.

What we can do is to graph the *probabilities for the values* of the random variable, i.e. the probability mass function.

For the random variable X of Example 9.5, Figure 7.1 shows a graph of $\mathbf{P}(X = k)$ versus k for $n = 30$ and $p = 1/3$.

In Example 8.7 we considered a slightly different random variable. Here $L_{120,30,40}$ is the number of red marbles in a set of 30 marbles randomly selected from a bowl containing 120 marbles, when 40 of the marbles in the bowl are red, so $L_{120,30,40}$ has a hypergeometric distribution. The graph of $\mathbf{P}(L_{120,30,40} = i)$ versus i was given in Figure 8.2b.

9.5 Expressing the distribution of X using a density on the real line

If X is a random variable which does not have a finite range, it may not be obvious how to describe its distribution. How can we picture the distribution? In the case of a finite-range random variable, we were able to picture the probability mass function as describing lumps of "probability mass" located at the values of the random variable. For a general random variable X, one might still have a vague picture of a pile of probability mass lying on the real line, even if there are no lumps. Just as before, we would say that the amount of probability mass lying on a set S gives us the probability that the value of X lies in the set S.

This picture has a precise mathematical description if it happens that the distribution of X can be described using a probability density. We'll state that in the present section. Example 9.15 in the next section will put such densities to work, and show how they simplify computations and clarify our thinking.

Probably densities were defined using equation (3.7) of Section 3.4. Readers may wish to review that definition, as well as Remark 3.7. The general

definition says that a probability distribution has a probability density f if the probability of every event A is given by the integral of f over A. Of course in this section we are concerned with a particular kind of distribution, namely the distribution of a real-valued random variable X, and so the density is defined on the real line.

Definition 9.11 (Density of the distribution of a real-valued random variable). The distribution of a real-valued random variable X is described by a probability density h on $\mathbb{R}$ if the probability that the value of X lies in a set S is given by the integral of h over S, for subsets S of the real line.

In other words,

$$\mathbf{P}(X \in S) = \int_S h, \tag{9.12}$$

for subsets S of the real line.

(To be more precise mathematically, equation (9.12) holds for every set S that we would ever use to describe an event. See Section 9.7 for another comment on this.)

In equation (9.12), the integral of h over the set S is written as $\int_S h$. This is the modern notation for integration over a set, as in equation (3.10) of Section 3.5.

Although the general concept of integration over a set is not difficult (see Definition 3.6), we are more familiar with the special case of integrating over intervals, using calculus notation. When S is the interval $[a, b]$,

$$\int_S h = \int_a^b h = \int_a^b h(x)\, dx. \tag{9.13}$$

Remark 9.12 (Intervals are enough). In Remark 3.7 it is asserted that if an equation like (9.12) is valid when S is an interval, then we are guaranteed that it holds for all subsets S of the real line. So if you want to check that some function h is the correct density for the distribution of X, it's enough to check that

$$\mathbf{P}(X \in J) = \int_J h \tag{9.14}$$

for all intervals J of the real line. When solving exercises, we often work with intervals.

Example 9.13 (Extending a given density to the whole line). Consider the experiment described in Exercise 3.8. We are choosing a point randomly from the interval $[0, 3]$. Let X be the random variable that gives the location of the chosen point. The statement of Exercise 3.8 implies that for any subset S of $[0, 3]$,

$$\mathbf{P}(X \in S) = \int_S f, \tag{9.15}$$

where the probability density f on $[0, 3]$ is given by

$$f(t) = \frac{2}{9}(3 - t) \tag{9.16}$$

for $t \in [0, 3]$. (See Figure 3.5.)

Is f a probability density for the distribution of X, in the sense of Definition 9.11?

Well, almost. There is one extra condition in Definition 9.11. For convenience, a probability density for the distribution of a real-valued random variable should be defined everywhere on the real line. The function f is only defined on $[0, 3]$, and in Exercise 3.8, equation (9.15) is only assumed to hold for subsets S of $[0, 3]$. Sometimes we may find ourselves dealing with points which are outside the interval $[0, 3]$. We want to handle such situations smoothly, without extra steps.

For example, suppose someone asks us to find $\mathbf{P}(X \in [1, 7])$. How can we do that? Well, notice that by definition X never takes values outside $[0, 3]$. So the event that X takes a value in $[1, 7]$ is exactly the same as the event that X takes a value in $[1, 3]$. Thus

$$\mathbf{P}(X \in [1, 7]) = \mathbf{P}(X \in [1, 3]) = \int_1^3 f(t)\, dt. \tag{9.17}$$

That wasn't hard, but after looking at equation (9.17), it seems sensible to define a function h everywhere on the real line, as follows:

$$h(x) = \begin{cases} f(x) & \text{if } x \in [0, 3], \\ 0 & \text{otherwise.} \end{cases} \tag{9.18}$$

See Figure 9.1.

With definition of h, equation (9.17) is exactly the same as the statement that

$$\mathbf{P}(X \in [1,7]) = \int_1^7 h(t)\,dt. \tag{9.19}$$

Equation (9.19) suggests that h is a correct probability density function for the distribution of X.

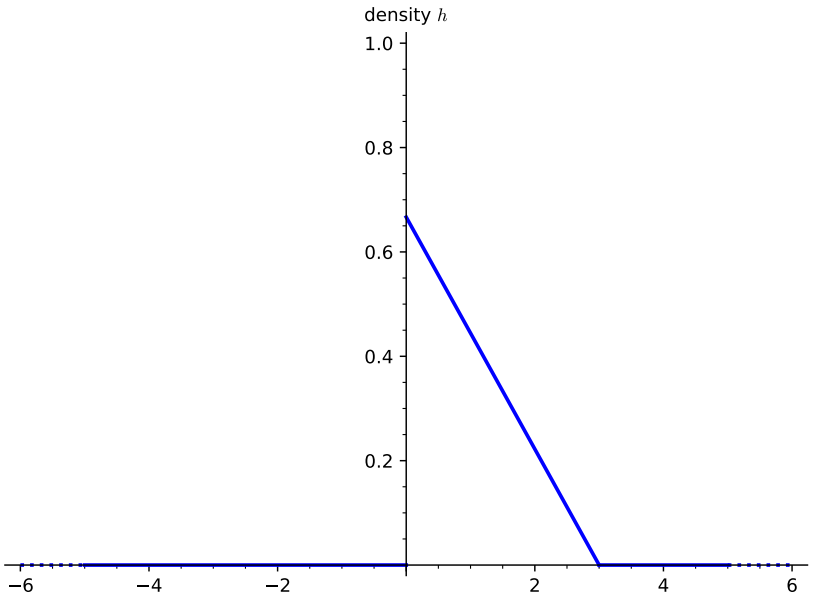


Figure 9.1: The probability density h extends the density f on $[0,3]$ that was shown in Figure 3.5.

Let's check that, by running the same argument with the set $[1,7]$ replaced by any subset S of $\mathbb{R}$. Since X never takes values outside $[0,3]$, the event $\{X \in S\}$ is exactly the same as the event $\{X \in S \cap [0,3]\}$. Hence

$$\mathbf{P}(X \in S) = \mathbf{P}(X \in S \cap [0,3]) = \int_{S\cap[0,3]} f(t)\,dt = \int_{S\cap[0,3]} h(t)\,dt. \tag{9.20}$$

Since h is equal to f on $[0,3]$ and is equal to zero everywhere else,

$$\int_{S\cap[0,3]} h(t)\,dt = \int_S h(t)\,dt.$$

Thus equation (9.20) is equivalent to the statement that

$$\mathbf{P}(X \in S) = \int_S h(t)\,dt. \tag{9.21}$$

By Definition 9.11, equation (9.21) says that h is a probability density for the distribution of X. We can use equation (9.21) conveniently for any set S we encounter, without fussing over whether S is or is not a subset of $[0, 3]$.

Remark 9.14 (Extending densities in general). The situation of Example 9.13 is not uncommon. Let D be a set which contains the range of some random variable X. Frequently we are given a probability density f on D, such that $\mathbf{P}(X \in S) = \int_S f$ for subsets S of D. If we wish to have an official probability density for the distribution of X, we can obtain that by *extending* f to a function h on the whole real line, and defining h be zero outside D.

In this case we might describe the situation in words by saying: "the distribution of X is given by a density f on D, and is zero outside D". If f is constant on D, we might also say "the distribution of X is uniform on D, and is zero everywhere else".

Exercise 9.4. Let X be a random variable whose distribution has a density h which is equal to a constant on $[3, 11]$ and is equal to zero elsewhere. Find $\mathbf{P}(1 \leq X \leq 5)$.

[Solution]

9.6 Random variables as a tool for thinking

When modeling a real-world problem, random variables occur naturally, and we naturally analyze the problem in terms of random variables.

Example 9.15. Recall Exercise 4.12. There we consider an experiment with two steps: first a fair coin is tossed. Then, if the result of the toss is a head,

in step two a point is chosen from $[0, 3]$, with no point favored. If the result of the coin toss is a tail, in step two a point is chosen from $[2, 4]$, with no point favored.

Our goal in Exercise 4.12 was to find the probability that the chosen point lies in a given subinterval J of $[0, 4]$.

Let X be the point chosen in step two. Then X is a random variable, and we can restate the goal of the problem as: find $\mathbf{P}(X \in J)$. This leads us to consider trying to obtain a *probability density* h for the distribution of X.

First, let's think about conditional probabilities. If the coin toss gives a head, the second step of the experiment consists of choosing a point from $[0, 3]$, with a probability distribution which is uniform on $[0, 3]$. Thus, conditional on obtaining a head, we know by Exercise 3.5 that the distribution of X has a density h_1 which is given by

$$h_1(x) = \begin{cases} \frac{1}{3} & \text{if } x \in [0, 3], \\ 0 & \text{otherwise.} \end{cases} \tag{9.22}$$

To say that h_1 is a density for the distribution of X conditional on H means that for all S,

$$\mathbf{P}(X \in S \,|\, H) = \int_S h_1. \tag{9.23}$$

Similarly, if the coin toss gives a tail, the second step of the experiment consists of choosing a point from $[2, 4]$. Conditional on obtaining a tail, the distribution of X has a density h_2 which is given by

$$h_2(x) = \begin{cases} \frac{1}{2} & \text{if } x \in [2, 4], \\ 0 & \text{otherwise.} \end{cases} \tag{9.24}$$

To say that h_2 is a density for the distribution of X conditional on T means that for all S,

$$\mathbf{P}(X \in S \,|\, T) = \int_S h_2. \tag{9.25}$$

How can we combine the conditional densities h_1, h_2 to find h?

Just as in the solution to the original form of Exercise 4.12, we can use the Law of Total Probability (Theorem 4.6). For any set S, applying the Law of Total Probability to the event $\{X \in S\}$ gives

$$\mathbf{P}(X \in S) = \mathbf{P}(H)\mathbf{P}(X \in S \,|\, H) + \mathbf{P}(T)\mathbf{P}(X \in S \,|\, T), \tag{9.26}$$

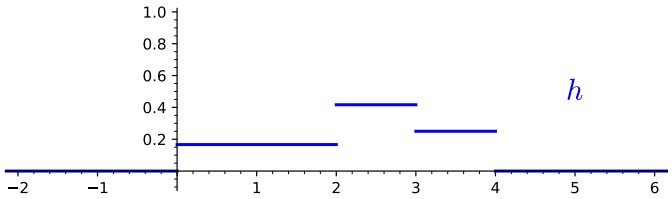


Figure 9.2: h is a density on $\mathbb{R}$ for the distribution of X, where X is chosen from overlapping intervals.

where H, T are the events that the coin toss gives a head or a tail.

Probability densities are not probabilities, but they are closely related to probabilities. Looking at equation (9.26) makes us think that we will get a valid density h if we define h by a nice neat equation:

$$h = \mathbf{P}(H)h_1 + \mathbf{P}(T)h_2. \tag{9.27}$$

And this is correct! Just integrate the right side of equation (9.27) over a set S, use equations (9.23) and (9.25), and you will obtain the right side of equation (9.26). Thus the integral of h over S gives us $\mathbf{P}(S)$.

A probability density is correct if it gives the correct probabilities when you integrate it, so the function h defined by equation (9.27) is a valid density for the distribution of X.

For this experiment, $\mathbf{P}(H) = \mathbf{P}(T) = 1/2$, and equations (9.22) and (9.24) give us h_1 and h_2. Substituting these values into Equation (9.27),

$$h(x) = \begin{cases} \frac{1}{2}\,\frac{1}{3} & \text{if } x \in [0,2), \\ \frac{1}{2}\,\frac{1}{3} + \frac{1}{2}\,\frac{1}{2} & \text{if } x \in [2,3), \\ \frac{1}{2}\,\frac{1}{2} & \text{if } x \in [3,4], \\ 0 & \text{otherwise.} \end{cases} \tag{9.28}$$

A graph of h is shown in Figure 9.2.

Please check that integrating h gives all the information obtained in Cases (i), (ii), (iii), (iv) of the solution for Exercise 4.12. Of course the density h contains much more information, and we can display the graph of h!

9.7 A technical point about sets

We mentioned in Section 9.1 that Definition 9.1 omitted a technicality. For those who are interested, here is a remark about that.

Remark 9.16 (Measurable sets). Recall that Definition 2.2 stated that in any probability model, some subsets of a sample space are designated as events. This definition did not say that *every* subset of the sample space is an event. And, in fact, a complete description of a mathematical probability model includes an extra requirement, something like this: every event must be a "measurable set".

What does the mathematical term "measurable" mean in this context? It has a special meaning here. It refers to the subsets A of the sample space for which a probability value $\mathbf{P}(A)$ can be defined. As a typical example, consider the real numbers. Roughly speaking, a measurable set of real numbers is any set which has an explicit description. The term "explicit description" is used rather generously, since it includes any mathematical construction using an infinite sequence of set operations on intervals, or an infinite sequence of infinite sequences of set operations, and so on, forever. Any set that could conceivably be used in our applications of mathematics is a measurable set.

An optimistic person might conclude from these statements that every subset of the real line is measurable, but sadly this is not the case. It can be shown mathematically that there must exist subsets of the real line which are nonmeasurable. So the best we can say is that any set which is "of interest" is measurable.

One might philosophize that having nonmeasurable subsets lurking in the background is part of the price that we pay for using a powerful abstraction like the real numbers.

When studying advanced probability, it is necessary at times to check that the mathematical theorems of probability can be applied without using nonmeasurable sets. Definition 9.1 would then be slightly enlarged, to spell out the technical requirements that a random variable must satisfy. However, in applications such requirements are always met, and we will not take time to discuss measurable and nonmeasurable sets further. Given any real-valued function X on a sample space, you may simply take it for granted that X is a valid random variable. And if S is any subset of the real line that we are interested in, then you may take it for granted that $\{X \in S\}$ is an event.

In other words, readers may safely banish the subject of measurable sets

from their minds. However, from the standpoint of mathematical rigor, we are using the following convention:

> Any statement about sets, such as "for all sets S", or "for any set S", may actually mean "for all measurable sets S", or "for any measurable set S". And the difference has no practical significance.

This convention applies, for example, to equation (9.7) and equation (9.12).

9.8 Solutions for Chapter 9

Solution (Exercise 9.1). By definition, $\tilde{X}(\omega) = 1$ if $0 \leq \omega < p$, and $\tilde{X}(\omega) = 0$ if $p \leq \omega \leq 1$.

Thus

$$\left\{\tilde{X} \in S\right\} = \left\{\omega : \ \tilde{X}(\omega) = 0\right\} = [p, 1].$$

Solution (Exercise 9.2).

- X is the number that shows on the die when it comes to rest. The range of X is $\{1, 2, 3, 4, 5, 6\}$.

  Let's take the sample space Ω to be the six-point set $\{1, 2, 3, 4, 5, 6\}$, so that $X(\omega) = \omega$ for each $\omega \in \Omega$.

  From the definition of X, $A = \{5, 6\}$ and $B = \{2, 4, 6\}$.

- By additivity,

$$\mathbf{P}(B) = \mathbf{P}(\{2\}) + \mathbf{P}(\{4\}) + \mathbf{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Omitting the sample space Let's repeat the same argument again, this time without using an explicit sample space. There is really no need to define a sample space, as long as we understand the behavior of X.

Each possible value of X has the same probability, so we can immediately say that

$$\mathbf{P}(X = 1) = \mathbf{P}(X = 2) = \mathbf{P}(X = 3) = \mathbf{P}(X = 4) = \mathbf{P}(X = 5) = \mathbf{P}(X = 6).$$

These numbers add to one, so $\mathbf{P}(X = i) = 1/6$ for $i = 1, \ldots, 6$.

If the value of X is even, then the value of X is one of the numbers $2, 4, 6$. Hence

$$\{X \text{ is even}\} = \{X = 2\} \cup \{X = 4\} \cup \{X = 6\}.$$

By the additivity of probability,

$$\mathbf{P}(X \text{ is even}) = \mathbf{P}(X = 2) + \mathbf{P}(X = 4) + \mathbf{P}(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

With practice, reasoning about events in terms of the values of a random variable will seem very natural.

Solution (Exercise 9.3).

(i) We must show that $\{X \in W\}$ is the union of the sets $\{X \in W_1\}, \ldots, \{X \in W_k\}$.

For any sets A and B, one can prove that $A = B$ by showing two facts: first, that every member of A is a member of B, and second, that every member of B is a member of A. Thus we consider two cases here.

(1.) Suppose that ω is in the union of the sets $\{X \in W_1\}, \ldots, \{X \in W_k\}$. Then for some i, $\omega \in \{X \in W_i\}$, and so $X(\omega) \in W_i$, and thus $X(\omega) \in W$, and so $\omega \in \{X \in W\}$.

(2.) Suppose that $\omega \in \{X \in W\}$. If $X(\omega) \in W$, then $X(\omega) \in W_i$ for some i, and so $\omega \in \{X \in W_i\}$, and thus ω is in the union of the sets $\{X \in W_1\}, \ldots, \{X \in W_k\}$.

Facts 1. and 2. show that $\{X \in W\}$ is the union of the sets $\{X \in W_1\}, \ldots, \{X \in W_k\}$, as claimed.

(ii) Let $W_1, \ldots, W_k$ be disjoint sets. We claim that the sets $\{X \in W_1\}, \ldots, \{X \in W_k\}$ are disjoint.

To see that, suppose that for some sample point ω, we have $\omega \in \{X \in W_i\}$ and $\omega \in \{X \in X_j\}$. Then $X(\omega) \in W_i$ and $X(\omega) \in W_j$. Since the sets $W_1, \ldots, W_k$ are assumed to be disjoint, it must be the case that $i = j$.

Thus for $i \neq j$, $\{X \in W_i\}$ and $\{X \in W_j\}$ have no points in common, i.e. they are disjoint, as claimed.

Let W be the union of the sets $W_1, \ldots, W_k$. By part (i), $\{X \in W\}$ is the union of the sets $\{X \in W_1\}, \ldots, \{X \in W_k\}$, i.e.

$$\{X \in W\} = \{X \in W_1\} \cup \ldots \cup \{X \in W_k\}.$$

Using additivity, equation (9.7) holds.

Solution (Exercise 9.4). We are told that h is constant on $[3, 11]$. Let c denote the value of h on $[3, 11]$. Since h is a probability density on $\mathbb{R}$,

$$\int_{-\infty}^{\infty} h = 1.$$

Since $h = 0$ outside $[3, 11]$,

$$\int_{-\infty}^{\infty} h = \int_{3}^{11} h = c(11 - 3) = 8c.$$

Thus $c = 1/8$.

Let $J = [1, 5]$. We're asked to find $\mathbf{P}(X \in J)$. Since h is a density for the distribution of X,

$$\mathbf{P}(X \in J) = \int_{J} h = \int_{1}^{5} h.$$

Since $h = 0$ outside $[3, 11]$,

$$\mathbf{P}(X \in J) = \int_{3}^{5} h = \int_{3}^{5} \frac{1}{8} = \frac{1}{4}.$$

Chapter 10

Expected values, finite range case

10.1 Expected value defined

We will define expected value in this chapter for the case of a random variable with finite range, and then establish the main properties of expected values.

When more general random variables are studied, the definition of expected value will have to be appropriately extended. However the properties of expected value will remain unchanged, for the most part.

Example 10.1 (Average payoff). Consider tossing an unfair coin. If the result is a head, we say you have success. Let the probability of success be 3/5.

Suppose you toss the coin 1000 times, and for each success you receive 2 dollars. For failure you get nothing. What is the average amount of money that you would expect to earn *per toss*?

The amount you actually receive on any given toss might be called the "payoff". So we are asking for the average payoff.

Notice that the payoff on any given toss is determined by the outcome of the toss. Thus it is a *function* of the physical result of the toss, and so it is a random variable in the physical sense. We want to know the average value of this random variable over a sequence of repeated tosses.

If we choose a mathematical sample space to represent the physical experiment, then the payoff is represented by a mathematical function whose

domain is the sample space, so it is a random variable in the mathematical sense. But let's think physically for a moment.

In order to calculate the average payoff, think about the frequency of successes. If you toss the coin 1000 times, it is likely that your success frequency will be approximately 3/5. Since $3/5 \times 1000 = 600$, approximately 600 of the tosses will result in success. Thus you expect to earn approximately 1200 dollars over the whole sequence of trials, so you expect to earn approximately 1.20 per toss. This is your "average payoff".

In Example 10.1, we have just given a theoretical estimate for the average payoff in repeated tosses, without actually performing any tosses. That is certainly simpler than doing repeated experiments! We call this theoretical estimate the *expected value* for the payoff random variable.

The expected value of a random variable is only a single number. But it tells us something about all the possible values of the random variable, taken together. This is a new idea.

The theoretical approach to finding an expected value is not just simpler than the experimental alternative. It may also help us to understand the experiment situation which is being studied.

Now we will give a precise mathematical definition for expected value, for the case of a random variable with finite range.

Definition 10.2 (Expected value, finite range case). Let X be a random variable. Suppose that the range of X is equal to $\{x_1, \ldots, x_k\}$, where $x_1, \ldots, x_k$ are distinct numbers.

The expected value of X, denoted by $\mathbf{E}[X]$, is defined by

$$\mathbf{E}[X] = \sum_{i=1}^{k} x_i \mathbf{P}(X = x_i). \tag{10.1}$$

In other books, $\mathbf{E}[X]$ is often written as $\mathbf{E}X$.

The expected value of X is also called the **expectation** of X or the **mean** of X. A random variable with expected value zero is often called a *mean zero* random variable.

Occasionally it is helpful to have a notation which explicitly states which probability is being used to calculate the expected value. The expected value

of X using probability set-function $\mathbf{P}$ is then denoted by

$$\mathbf{E}_{\mathbf{P}}[X]. \tag{10.2}$$

If a different probability set-function $\mathbf{Q}$ is used, then the corresponding expected value is denoted by $\mathbf{E}_{\mathbf{Q}}[X]$, and so on.

The expected value of X is defined by the sum in equation (10.1). This sum is said to be a *weighted average* (Definition A.2). It is a weighted average of the possible values of X, in which the weight of each value x_i is the probability $\mathbf{P}(X = x_i)$ with which it occurs. Readers who have not used weighted averages may find it worthwhile to work through some exercises in Appendix A.

Example A.3 shows we can picture the expected value of X as the *center of mass* of the distribution, when the distribution is represented as lumps of probability mass located at the values of X.

Since expected value of X is defined in terms of the values of X and their probabilities, it is determined by the distribution of X . The distribution of a random variable is a real and testable physical property, so the expected value is a real and testable physical property. Expectation can be calculated using any valid sample space representation that you like, but the value must be the same for any valid sample space.

Exercise 10.1 (One-toss payoff). In Example 10.1, consider just tossing the coin once.

If the sample space for one toss is $\Omega = \{0, 1\}$, the payoff function Y for one toss is simply defined by $Y(1) = 2$, $Y(0) = 0$.

Use Definition 10.2 to find $\mathbf{E}[Y]$.

[Solution]

Exercise 10.2 (Distinct values). In Definition 10.2, why are the numbers $x_1, \ldots, x_n$ required to be distinct?

[Solution]

Exercise 10.3 (Order of values is irrelevant). In Definition 10.2, $x_1, \ldots, x_k$ is a list of the distinct values in the range of the random variable. Explain why the order in which we list the values does not matter.

[Solution]

Is it important to check our general formulas, to make sure they are right? Well, somebody should certainly check. Expectation is such an important concept that it seems worthwhile to check that every property we need is a consequence of the definitions. We won't always take time to do this, but in the proof of the next lemma we will.

Lemma 10.3 (Single event expectation). Let A be an event and let c be a real number. Let X be the random variable defined by $X(\omega) = c$ if $\omega \in A$, $X(\omega) = 0$ otherwise. Then

$$\mathbf{E}[X] = c\,\mathbf{P}(A). \tag{10.3}$$

Proof. **Case 1** If A is the empty set, then $X(\omega) = 0$ for all ω, so the range of X is $\{0\}$.

Then by definition $\mathbf{E}[X] = 0 \cdot \mathbf{P}(X = 0) = 0 = c\,\mathbf{P}(A)$, so equation (10.3) holds.

Case 2 If $c = 0$ then again the range of X is $\{0\}$, and equation (10.3) holds as in Case 1.

Case 3 If $A = \Omega$, then the range of X is $\{c\}$, and by definition $\mathbf{E}[X] = c \cdot \mathbf{P}(X = c) = c \cdot \mathbf{P}(\Omega)$, so equation (10.3) holds.

Case 4 The only remaining case is that $A \neq \emptyset$ and $A \neq \Omega$, with $c \neq 0$.

Then the range of X consists of the distinct points $0, c$.

Then by definition $\mathbf{E}[X] = c \cdot \mathbf{P}(A) + 0 \cdot \mathbf{P}(A^c) = c \cdot \mathbf{P}(A)$. Thus equation (10.3) holds.

□

An important consequence of Lemma 10.3: for any constant c,

$$\mathbf{E}[c] = c, \tag{10.4}$$

where $\mathbf{E}[c]$ denotes the expected value of the random variable which is equal to c for all outcomes.

Hmm, we looked at a lot of cases when proving Lemma 10.3. The next exercise would have saved some work!

Exercise 10.4 (Unused values in the definition are ok). Let X be a random variable. Let $y_1, \ldots, y_n$ be distinct numbers, such that every nonzero number in the range of X is included in the list $y_1, \ldots, y_n$. Prove that

$$\mathbf{E}[X] = \sum_{i=1}^{n} y_i \mathbf{P}(X = y_i). \tag{10.5}$$

[Solution]

Exercise 10.5. Use Exercise 10.4 to give a shorter proof of Lemma 10.3.
[Solution]

Example 10.4 (One coin toss). Let X be as in Example 9.2. Notice that X is the number of successes obtained in the coin toss (either 0 or 1).

Let A be the event that the toss gives success. Then $X(\omega) = 1$ when $\omega \in A$, and $X(\omega) = 0$ otherwise.

Applying Exercise 10.4, $\mathbf{E}[X] = 1 \cdot \mathbf{P}(X = 1) + 0 \cdot \mathbf{P}(X = 0) = p$.

Example 10.5 (One roll of a die). We deal with the result of rolling a die (Example 2.15) similarly to Example 10.4. Let X be the number obtained by rolling the die, so that the range of X is $\{1, 2, 3, 4, 5, 6\}$.

By definition, $\mathbf{E}[X] = 1 \cdot \mathbf{P}(X = 1) + 2 \cdot \mathbf{P}(X = 2) + 3 \cdot \mathbf{P}(X = 3) + 4 \cdot \mathbf{P}(X = 4) + 5 \cdot \mathbf{P}(X = 5) + 6 \cdot \mathbf{P}(X = 6)$, i.e.

$$\mathbf{E}[X] = \sum_{i=1}^{6} i \, \mathbf{P}(X = i).$$

Lemma 10.6 (Expectation of a scaled random variable). Let X be a random variable and let c be a number. Then

$$\mathbf{E}\left[c\,X\right] = c\,\mathbf{E}\left[X\right]. \tag{10.6}$$

Proof. The statement is true in general, but we only consider the case of a finite-range random variable here.

If $c = 0$, then $c\,X$ is the zero random variable. Using Lemma 10.3 (or the definition of expected value), we know that $\mathbf{E}\left[c\,X\right] = 0$, so we certainly have $\mathbf{E}\left[c\,X\right] = c\,\mathbf{E}\left[X\right]$.

From now on suppose that $c \neq 0$.

Let $x_1, \ldots, x_k$ be the distinct numbers in the range of X. Since $c \neq 0$, we have $X = x_i$ if and only if $c\,X = c\,x_i$. Hence the range of $c\,X$ is the set $\{c\,x_1, \ldots, c\,x_k\}$, and the numbers $c\,x_1, \ldots, c\,x_k$ are distinct. By the definition of expected value,

$$\mathbf{E}\left[c\,X\right] = c\,x_1\mathbf{P}(c\,X = c\,x_1) + \ldots + c\,x_k\mathbf{P}(c\,X = c\,x_k).$$

But $\mathbf{P}(c\,X = c\,x_i) = \mathbf{P}(X = x_i)$ (it's the same event), so

$$\mathbf{E}\left[c\,X\right] = c\,x_1\mathbf{P}(X = x_1) + \ldots + c\,x_k\mathbf{P}(X = x_k) = c\,\mathbf{E}\left[X\right].$$

□

The property of expectation stated in Lemma 10.6 is very simple, but it's important. Just to have a name for this property, we'll call it the **scaling property**. This is not a standard mathematical term, but it fits, since Lemma 10.6 says that if we scale (up or down) all the values of X by a factor c, then we scale $\mathbf{E}\left[X\right]$ by the same factor.

The next exercise is an example for an upcoming theorem, Theorem 10.8. However, it is instructive to solve it directly here.

Exercise 10.6 (Finding expected value using cases). Consider the number wheel described in Exercise 2.13.

Imagine a game in which the wheel is spun. Let Z be the number at which the wheel stops. Then Z is a random variable, and the possible values

for Z are $0, 1, 2, \ldots, 100$. We assume that each of these numbers occurs with the same probability.

In this game, a payoff is given, based on the number where the wheel stops, i.e. based on the value of Z.

Let the payoff be called X.

The rules are as follows:

- If $Z = 0$, then $X = 0$.
- If $Z = 100$, $X = 5$.
- If Z is even and less than 100, then $X = 2$.
- If Z is odd, then $X = 1$.

Thus $X = \varphi(Z)$, where φ is defined in the obvious way:

$$\begin{aligned} &\varphi(0) = 0, \\ &\varphi(100) = 5, \\ &\varphi(i) = 2 \text{ if } i \text{ is even and less than } 100, \\ &\varphi(i) = 1 \text{ if } i \text{ is odd.} \end{aligned} \tag{10.7}$$

(i) Find $\mathbf{E}[X]$, using the definition of expected value.

(ii) Show that

$$\mathbf{E}[X] = \sum_{i=0}^{100} \varphi(i)\mathbf{P}(Z = i). \tag{10.8}$$

Does the statement of equation (10.8) feel right to you? Since $\varphi(i)$ is the value of X when $Z = i$, this equation says that $\mathbf{E}[X]$ is equal to a weighted sum of values of X, but it's not the same sum which is used in the definition of $\mathbf{E}[X]$. Instead it's a sum over cases, where each case is given by the value of Z. The weight of each case is the probability of the case.

[Solution]

10.2 Expected value by cases

Please think about the statement of the next theorem until it seems reasonable. The proof is optional, but it is not hard to understand the basic step: grouping similar terms in a sum over cases, where each term is a value times a probability.

Theorem 10.7 (Expectation by cases). Let $D_1, \ldots, D_k$ be disjoint events in some model, such that $D_1 \cup \ldots \cup D_k = \Omega$.

Let $v_1, \ldots, v_k$ be numbers, and let X be a random variable such that $X(\omega) = v_i$ at every point ω in D_i. (Thus X has value v_i whenever event D_i happens!)

Then

$$\mathbf{E}[X] = \sum_{i=1}^{k} v_i \mathbf{P}(D_i). \tag{10.9}$$

See Figure 10.1.

Proof. Let $x_1, \ldots, x_m$ be a list of the distinct numbers in the range of X.

If an event D_i is nonempty, then it contains at least one point ω. By assumption, $X(\omega) = v_i$. Thus for every nonempty D_i, v_i is a point in the range of X.

For each i, if D_i happens to be empty, change v_i to one of the values $x_1, \ldots, x_m$. Such changes clearly make no difference to the sum in equation (10.9), so they don't affect the truth of the theorem.

Now we can say that the numbers $v_1, \ldots, v_k$ are a list of the values $x_1, \ldots, x_m$, possibly with repetitions, as Figure 10.1 illustrates.

Since $D_1 \cup \ldots \cup D_k = \Omega$, for every x_j there must be at least one event D_i such that $v_i = x_j$.

We can choose the labels for the numbers $x_1, \ldots, x_m$ so that $x_1 < x_2 < \cdots < x_m$.

The order in which we write the events D_i makes no difference in equation (10.9). For convenience, relabel the events and associated values so that $v_1 \leq v_2 \leq \ldots \leq v_{k-1} \leq v_k$.

For every j, let i_j be the largest index i such that $v_i = x_j$. Because of the ordering of the values, i_m must be k.

The following picture illustrates the general situation.

The top row is simply $D_1, \ldots, D_k$, in order, but grouped.

$$\begin{array}{cccccccccc} D_1 & \cdots & D_{i_1} & D_{i_1+1} & \cdots & D_{i_2} & \cdots\cdots & D_{i_{m-1}+1} & \cdots & D_{i_m} \\ \underbrace{v_1 = \cdots = v_{i_1}}_{x_1} & < & & \underbrace{v_{i_1+1} = \cdots = v_{i_2}}_{x_2} & < & & \cdots\cdots < & \underbrace{v_{i_{m-1}+1} = \cdots = v_{i_m}}_{x_m} & . & \end{array} \quad (10.10)$$

The events in Figure 10.1 are grouped in a similar way.

We see that

$$\{X = x_1\} = D_1 \cup \cdots \cup D_{i_1},$$

and for $j = 2, \ldots, m$,

$$\{X = x_j\} = D_{i_{j-1}+1} \cup \cdots \cup D_{i_j}, \quad (10.11)$$

Define $i_0 = 0$. Then equation (10.11) holds for all $j = 1, \ldots, m$, which saves writing.

Since the events D_i are disjoint, for every j we have

$$\mathbf{P}(X = x_j) = \mathbf{P}(D_{i_{j-1}+1}) + \ldots + \mathbf{P}(D_{i_j})$$

By definition,

$$\begin{aligned} \mathbf{E}[X] &= \sum_{j=1}^{m} x_j \mathbf{P}(X = x_j) = \sum_{j=1}^{m} x_j \left(\mathbf{P}(D_{i_{j-1}+1}) + \ldots + \mathbf{P}(D_{i_j})\right) \\ &= \sum_{j=1}^{m} \left(x_j \mathbf{P}(D_{i_{j-1}+1}) + \ldots + x_j \mathbf{P}(D_{i_j})\right) \\ &= \sum_{j=1}^{m} \left(v_{i_{j-1}+1} \mathbf{P}(D_{i_{j-1}+1}) + \ldots + v_{i_j} \mathbf{P}(D_{i_j})\right) \\ &= \sum_{i=1}^{k} v_i \mathbf{P}(D_i). \end{aligned}$$

□

In the proof of Theorem 10.7, did we really need to relabel the events D_i and values v_i so that $v_1 \leq v_2 \leq \ldots \leq v_{k-1} \leq v_k$?

We didn't do that in Figure 10.1, did we? And no, we don't really need the relabelling step. But if we didn't do that, we wouldn't be able to display the general grouping picture given in equation (10.10).

You can still *talk* about the grouping, though, by defining sets of indices: you would define N_j be the set of indices i such that $v_i = x_j$. Then, instead of saying:

$$\mathbf{P}(X = x_j) = \mathbf{P}(D_{i_{j-1}+1}) + \ldots + \mathbf{P}(D_{i_j}),$$

you would say:

$$\mathbf{P}(X = x_j) = \sum_{i \in N_j} \mathbf{P}(D_i),$$

and run the same argument.

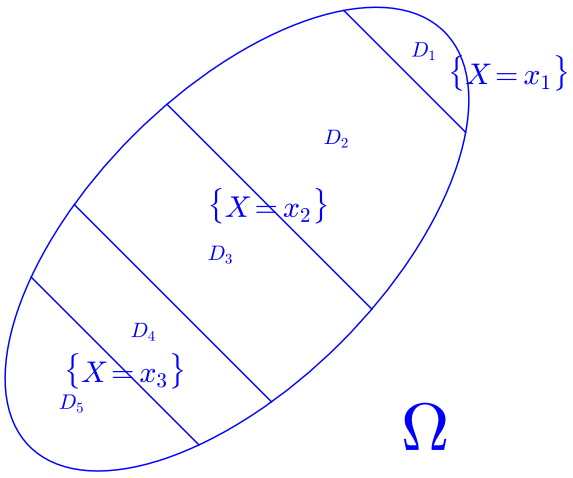


Figure 10.1: For Theorem 10.7. Here $v_1 = x_1$, $v_2 = v_3 = x_2$, and $v_4 = v_5 = x_3$, where x_1, x_2, x_3 are distinct. $\{X = x_1\} = D_1$, $\{X = x_2\} = D_2 \cup D_3$, $\{X = x_3\} = D_4 \cup D_5$.

Exercise 10.6 is an example for the next theorem.

Theorem 10.8 (Expectation of a function of a finite-range random variable). Let Y be a random variable on a sample space Ω. Let the distinct values in the range of Y be $y_1, \ldots, y_k$. In this theorem there is no need to assume that $y_1, \ldots, y_k$ are numbers. They can be anything.

Let φ be any real-valued function whose domain includes $y_1, \ldots, y_k$. Then

$$\mathbf{E}\left[\varphi(Y)\right] = \sum_{i=1}^{k} \varphi\left(y_i\right) \mathbf{P}\left(Y = y_i\right). \tag{10.12}$$

Proof. Let $D_i = \{Y = y_i\}$, let $v_i = \varphi\left(y_i\right)$, and apply Theorem 10.7. □

Exercise 10.7. Suppose that the distribution of X is uniform on the points $\{-2, 1, 0, 1, 2\}$. Find $\mathbf{E}\left[X^2\right]$ in two ways: from the definition and using Theorem 10.8.

[Solution]

Example 10.9 (Expectations on finite sample spaces). Let X be a random variable defined on a finite sample space Ω. Let the distinct sample points be $\omega_1, \ldots, \omega_n$.

In Theorem 10.7, let $D_i = \{\omega_i\}$. Equation (10.9) gives us a pleasantly simple formula for expected value:

$$\mathbf{E}\left[X\right] = \sum_{i=1}^{n} X(\omega_i)\mathbf{P}(\{\omega_i\}). \tag{10.13}$$

Of course the values $X(\omega_i)$ might not be distinct. But as usual that's ok, we only give each outcome its own probability weight, so there is no "double-counting".

10.3 The frequency interpretation for expectation

Here is a general statement of the key fact linking expected value to the real world. We have already seen this in Example 10.1.

Probability Fact 10.1 (The frequency interpretation of expected value). Let X be a random variable defined in terms of an experiment. If the experiment is repeated many times, the theoretical expected value of the corresponding mathematical random variable is likely to be approximately equal to *average experimental value* of the random variable X obtained from the repeated experiments.

Justifying the frequency interpretation for expected values We cannot give a rigorous proof of a practical statement, but we will show that for a random variable with finite range this rule is a direct consequence of the frequency interpretation of probability. A similar argument was already used in Example 10.1.

Let X be a random variable with finite range. Suppose that the range of X consists of the distinct numbers $x_1, \ldots, x_k$.

By definition,

$$\mathbf{E}[X] = x_1\mathbf{P}(X = x_1) + \ldots + x_n\mathbf{P}(X = x_n).$$

Suppose that X represents a physical random variable in some experiment. Consider a sequence of N repetitions of the experiment.

Let M_i be the number of those experiments for which the value of X is equal to x_i. Then the average experimental value $\bar{x}$ for X is given by

$$\bar{x} = \frac{1}{N}(\text{sum of all measured values}) = \frac{1}{N}\sum_{i=1}^{k} x_i M_i = \sum_{i=1}^{k} x_i \frac{M_i}{N}.$$

The frequency interpretation for *probability* says that for large N, it is very likely that $M_i/N \approx \mathbf{P}(X = x_i)$. Applying this approximation to every term in the sum for $\bar{x}$,

$$\bar{x} \approx \sum_{i}^{k} x_i \mathbf{P}(X = x_i) = \mathbf{E}[X]. \tag{10.14}$$

Equation (10.14) expresses Rule 10.1, so we have justified this rule.

A traditional name for Rule 10.1 is "the Law of Large Numbers". Some corresponding mathematical properties of expected value are given in two

well-known theorems called "the Weak Law of Large Numbers" and "the Strong Law of Large Numbers". These are mathematical statements. The law of large numbers expressed here in Rule 10.1 is a practical statement, not a mathematical theorem.

10.4 Additivity of expectation

The frequency interpretation of expected value provides a strong connection between theoretical calculations and experimental results. We can use the physical interpretation of expected value to tell us what the mathematical properties must be.

For example, consider two physical random variables X and Y which are defined for the same experiment. The measured value of $X + Y$ is, by definition, the sum of the value of X and value of Y. By the frequency interpretation, the average of the measured values of $X+Y$, over sufficiently many repeated experiments, is approximately $\mathbf{E}[X + Y]$. And the average value for $X+Y$ is equal to the sum of the average value for X and the average value for Y. This frequency argument leaves us in no doubt that *additivity* must hold for mathematical expected values:

$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]. \tag{10.15}$$

Equation (10.15) is confirmed with the formal proof in Lemma 10.10, given below. Additivity actually holds for expectations of all random variables (see the statement of Theorem 14.9).

Lemma 10.10 (Additivity of expectation). Let X and Y be finite-range random variables defined for the same probability model. Then

$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]. \tag{10.16}$$

Proof. Let $x_1, \ldots, x_n$ be the distinct numbers in the range of X, and let $y_1, \ldots, y_m$ be the distinct numbers in the range of Y.

Let $D_{ij} = \{X = x_i, Y = y_j\}$.

By Theorem (10.7),

$$\mathbf{E}[X] = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i \mathbf{P}(D_{ij}) \tag{10.17}$$

$$\mathbf{E}[Y] = \sum_{i=1}^{n} \sum_{j=1}^{m} y_j \mathbf{P}(D_{ij}) \tag{10.18}$$

and

$$\mathbf{E}[X+Y] = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i + y_j) \mathbf{P}(D_{ij}). \tag{10.19}$$

The right side of equation (10.19) is the sum of the right sides of equations (10.17) and (10.18). Hence $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.

□

Exercise 10.8. If you love the frequency interpretation, write out a careful derivation of the additivity property for physical random variables, using the frequency interpretation.

[Solution]

Remark 10.11 (Using multiple indices). In the proof of Lemma 10.10, does it seem strange to apply Theorem (10.7) to a situation in which the disjoint sets are described by multiple indices i, j?

It is important to see that this is ok. Notice that the properties of the sets $D_1, \ldots, D_k$ in Theorem (10.7), such as disjointness and having union equal to the whole space, do not depend on how the sets $D_1, \ldots, D_k$ are labelled.

Also, the sum in equation (10.9) would not be changed if we listed the sets $D_1, \ldots, D_k$ in a different way, provided that we included all of the sets and did not list any of them more than once.

So the way we label our indices doesn't matter.

Here's some useful terminology.

Definition 10.12 (Linear operations). Consider any set of mathematical elements such that "addition" and "multiplication by a number" make sense. Examples: the set of coordinate vectors in $\mathbb{R}^n$, the set of functions on an interval, the set of random variables on a sample space.)

An operation on such elements is said to be a *linear operation* if it preserves addition and also preserves multiplication by a number. More precisely, a linear operation is an operation with the following properties:

(i) **The Additivity Property**. The result of applying the operation to a sum of elements is equal to the sum of the results of applying the operation to each term separately.

(ii) **The Scaling Property**. The result of multiplying an element by a constant and then applying the operation is the same as the result of applying the operation first, and multiplying by the number afterwards.

Linearity is a handy term, and we often use it. The rules of calculus tell us that integration of functions is an example of a linear operation. By Lemma 10.10 and 10.6, the operation of taking expected value is a linear operation. The next lemma records this fact for future reference.

Lemma 10.13 (Expectation is linear). Taking expected value is a linear operation, i.e.

(i)
$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$$

and

(ii)
$$\mathbf{E}[cX] = c\mathbf{E}[X].$$

10.5 Using linearity to find expectations

We will use linearity so much that it will seem instinctive. Here are a few examples in which linearity plays a role in finding expectations.

10.5.1 Expected number of successes for Bernoulli trials

As in Example 9.5, let S_n be the total number of successes in n Bernoulli trials with success probability p. Then S_n has a binomial distribution. We wish to find $\mathbf{E}[S_n]$.

Before we perform this calculation, let's try to make it seem a little impressive. Remember that n could be huge, a million or a trillion. If we really care about the result, our method had better be right. Definitions and proofs are what gives us the confidence to produce numbers in situations where even computers would be too slow.

Method 1: using additivity Let $X_i = 1$ on A_i and $X_i = 0$ on A_i^c. (Thus X_i is the indicator function for the event A_i, as defined in Definition 11.1.) We have already observed that X_i is the number of successes obtained in trial i (either 0 or 1). From the definition of S_n, it follows that $S_n = X_1 + \ldots + X_n$. By additivity, $\mathbf{E}[S_n] = \mathbf{E}[X_1] + \ldots + \mathbf{E}[X_n]$. From the definition of expectation, $\mathbf{E}[X_i] = 1 \cdot \mathbf{P}(X_i = 1) + 0 \cdot \mathbf{P}(X_i = 0) = p$. Hence

$$\mathbf{E}[S_n] = np. \tag{10.20}$$

Notice that $\mathbf{E}[S_n]$ is exactly what we would immediately compute from the frequency interpretation, which is the basis of the common sense reasoning used in Example 10.1.

Exercise 10.9 (Method 2 for expected number of heads). Method 2 is what you use when you don't remember that expectation is additive. It is perhaps unnecessary to add that this is *not* the right approach. Nevertheless, we can learn from it.

By equation (9.3), $\mathbf{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Calculate $\mathbf{E}[S_n]$ again, this time using Definition 10.2 and this formula.

As in Example 10.1, we can *guess* ahead of time that $\mathbf{E}[S_n] = np$. So if we see a factor of np in the algebra we should hang onto it in the calculation.

Finishing the calculation will verify our guess.

[Solution]

10.5.2 Expected value of a hypergeometric random variable

Consider the experiment of Exercise 8.4. In that experiment, we have a bowl which contains N marbles. A subset of n marbles is selected, with no marble favored. There are K red marbles in the bowl, the others being green. Let

the random variable $L_{N,K,n}$ be the number of red marbles selected. We wish to find $\mathbf{E}[L_{N,K,n}]$.

$L_{N,K,n}$ has hypergeometric distribution with parameters N, K, n (Definition 9.10), and our calculation for the expected value applies to any such random variable.

Since linearity worked so well for coin-tossing, it's the natural method to try here. And it works. But we need to set up the problem, and use all the information we have. The tricks we use here are worth noting!

Method 1 for $\mathbf{E}[L_{N,K,n}]$: using additivity Assign each marble an identification number, so that the marbles are numbered from 1 to N. For convenience, let marbles $1, \ldots, K$ be the red ones.

Let $X_\ell = 1$ if marble ℓ is selected, $X_\ell = 0$ otherwise. Since no marble is favored, $\mathbf{E}[X_\ell]$ is the same for every ℓ.

Incidentally, the experiment was defined as choosing a subset of n marbles. But it's ok to focus on what happens to a particular marble. From the definitions,

$$L_{N,K,n} = X_1 + \ldots + X_K. \tag{10.21}$$

By linearity,

$$\mathbf{E}[L_{N,K,n}] = \mathbf{E}[X_1] + \ldots + \mathbf{E}[X_K] = K\mathbf{E}[X_1]. \tag{10.22}$$

To find $\mathbf{E}[X_1]$ with minimal work, note that from the description of the experiment there are always exactly n marbles selected. Hence

$$X_1 + \ldots + X_N = n, \tag{10.23}$$

always.

Mathematical expectation has been proven to be linear. So now you can go ahead and take the expected value of equation (10.23). This gives

$$n = \mathbf{E}[X_1] + \ldots + \mathbf{E}[X_N] = N\mathbf{E}[X_1].$$

This gives $\mathbf{E}[X_1] = n/N$, and hence by equation (10.22) we have

$$\mathbf{E}[L_{N,K,n}] = \frac{n\,K}{N}, \tag{10.24}$$

so we are done.

Remark 10.14 (Special cases). Fact 1 Incidentally, since we showed that $\mathbf{E}[X_1] = n/N$, and since X_i is always either 1 or 0, equation (10.1) tells us that $\mathbf{E}[X_i] = \mathbf{P}(X_i = 1)$. Thus our work has shown that the probability of any particular marble being selected is exactly n/N.

Fact 2 Also, by equation (10.24) we know that $\mathbf{E}[L_{N,K,1}] = \frac{n\,K}{N} = K/N$. Since $L_{N,K,1}$ is always either 1 or 0, we know that the probability that a single selected marble lies in the target set is K/N. This probability agrees with the statement of Theorem 2.22.

Do you think that Fact 1 and Fact 2 are really the same statement? In Fact 1, we have a particular element, and randomly choose n points. In Fact 2, we have a particular set of K element, and randomly choose a one point. But since the particular point could be any point, and the particular set could be any set, it seems that in both cases we might as well say that we have a random set and a random point, chosen independently, and we are finding the probability that the random point is an element in the random set.

Method 2 for $\mathbf{E}[L_{N,K,n}]$: direct calculation By "direct calculation" we mean something like the method of Exercise 10.9. This is feasible, and a calculation is given next. You should definitely skip it as long as you agree that this is not the easy approach!

As in Definition 9.10, the range of $L_{N,K,n}$ is the set of all i such that equation (8.13) holds.

By equation (9.11), for each i in the range of $L_{N,K,n}$ we have

$$\mathbf{P}(L_{N,K,n} = i) = \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}.$$

Hence by definition:

$$\mathbf{E}[L_{N,K,n}] = \sum_i^* i\mathbf{P}(L_{N,K,n} = i) = \sum_i^* i\,\frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} = \sum_{i>0}^* \frac{i\binom{K}{i}\binom{N-K}{N-i}}{\binom{N}{n}},$$

where we write $\sum_i^*$ to mean a sum over the indices i in the range of $L_{N,K,n}$,

and we write $\sum_{i>0}^{*}$ to mean a sum over the nonzero indices i in the range of $L_{N,K,n}$.

Using equation (8.8),

$$\mathbf{E}\left[L_{N,K,n}\right] = \sum_{i>0}^{*} \frac{i\frac{K}{i}\binom{K-1}{i-1}\binom{N-K}{n-i}}{\frac{N}{n}\binom{N-1}{n-1}} = \frac{Kn}{N}\sum_{i>0}^{*} \frac{\binom{K-1}{i-1}\binom{N-K}{n-i}}{\binom{N-1}{n-1}}. \tag{10.25}$$

Sincd $N-K = (N-1)-(K-1)$ and $n-i = (n-1)-(i-1)$, this gives

$$\mathbf{E}\left[L_{N,K,n}\right] = \frac{Kn}{N}\sum_{i>0}^{*} \frac{\binom{K-1}{i-1}\binom{(N-1)-(K-1)}{(n-1)-(i-1)}}{\binom{N-1}{n-1}}. \tag{10.26}$$

By equation (8.13), the nonzero indices i in the range of $L_{N,K,n}$ are those i such that

$$1 \le i,\ K-i \le N-n,\ i \le K,\ \text{and } i \le n. \tag{10.27}$$

Let $L_{N-1,K-1,n-1}$ denote a hypergeometric random variable with parameters $N-1, K-1, n-1$.

Let $\ell = i-1$. Then equation (10.27) says that

$$0 \le \ell,\ (K-1)-\ell \le (N-1)-(n-1),\ \text{and } \ell \le n-1.$$

This is exactly the statement that ℓ is in the range of $L_{N-1,K-1,n-1}$. Hence

$$\sum_{i>0}^{*} \frac{\binom{K-1}{i-1}\binom{(N-1)-(K-1)}{(n-1)-(i-1)}}{\binom{N-1}{n-1}} = \sum_{\ell}^{**} \frac{\binom{K-1}{\ell}\binom{(N-1)-(K-1)}{(n-1)-\ell}}{\binom{N-1}{n-1}}, \tag{10.28}$$

where we write $\sum_{\ell}^{**}$ to mean a sum over the indices ℓ in the range of $L_{N-1,K-1,n-1}$,

Using equation (9.11) with N, K, n replaced by $N-1, K-1, n-1$, equation (10.28) says that

$$\sum_{i>0}^{*} \frac{\binom{K-1}{i-1}\binom{(N-1)-(K-1)}{(n-1)-(i-1)}}{\binom{N-1}{n-1}} = \sum_{\ell}^{**} \mathbf{P}(L_{N-1,K-1,n-1} = \ell).$$

This is the sum of the probabilities of the values of $L_{N-1,K-1,n-1}$ over all possible values, so the sum is equal to one! By equation (10.26), $\mathbf{E}\left[L_{N,K,n}\right] = Kn/N$.

10.5.3 Reflection symmetry

If x is a point on the real line we will say that the point $-x$ is the mirror image of x under reflection in the origin.

Consider a physical random variable X such that for any possible value of X, the negative of that value is just as likely to occur. Over many experiments, the positive and negative values of this random variable will tend to cancel out. In the long run, the average should be close to zero. By the Frequency Interpretation of Expected Value, it must be true that $\mathbf{E}[X] = 0$.

The next exercise asks you to give a more precise argument to show that $\mathbf{E}[X] = 0$.

Exercise 10.10. Let X be a random variable with finite range. Let $a_1, \ldots, a_k$ be distinct positive numbers, and suppose that the nonzero range of X is the set of numbers $a_1, -a_1, a_2, -a_2, \ldots, a_k - a_k$. In addition, suppose that for each $i = 1, \ldots, k$,

$$\mathbf{P}(X = -a_i) = \mathbf{P}(X = a_i). \tag{10.29}$$

Use Definition 10.2 to show that $\mathbf{E}[X] = 0$. As usual, Exercise 10.4 is convenient in applying the definition of expected value.

[Solution]

Exercise 10.10 uses mathematical reasoning which is close to the physical picture. But general mathematical arguments can be more powerful, as in the next lemma.

Lemma 10.15 (Reflection symmetry gives mean zero). Let X be a random variable such that X and $-X$ have the same distribution.

If $\mathbf{E}[X]$ exists, then $\mathbf{E}[X] = 0$.

Proof. The expected value of any random variable is determined by its distribution, and for this particular random variable X it is assumed that X and $-X$ have the same distribution. Therefore $\mathbf{E}[X] = \mathbf{E}[-X]$.

Using linearity of expectation, $\mathbf{E}[-X] = -\mathbf{E}[X]$, and so $2\mathbf{E}[X] = 0$.

□

Does Lemma 10.15 give Exercise 10.15 as a special case? Sure! Saying that $\mathbf{P}(X = -a_i) = \mathbf{P}(X = a_i)$ is the same as saying that $\mathbf{P}(-X = a_i) = \mathbf{P}(X = a_i)$, and so the assumption of Exercise 10.15 is equivalent to the assumption that X and $-X$ have the same distribution.

Exercise 10.11. Let X be a random variable whose range is exactly the set S of integers i with $-1000000 \leq i \leq 1000000$. Assume that the distribution of X is uniform on S. Find $\mathbf{E}[X]$ in two ways:

(i) By calculation using the definition of expected value, and

(ii) using Lemma 10.15.

[Solution]

Exercise 10.12. Let X be a random variable whose range is exactly the set S of integers i with $0 \leq x \leq 1000$. Assume that the distribution of X is uniform on S.

(i) Find the distribution of $X - 500$.

(ii) Find $\mathbf{E}[X]$.

[Solution]

10.6 Monotonicity of expectations

Exact values are often not available, so we need to be able to deal with estimates and inequalities.

Lemma 10.16 (Monotonicity of expectations). Let X and Y be random variables with finite range, such that $X(\omega) \leq Y(\omega)$ for all sample points ω. Then

$$\mathbf{E}[X] \leq \mathbf{E}[Y]. \tag{10.30}$$

Proof. By assumption, $Y - X$ is a nonnegative random variable, i.e. all values are nonnegative.

The definition of expected value in the finite range case shows that $\mathbf{E}[Y - X] \geq 0$.

Since $Y = X + (Y - X)$, additivity tells us that

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

Since $\mathbf{E}[Y - X] \geq 0$, we are done.

□

Exercise 10.13. Give a derivation of the monotonicity property for physical random variables, using the frequency interpretation.

[Solution]

Example 10.17 ($\mathbf{E}[X]$ and $\mathbf{E}[|X|]$). Let X be a finite-range random variable, and let $x_1, \ldots, k$ be the distrinct numbers in the range of X. Using the definition of expected value and the triangle inequality (Appendix B),

$$|\,\mathbf{E}[X]\,| = \left| \sum_{i=1}^{k} x_i \mathbf{P}(X = x_i) \right| \leq \sum_{i=1}^{\infty} |\,x_i \mathbf{P}(X = x_i)\,| = \sum_{i=1}^{\infty} |x_i|\,\mathbf{P}(X = x_i).$$

The numbers $|x_1|, \ldots, |x_k|$ may not be distinct, if X happens to have both positive and negative values. But using Theorem 10.12 we see that

$$\sum_{i=1}^{\infty} |x_i|\,\mathbf{P}(X = x_i) = \mathbf{E}[\,|X|\,].$$

So we have proved an interesting inequality:

$$|\,\mathbf{E}[X]\,| \leq \mathbf{E}[\,|X|\,]. \tag{10.31}$$

A Better Proof for equation (10.31)

Of course, our proof used the finite-range property for X. But the inequality is true for general expected values. Furthermore, there is actually a slick way to derive it, just using general properties: linearity and monotonicity:

Note that $X \leq |X|$. Hence, by monotonicity, $\mathbf{E}[X] \leq \mathbf{E}[\,|X|\,]$.

But we also have $-X \leq |-X| = |X|$. So, by monotonicity, $\mathbf{E}[-X] \leq \mathbf{E}[\,|X|\,]$. Then, by linearity, $-\mathbf{E}[X] \leq \mathbf{E}[\,|X|\,]$.

One of the numbers $\mathbf{E}[X]$, $-\mathbf{E}[X]$ must be equal to $|\,\mathbf{E}[X]\,|$. And each of these numbers is less than or equal to $\mathbf{E}[\,|X|\,]$. So we have shown that equation (10.31) holds in general.

10.7 General random variables

We won't take time to define expected value for general mathematical random variables carefully in this book, but later we will use mathematical expectation for lots of random variables that do not have finite range. Expectation can be defined for any bounded random variable, and for unbounded random variables that are not too big.

A bounded random variable is a random variable which is a bounded function on the sample space.

You could probably guess the definition of a bounded function, but we'll state it carefully anyway.

Definition 10.18 (Bounded functions). A function f on any set is said to be bounded if there is some number c such that $|f(x)| \leq c$ holds for every x in the domain of f.

For unbounded random variables, sometimes the expected value exists, and sometimes it doesn't.

Remark 10.19 (Linearity for expectation of general random variables). The linearity property holds for bounded random variables, as we would hope. For unbounded random variables, the linear property comes with a little bit of "fine print", since expected values might not exist. So the correct way to state additivity of expectation in the general case will be to say that *if* $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ exist, then $\mathbf{E}[X+Y]$ exists and equation (10.16) holds (see Theorem 14.9). Similarly, the general version of the scaling property says that *if* $\mathbf{E}[X]$ exists then $\mathbf{E}[c\,X]$ exists and equation (10.6) holds.

That seems easy to remember.

We'll often stick to finite-range random variables when we want to give a careful derivation of some fact. But much of what is true for finite-range random variables is true in general.

Example 10.20 (Expectations with a density on the sample space). Just so readers can see an example of calculating expected values using a different method from that given in equation (10.1), suppose that the sample space Ω is a continuous interval $[s, t]$ of the real line, as discussed in Chapter 3. Assume that probabilities are given by a uniform distribution on $[s, t]$ (Definition 3.2).

As in Exercise 3.5, the uniform distribution on $[s, t]$ is given by a constant density, say $f = c$. And since $\mathbf{P}(\Omega) = 1$ must hold, we need to have $\int_s^t f = 1$, and so $c = 1/(t - s)$.

In that situation, if X is a random variable on the sample space, it turns out that the correct formula for $\mathbf{E}[X]$ is:

$$\mathbf{E}[X] = \int_s^t X(u)c\,du. \tag{10.32}$$

In other words, here we find $\mathbf{E}[X]$ by integrating its value over the sample space.

More generally, if the distribution on $\Omega = [s, t]$ is given by a density function f (as in Definition 3.4), the correct formula for $\mathbf{E}[X]$ is:

$$\mathbf{E}[X] = \int_s^t X(u)f(u)\,du. \tag{10.33}$$

Not surprising, just different from finite-range case.

For more discussion of finding expectations using densities, see Section 15.3.

10.8 Solutions for Chapter 10

Solution (Exercise 10.1). The range of Y is $\{0, 2\}$, while $\mathbf{P}(Y = 0) = 2/5$ and $\mathbf{P}(Y = 2) = 3/5$.

By definition,

$$\mathbf{E}[Y] = \frac{2}{5}\,0 + \frac{3}{5}\,2 = 1.2$$

Solution (Exercise 10.2). Each value v in the range should contribute to the expected value $\mathbf{E}[X]$. The contribution is given by the term $v\,\mathbf{P}(X = v)$ in the sum which defines $\mathbf{E}[X]$.

Physically, the importance of a value v for the expected value should depend on the probability that X is equal to v. If a value appears twice in the sum, then its contribution to the sum is doubled. This is not consistent with actual importance of the value.

Solution (Exercise 10.3). Definition 10.2 shows that $\mathbf{E}[X]$ is given by a weighted sum of terms.

Addition is commutative! (Yes, I've been waiting for a chance to say that.) Changing the order of the terms in a sum does not change the a value of the sum.

Solution (Exercise 10.4). Let $x_1, \ldots, x_k$ be a list of the distinct elements in the range of X. By definition,

$$\mathbf{E}[X] = \sum_{j=1}^{k} x_j \mathbf{P}(X = x_j).$$

To show that equation (10.5) holds, we need to compare two sums, and see if they are equal:

$$\sum_{j=1}^{k} x_j \mathbf{P}(X = x_j) \stackrel{?}{=} \sum_{i=1}^{n} y_i \mathbf{P}(X = y_i).$$

The order of the terms in a sum does not matter.

Suppose that a value y_i is not in the range. Then $\mathbf{P}(X = y_i) = 0$, and the term $y_i \mathbf{P}(X = y_i) = 0$, so that term contributes nothing to the sum on the right. We can throw away any term like that from the sum on the right.

Suppose that 0 is in the range. Then $0 = x_j$ for some j. The term $x_j \mathbf{P}(X = x_j) = 0$, so that term contributes nothing to the sum on the left, and we can throw it away from the sum on the left. For the same reason, if $y_i = 0$ for some i, we can throw away the term $y_i \mathbf{P}(X = y_i)$.

After all this throwing away, the remaining sum on the left will have the same terms as the remaining sum on the right, possibly in a different order. So the sums are indeed equal.

Solution (Exercise 10.5). Apply Exercise 10.4 with $n = 1$ and $y_1 = c$.

Solution (Exercise 10.6).

(i) From the assumptions, the range of X is $\{0, 1, 2, 5\}$, and

$$\begin{aligned}
&\{X = 0\} = \{Z = 0\}, \\
&\{X = 1\} = \{Z = 1\} \cup \{Z = 3\} \cup \ldots \cup \{Z = 99\}, \\
&\{X = 2\} = \{Z = 2\} \cup \{Z = 4\} \cup \ldots \cup \{Z = 98\}, \\
&\{X = 5\} = \{D = 100\}.
\end{aligned} \tag{10.34}$$

Hence the distribution of X is given by:

$$\begin{aligned}
&\mathbf{P}(X = 0) = \mathbf{P}(Z = 0) = \frac{1}{101}, \\
&\mathbf{P}(X = 1) = \mathbf{P}(Z = 1) + \mathbf{P}(Z = 3) + \ldots + \mathbf{P}(Z = 99) = \frac{50}{101}, \\
&\mathbf{P}(X = 2) = \mathbf{P}(Z = 2) + \mathbf{P}(Z = 4) + \ldots + \mathbf{P}(Z = 98) = \frac{49}{101}, \\
&\mathbf{P}(X = 5) = \mathbf{P}(Z = 100) = \frac{1}{101}.
\end{aligned} \tag{10.35}$$

By definition,

$$\begin{aligned}
\mathbf{E}[X] &= 0 \cdot \mathbf{P}(X = 0) + 1 \cdot \mathbf{P}(X = 1) + 2 \cdot \mathbf{P}(X = 2) + 5 \cdot \mathbf{P}(X = 5) \\
&= 0 \cdot \frac{1}{101} + 1 \cdot \frac{50}{101} + 2 \cdot \frac{49}{101} + 5 \cdot \frac{1}{101} = \frac{153}{101}.
\end{aligned} \tag{10.36}$$

(ii) By equation (10.35),

$$\begin{aligned}
&0 \cdot \mathbf{P}(X = 0) = 0 \cdot \mathbf{P}(Z = 0), \\
&1 \cdot \mathbf{P}(X = 1) = 1 \cdot \mathbf{P}(Z = 1) + 1 \cdot \mathbf{P}(Z = 3) + \ldots + 1 \cdot \mathbf{P}(Z = 99), \\
&2 \cdot \mathbf{P}(X = 2) = 2 \cdot \mathbf{P}(Z = 2) + 2 \cdot \mathbf{P}(Z = 4) + \ldots + 2 \cdot \mathbf{P}(Z = 98), \\
&5 \cdot \mathbf{P}(X = 5) = 5 \cdot \mathbf{P}(Z = 100).
\end{aligned} \tag{10.37}$$

Since φ gives the value of X in terms of the value of Z, we can rewrite equation (10.37) as:

$$\begin{aligned}
&0 \cdot \mathbf{P}(X = 0) = \varphi(0) \cdot \mathbf{P}(Z = 0), \\
&1 \cdot \mathbf{P}(X = 1) = \varphi(1) \cdot \mathbf{P}(Z = 1) + \varphi(3) \cdot \mathbf{P}(Z = 3) + \ldots + \varphi(99) \cdot \mathbf{P}(Z = 99), \\
&2 \cdot \mathbf{P}(X = 2) = \varphi(2) \cdot \mathbf{P}(Z = 2) + \varphi(4) \cdot \mathbf{P}(Z = 4) + \ldots + \varphi(98) \cdot \mathbf{P}(Z = 98), \\
&5 \cdot \mathbf{P}(X = 5) = \varphi(100) \cdot \mathbf{P}(Z = 100).
\end{aligned} \tag{10.38}$$

If you add up all the equations in statement (10.38), you obtain:

$$\mathbf{E}[X] = \sum_{i=0}^{100} \varphi(i)\mathbf{P}(Z = i),$$

which is equation (10.8).

To phrase this differently: the proof of equation (10.8) is just a matter of *grouping* the terms in the sum, in order to obtain equation (10.36).

Solution (Exercise 10.7).

From the definition of expected value

$$\mathbf{E}\left[X^2\right] = 0 \cdot \mathbf{P}(X^2 = 0) + 1 \cdot \mathbf{P}(X^2 = 1) + 4 \cdot \mathbf{P}(X^2 = 4) = \frac{2}{5} + 4 \cdot \frac{2}{5} = 2.$$

Using Theorem 10.8

$$\mathbf{E}\left[X^2\right] = (-2)^2 \cdot \frac{1}{5} + (-1)^2 \cdot \frac{1}{5} + 0 \cdot \frac{1}{5} + 1^2 \cdot \frac{1}{5} + 2^2 \cdot \frac{1}{5} = \frac{4}{5} + \frac{1}{5} + \frac{1}{5} + \frac{4}{5} = 2.$$

Solution (Exercise 10.8). Consider a long sequence of N repeated experiments. Let the measured values of X in these experiments be $x_1, \ldots, x_N$ and let the measured values of Y in these experiments by $y_1, \ldots, y_N$. Then the measured results for $X + Y$ are $x_1 + y_1, \ldots, x_N + y_N$. The corresponding experimental averages are:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,\ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,\ \overline{x+y} = \frac{1}{N}\sum_{i=1}^{N} (x_i + y_i)\,. \tag{10.39}$$

Of course

$$\frac{1}{N}\sum_{i=1}^{N} (x_i + y_i) = \frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} y_i,$$

so

$$\overline{x+y} = \bar{x} + \bar{y}. \tag{10.40}$$

The frequency interpretation for expected value tells us that for large N, $\bar{x} \approx \mathbf{E}[X]$, $\bar{y} \approx \mathbf{E}[Y]$ and $\overline{x+y} \approx \mathbf{E}[X+Y]$. Since these approximations can be made as precise as we like by taking a large number of repetitions, equation (10.40) implies $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.

Solution (Exercise 10.9). By definition

$$\mathbf{E}\left[S_n\right] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

The $k = 0$ term is zero, so

$$\mathbf{E}\left[S_n\right] = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

By equation (8.8)

$$\mathbf{E}\left[S_n\right] = \sum_{k=1}^{n} k \frac{n}{k} \binom{n-1}{k-1} p^k (1-p)^{n-k} = \sum_{k=1}^{n} kp \frac{n}{k} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)}.$$

Letting $i = k - 1$,

$$\mathbf{E}\left[S_n\right] = np \sum_{i=0}^{n-1} \binom{n-1}{i} p^i (1-p)^{(n-1)-i}. \tag{10.41}$$

Let S_{n-1} denote the number of heads obtained in $n-1$ coin tosses, when the coin has success probability p. Equation (10.41) says that

$$\mathbf{E}\left[S_n\right] = np \sum_{i=0}^{n-1} \mathbf{P}(S_{n-1} = i).$$

Since the range of S_{n-1} is $\{0, 1, \ldots, n-1\}$,

$$\sum_{i=0}^{n-1} \mathbf{P}(S_{n-1} = i) = 1.$$

Hence $\mathbf{E}\left[S_n\right] = np$.

Solution (Exercise 10.10). By Exercise 10.4,

$$\mathbf{E}\left[X\right] = a_1 \mathbf{P}(X = a_1) + \ldots + a_k \mathbf{P}(X = a_k) - a_1 \mathbf{P}(X = -a_1) - \ldots - a_k \mathbf{P}(X = -a_k) = 0.$$

Solution (Exercise 10.11).

Method 1

$$\text{range of } X \ = \{-10000000, -999999, \ldots, 999999, 1000000\},$$

and all these points have equal probability. There are 2000001 points in the range. Hence, by definition,

$$\mathbf{E}[X] = \sum_{i=-1000000}^{1000000} i \ \frac{1}{2000001} = \left(\sum_{i=-1000000}^{-1} i \ \frac{1}{2000001} \right) + 0 \ \frac{1}{2000001} + \sum_{i=1}^{1000000} i \ \frac{1}{2000001}.$$

Thus

$$\mathbf{E}[X] = \left(\sum_{i=-1000000}^{-1} i \ \frac{1}{2000001} \right) + \sum_{i=1}^{1000000} i \ \frac{1}{2000001}.$$

Let $j = -i$ in the first of these two sums. Then that sum becomes

$$- \sum_{j=1}^{10000000} j \frac{1}{20000001},$$

and this term cancels with the second sum, so $\mathbf{E}[X] = 0$.

Method 2 Again we note that

$$\text{range of } X \ = \{-10000000, -999999, \ldots, 999999, 1000000\},$$

All points in the range have the same probability, and if j is in the range then so is $-j$.

Since $\mathbf{P}(-X = j) = \mathbf{P}(X = -j) = \mathbf{P}(X = j)$, it follows that X and $-X$ have the same distribution. Hence by Lemma 10.15, $\mathbf{E}[X] = 0$.

Solution (Exercise 10.12). We see that

$$\text{range of } X \ = \{0, 1, \ldots, 1000\}.$$

All these values have the same probability, so

$$\mathbf{P}(X = j) = \frac{1}{1001}$$

Let $Y = X - 500$.

range of $Y \ = \{0 - 500, 1 - 500, \ldots, 1000 - 500\} = \{-500, -499, \ldots, 499, 500\}$.

Hence

Fact 1 Y and $-Y$ have the same range.

We also see that for each i in the range of Y,

$$\mathbf{P}(Y = i) = \mathbf{P}(X - 500 = i) = \mathbf{P}(X = i + 500) = \frac{1}{1001}.$$

Hence

Fact 2 All the points in the range of Y have the same probability.

Facts 1 and 2 imply that $\mathbf{P}(Y = i) = \mathbf{P}(Y = -i) = \mathbf{P}(-Y = i)$ for all i in the range of Y. Thus the distributions of Y and $-Y$ are the same.

By Lemma 10.15, $\mathbf{E}[Y] = 0$.

That is, $\mathbf{E}[X - 500] = 0$.

Now we use linearity again. Since expectation is a linear operation, $\mathbf{E}[X] - \mathbf{E}[500] = 0$, i.e. $\mathbf{E}[X] = 500$.

Solution (Exercise 10.13). Consider a long sequence of N repeated experiments. Let the measured values of X in these experiments be $x_1, \ldots, x_N$ and let the measured values of Y in these experiments by $y_1, \ldots, y_N$.

By assumption, $x_i \leq y_i$ for every i.

The corresponding experimental averages are:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{10.42}$$

Since

$$\frac{1}{N}\sum_{i=1}^{N} x_i \leq \frac{1}{N}\sum_{i=1}^{N} y_i.$$

we have $\bar{x} \leq \bar{y}$.

Taking N larger and larger gives averages $\bar{x}$ and $\bar{y}$ which approximate $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ as precisely as we like. Hence we must have $\mathbf{E}[X] \leq \mathbf{E}[Y]$.

Chapter 11

More properties of expected value

11.1 Indicator Functions

In this section we introduce a simple notation which is useful when writing expressions involving expectations or integrals.

Definition 11.1 (Indicator function of a set). Let a set S be given, and let A be any subset of S. We define the **indicator function** of A, denoted by $\mathbf{1}_A$, as follows.

$\mathbf{1}_A$ is a function on S, and for any $x \in S$:

$$\mathbf{1}_A(x) = \begin{cases} 1 \text{ if } x \in A, \\ 0 \text{ otherwise.} \end{cases} \tag{11.1}$$

It should be emphasized that indicator functions are a general idea, defined for subsets of *any* set, not just sample spaces or subsets of the real line. You can picture $\mathbf{1}_A(x)$ as a signal light which comes on when x is a member of A.

Please check that from the definition,

$$\mathbf{1}_A = \mathbf{1}_B \iff A = B. \tag{11.2}$$

Here we use $\iff$ to mean "if and only if" (i.e. "implies" in both directions).

Lemma 10.3 can be expressed using indicator functions. It says that:

$$\mathbf{E}\left[c\,\mathbf{1}_A\right] = c\,\mathbf{P}(A). \tag{11.3}$$

In particular we have the fundamental equation connecting probability and expected value:

$$\mathbf{E}\left[\mathbf{1}_A\right] = \mathbf{P}(A). \tag{11.4}$$

It seems to be easier to manipulate numbers than sets, so it can be profitable to translate set statements into indicator statements.

Exercise 11.1 (Basic indicator facts). Prove all the following facts:

$$\mathbf{1}_A^2 = \mathbf{1}_A, \tag{11.5}$$
$$\mathbf{1}_{A^c} = 1 - \mathbf{1}_A, \tag{11.6}$$
$$\mathbf{1}_{A\cap B} = \min\left(\mathbf{1}_A, \mathbf{1}_B\right) = \mathbf{1}_A\mathbf{1}_B, \tag{11.7}$$
$$\mathbf{1}_{A\cup B} = \max\left(\mathbf{1}_A, \mathbf{1}_B\right). \tag{11.8}$$

[Solution]

As a rather trivial example, note that using equation (11.6) twice we have

$$\mathbf{1}_{(A^c)^c} = 1 - \mathbf{1}_{A^c} = 1 - (1 - \mathbf{1}_A) = \mathbf{1}_A.$$

This gives another derivation of equation (2.24), which says that $(A^c)^c = A$.

Exercise 11.2 (Indicator of a disjoint union). Suppose that A, B are any subsets of a given set S. Show that

$$A \text{ and } B \text{ are disjoint} \iff \mathbf{1}_{A\cup B} = \mathbf{1}_A + \mathbf{1}_B. \tag{11.9}$$

[Solution]

Here's a useful fact about numbers.

Exercise 11.3 (Sum equals max plus min). Show that for any real numbers t, u,

$$t + u = \max(t, u) + \min(t, u). \tag{11.10}$$

[Solution]

As a consequence of Exercise 11.3 and equations (11.7) and (11.8), we have

$$\mathbf{1}_{A\cup B} + \mathbf{1}_{A\cap B} = \mathbf{1}_A + \mathbf{1}_B. \tag{11.11}$$

Rewriting equation (11.11) as

$$\mathbf{1}_{A\cup B} = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_{A\cap B}, \tag{11.12}$$

taking expectations of both sides, and then by applying equation (11.4), we obtain Theorem 2.24, the Inclusion-Exclusion formula.

You may remember that in the original proof for inclusion-exclusion, we used the trick of breaking up events into disjoint pieces. That seemed useful, but we don't seem to be using that trick with this approach. Or are we? Maybe the pieces are the one-point sets in the sample space.

Since $\mathbf{1}_{A\cap B}$ is the zero function if and only if $A \cap B$ is the empty set, equation (11.11) gives us equation (11.9) as a special case.

Note that using equation (11.7) we can rewrite equation (11.12) as

$$\mathbf{1}_{A\cup B} = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A\,\mathbf{1}_B. \tag{11.13}$$

Exercise 11.4. One can generalize Theorem 2.24 to the case of n sets. The usual proof using set operations has two steps. In the first step, one guesses the correct formula in some way. In the second step, one proves the conjectured formula by induction.

Instead of using that approach, derive the correct formula for the case $n = 3$, and prove it at the same time, by applying equation (11.13) twice.

[Solution]

Remark 11.2 (Subadditivity for indicators). We used the Inclusion-Exclusion formula to prove subadditivity, Theorem 2.25. But now that we have indicator functions, it seems more direct just to note an obvious subadditivity fact for indicator functions: for any events $D_1, \ldots, D_k$, if D is the union of these events, then

$$\mathbf{1}_D \leq \sum_{j=1}^{k} \mathbf{1}_{D_j}. \tag{11.14}$$

To prove equation (11.14), just evaluate both sides for a sample point ω, as follows.

If $\omega \in D$ then $\omega \in D_j$ for at least one j. Thus the left side of the equation is one, and the right side is greater than or equal to one.

And if ω is not in D, then the left side of the equation is zero, and the right cannot be negative.

So equation (11.14) holds. Now take expectation of both sides of the equality in this equation. Using monotonicity and linearity of expectation, and equation (11.4), you will produce Theorem 2.25.

The next exercise is important for understanding how our concepts fit together.

Exercise 11.5 (Random variable as a sum of constants times indicators). Let X be a random variable with finite range. Let $x_1, \ldots, x_k$ be the distinct values in the range of X.

(i) Explain why

$$X = x_1 \mathbf{1}_{\{X=x_1\}} + \ldots + x_k \mathbf{1}_{\{X=x_k\}}. \tag{11.15}$$

(ii) Show that linearity of expectation and equation (11.4) imply equation (10.1), which is the defining formula for $\mathbf{E}[X]$.

Thus for finite range random variables, linearity of expectation and equation (11.4) imply everything about expected values.

[Solution]

Remark 11.3 (Integral over a set using indicator notation). In calculus we are very familiar with the idea of integrating a function over a set, usually when the set is an interval.

In equation (3.16) we gave the general definition for integrating a function over a set. The function g in equation (3.16) is easily seen to be equal to $\mathbf{1}_A f$, so $\int_A f$ can be conveniently expressed using indicator functions:

$$\int_A f = \int \mathbf{1}_A f. \tag{11.16}$$

We'll sometimes use this notation later, for convenience.

Example 11.4 (Additivity for integration). Back in Section 3.5 we mentioned that for disjoint sets, the integral over the union is the sum of the integrals over the disjoint sets making up the union (equation (3.14)): i.e. if $A = D_1 \cup D_2$, where D_1, D_2 are disjoint, then

$$\int_A f = \int_{D_1} f + \int_{D_2} f. \tag{11.17}$$

This follows from the definition of integration over a set. We can express the argument very neatly by using indicator function notation and equation (11.16). Equation (11.9) tells us that

$$\mathbf{1}_A = \mathbf{1}_{D_1} + \mathbf{1}_{D_2}.$$

Since integration is an additive operation, integrating this equation gives equation (11.17).

Exercise 11.6 (Writing a random variable using cases). Let $D_1, \ldots, D_n$ be events for some probability model. Suppose that:

(a) The events $D_1, \ldots, D_n$ are disjoint, and

(b) The union of $D_1, \ldots, D_n$ is the whole sample space Ω.

Let Z be a random variable such that Z is constant on each set D_i (as in Figure 10.1). For each i, let v_i be a number such that $X(\omega) = v_i$ for every $\omega \in D_i$. (Thus if D_i is empty, v_i can be any number.)

Under these assumptions, prove that

$$Z = v_1 \mathbf{1}_{D_1} + \ldots + v_n \mathbf{1}_{D_n}. \tag{11.18}$$

[Solution]

11.2 Expectation over a set

We defined integration over a set in equation (3.16). In this section we define a similar concept for expected value. The idea is simple but convenient.

Definition 11.5 (Expectation over a subset of the sample space). Let a probability model be given with sample space Ω. For any real-valued random variable and any event A, define the expectation of X over A by

$$\text{expectation of } X \text{ over } A \ = \mathbf{E}\left[Z\right], \tag{11.19}$$

where

$$Z(\omega) = \begin{cases} X(\omega) & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{11.20}$$

Since this definition is intended to apply to general random variables, we have to mention that equation (11.19) is the definition of the expectation of X over A, *if* $\mathbf{E}\left[Z\right]$ exists. If $\mathbf{E}\left[Z\right]$ does not exist, the the expectation of X over A is undefined. Of course if the range of X is finite, the range of Z is finite, so $\mathbf{E}\left[Z\right]$ certainly exists, and there is no problem.

Indicator function notation (Definition 11.1), gives us a handy way to write expectation over a set:

$$\text{expectation of } X \text{ over A} \ = \mathbf{E}\left[\mathbf{1}_A X\right]. \tag{11.21}$$

This definition applies to any random variable, although in the present chapter we will only study its properties in the case of random variables with finite range.

Expectation over a set is especially useful when dealing with combining several expectations obtained under differing assumptions. The key concept for that purpose is called *conditional expectation* and it deserves its own section.

11.3 Conditional expectation

The following definition holds for general random variables, not just random variables with finite range.

Definition 11.6 (Conditional expectation). Let X be a random variable on a sample space Ω, and let A be an event with $\mathbf{P}(A) > 0$. The conditional expectation of X given A, denoted by $\mathbf{E}[X \mid A]$, is defined by

$$\mathbf{E}[X|A] = \frac{\mathbf{E}[\mathbf{1}_A X]}{\mathbf{P}(A)}. \tag{11.22}$$

Equation (11.22) is a convenient mathematical formula for conditional expectation, but the physical meaning of conditional expectation is better expressed in the following lemma. Like Definition 11.6, this lemma holds for all random variables, not just random variables with finite range.

Lemma 11.7 (Conditional expection uses conditional probabilities). Define the conditional probability set-function $\tilde{\mathbf{P}}$ by

$$\tilde{\mathbf{P}}(D) = \mathbf{P}(D \mid A) \tag{11.23}$$

for any event D. The definition of $\tilde{\mathbf{P}}$ says that it is the probability distribution which incorporates additional knowledge, namely that event A has occurred.

Then: the conditional expectation of X given A, which was defined in equation (11.22), is equal to the expected value of X using $\tilde{\mathbf{P}}$ instead of $\mathbf{P}$.

When we are using $\tilde{\mathbf{P}}$ as our probability set-function we can denote the expectation of X by $\mathbf{E}_{\tilde{\mathbf{P}}}[X]$. Thus Lemma 11.7 can be stated compactly as:

$$\mathbf{E}[X|A] = \mathbf{E}_{\tilde{\mathbf{P}}}[X]. \tag{11.24}$$

And this equation expresses the fact that conditional expectation really does mean what its name suggests.

Proof. For the proof we assume that X has a finite range.

Let $x_1, \ldots, x_k$ be a list of the distinct values in the range of X.

By Exercise 11.5,

$$X = \sum_{i=1}^{k} x_i \mathbf{1}_{\{X=x_i\}}.$$

Let A be an event with $\mathbf{P}(A) > 0$. Then

$$\mathbf{1}_A X = \sum_{i=1}^{k} x_i \mathbf{1}_A \mathbf{1}_{\{X=x_i\}}.$$

Using equation (11.7), this says that

$$\mathbf{1}_A X = \sum_{i=1}^{k} x_i \mathbf{1}_{A\cap\{X=x_i\}}.$$

Taking expected value of both sides of the equation, and using equation (11.4),

$$\mathbf{E}[\mathbf{1}_A X] = \sum_{i=1}^{k} x_i \mathbf{P}(A \cap \{X = x_i\}),$$

so

$$\frac{\mathbf{E}[\mathbf{1}_A X]}{\mathbf{P}(A)} = \sum_{i=1}^{k} x_i \frac{\mathbf{P}(A \cap \{X = x_i\})}{\mathbf{P}(A)} = \sum_{i=1}^{k} x_i \tilde{\mathbf{P}}(\{X = x_i\}).$$

By the mathematical definition, the left side of this equation is $\mathbf{E}[X|A]$.

The right side of the equation is equal to $\mathbf{E}_{\tilde{\mathbf{P}}}[X]$ by the definition of expected value.

This proves equation (11.24). □

Our proof justified Lemma (11.7) for the special case of a random variable X with finite range, but remember that this lemma holds for all random variables X.

Conditional probabilities may be simpler to use than the original probability distribution, since they permit us to break up a calculation into cases. In particular, we have the Law of Total Expectation, which generalizes the Law of Total Probability (Theorem 4.6).

Theorem 11.8 (Law of Total Expectation). Let $D_1, \ldots, D_k$ be disjoint events with union D, and let X be a random variable such that $\mathbf{E}[X]$ exists. Then

$$\mathbf{E}[\mathbf{1}_D X] = \sum_{i=1}^{k} \mathbf{P}(D_i)\, \mathbf{E}[X|\, D_i], \tag{11.25}$$

where for each i if $\mathbf{P}(D_i) = 0$ we replace $\mathbf{E}[X|\, D_i]$ by any number we like.

If $D = \Omega$, this becomes

$$\mathbf{E}[X] = \sum_{i=1}^{k} \mathbf{P}(D_i)\, \mathbf{E}[X|\, D_i]. \tag{11.26}$$

Exercise 11.7. Let $D_1, \ldots, D_k$ be disjoint events with union D. Prove that

$$\mathbf{1}_D = \sum_{i=1}^{k} \mathbf{1}_{D_i}. \tag{11.27}$$

This equation generalizes equation (11.9), of course.

[Solution]

Exercise 11.8 (Proof of the theorem). Prove Theorem 11.8.

Equation (11.22) is convenient for this purpose. [Solution]

Example 11.9 (A simple model with two cases). Consider the following very simple game.

A player tosses a fair coin. If the toss gives a head, the player wins ten dollars, and the game ends. Otherwise, a fair die is rolled, and the player wins the number of dollars shown on the die. Let X be the amount that the player wins. We wish to find $\mathbf{E}[X]$.

We will actually calculate $\mathbf{E}[X]$ twice, and compare the two approaches.

Method 1

What should be the sample space for this calculation? A reasonable choice is to take the set Ω to consist of a symbol H together with the numbers $1, 2, 3, 4, 5, 6$.

The symbol H represents the outcome for which the coin toss results in a head. The number i represents the outcome for which the coin toss results in a tail *and* then score on the die roll is i.

The distribution $\mathbf{P}$ for this sample space is obtained as follows.

By the description of the experiment, $\mathbf{P}(H) = \frac{1}{2}$.

By the multiplied-through version of the conditional probability formula (equation (4.2)),

$$\mathbf{P}(\{i\}) = \mathbf{P}(H^c)\mathbf{P}(\{i\} \mid H^c).$$

In this equation we have physical probabilities, since the abstract model is still being defined. We know the probabilities for a fair die roll, and we know the roll of the die is not affected by the coin toss, so

$$\mathbf{P}(\{i\} \mid H^c) = \frac{1}{6}.$$

Thus in our model we should define

$$\mathbf{P}(\{i\}) = \frac{1}{2}\frac{1}{6} = \frac{1}{12} \text{ for } i = 1, 2, 3, 4, 5, 6.$$

Using the definition of $\mathbf{E}[X]$ gives

$$\mathbf{E}[X] = \frac{1}{2}10 + \sum_{i=1}^{6} \frac{1}{12} i = \frac{1}{2}10 + \frac{1}{12}(1 + 2 + 3 + 4 + 5 + 6).$$

Method 2

Let's start the problem again, and apply the Law of Total Expectation, Theorem 11.8.

$$\mathbf{E}[X] = \mathbf{P}(H)\mathbf{E}[X|\, H] + \mathbf{P}(H^c)\mathbf{E}[X|\, H^c]. \tag{11.28}$$

Given H, $X = 10$, so

$$\mathbf{E}[X|H] = 10.$$

The physical description of the situation given H^c is that we are rolling a fair die, and X is the number shown on the die. We also know by equation (11.24) that $\mathbf{E}[X|H^c]$ is equal to the expected value of X in this situation. So we have a little self-contained problem, which we know how to solve: finding the expected value for one roll of a fair die. Thus

$$\mathbf{E}[X|H^c] = \frac{1}{6}(1+2+3+4+5+6).$$

Since $\mathbf{P}(H^c) = 1/2 = \mathbf{P}(H)$, substituting in equation (11.28) gives:

$$\mathbf{E}[X] = \frac{1}{2}10 + \frac{1}{2}\left(\frac{1}{6}(1+2+3+4+5+6)\right).$$

Either of the two methods of calculating $\mathbf{E}[X]$ seems equally easy in the present example, but we can already notice two significant benefits of using the Law of Total Expectation.

(i) The problem is decomposed in a natural way into simpler problems, which are "self-contained problems", i.e. problems which can be considered separately.

(ii) There is no need to define a sample space for the original problem.

These benefits are more significant in complex problems.

Exercise 11.9. A player has a fair die and a coin with success probability 1/5. In the first stage of the experiment, the player rolls the die once. Let k be the number obtained. The player then tosses the coin k times. Find the expected number of successes obtained in this experiment.

[Solution]

11.4 Solutions for Chapter 11

Solution (Exercise 11.1).

Proving equation (11.5) For any set A, the only possible values for $\mathbf{1}_A$ are 0 and 1, Since $0^2 = 0$ and $1^2 = 1$, equation (11.5) follows.

Proving equation (11.6) From the definitions, $\mathbf{1}_{A^c}(t) = 0$ exactly when $\mathbf{1}_A(t) = 1$, and $\mathbf{1}_{A^c}(t) = 1$ exactly when $\mathbf{1}_A(t) = 0$. This is equivalent to saying that $\mathbf{1}_{A^c}(t) = 1 - \mathbf{1}_A(t)$, so equation (11.6) holds.

Proving equation (11.7)

When $t \in A \cap B$, $\mathbf{1}_{A\cap B} = 1$, and both of the statements $\mathbf{1}_A(t) = 1$, $\mathbf{1}_A(t) = 1$ hold. Hence $\mathbf{1}_{A\cap B}(t) = \min(\mathbf{1}_A(t), \mathbf{1}_B(t))$.

When $t \notin A \cap B$, certainly $\mathbf{1}_{A\cap B}(t) = 0$ by definition. Also at least one of the statements $t \notin A$, $t \notin B$ holds. Thus at least one of the statements $\mathbf{1}_A(t) = 0$, $\mathbf{1}_B(t) = 0$, holds, so $\min(\mathbf{1}_A(t), \mathbf{1}_B(t)) = 0 = \mathbf{1}_{A\cap B}(t)$.

This proves the first equality in equation (11.7) for all possible cases.

When $t \in \{0, 1\}$ and $u \in \{0, 1\}$, we see by checking cases that $t\,u = \min(t, u)$. This proves the remaining equality in equation (11.7).

Proving equation (11.8) When $t \in A \cup B$, $\mathbf{1}_{A\cup B} = 1$, and at least one of the statements $\mathbf{1}_A(t) = 1$, $\mathbf{1}_A(t) = 1$ hold. Hence $\mathbf{1}_{A\cup B}(t) = \max(\mathbf{1}_A(t), \mathbf{1}_B(t))$.

When $t \notin A \cup B$, certainly $\mathbf{1}_{A\cup B}(t) = 0$ by definition. Also both of the statements $t \notin A$, $t \notin B$ holds. Thus both of the statements $\mathbf{1}_A(t) = 0$, $\mathbf{1}_B(t) = 0$, holds, so $\max(\mathbf{1}_A(t), \mathbf{1}_B(t)) = 0 = \mathbf{1}_{A\cap B}(t)$.

This proves equation (11.8) for all possible cases.

Proving equation (11.11) Since $\mathbf{1}_{A\cup B} = \max(\mathbf{1}_A, \mathbf{1}_B)$ and $\mathbf{1}_{A\cap B} = \min(\mathbf{1}_A, \mathbf{1}_B)$, this proves equation (11.11).

Solution (Exercise 11.2).

$\Longrightarrow$**:** Suppose that A and B are disjoint subsets of the given set S.

Let x be a point in S.

If $x \in A$ then $x \in A \cup B$ and $x \notin B$. Thus $\mathbf{1}_A(x) = 1$, $\mathbf{1}_{A\cup B} = 1$ and $\mathbf{1}_B = 0$. Since $1 = 1 + 0$, equation (11.9) holds.

Similarly equation (11.9) holds if $x \in B$.

The remaining case is the case that $x \in (A \cup B)^c$. In this case $\mathbf{1}_{A\cup B}(x) = 0 = \mathbf{1}_A(x) = \mathbf{1}_B(x)$. Since $0 = 0 + 0$, equation (11.9) holds in this case also.

$\Longleftarrow$ **:** Suppose that equation (11.9) holds. If there were a point $x \in A \cap B$, for that point we would have $\mathbf{1}_{A\cup B}(x) = 1$, $\mathbf{1}_A(x) = 1$ and $\mathbf{1}_B(x) = 1$. Since $1 \neq 1 + 1$, equation (11.9) would not hold at x.

Since equation (11.9) does hold, we conclude that $A \cap B$ is empty.

Solution (Exercise 11.3). If $t \neq u$, then one of these two numbers is the max, and the other is the min. Thus $t + u = \max(t, u) + \min(t, u)$.

On the other hand, if $t = u$, then both number are equal to the max and both are equal to the min. Hence once again we have $t + u = \max(t, u) + \min(t, u)$.

This proves equation (11.10).

Solution (Exercise 11.4). Using equation (11.13),

$$\mathbf{1}_{A\cup B\cup C} = \mathbf{1}_{(A\cup B)\cup C} = \mathbf{1}_{A\cup B} + \mathbf{1}_C - \mathbf{1}_{A\cup B}\,\mathbf{1}_C.$$

Applying equation (11.13) to $\mathbf{1}_{A\cup B}$,

$$\begin{aligned}\mathbf{1}_{A\cup B\cup C} &= \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A\,\mathbf{1}_B + \mathbf{1}_C - (\mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A\mathbf{1}_B)\,\mathbf{1}_C \\ &= \mathbf{1}_A + \mathbf{1}_B + \mathbf{1}_C - \mathbf{1}_A\,\mathbf{1}_B - \mathbf{1}_A\mathbf{1}_C - \mathbf{1}_B\mathbf{1}_C + \mathbf{1}_A\mathbf{1}_B\mathbf{1}_C.\end{aligned}$$

Using equation (11.7), this says that

$$\mathbf{1}_{A\cup B\cup C} = \mathbf{1}_A + \mathbf{1}_B + \mathbf{1}_C - \mathbf{1}_{A\cap B} - \mathbf{1}_{A\cap C} - \mathbf{1}_{B\cap C} + \mathbf{1}_{A\cap B\cap C}.$$

Taking expectations of both sides, and then applying equation (11.4), we obtain

$$\begin{aligned}\mathbf{P}(A \cup B \cup C) = {} & \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) \\ & - \mathbf{P}(A \cap B) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) \\ & + \mathbf{P}(A \cap B \cap C).\end{aligned} \tag{11.29}$$

This is the generalization of equation (2.14) of Theorem 2.24, to the case of three events.

Notice that the plus and minus signs alternate, depending on the number of sets in the intersection. This pattern holds for all n.

Solution (Exercise 11.5).

(i) For any outcome ω, we must show that

$$X(\omega) = x_1\mathbf{1}_{\{X=x_1\}}(\omega) + \ldots + x_k\mathbf{1}_{\{X=x_k\}}(\omega). \tag{11.30}$$

Let x_j be the value of $X(\omega)$. Then $\omega \in \{X = x_j\}$.

On the right side of equation (11.30), $\mathbf{1}_{\{X=x_i\}}(\omega) = 0$ unless $i = j$. In that case, $\mathbf{1}_{\{X=x_i\}}(\omega) = 1$.

Thus the only surviving term on the right side of the equation is $x_j \cdot 1 = x_j$. This equals the left side, so we are done.

(ii) Taking expected value of both sides of equation (11.30) gives

$$\mathbf{E}[X] = x_1 \mathbf{E}\left[\mathbf{1}_{\{X=x_1\}}\right] + \ldots + x_k \mathbf{E}\left[\mathbf{1}_{\{X=x_k\}}\right].$$

Applying equation (11.4) to each term on the right side of this equation gives

$$\mathbf{E}[X] = x_1 \mathbf{P}(\{X = x_1\}) + \ldots + x_k \mathbf{P}(\{X = x_k\}).$$

This is equation (10.1).

Solution (Exercise 11.6). For any outcome ω, we must show that

$$Z(\omega) = v_1 \mathbf{1}_{D_1}(\omega) + \ldots + v_n \mathbf{1}_{D_n}(\omega). \tag{11.31}$$

Suppose that $\omega \in D_j$ (it has to be somewhere).

On the right side of equation (11.31), $\mathbf{1}_{D_i}(\omega) = 0$ unless $i = j$. In that case, $\mathbf{1}_{D_i}(\omega) = 1$.

Thus the only surviving term on the right side of the equation is $v_j \cdot 1 = v_j$. This equals the left side, so we are done.

Solution (Exercise 11.7).

Method 1 The argument here is similar to the solution to Exercise 11.6.

One checks directly that equation (11.27) holds by verifying that it holds for every ω.

It holds when $\omega \in D_i$ for some i, using disjointness, and it holds when ω is not a member of any D_i, since both sides of the equation are zero in that case.

Method 2 When $n = 1$, the equation is obvious.

When $n = 2$, the equation is equivalent to equation (11.9).

To obtain the equation for general n, use The Old Induction Trick for generalizing from 2 to n, (Exercise 2.23).

Solution (Exercise 11.8). By equation (11.22),

$$\mathbf{E}\left[X \middle| D_i\right] = \frac{\mathbf{E}\left[\mathbf{1}_{D_i} X\right]}{\mathbf{P}(D_i)} \text{ for each } i.$$

That is,

$$\mathbf{P}(D_i)\mathbf{E}\left[X \middle| D_i\right] = \mathbf{E}\left[\mathbf{1}_{D_i} X\right] \text{ for each } i.$$

Thus equation (11.25) is exactly the statement that

$$\mathbf{E}\left[X\mathbf{1}_D\right] = \sum_{i=1}^{k} \mathbf{E}\left[X\mathbf{1}_{D_i}\right].$$

By linearity of expectation, this will be true if

$$X\mathbf{1}_D = \sum_{i=1}^{k} X\mathbf{1}_{D_i}.$$

And this last equation holds, since equation (11.27) says that

$$\mathbf{1}_D = \sum_{i=1}^{k} \mathbf{1}_{D_i}.$$

Solution (Exercise 11.9). Let D_k be the event that the result of rolling the die is the number k.

Let X be the number of successes when tossing the coin. By equation (11.26),

$$\mathbf{E}\left[X\right] = \sum_{k=1}^{6} \mathbf{P}(D_k)\mathbf{E}\left[X \middle| D_k\right].$$

Since the die is fair, $\mathbf{P}(D_k) = 1/6$ for all k.

To find $\mathbf{E}\left[X \middle| D_k\right]$, think of a simple little self-contained experiment, namely tossing a coin k times. We know from previous work that the expectation is kp, where p is the success probability of the coin. Thus

$$\mathbf{E}\left[X \middle| D_k\right] = k\left(\frac{1}{5}\right).$$

Hence

$$\mathbf{E}\left[X\right] = \frac{1}{6}\frac{1}{5} + \frac{1}{6}\frac{2}{5} + \frac{1}{6}\frac{3}{5} + \frac{1}{6}\frac{4}{5} + \frac{1}{6}\frac{5}{5} + \frac{1}{6}\frac{6}{5} = \frac{1+2+3+4+5+6}{30} = \frac{21}{30} = \frac{7}{10}.$$

Chapter 12

Independent random variables, first applications

12.1 Two independent random variables

Definition 12.1 (Independence for physical random variables). Let X and Y be physical random variables defined for the same experiment. We say that X and Y are independent if every event defined in terms of the values of X is independent of every event defined in terms of the values of Y.

As usual we can express an independence statement in terms of information. For independent random variables, information about the observed value of X tells us nothing about the observed value of Y.

Definition 12.1 is a statement about physical random variables, not mathematical random variables. Here is a definition of independence for mathematical random variables.

Definition 12.2 (Independence for mathematical random variables). Let X and Y be real-valued random variables for some probability model. We say that X and Y are independent if, for any subsets S and T of $\mathbb{R}$, the events $\{X \in S\}$ and $\{Y \in T\}$ are independent.

Definition 12.2 applies to all mathematical random variables, not just those with finite range. Probability theory uses lots of mathematical random variables with infinite range, but in this chapter we will focus on the finite range case.

In the special case of mathematical random variables with finite range, the next lemma tells us that we can check independence by considering events of the form $\{X = x\}$ and $\{Y = y\}$. This is simpler than using Definition 12.2.

Lemma 12.3 (Checking independence for finite-range random variables). Let X and Y be finite range random variables. Then the following statements are equivalent.

(i) X and Y are independent random variables.

(ii) For every number x in the range of X, and every number y in the range of Y,

$$\{X = x\}, \{Y = y\} \text{ are independent events.} \tag{12.1}$$

Proof. **(i) $\implies$ (ii):** Assume that X and Y are independent random variables. Let $S = \{x\}$ and let $T = \{y\}$.

Since $\{X = x\} = \{X \in S\}$ and $\{Y = y\} = \{Y \in T\}$, Definition 12.2 gives equation (12.1).

(ii) $\implies$ (i): Assume condition (ii) holds.

Let S and T be any sets of real numbers. Let $c_1, \ldots, c_k$ list the distinct numbers in the range of X which are members of S. Let $d_1, \ldots, d_\ell$ list the distinct numbers in the range of Y which are members of T.

The statement that $X \in S$ is exactly the statement that one of the events $\{X = c_i\}$ has occurred. The statement that $Y \in T$ is exactly the statement that one of the events $\{Y = d_j\}$ has occurred.

By condition (ii), knowing that an event $\{X = c_i\}$ has occurred does not change our opinion about any event $\{Y = d_j\}$. This suggests that knowing that $\{X \in S\}$ has occurred should not change our opinion about $\{Y \in T\}$. Thus we expect that $\{X \in S\}$ and $\{Y \in T\}$ should be independent.

To argue more formally, we note that for any sample point ω, if $X(\omega) \in S$ then $X(\omega)$ must be equal to some c_i and if $Y(\omega) \in T$ then $Y(\omega)$ must be

equal to some d_j. Thus

$$\{X \in S\} = \bigcup_{i=1}^{k} \{X = c_i\}\,,\ \{Y \in T\} = \bigcup_{j=1}^{\ell} \{Y = d_j\}\,. \tag{12.2}$$

Similarly, if $\omega \in \{X \in S\} \cap \{Y \in T\}$, then $X(\omega) = x_i$ for some i *and* $Y(\omega) = y_j$ for some j. Thus

$$\{X \in S\} \cap \{Y \in T\} = \bigcup_{ij} \{X = x_i\} \cap \{Y = y_j\}\,.$$

Hence

$$\begin{aligned}\mathbf{P}(\{X \in S\} \cap \{Y \in T\}) &= \sum_{ij} \mathbf{P}(\{X = x_i\} \cap \{Y = y_j\}) \\ &= \sum_{ij} \mathbf{P}(X = x_i)\mathbf{P}(Y = y_j).\end{aligned}$$

Using the distributive law,

$$\begin{aligned}\sum_{ij} \mathbf{P}(X = x_i)\mathbf{P}(Y = y_j) &= \left(\sum_{i=1}^{k} \mathbf{P}\left(X = c_i\right)\right)\left(\sum_{j=1}^{\ell} \mathbf{P}\left(Y = d_j\right)\right) \\ &= \mathbf{P}(X \in S)\mathbf{P}(Y \in T).\end{aligned}$$

Thus we have shown that

$$\mathbf{P}(\{X \in S\} \cap \{Y \in T\}) = \mathbf{P}(X \in S)\mathbf{P}(Y \in T).$$

We have shown that for any subsets S and T of $\mathbb{R}$, the events $\{X \in S\}$ and $\{Y \in T\}$ are independent. Then Definition 12.2 tells us that condition (i) holds.

□

Statement (i) of Lemma 12.3 and statement (ii) of Lemma 12.3 are logically equivalent, for finite-range random variables. However: statement (i) of Lemma 12.3 is convenient for applying independence to a physical situation, while statement (ii) is convenient for showing that independence holds.

Example 12.4. Consider a two-step experiment, in which a fair coin is tossed twice.

Let $X_i = 1$ if toss i gives success, and $X_i = 0$ otherwise.

In this experiment, let H_1 be the event that the first toss results in a head, and let H_2 be the event that the second toss results in a head. Then $\mathbf{P}(H_1) = 1/2$, $\mathbf{P}(H_2) = 1/2$ and, as noted in Example 5.4, H_1, H_2 are independent. Furthermore, it is also noted in Example 5.4 that H_1, H_2^c, $H_1^c H_2$, and H_1^c, H_2^c are independent pairs.

Since $\{X_i = 1\} = H_i$ and $\{X_i = 0\} = H_i^c$, we conclude using Lemma 12.3, that X_1 and X_2 are independent!

Remark 12.5 (The naturalness of random variables). In Example 12.4, notice how the single statement that X_1, X_2 are independent random variables immediately conveys four statements about independent events: H_1, H_2 are independent, H_1, H_2^c are independent, H_1^c, H_2 are independent, and H_1^c, H_2^c are independent.

Example 12.6. Let's think about another random variable connected with the coin-tossing experiment in Example 12.4: let $X_3 = X_1 X_2$.

We will show that X_1, X_3 are *not* independent. One way to do that is to think about information. Suppose someone tells you that $X_1 = 0$. Do you know the value of X_3? Heck yes! X_3 is zero! So $\mathbf{P}(X_3 = 0 \,|\, X_1 = 0) = 1$.

In contrast to the case that $X_1 = 0$, if someone tells you that $X_1 = 1$, then the value of X_3 is just the value of X_2. Since X_1, X_2 are independent, there are still two possible values for X_3 in this case, and indeed

$$\mathbf{P}(X_3 = 0 \,|\, X_1 = 1) = \mathbf{P}(X_2 = 0 \,|\, X_1 = 1) = \frac{1}{2}. \tag{12.3}$$

Thus

$$\mathbf{P}(X_3 = 0 \,|\, X_1 = 0) \neq \mathbf{P}(X_3 = 0 \,|\, X_1 = 1). \tag{12.4}$$

So knowledge about the value of X_1 can definitely change our opinion about the value of X_3, and thus X_1, X_2 are not independent random variables.

More formally, since equation (12.4) holds, part (ii) of Exercise 5.10 tells us that $\{X_1 = 0\}, \{X_3 = 0\}$ are not independent events. The definition of

independence for random variables then tells us that X_1, X_3 are not independent.

Exercise 12.1. In the setting of Example 12.4, let Y_i represent a "payoff" connected with this experiment. The rule is that $Y_i = 5$ if toss i gives success, and $Y_i = -5$ otherwise.

Y_1 and Y_2 are independent random variables, for the same reason that that X_1 and X_2 in Example 12.4 are independent random variables.

Let $Y_3 = Y_1 Y_2$.

Prove that Y_1, Y_3 are independent.

[Solution]

We are concentrating on finite-range random variables at the moment. But for future reference, here's a criterion that saves work when checking independence for general random variables.

Lemma 12.7 (Intervals are sufficient). Real-valued random variables X, Y are independent if for all intervals $[a, b], [c, d]$, the events $\{X \in [a, b]\}$ and $\{Y \in [c, d]\}$ are independent.

The proof depends on technicalities and is omitted.

12.2 Independent indicators

Lemma 12.8 (Sets are independent if and only if their indicators are independent). Let A, B be events in some model. Then the indicator functions $\mathbf{1}_A, \mathbf{1}_B$ independent are independent if and only if A, B are independent.

Nothing surprising here if you think about information. Knowing whether or not A occurred is exactly the same as knowing the value of $\mathbf{1}_A$, and knowing

whether or not B occurred is exactly the same as knowing the value of $\mathbf{1}_B$. The proof is just a matter of checking that the definitions mean what you think they mean.

Proof. Let $\mathbf{1}_A, \mathbf{1}_B$ are independent random variables.

Then for any number x in the range of X, and any number y in the range of Y, $\{X = x\}$ and $\{Y = y\}$ are independent.

In particular, $\{X = 1\}, \{Y = 1\}$ are independent. That is, A, B are independent.

Conversely, suppose that A, B are independent.

By Lemma 5.6, the independence of A, B implies the following facts:

- A, B are independent,
- A, B^c are independent,
- A^c, B are independent, and
- A^c, B^c are independent.

These statements say that:

- $\{\mathbf{1}_A = 1\}, \{\mathbf{1}_B = 1\}$ are independent,
- $\{\mathbf{1}_A = 1\}, \{\mathbf{1}_B = 0\}$ are independent,
- $\{\mathbf{1}_A = 0\}, \{\mathbf{1}_B = 1\}$ are independent, and
- $\{\mathbf{1}_A = 0\}, \{\mathbf{1}_B = 0\}$ are independent.

We have shown that for every x in the range of $\mathbf{1}_A$, and every y in the range of $\mathbf{1}_B$, $\{\mathbf{1}_A = x\}, \{\mathbf{1}_B = y\}$ are independent.

Thus by Lemma 12.3, $\mathbf{1}_A, \mathbf{1}_B$ are independent.

□

12.3 Functions of independents

Let's review notations from calculus:

Definition 12.9 (Compositions of functions). Let f and g be any functions. Suppose that for any point t in the domain of g, $g(t)$ is in the domain of f. Then $f(g(t))$ is defined, and we can define a new function h by $h(t) = f(g(t))$. We will often denote this function simply as $f(g)$. This notation is only a shorthand for the function which sends t to $f(g(t))$, but it seems to convey the meaning clearly.

People sometimes refer to the function $f(g)$ using words, as "the composition of f *with* g". However, it's safer to write the composition symbolically, since someone might interpret the same phrase as meaning $g(f)$.

Another notation for the composition of functions is $f \circ g$. Thus $f \circ g$ and $f(g)$ mean the same thing, and

$$f \circ g(t) = f(g(t)). \tag{12.5}$$

We can also use this notation when more functions are involved. For example, $f \circ g \circ h$ is the function $f(g(h)))$.

Let X and Y be real-valued physical random variables which measure quantities for the same experiment. Suppose that X and Y are independent physical random variables. Let φ and θ be functions on $\mathbb{R}$. Since $\varphi(X)$ is determined by X, any information given by $\varphi(X)$ is also information about X. Similarly any information given by $\theta(Y)$ is also information about Y.

By assumption, information about X does not change your opinion concerning information about Y. So information about $\varphi(X)$ does not change your opinion concerning information about $\theta(Y)$. Hence $\varphi(X)$ and $\theta(Y)$ must be independent.

The mathematical version of this physical statement is expressed in much the same way. The proof is just a matter of using the definitions carefully, so it may not be a high priority for readers. The physical meaning of the lemma is important, of course.

Lemma 12.10 (Functions of independents are independent). Suppose X and Y are independent real-valued random variables for some probability model, and φ and θ are functions on $\mathbb{R}$. Then $\varphi(X)$ and $\theta(Y)$ are independent.

Proof. We will use Definition 12.2. Let X and Y be real-valued random variables for some probability model. Let S and T be subsets of $\mathbb{R}$. We must show that the events $\{\varphi(X) \in S\}$ and $\{\theta(Y) \in T\}$ are independent.

Let

$$G = \{z : \ z \in \mathbb{R}, \ \varphi(z) \in S\}. \tag{12.6}$$

To say that $\varphi(X(\omega)) \in S$ is logically equivalent to saying that $X(\omega) \in G$. Thus

$$\{\varphi(X) \in S\} = \{X \in G\}. \tag{12.7}$$

Similarly, let $H = \{z : \ z \in \mathbb{R}, \theta(z) \in T\}$. Then

$$\{\theta(Y) \in T\} = \{Y \in H\}. \tag{12.8}$$

Definition 12.2 tells us that $\{X \in G\}, \{Y \in H\}$ are independent events. Thus $\{\varphi(X) \in S\}, \{\theta(Y) \in T\}$ are independent events.

Since S, T were any subsets of $\mathbb{R}$, $\varphi(X), \theta(Y)$ are independent by Definition 12.2.

□

The following exercise is a simple test of Lemma 12.10.

Exercise 12.2. Let X, Y be finite range real-valued random variables, and suppose that X, Y are independent.

Let $G = 5X$ and let $H = 16Y$. Are G, H independent? Sure they are, it seems physically obvious.

But let's check that. We could appeal to Lemma 12.10, and that would work even if X and Y did not have finite range. But it seems instructive to use a more basic argument.

So: using condition (ii) of Lemma 12.3, and without using Lemma 12.10, show that G, H are independent.

[Solution]

12.4 Expectation of a product

The next theorem extends the multiplicative property from independent events to independent random variables.

Theorem 12.11 (Expectation of a product of independents). Let X and Y be independent random variables defined for the same probability model. Assume that $\mathbf{E}[X], \mathbf{E}[Y]$ exist. Then $\mathbf{E}[XY]$ exists, and

$$\mathbf{E}[XY] = \mathbf{E}[X]\ \mathbf{E}[Y]. \tag{12.9}$$

Proof. The theorem holds for general random variables, but we will only write down a proof for the finite-range case.

We can follow the pattern of the proof of Lemma 10.10.

Let $x_1, \ldots, x_n$ be the distinct numbers in the range of X, and let $y_1, \ldots, y_m$ be the distinct numbers in the range of Y.

Let $D_{ij} = \{X = x_i, Y = y_j\}$.

By Theorem (10.7),

$$\mathbf{E}[XY] = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i\, y_j\, \mathbf{P}(D_{ij}). \tag{12.10}$$

Up to this point we have not used the assumption that X, Y are independent. This tells us that

$$\mathbf{P}(X = x_i, Y = y_j) = \mathbf{P}(X = x_i)\mathbf{P}(Y = y_j).$$

Hence

$$\mathbf{E}[XY] = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i\, y_j \mathbf{P}(X = x_i)\mathbf{P}(Y = y_j). \tag{12.11}$$

Using the distributive law, we see that

$$\mathbf{E}[XY] = \left(\sum_{i=1}^{n} x_i \mathbf{P}(X = x_i)\right)\left(\sum_{j=1}^{m} y_j \mathbf{P}(Y = y_j)\right). \tag{12.12}$$

Thus $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$.

□

Exercise 12.3. Let X and Y be independent random variables with finite range, such that $\mathbf{E}[X^3]$ exists and $\mathbf{E}[Y^8]$ exists. Prove that $\mathbf{E}[X^3Y^8] = \mathbf{E}[X^3]\,\mathbf{E}[Y^8]$.

[Solution]

Independence of random variables is a powerful tool in analyzing the behavior of probability models.

12.5 Independence for a sequence of random variables

Just as in the case of independence for events, we can consider a (possibly long) sequence of random variables defined on the sample space of some experiment. Here's the general definition.

Definition 12.12 (Independent sequences of random variables). Let $X_1, \ldots, X_n$ be real-valued random variables which are defined on the same sample space. The random variables $X_1, \ldots, X_n$ are said to be independent if the following holds.

For any subsets $D_1, \ldots, S_n$ of $\mathbb{R}$,

$$\mathbf{P}(X_1 \in S_1, \ldots, X_n \in S_n) = \mathbf{P}(X_1 \in S_1) \cdots \mathbf{P}(X_n \in S_n). \tag{12.13}$$

Take a moment to check that this definition agrees with Definition 12.2 when $n = 2$!

When the random variables happen to have finite range, things are simpler, much as in Lemma 12.3.

Lemma 12.13 (Checking independence for a sequence of finite-range random variables). Let $X_1, \ldots, X_n$ be finite range random variables. Then the following statements are equivalent.

(i) $X_1, \ldots, X_n$ are independent.

(ii) For every sequence of numbers $x_1, \ldots, x_n$, where x_i is in the range of X_i,

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbf{P}(X_1 = x_1) \cdots \mathbf{P}(X_n = x_n). \tag{12.14}$$

The properties of independent sequences of random variables are similar to the properties of two independent random variables. Physical experience continues to be a reliable guide, and we will cheerfully write down mathematical equations without proofs, based on our ideas about physical independence.

Exercise 12.4 (Maximum of independent). Let $X_1, \ldots, X_n$ be an independent sequence.

Suppose that each random variable X_j has a uniform distribution on $\{1, \ldots, 10\}$. That is, suppose $\mathbf{P}(X_j = i) = 1/10$ for $i = 1, \ldots, 10$.

Let M be the maximum of $X_1, \ldots, X_n$. Find $\mathbf{P}(M \leq 4)$.

[Solution]

Exercise 12.5 (Minimum of independent). In the setting of Exercise 12.4, let m be the minimum of $X_1, \ldots, X_n$. Find $\mathbf{P}(m > 4)$.

[Solution]

We could have fun proving independence properties based on the definition. For example,

$$A_1, \ldots, A_n \text{ are independent} \iff \mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n} \text{are independent.} \quad (12.15)$$

However, it seems better to just assume facts like that, and keep going.

12.6 Random walk

Sequences of independent random variables occur in many situations, in areas such as economics, physics and biology.

In this section we present a simple example.

A bug is moving around randomly on the integers. The bug is sitting on the origin initially.

The movement is as follows.

Every second, a fair coin is tossed. The first toss takes place at time 1, and every second thereafter. Each toss results in a "step" by the bug, as follows:

- If the bug is on integer k, a successful toss (a head) makes the bug jump instantly to $k+1$, and
- if the bug is on integer k, a failure makes the bug jump instantly to $k-1$.

This type of mathematical motion is called "random walk", or more specifically, "simple symmetric random walk". (The word "simple" refers to the fact that the bug can only jump a distance of one unit. The word "symmetric" is used because the bug does not favor right or left.)

Since the bug changes direction frequently, it is moving *very inefficiently.* A basic question: how far from the origin is the bug likely to be, after n steps?

We can start by thinking about a more abstract model for the bug's motion. Let X_i be a random variable that represents the result of the coin toss at time i. We will represent success by 1 and failure by -1, so that

$$\mathbf{P}(X_i = 1) = \frac{1}{2} \text{ and } \mathbf{P}(X_i = -1) = \frac{1}{2}.$$

Our physical picture tells us that in the mathematical model, $X_1, \ldots, X_n$ is an independent sequence of random variables.

Define $S_0 = 0$, and let

$$S_n = X_1 + \ldots + X_n \text{ for each } n = 1, 2, \ldots. \tag{12.16}$$

S_0 is the location of the bug at time 0. At time 1, the bug has just taken a single step, and is now at $S_1 = X_1$. At time 2, the bug has taken its second step, and is now at $S_1 = X_1 + X_2$. And so on.

There are powerful techniques for analyzing random walk. In the present chapter we will just consider one fact, which is a consequence of Theorem 12.11:

Lemma 12.14 (Random walk squared distance).

$$\mathbf{E}\left[S_n^2\right] = n. \tag{12.17}$$

Proof. Note first that, from the distribution, $\mathbf{E}[X_i] = 0$ for every step. So (by linearity) $\mathbf{E}[S_n] = 0$. That doesn't help us much.

Notice that the range of S_n contains quite a few points. It's easy to see that the largest possible value in the range is n (when every step is success), and the smallest possible value is $-n$ (when every step is failure).

A little more thought will convince you that when n is even, the range consists of all even numbers from $-n$ to n, and when n is odd, the range consists of all odd numbers from $-n$ to n.

The bottom line is that it won't be easy to find $\mathbf{E}[S_n^2]$ directly from the definition of expected value.

However we can expand S_n^2.

$$\mathbf{E}\left[S_n^2\right] = \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j\right)\right].$$

Thus

$$\mathbf{E}\left[S_n^2\right] = \mathbf{E}\left[\sum_{i,j=1}^{n} X_i X_j\right]. \tag{12.18}$$

Equation (12.18) is the usual distributive law manipulation. If it looks strange, please write out the case $n = 3$ to see what is going on!

Additivity of expectation then gives

$$\mathbf{E}\left[S_n^2\right] = \sum_{i,j=1}^{n} \mathbf{E}\left[X_i X_j\right].$$

Now comes the key point. For $i \neq j$, X_i, X_j are independent random variables. In that case Theorem 12.11 tells us that

$$\mathbf{E}[X_i X_j] = \mathbf{E}[X_i]\,\mathbf{E}[X_j] = 0.$$

Thus, keeping only the surviving terms on the right, we have:

$$\mathbf{E}\left[S_n^2\right] = \sum_{i=1}^{n} \mathbf{E}\left[X_i^2\right].$$

Much simpler! The values of X_i are -1 or 1. So $X_i^2 = 1$, always. Hence $\mathbf{E}[X_i^2] = 1$, and we have $\mathbf{E}[S_n^2] = n$, as claimed.

□

Let's pause to admire equation (12.17) for a moment. The largest possible value for S_n^2 is n^2. When n is large, the average value of S_n^2 is *much* smaller than n^2. So equation (12.17) is telling us that the distribution of S_n does not put much weight near the extreme values of S_n.

We would like to make that last statement more precise.

12.7 The Markov Inequality

To extract a little more information from equation (12.17), one can use an inequality (you will do that in Exercise 12.6). The inequality you will use is known as the Markov inequality. Despite its simplicity, the Markov inequality is useful in many situations.

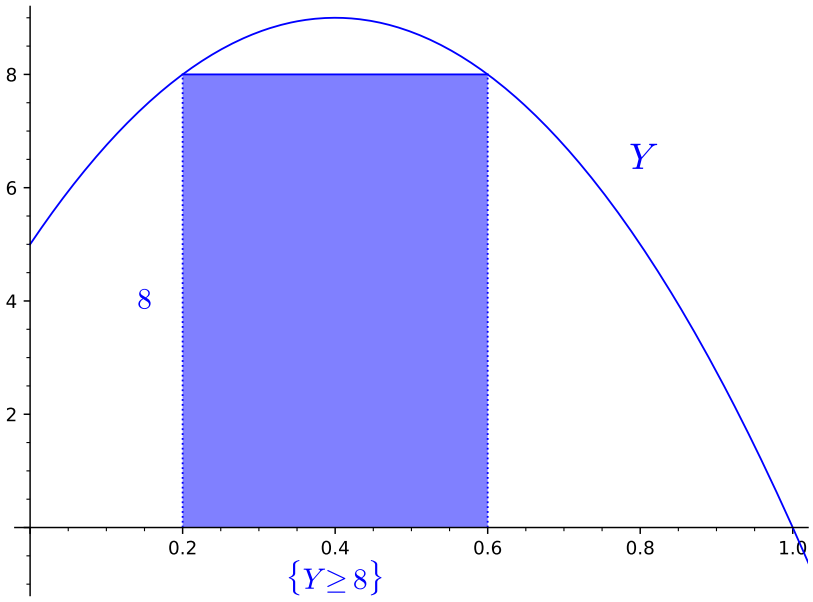


Figure 12.1: Sample space $[0, 1]$, uniform probability

Lemma 12.15 (The Markov Inequality). Let Y be a nonnegative random variable such that $\mathbf{E}[Y]$ exists. For any number α,

$$\alpha \mathbf{P}(Y \geq \alpha) \leq \mathbf{E}[Y]. \tag{12.19}$$

See Figure 12.1

This lemma applies to general random variables. (If Y has finite range then of course the assumption that $\mathbf{E}[Y]$ exists is automatically true.)

The proof of the lemma given here is also general, since it only uses properties of expectation which always hold.

Proof. Let $A = \{Y \geq \alpha\}$. We claim that

$$\alpha \mathbf{1}_A \leq Y \tag{12.20}$$

holds everywhere on the sample space.

Indeed, consider a sample point ω.

If $\omega \in A$, then by the definition of A we must have $\alpha \leq Y(\omega)$. Since $\mathbf{1}_A(\omega) = 1$, that is exactly what equation (12.20) asserts.

On the other hand, if $\omega \notin A$, then $\mathbf{1}_A(\omega) = 0$, and $0 \leq Y(\omega)$ holds since Y is assumed to be nonnegative.

Thus equation (12.20) holds everywhere.

Since expectation is monotone,

$$\mathbf{E}[\alpha \mathbf{1}_A] \leq \mathbf{E}[Y].$$

And

$$\mathbf{E}[\alpha \mathbf{1}_A] = \alpha \mathbf{E}[\mathbf{1}_A] = \alpha \mathbf{P}(A),$$

using linearity for the first equality and equation (11.4) for the second equality.

□

Figure 12.1 illustrates the Markov inequality when the sample space is $[0, 1]$ and $\mathbf{P}$ is the uniform distribution on $[0, 1]$. We consider a continuous sample space because it's easier to draw the graph of a random variable in that situation. In Figure 12.1, $\alpha = 8$, and the blue area is $\alpha \mathbf{P}(Y \geq \alpha)$. As in Example 10.20, $\mathbf{E}[Y] = \int_0^1 Y(u)\, du$, and so we can picture the expected value easily in this case: $\mathbf{E}[Y]$ is equal to the entire area under the curve.

Exercise 12.6 (Searching for Charlie). Your pet bug escaped from the origin at time 0, and has undoubtedly been performing simple symmetric random walk ever since then. $10,000$ seconds have elapsed since Charlie started

roaming. You are distressed and searching frantically. Use the Markov inequality and equation (12.17) to estimate the probability that Charlie at least 500 units away from the origin.

[Solution]

Exercise 12.7. Let X be a random variable such that $\mathbf{E}\left[e^X\right] = 5$, and let $\beta > 0$ be a number.

Find an upper bound estimate for $\mathbf{P}(X \geq \beta)$. Your estimate should be in the form:

$$\mathbf{P}(X \geq \beta) \leq \text{ something.}$$

[Solution]

Exercise 12.8. Is Lemma 12.15 an interesting statement when $\alpha \leq 0$?

[Solution]

Exercise 12.9. Let X be a nonnegative random variable such that $\mathbf{E}[X]$ exists, and let α be any number.

Suppose someone asks you to find an upper bound for $\mathbf{P}(X > \alpha)$.

The Markov inequality gives you an upper bound for $\mathbf{P}(X \geq \alpha)$, not $\mathbf{P}(X > \alpha)$. Can you use the Markov inequality to get what you need?

[Solution]

Exercise 12.10 (Using $\mathbf{E}[f(X)]$). Let X be a random variable. Suppose that f is a nonnegative function, which is strictly increasing on the range of X.

You are given that $\mathbf{E}[f(X)]$ exists, and $\mathbf{E}[f(X)] = c$, for some number c.

Let $\beta > 0$ be a number. Find an upper bound estimate for $\mathbf{P}(X \geq \beta)$, in terms of β, f, and c.

[Solution]

Remark 12.16 (The "infinite expectation" convention). Let X be a general nonnegative random variable. When $\mathbf{E}[X]$ does not exist we often say that $\mathbf{E}[X] = \infty$. Since ∞ is not a number, this is only a convention, but it can be convenient. Using this convention, we can assert that the Markov inequality (equation (12.19)) holds for *every* nonnegative random variable Y, whether or not $\mathbf{E}[Y]$ exists.

12.8 The effect of independent steps

Let's take a moment to consider the idea behind equation (12.17). There we considered independent steps X_i, where $\mathbf{P}(X_i = 1) = \mathbf{P}(X_i = -1) = 1/2$. S_n is the position after n steps, so $S_n = X_1 + \ldots + X_n$.

The maximum possible value for S_n^2 is n^2, but we showed in equation (12.17) that $\mathbf{E}[S_n^2]$ is only equal to n. The general explanation for the smallness of $\mathbf{E}[S_n^2]$ is cancellation, since negative steps happen approximately as often as positive steps. Calculating $\mathbf{E}[S_n^2]$ gave us a measure of how much cancellation takes place. But it should be emphasized that $\mathbf{E}[S_n^2]$ is only one number. In Chapter 18 we will discuss the Central Limit Theorem (Theorem 18.14), which provides much more detailed information about the distribution of S_n after n independent steps have taken place.

12.9 Solutions for Chapter 12

Solution (Exercise 12.1).

There is more than one way to write down a solution. Exercise 5.10 seemed to match our thinking in Example 12.4 so we'll go with that here.

Notice that the range of Y_1 is the set $\{5, -5\}$ and the range of Y_3 is the set $\{25, -25\}$.

If $Y_1 = 5$, then $Y_3 = 5Y_2$. Hence

$$\mathbf{P}(Y_3 = 25 \,|\, Y_1 = 5) = \mathbf{P}(Y_2 = 5 \,|\, Y_1 = 5) = \frac{1}{2}.$$

Similarly,

$$\mathbf{P}(Y_3 = 25 \,|\, Y_1 = -5) = \mathbf{P}(Y_2 = -5 \,|\, Y_1 = -5) = \frac{1}{2}.$$

Since $\{Y_3 = 25\}^c = \{Y_3 = -25\}$, part (i) of Exercise 5.10 tells us that $\{Y_1 = 5\}, \{Y_3 = 25\}$ are independent.

We could repeat this argument for other cases, and conclude that we also have that $\{Y_1 = 5\}, \{Y_3 = -25\}$ are independent, $\{Y_1 = -5\}, \{Y_3 = 25\}$ are independent, and $\{Y_1 = -5\}, \{Y_3 = -25\}$ are independent.

Or we could appeal to Lemma 5.6 to obtain the same facts.

At any rate, we now know that for each of the events $\{Y_1 = 5\}, \{Y_1 = -5\}$ is independent of each of the sets $\{Y_3 = 25\}, \{Y_3 = -25\}$.

Hence, by Lemma 12.3, Y_1, Y_3 are independent.

Solution (Exercise 12.2). Let $x_1, \ldots, x_k$ be the distinct numbers in the range of X, and let $y_1, \ldots, y_\ell$ be the distinct numbers in the range of Y.

Then $5x_1, \ldots, 5x_k$ are the distinct numbers in the range of G, and $16y_1, \ldots, 16y_\ell$ are the distinct numbers in the range of H.

For any i, j,

$$\{G = 5x_i\} = \{X = x_i\}, \text{ and } \{H = 16y_j\} = \{Y = y_j\}.$$

Since X, Y are assumed to be independent, $\{X = x_i\}, \{Y = y_j\}$ are independent events. That is, $\{G = 5x_i\}, \{H = 16y_j\}$ are independent events. Since this is true for every value $5x_i$ of G and every value $15y_j$ of H, by Lemma 12.3 G, H are independent random variables.

Solution (Exercise 12.3). This exercise is just checking that you noticed Lemma 12.10.

By that lemma, X^3, Y^8 are independent random variables. Then we can use Theorem 12.11.

Solution (Exercise 12.4). Since M is the maximum, to say that $M \leq 4$ is the same as saying that $X_i \leq 4$ holds for every $i = 1, \ldots, n$. Thus

$$\{M \leq 4\} = \{X_1 \leq 4\} \cap \ldots \cap \{X_n \leq 4\}.$$

Since the events $\{X_1 \leq 4\}, \ldots, \{X_n \leq 4\}$ are independent,

$$\mathbf{P}(M \leq 4) = \mathbf{P}(X_1 \leq 4) \cdots \mathbf{P}(X_n \leq 4) = \frac{4}{10} \cdots \frac{4}{10} = \left(\frac{4}{10}\right)^n.$$

Solution (Exercise 12.5). Since m is the minimum, to say that $m > 4$ is the same as saying that $X_i > 4$ holds for every $i = 1, \ldots, n$. Thus

$$\{m > 4\} = \{X_1 > 4\} \cap \ldots \cap \{X_n > 4\}.$$

Since the events $\{X_1 > 4\}, \ldots, \{X_n > 4\}$ are independent,

$$\mathbf{P}(M > 4) = \mathbf{P}(X_1 > 4) \cdots \mathbf{P}(X_n > 4) = \frac{6}{10} \cdots \frac{6}{10} = \left(\frac{6}{10}\right)^n.$$

Solution (Exercise 12.6). Let S_n be the random variable defined by equation (12.16). Thus S_n is Charlie's location after n steps.

Let $n = 10000$.

We would like to estimate $\mathbf{P}(|S_n| \geq 500)$.

That is the same as $\mathbf{P}(S_n^2 \geq 250000)$.

Using the Markov Inequality,

$$250000\mathbf{P}(S_n^2 \geq 250000) \leq \mathbf{E}\left[S_n^2\right] = 10000.$$

Hence

$$\mathbf{P}(|S_n| \geq 500) \leq \frac{10000}{250000} = 1/25.$$

(With more work, one can get a much sharper estimate. But this inequality is already interesting.)

Solution (Exercise 12.7). We are told that $\mathbf{E}\left[e^X\right] = 5$, so we know how to apply the Markov Inequality to e^X.

What does $\{X \geq \beta\}$ look like in terms of e^X?

The exponential function is a strictly increasing function, isn't it? So the statement that $X \geq \beta$ is exactly equivalent to the statement that $e^X \geq e^\beta$.

So $\{X \geq \beta\} = \left\{e^X \geq e^\beta\right\}$. And, using the Markov Inequality (with Y replaced by e^X and α replaced by e^β),

$$e^\beta \mathbf{P}(e^X \geq e^\beta) \leq \mathbf{E}\left[e^X\right] = 5.$$

Thus

$$\mathbf{P}(X \geq \beta) \leq \frac{5}{e^\beta}.$$

That finishes the problem.

Solution (Exercise 12.8). Since $\mathbf{P}(Y \geq \alpha) \geq 0$, for $\alpha \leq 0$ we always have $\alpha\mathbf{P}(Y \geq \alpha) \leq 0$.

Since Y is assumed to be nonnegative, $\mathbf{E}[Y] \geq 0$.

So the statement that $\alpha\mathbf{P}(Y \geq \alpha) \leq \mathbf{E}[Y]$ is rather obvious.

Solution (Exercise 12.9). If $X > \alpha$ then certainly $X \geq \alpha$. In other words, the statement "$X(\omega) > \alpha$" is a stronger statement than "$X(\omega) \geq \alpha$". Hence $\{X > \alpha\} \subset \{X \geq \alpha\}$.

So we always have $\mathbf{P}(X > \alpha) \leq \mathbf{P}(X \geq \alpha)$. And so the upper bound estimate for $\mathbf{P}(X \geq \alpha)$ is already an upper bound estimate for $\mathbf{P}(X > \alpha)$.

Solution (Exercise 12.10). Applying the Markov inequality to $f(X)$, we have

$$\alpha \mathbf{P}(f(X) \geq \alpha) \leq \mathbf{E}\,[f(X)] = c. \tag{12.21}$$

Since f is strictly increasing on the range of X, to say that $X(\omega) \geq \beta$ is exactly equivalent to saying that $f(X(\omega)) \geq f(\beta)$. Hence

$$\mathbf{P}(X \geq \beta) = \mathbf{P}(\, f(X) \geq f(\beta)\,).$$

Let $\alpha = f(\beta)$. Then equation (12.21) says that

$$f(\beta)\mathbf{P}(X \geq \beta) \leq c.$$

Whenever $f(\beta) > 0$, we can write this as

$$\mathbf{P}(X \geq \beta) \leq \frac{1}{f(\beta)} c.$$

Chapter 13

Waiting times

13.1 Waiting for the first head, with a deadline

We have mainly worked with mathematical random variables that have finite range. Now we are going to broaden our view. The present section may suggest why this is desirable.

Consider tossing a coin n times. Let p be the success probability for the coin, where as usual by success for a toss we mean that the toss results in a head. Let $q = 1 - p$.

Let's study the random variable T_n, which we define as the time of the first success in the sequence of n tosses, if success ever occurs. Otherwise let $T_n = n$.

We might imagine that n is our deadline, and we shut down the experiment at time n if there has been no success by that time.

Our first goal is to write down the distribution of T_n.

Let A_i be the event that toss i gives success. Since the results of the tosses are assumed to be physically independent, the events $A_1, \ldots, A_n$ are mathematically independent in the sense of Definition 7.7.

By assumption, $\mathbf{P}(A_i) = p$. For any $k = 1, \ldots, n-1$, $T_n > k$ means that no success occurred on tosses 1 through k. Thus

$$\{T_n > k\} = A_1^c \cap \ldots \cap A_k^c, \tag{13.1}$$

so

$$\mathbf{P}(T_n > k) = q^k. \tag{13.2}$$

Since $q^0 = 1$, the same equation holds for $k = 0$. (As usual, interpret 0^0 as 1 to include the case $p = 1$ and $k = 0$.)

That gives us the value of $\mathbf{P}(T_n > k)$. If you want $\mathbf{P}(T_n = k)$, note that for $1 \leq k < n$,

$$\{T_n = k\} = \{T_n > k-1\} - \{T_n > k\}.$$

Thus for $1 \leq k < n$,

$$\mathbf{P}(T_n = k) = q^{k-1} - q^k = q^{k-1}p. \tag{13.3}$$

Please check that we can obtain the same probability by noting that $\{T_n = k\} = A_1^c \cap \ldots \cap A_{k-1}^c \cap A_k$, and using independence!

We have found $\mathbf{P}(T_n = k)$ for $k < n$. To get $\mathbf{P}(T_n = n)$, we go back to the description of the experiment.

Rember that we shut down the experiment by time n, whether or not success has been achieved.

Thus $\{T_n = n\} = \{T_n > n-1\}$, so

$$\mathbf{P}(T_n = n) = q^{n-1}. \tag{13.4}$$

Combining these facts gives:

Lemma 13.1 (Distribution of T_n).

$$\mathbf{P}(T_n = k) = q^{k-1}p \text{ for } 1 \leq k < n;\ \mathbf{P}(T_n = n) = q^{n-1}. \tag{13.5}$$

$\mathbf{P}(T = k) = 0$ otherwise.

Exercise 13.1. As a check, please verify in some way that the values we have found in equation (13.5) for $\mathbf{P}(T_n = 1), \ldots, \mathbf{P}(T_n = n)$ actually add up to one.

[Solution]

Let's find $\mathbf{E}[T_n]$.

Using Definition 10.2,

$$\mathbf{E}[T_n] = \sum_{k=1}^{n-1} kq^{k-1}p + nq^{n-1}. \tag{13.6}$$

If $p = 0$, the formula gives $\mathbf{E}[T_n] = n$. We can check this value directly from the definition of the experiment. When $p = 0$ the probability that a head ever occurs is zero. By the definition of the experiment, if a head never occurs then $T_n = n$. Thus when $p = 0$, $\mathbf{P}(T_n = n) = 1$. Hence $\mathbf{E}[T_n] = n$ is correct.

From now on assume that $p > 0$.

We need two tricks to evaluate the sum in the formula for $\mathbf{E}[T_n]$.

First, recall how we find the sum of a **finite geometric series**. Let $s_n = 1 + q + q^2 + \ldots + q^n$, where for a moment we allow q to stand for any number. The first trick is to multiply by q.

This gives $q\, s_n = q + q^2 + \ldots + q^{n+1} = s_n - 1 + q^{n+1}$. Solving for s_n when $q \neq 1$ gives the familiar formula for the sum of a finite geometric series:

$$s_n = 1 + q + q^2 + \ldots + q^n = \frac{1 - q^{n+1}}{1 - q}. \tag{13.7}$$

The second trick is to differentiate the expressions in equation (13.7) with respect to q. This gives

$$0 + 1 + 2q + \ldots + nq^{n-1} = \frac{d}{dq}\frac{1 - q^{n+1}}{1 - q},$$

i.e.

$$\sum_{k=1}^{n} kq^{k-1} = \frac{1 - q^{n+1}}{(1 - q)^2} - \frac{(n + 1)q^n}{1 - q} = \frac{1 - q^{n+1}}{p^2} - \frac{(n + 1)q^n}{p}. \tag{13.8}$$

Replacing n by $n-1$ in equation (13.8), and substituting in equation (13.6),

$$\mathbf{E}[T_n] = \frac{1 - q^n}{p} - nq^{n-1} + nq^{n-1},$$

so

$$\mathbf{E}[T_n] = \frac{1 - q^n}{p}. \tag{13.9}$$

Exercise 13.2. Derive equation (13.9) in a different way, without using the differentiation trick.

First find $\mathbf{E}[T_{n+1}] - \mathbf{E}[T_n]$. Then show that

$$\mathbf{E}[T_n] = 1 + q + \ldots + q^{n-1}. \tag{13.10}$$

[Solution]

Since $q < 1$, $\lim_{n\to\infty} q^n = 0$. Hence for large n,

$$\mathbf{E}[T_n] \approx 1/p, \tag{13.11}$$

which is a tidier expression for $\mathbf{E}[T_n]$, although now it is only an approximation. We note that $1/p$ grows larger as p becomes smaller, which is completely reasonable, since it is harder to obtain a head when p is smaller, and thus it should take longer.

Equation (13.11) approximates one number by another. Can we think of this approximation as arising from a new model?

13.2 Time of first success in ∞ trials

The simplicity of equation (13.11) suggests that we might gain a bit of elegance by replacing a probability model with large n with a probability model in which $n = \infty$, that is, a model in which the coin tossing goes on forever. (By accepting a more complex concept we obtain a simpler calculation. Conceptual thinking is our human strength, so this seems like a good strategy in general.)

We will not try to use a sample space to build a rigorous mathematical model for infinitely many coin tosses. This is possible, and is routine in advanced courses, but it requires significant technicalities. Here we will simply use the rules of probability theory to calculate physically relevant numbers.

We are studying how long it takes to obtain the first head. For that purpose, thinking about an infinite number of tosses seems like a reasonable idealization. After all, in the physical picture there is not a natural limit on the number of tosses. And the time of first success does not depend at all on what happens in coin tosses *after* the first success.

Let T denote the time of the first head, in an infinite sequence of coin tosses, if a head is ever obtained. The mathematical random variable T is often simply referred to as the *waiting time* for the first success.

Of course, if a success is never obtained, we need a way to record that fact. So:

$$\text{By definition, if success is never obtained, } T = \infty. \tag{13.12}$$

Notice that if $p = 0$ we will never obtain a head, so the probability that $T = \infty$ is one, and there is really nothing else to say about this situation. From now on assume $p > 0$. Let $q = 1 - p$.

We showed in equation (13.3) that $\mathbf{P}(T_n > k) = q^k$, for $k = 0, \ldots, n-1$. The same argument shows that here we have

$$\mathbf{P}(T > k) = q^k \tag{13.13}$$

for every $k = 0, 1, \ldots$.

And $\mathbf{P}(T = k) = \mathbf{P}(T > k-1) - \mathbf{P}(T > k) = q^{k-1} - q^k$. Thus

Lemma 13.2 (Distribution of T).

$$\mathbf{P}(T = k) = q^{k-1}p \text{ for } 1 \leq k < \infty. \tag{13.14}$$

$\mathbf{P}(T = k) = 0$ otherwise.

Notice that $\{T = \infty\} \subset \{T > k\}$ for every k. Thus $\mathbf{P}(T = \infty) \leq q^k$ for every k. Since we are restricting attention now to the case that $q < 1$, $q^k \to 0$ as $k \to \infty$. So

$$\mathbf{P}(T = \infty) = 0 \text{ when } p > 0. \tag{13.15}$$

Definition 13.3 (The geometric distribution). The distribution of T given by equation (13.14) is usually called the *geometric distribution, with parameter p*.

The waiting time T has no direct physical meaning, since no real experiment goes on forever. A mathematical random variable with a direct physical interpretation is T_n, and T is one step further away from the physical world. However, part of the usefulness of the mathematical model for infinitely many tosses is that we can *almost* picture it. And so we can still be guided by reality as we use it.

We notice that $\mathbf{P}(T > k) \to 0$ rapidly ("exponentially fast" or "geometrically fast") as $k \to \infty$. This suggests that calculations using T should give us good approximations to the results of calculations with T_n.

Exercise 13.3 (The memoryless property of the geometric distribution). Someone is tossing a coin repeatedly, and waiting for the first success. The success probability is p, where $0 < p < 1$. Let T be the number of tosses

needed to obtain the first success. Let $\mathbf{P}$ be the distribution based on the knowledge that the tosser has, at the time when the sequence of tosses starts. Then $\mathbf{P}(T = n)$ is given by the geometric distribution with parameter p.

Now consider the viewpoint of a spectator who comes upon the tosser after n tosses have been made. The spectator learns that up to this time no success has been obtained.

The spectator decides to wait until the first success. Thus the spectator will wait for $T - n$ *additional* tosses. Based on the knowledge that the tosser (and the spectator) have at that moment, the probability that $T - n = m$ is given by

$$\mathbf{P}\left(T - n = m \,|\, T > n\right) = \mathbf{P}(T > n + m \,|\, T > n).$$

Calculate this conditional distribution for $T - n$, and show that

$$\mathbf{P}(T > n + m \,|\, T > n) = \mathbf{P}(T > m). \tag{13.16}$$

This is called the memoryless property of the geometric distribution. It shows that knowing how long you have already waited for success is not helpful in estimating the *additional time* that you will have to wait.

[Solution]

What about the expectation of T? We have not defined expected values for random variables which do not have finite range. But we can easily guess the right definition. Simply replace the usual sum by an infinite series. (See Definition 14.6 for a formal statement.)

Thus

$$\mathbf{E}\left[T\right] = \sum_{k=0}^{\infty} k\, \mathbf{P}(T = k). \tag{13.17}$$

The term with $k = 0$ contributes nothing, of course. So

$$\mathbf{E}\left[T\right] = \sum_{k=1}^{\infty} k\, \mathbf{P}(T = k) = \sum_{k=1}^{\infty} k q^{k-1} p. \tag{13.18}$$

By definition,

$$\sum_{k=1}^{\infty} k\, q^{k-1} p = \lim_{n\to\infty} \sum_{k=1}^{n} k q^{k-1} p.$$

Using equation (13.8)

$$\sum_{k=1}^{\infty} k\, q^{k-1} p = \lim_{n\to\infty} \left(\frac{1-q^{n+1}}{p} - (n+1)q^n \right).$$

We have assumed in this discussion that $p > 0$, so $q < 1$. Then $q^n \to 0$ more rapidly than $n \to \infty$, so we have both $(n+1)q^n \to 0$ and $q^{n+1} \to 0$. (One can use a calculus trick based on the Ratio Test to prove these statements.) Thus

$$\lim_{n\to\infty} \left(\frac{1-q^{n+1}}{p} - (n+1)q^n \right) = \frac{1}{p}.$$

We have shown:

Lemma 13.4 (Expectation of a waiting time). Let T have the geometric distribution, given in equation (13.14), with $p > 0$. Then

$$\mathbf{E}[T] = \frac{1}{p}. \tag{13.19}$$

Letting $n \to \infty$ in equation (13.9) gives

$$\lim_{n\to\infty} \mathbf{E}[T_n] = \mathbf{E}[T]. \tag{13.20}$$

This limiting agreement increases our confidence that T is a useful approximation to T_n.

Remark 13.5 (The geometric series). The **sum of a geometric series** is a standard calculus fact:

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \text{ for all } x \in (-1, 1). \tag{13.21}$$

Remember that x^0 means 1 in this series, even for $x = 0$. In other words, the series is $1 + x + x^2 + x^3 + \ldots$.

To recall why this equation (13.21) holds, remember first that the Ratio Test shows the series in equation (13.21) converges, for $|x| < 1$. Let s be the

sum of this series. The same manipulation used in equation (13.7) applies here:

$$xs = s - 1, \text{ and hence } s = \frac{1}{1-x}.$$

We'll use equation (13.21) in the next examples.

Example 13.6. Here's an alternative derivation of equation (13.19). First, note that

$$\mathbf{E}[T] = \sum_{k=1}^{\infty} kq^{k-1}p = p\left(1 + 2q + 3q^2 + \ldots\right). \tag{13.22}$$

Then:

$$\begin{aligned} 1 + 2q + 3q^2 + \ldots = {} & 1 + q + q^2 + q^3 + \ldots \\ & + 0 + q + q^2 + q^3 + \ldots \\ & + 0 + 0 + q^2 + q^3 + \ldots \\ & + 0 + 0 + 0 + q^3 + \ldots \\ & \vdots \end{aligned} \tag{13.23}$$

Adding the columns in equation (13.23) shows why the equation holds. This is not a rigorous argument, but it is convincing.

Hence

$$\begin{aligned} 1 + 2q + 3q^2 + \ldots = {} & \tfrac{1}{1-q} \\ & + \tfrac{q}{1-q} \\ & + \tfrac{q^2}{1-q} \\ & + \tfrac{q^3}{1-q} \\ & \vdots \end{aligned}.$$

Thus

$$\begin{aligned} 1 + 2q + 3q^2 + \ldots &= \left(1 + q + q^2 + q^3 + \ldots\right)\left(\frac{1}{1-q}\right), \\ &= \left(\frac{1}{1-q}\right)\left(\frac{1}{1-q}\right) \end{aligned}$$

By equation (13.22),

$$\mathbf{E}\,[T] = p\left(\frac{1}{1-q}\right)^2 = p\frac{1}{p^2} = \frac{1}{p}.$$

Exercise 13.4 (One more derivation). Suppose you remember the formula in equation (13.21) from your good old calculus days. You also remember from calculus that a convergent power series in x can be differentiated term by term inside its interval of convergence. The same derivation trick that gave us equation (13.7) can be applied directly, to show:

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2} \tag{13.24}$$

for $x \in (-1, 1)$.

Check this, please. Then use equation (13.24) to obtain equation (13.19). [Solution]

Remark 13.7 (A sample space describing an infinite sequence of coin tosses). Suppose we have a sample space which represents everything that can happen in in an infinite sequence of trials. In this situation each sample point represents a very large amount of information!

Perhaps a sample point would be an infinite sequence $(x_1, x_2, \ldots)$, where $x_i = 1$ if trial i gives success, and $x_i = 0$ otherwise. If the model represents infinitely many tosses of a fair coin, what is the probability of the one-point set containing a single sample point? Consider, say $A = \{\,(x_1, x_2, \ldots)\,\}$. Imitating the approach for finitely many tosses, we would say that

$$\mathbf{P}(A) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \ldots.$$

In this way we are soon forced to conclude that $\mathbf{P}(A) = 0$, for *every* sample point. Since the probability of the whole sample space must be equal to one, a rigorous definition of the probability distribution on this sample space is going to require some technicalities. And that is why we will not take time to provide such a definition.

13.3 Solutions for Chapter 13

Solution (Exercise 13.1). Let v_k^n be the value calculated in equation (13.5) for $\mathbf{P}(T_n = k)$.

Then $v_k^n = q^{k-1}p$ for $k = 1, \ldots, n-1$, $v_n^n = q^{n-1}$ and $v_k^n = 0$ for all other k.

In this problem we are asked to show that

$$v_1^n + \ldots + v_n^n = 1. \tag{13.25}$$

Looking at the formula for v_k^n, or perhaps looking back at equation (13.3), we notice that for $k = 1, \ldots, n-1$ we have

$$v_k^n = q^{k-1} - q^k. \tag{13.26}$$

So

$$v_1^n + \ldots + v_{n-1}^n = (1-q) + (q-q^2) + \ldots + (q^{n-2} - q^{n-1}).$$

This is a good old *telescoping sum*. Cancelling out adjacent positive and negative terms we see that

$$v_1^n + \ldots + v_{n-1}^n = (1 - q^{n-1}).$$

Since $v_n^n = q^{n-1}$, it follows that

$$v_1^n + \ldots + v_n^n = 1.$$

Solution (Exercise 13.2). From the definitions, $T_1 = 1$.

Notice that by the definition of T_n, if success occurs by time n or earlier, then $T_{n+1} = T_n$.

If success does not occur by time n, then necessarily $T_{n+1} = n+1$ and $T_n = n$.

Thus $T_{n+1} - T_n$ is either 1 or 0, and $\mathbf{P}(T_{n+1} - T_n = 1)$ is the probability of no success by time n, i.e. q^n.

Hence

$$\mathbf{E}[T_{n+1}] - \mathbf{E}[T_n] = \mathbf{E}[T_{n+1} - T_n] = q^n. \tag{13.27}$$

Since

$$\mathbf{E}[T_n] - \mathbf{E}[T_1] = (\mathbf{E}[T_2] - \mathbf{E}[T_1]) + (\mathbf{E}[T_3] - \mathbf{E}[T_2]) + \ldots + (\mathbf{E}[T_n] - \mathbf{E}[T_{n-1}]),$$

we have

$$\mathbf{E}\left[T_n\right] - \mathbf{E}\left[T_1\right] = q + q^2 + \ldots + q^{n-1}.$$

Since obviously $\mathbf{E}\left[T_1\right] = 1$,

$$\mathbf{E}\left[T_n\right] = 1 + q + \ldots + q^{n-1},$$

as claimed. By the formula in equation (13.7) for the sum of a finite geometric series, this agrees with equation (13.9).

One could also use induction and equation (13.27) to verify equation (13.9), of course.

Solution (Exercise 13.3). By the conditional probability formula and equation (13.13),

$$\begin{aligned}\mathbf{P}(T > n+m \,|\, T > n) &= \frac{\mathbf{P}(\{T > n+m\} \cap \{T > n\})}{\mathbf{P}(T > n)} \\ &= \frac{\mathbf{P}(T > n+m)}{\mathbf{P}(T > n)} = \frac{q^{n+m}}{q^n} = q^m.\end{aligned}$$

The second equality holds because $\{T > n+m\} \subset \{T > n\}$, and so

$$\{T > n+m\} \cap \{T > n\} = \{T > n+m\}.$$

Solution (Exercise 13.4).

$$\frac{d}{dx}\sum_{k=0}^{\infty} x^k = \sum_{k=0}^{\infty} \frac{d}{dx}\, x^k = \sum_{k=1}^{\infty} \frac{d}{dx}\, x^k. \tag{13.28}$$

We dropped the $k = 0$ term here because x^0 is constant.

Equation (13.28) shows that

$$\frac{d}{dx}\sum_{k=0}^{\infty} x^k = \sum_{k=1}^{\infty} k x^{k-1}. \tag{13.29}$$

By equation (13.21),

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x},$$

so

$$\frac{d}{dx}\sum_{k=0}^{\infty} x^k = \frac{1}{(1-x)^2}. \tag{13.30}$$

Comparing equations (13.29) and (13.30) proves equation (13.24).

Now let's find $\mathbf{E}[T]$. Let $q = 1 - p$. By equation (13.14), $\mathbf{P}(T = k) = q^{k-1}p$, for $1 \leq k < \infty$, and $\mathbf{P}(T = k) = 0$ otherwise.

Thus

$$\mathbf{E}[T] = \sum_{k=1}^{\infty} k\mathbf{P}(T = k) = \sum_{k=1}^{\infty} kq^{k-1}p = p\sum_{k=1}^{\infty} kq^{k-1} = p\frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}.$$

Chapter 14

Random variables with countable range

14.1 Countable range

The waiting time T defined in Section 13.2 of Chapter 13 is a good example of a random variable which does not have finite range. Now we should discuss the general properties of random variables which are somewhat similar to T.

Readers can probably guess many of the details, but please give some time to this. Developing a feeling for the theoretical concepts will pay off in your later work in probability.

14.2 Countability

Definition 14.1 (Countable sets). A set is said to be "countable" if its elements can be listed in a finite or infinite sequence. A set which is not countable is said to be uncountable.

It should be emphasized that, by definition, a finite set is a countable set.

Remark 14.2 (Sizes of infinity). After reading Definition 14.1, it is natural to wonder if there even *is* such a thing as an uncountable set, since an infinite sequence seems to be the natural way to describe an infinite set! However, it was proved by Georg Cantor (1874) that the real line is uncountable, in the sense of Definition 14.1. In other words, given any sequence of

real numbers, there must always be real numbers which are left out, and do *not* appear in the sequence ([7]). Cantor's remarkable discovery showed that, from the standpoint of mathematics, there are indeed different sizes of infinity. Of course, out in the real world, life continued much as usual, despite this disturbing news.

Recall the definition of a bounded function (Definition 10.18). The waiting time T defined in Section 13.2 is definitely not a bounded function. However, we should be aware that a random variable which has infinite range can still be bounded. The random variable $1/T$ is an example.

14.3 Countable additivity

The theoretical properties of mathematical probability theory are simpler if we add two technical assumptions about mathematical probability models. Fortunately, these technical assumptions hold for any model that is commonly used.

Probability Assumption 14.1 (Union of an infinite sequence of abstract events). For any probability model that we use, whenever $A_1, A_2, \ldots$ is a sequence of abstract events, the union of these sets is also an abstract event.

When we consider probabilities for an infinite sequence of events, one more assumption will be made:

Probability Assumption 14.2 (Additivity for an infinite sequence of abstract events). For any probability model that we use, whenever $A_1, A_2, \ldots$ is a sequence of disjoint events with union A,

$$\mathbf{P}(A) = \sum_{j=1}^{\infty} \mathbf{P}\left(A_j\right). \tag{14.1}$$

The property described in equation (14.1) is usually referred to as *countable additivity*.

From now on, Assumption 14.1 and Assumption 14.2 will hold, even if we don't mention them.

An infinite sequence of abstract events has no direct physical meaning. Similarly the action of summing an infinite series of probabilities has no direct physical meaning. Thus Assumption 14.1 and Assumption 14.2 do not seem to contribute any physical insight to our probability models, despite their technical usefulness. So it is interesting that in practice, for any experiment we can always choose a valid probability model such that Assumptions 14.1 and 14.2 hold.

Where are Assumption 14.1 and Assumption 14.2 going to be used? Sometimes we add up an infinite sequence of probabilities, in the process of calculating a physically meaningful probability value. But our assumptions are also used behind the scenes, to guarantee that abstract events, probabilities, and expectations exist and have convenient properties.

Since we now assume countable additivity, we could generalize some earlier statements. Typically the generalization amounts to simply replacing a finite sum by the sum of an infinite series. We usually don't bother to state generalizations of this sort, but simply use them when and if they are needed.

Incidentally, countable additivity often holds in mathematical models for quantities other than probability, for example for abstract quantities that represent physical properties such as length, area, weight and displacement. Here's an example.

Example 14.3 (Chopping up the unit interval). Let $A_i = (1/2^{i+1}, 1/2^i]$ for $i = 0, 1, \ldots$.

It is easy to see that the sets A_i are disjoint, that the union of all the sets A_i is exactly equal to $(0, 1]$, and that the length of A_i is $1/2^{i+1}$.

Then

$$\sum_{i=0}^{\infty} \mathbf{length}(A_i) = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{i+1} = \frac{1}{2} \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{i} = \frac{1}{2} \frac{1}{1 - \frac{1}{2}} = 1.$$

Thus the sum of the lengths of the intervals A_i is equal to the length of the union of these intervals, verifying a particular case of countable additivity for length.

Example 14.4 (More general chopping). Let $b_0 < b$ be real numbers, and let $b_k < b$ be an increasing sequence of real numbers such that $b_k \to b$ (one often writes $b_k \nearrow b$ to indicate an increasing limit).

You can easily convince yourself that the intervals $[b_i, b_{i+1})$ are disjoint, and that

$$[b_0, b) = \bigcup_{i=0}^{\infty} [b_i, b_{i+1}).$$

Clearly

$$\textbf{length}([b_i, b_{i+1})) = b_{i+1} - b_i.$$

Notice that

$$\sum_{i=0}^{n} \textbf{length}([b_i, b_{i+1})$$
$$= \sum_{i=0}^{n} (b_{i+1} - b_i) = (b_1 - b_0) + (b_2 - b_1) + \ldots + (b_n - b_{n-1}) = b_n - b_0.$$

In other words, the sum *telescopes*.

Then

$$\sum_{i=0}^{\infty} \textbf{length}([b_i, b_{i+1})) = \lim_{n\to\infty} \sum_{i=0}^{n} \textbf{length}([b_i, b_{i+1}))$$
$$= \lim_{n\to\infty} (b_n - b_0) = b - b_0 = \textbf{length}([b_0, b)). \quad (14.2)$$

Thus the sum of the lengths of the intervals $[b_i, b_{i+1})$ is equal to the length of the union of these intervals, verifying another particular case of countable additivity for length.

In equation (14.2), why does the first equality hold? It is simply the definition of the sum of an infinite series. That definition seems to be based on our geometrical picture of the real line, and gives countable additivity for lengths of pieces of the real line.

Exercise 14.1. We have not bothered to describe a sample space which is an adequate model for an infinite sequence of trials (see Remark 13.7). Given

such a model, we can certainly define the mathematical waiting time T in Chapter 13, as a function on the sample space.

In this problem, consider *any* sample space on which T is defined, with the correct distribution. Following the physical meaning of the random variables, we would naturally define the random variables T_n of Chapter 13 on the same sample space, with $T_n = T$ when $T < n$ (success before time n) and $T_n = n$ otherwise (success at time n or later, or no success ever).

Thus:

$$\{T_n = n\} = \{T = n\} \cup \{T = n + 1\} \cup \ldots \cup \{T = \infty\}. \tag{14.3}$$

Using countable additivity, we must be able to calculate $\mathbf{P}(T_n = n)$ by finding the sum of the series of probabilities of events on the right side of equation (14.3). In this problem you will check this.

Without assuming countable additivity, perform the calculation of the sum of the series of probabilities of events on the right side of equation (14.3). Assume $p > 0$ for simplicity and use equation (13.14) to get the probabilities you need.

Check that the result agrees with equation (13.5). This shows that countable additivity holds for the events in equation (14.3).

[Solution]

Exercise 14.2. Let T be the mathematical waiting time in Chapter 13.

Find the probability that T is even, assuming countable additivity and summing a series.

[Solution]

Example 14.5 (Even and odd). In the situation of Exercise 14.2, Lemma 13.2 tells us that

$$\mathbf{P}(T = k) = q^{k-1}p \text{ for } 1 \leq k < \infty, \tag{14.4}$$

and $\mathbf{P}(T = k) = 0$ otherwise. You can use equation (14.4) to solve Exercise 14.2. But let's try to find the probability that T is even, without using equation (14.4) explicitly.

We have, using independence, and the idea that tossing a coin starts over again on every toss,

$$\begin{aligned}&\mathbf{P}(T = k+1) = \\ &\mathbf{P}(\{\text{failure on first toss }\} \cap \{ \text{ first success occurs on } k \text{ th toss after that }\}) \\ &\mathbf{P}(\{\text{failure on first toss }\})\mathbf{P}(\{ \text{ first success occurs on } k \text{ th toss after that }\}) \\ &\qquad = q\mathbf{P}(T = k). \quad (14.5)\end{aligned}$$

By countable additivity,

$$\mathbf{P}(T \text{ odd }) = \mathbf{P}(T = 1) + \mathbf{P}(T = 3) + \mathbf{P}(T = 5) + \dots. \quad (14.6)$$

Also

$$\begin{aligned}\mathbf{P}(T \text{ even}) &= \mathbf{P}(T = 2) + \mathbf{P}(T = 4) + \dots \\ &= q\mathbf{P}(T = 1) + q\mathbf{P}(T = 3) + q\mathbf{P}(T = 5) + \dots \quad (14.7) \\ &= q\mathbf{P}(T \text{ odd}).\end{aligned}$$

Hmm, did we really need to use countable additivity and equations (14.6) and (14.7), in order to obtain the conclusion that $\mathbf{P}(T \text{ even}) = q\mathbf{P}(T \text{ odd})$?

Well, yes and no.

Here's a streamlined version of the same argument. We have, using independence, and the idea that tossing a coin starts over again on every toss,

$$\begin{aligned}&\mathbf{P}(T \text{ even}) = \\ &\mathbf{P}(\{\text{failure on first toss }\} \cap \{ \text{ odd number of subsequent tosses give first success }\}) \\ &\mathbf{P}(\{\text{failure on first toss }\})\mathbf{P}(\{ \text{ odd number of subsequent tosses give first success }\}) \\ &\qquad = q\mathbf{P}(T \text{ odd}). \quad (14.8)\end{aligned}$$

We used our physical ideas about independence to obtain equation (14.5). And we cheerfully used the same physical ideas to write down equation (14.8). In future we will do that as needed, since we are fearless. But we should take note that an infinite number of tosses is a bit further away from physical reality than a finite number of tosses. The concept of countable additivity, even behind the scenes, seems to give us more confidence to scoff at danger and use physical arguments in such cases.

At any rate, using either equation (14.7) or equation (14.8), we have that

$$\mathbf{P}(T \text{ even}) + \mathbf{P}(T \text{ odd}) = q\mathbf{P}(T \text{ odd}) + \mathbf{P}(T \text{ odd}). \quad (14.9)$$

Since it must be true that $\mathbf{P}(T \text{ even}) + \mathbf{P}(T \text{ odd}) = 1$, we have

$$(1+q)\mathbf{P}(T \text{ odd}) = 1,$$

so

$$\begin{aligned} \mathbf{P}(T \text{ odd} &= \frac{1}{1+q}, \\ \mathbf{P}(T \text{ even}) &= 1 - \frac{1}{1+q} = \frac{q}{1+q}. \end{aligned} \tag{14.10}$$

14.4 Calculus review: summing an absolutely convergent series

We often have to consider the sum of an infinite series of nonnegative numbers, or, more generally, the sum of a series which converges absolutely. This will be the case when adding probability values, but can also happen in other situations. We can use some facts from calculus.

Series Property 14.1 (Convergence test). If the terms in the series are nonnegative, and partial sums are bounded, then the series converges.

Series Property 14.2 (Rearrangement property). If the series converges absolutely, then the terms in the series can be rearranged in any order without changing the sum of the infinite series.

The Properties 14.1 and 14.2 imply another useful calculus fact:

Series Property 14.3 (Exchanging the order of summmation). Let real numbers a_{ij} be given for all positive integers i and j. Suppose that

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |a_{ij}| < \infty. \tag{14.11}$$

Then

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}, \tag{14.12}$$

meaning that both sides of the equation are convergent, and they are equal.

We won't bother to write out the proof of equation (14.12), but exchanging the order of summation is an important trick.

14.5 Distributions for random variables with countable range

We stated the mathematical definition of the distribution of a general random variable X in Definition 9.7. When the random variable X has a finite range, equation (9.5) of Section 9.2 gives a simple formula for calculating $\mathbf{P}(X \in S)$, using the probability mass function for X. We can use a similar approach in when X has countable range.

Let $x_1, x_2, \ldots$ be a sequence of distinct values, which includes all the numbers in the range of X that are members of W. If the value of X is a member of W, then that value must be equal to one of the numbers in the sequence. Thus

$$\{X \in W\} = \{X = x_1\} \cup \{X = x_2\} \cup \ldots. \tag{14.13}$$

Since the values $x_1, x_2, \ldots$ are distinct, the sets $\{X = x_1\}, \{X = x_2\}, \ldots$ are disjoint. By the countable additivity of probability we have

$$\mathbf{P}(X \in W) = \mathbf{P}(X = x_1) + \mathbf{P}(X = x_2) + \ldots. \tag{14.14}$$

This has the same form as equation (9.5) but now the expression on the right can be either a finite sum or an infinite series.

Using the probability mass function $\boldsymbol{q}$ for X (defined in Definition 9.8), equation (14.14) can be rewritten as

$$\mathbf{P}(X \in W) = \boldsymbol{q}(x_1) + \boldsymbol{q}(x_2) + \ldots. \tag{14.15}$$

Thus the probability mass function for X characterizes the whole distribution of X.

14.6 Expected values: countable range case

Definition 14.6 (Expected value with countable range). Let X be a random variable whose range can be listed in a finite or infinite sequence of distinct values $x_1, x_2, \ldots$. When the sequence is finite, $\mathbf{E}[X]$ is defined in Definition 10.2. When the sequence is infinite, the expected value $\mathbf{E}[X]$ is defined by

$$\mathbf{E}[X] = \sum_{j=1}^{\infty} x_j \mathbf{P}(X = x_j), \tag{14.16}$$

but only in the case that the series converges **absolutely**, i.e. when

$$\sum_{j=1}^{\infty} |x_j| \, \mathbf{P}\left(X = x_j\right) < \infty. \tag{14.17}$$

Note that Definition 14.6 is the formula that we guessed (and then used) when calculating the expected value of the waiting time T for the first head (Lemma 13.4)! The expected value of this waiting time is an excellent example to ilustrate the general formula for expectation.

We see from Definition 14.6 that the expected value of a random variable is determined by its distribution.

Notice that Definition 14.6 agrees with Definition 10.2. That's what we want. We are interested in *extending* the definition of expected value to new situations, but we don't want to change the definition that we've already given earlier.

In Definition 14.6, the expected value of X exists if the series in equation (14.16) converges absolutely, and only in that case.

Remark 14.7 (The meaning of existence). To clarify our terminology, let us agree that if we say that $\mathbf{E}\left[X\right]$ exists, we mean that $\mathbf{E}\left[X\right]$ exists as real number, in the sense of Definition 14.6. Sometimes we might say " $\mathbf{E}\left[X\right]$ exists as a real number", just to avoid any possible misunderstanding.

For a *nonnegative* random variable X, if $\mathbf{E}\left[X\right]$ does not exist as a real number people sometimes say that $\mathbf{E}\left[X\right] = \infty$. That notation is helpful in showing what is going on, but, despite that notation, in this book we will *not* say that $\mathbf{E}\left[X\right]$ exists in that case.

In calculus, the comparison principle for infinite series tells us that a series which is dominated by a convergent series must itself be convergent. A similar comparison principle holds for integrals of unbounded functions, and for expected values of unbounded random variables.

Fact 14.8 (A comparison principle for unbounded random variables). $\mathbf{E}\left[X\right]$ exists if and only if $\mathbf{E}\left[\,|X|\,\right]$ exists. Furthermore, if $\mathbf{E}\left[X\right]$ exists and $|Y| \leq |X|$ everywhere then $\mathbf{E}\left[Y\right]$ exists also.

Fact 14.8 holds for general random variables, not just random variables with countable range. In the case of an unbounded random variable with countable range, we have defined expected value in terms of an infinite series, so Fact 14.8 is an immediate consequence of the comparison principle for infinite series. For general random variables, mathematical expected value is defined in a less direct way, but the comparison principle holds in the general case also.

14.7 Key properties

Let's collect some facts that hold for expectations.

Theorem 14.9 (Four key properties of expected values). Four key properties of expected value are valid for all random variables:

Linearity: If $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ exist then $\mathbf{E}[X+Y]$ exists,

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]. \tag{14.18}$$

This is the additive property of expectation.

Also, if $\mathbf{E}[X]$ exists then for any number c, $\mathbf{E}[cX]$ exists, and

$$\mathbf{E}[cX] = c\mathbf{E}[X]. \tag{14.19}$$

This is the scaling property of expectation.

Monotonicity: If $\mathbf{E}[X]$ exists and $\mathbf{E}[Y]$ exists, and if $X \leq Y$ holds everywhere, then

$$\mathbf{E}[X] \leq \mathbf{E}[Y]. \tag{14.20}$$

Expectation of an indicator This is equation (11.4), which says:

$$\mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A).$$

Comparison principle This is Fact 14.8, which says:

$\mathbf{E}[X]$ exists if and only if $\mathbf{E}[\,|X|\,]$ exists. Furthermore, if $\mathbf{E}[X]$ exists and $|Y| \leq |X|$ everywhere then $\mathbf{E}[Y]$ exists also.

It is important to keep in mind that these properties hold for all random variables, not just countable range random variables (see Theorem 15.2). The comparison principle is only needed for unbounded random variables.

Exercise 14.3. Consider a mathematical random variable X such that all the values of X are positive integers, and such that for some $c > 0$, $\mathbf{P}(X = n) = c/n^2$ for $n = 1, 2, \ldots$.

(i) Does such a mathematical random variable actually exist?

(ii) Assuming that X exists, does $\mathbf{E}[X]$ exist?

[Solution]

Exercise 14.4. Let b_j, $j = 0, 1, 2, \ldots$, be strictly increasing numbers in $[0, 1)$, with $b_0 = 0$. Suppose that $b_j \nearrow 1$ as $j \to \infty$.

It is clear that the intervals $[b_j, b_{j+1})$ are disjoint and have union equal to $[0, 1)$. You may use this fact in what follows. (Please sketch a picture if this fact is not clear!)

(i) Consider a probability model with sample space $[0, 1]$ and the uniform probability distribution. Let $X(t)$ be the length of the interval $[b_j, b_{j+1})$ which contains t. Write down a formula for $\mathbf{E}[X]$ as an infinite series of numbers.

(ii) Suppose that $b_j = 1 - \frac{1}{2^j}$, for $j = 0, 1, 2, \ldots$.

Calculate the exact numerical value of $\mathbf{E}[X]$ for the random variable in part (i).

[Solution]

Exercise 14.5. Consider the probability model described in part (i) of Exercise 14.4. Let Y be the random variable defined by

$$Y(t) = \frac{\sin(b_{j+1}) - \sin(b_j)}{b_{j+1} - b_j}$$

for every $t \in [b_j, b_{j+1})$. Calculate $\mathbf{E}[Y]$.

[Solution]

One of the most useful theoretical facts in Chapter 10 was Theorem 10.7. This theorem extends to random variables with countable range, with little change.

Theorem 14.10 (Expectation by cases). Let $D_1, D_2, \ldots$ be a sequence of disjoint events in some model, whose union is the whole sample space.

Let $v_1, v_2, \ldots$ be numbers, and let X be a random variable such that $X = v_i$ at every point of D_i.

Then

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} v_i \mathbf{P}(D_i), \tag{14.21}$$

in the sense that $\mathbf{E}[X]$ exists if and only the series on the right converges absolutely, and in this case equality holds.

The proof is similar to the proof of Theorem 10.7, replacing finite sums with sums of infinite series. Applying Theorem 14.10 gives the general formula for the expectation of a function of a random variable:

Theorem 14.11 (Expectation of a function of a countable-range random variable). Let Z be a countable-range random variable on a sample space Ω. We do *not* assume that Z is real-valued. The values of Z can be anything. Let the distinct values in the range of Z be $v_1, v_2, \ldots$.

Let φ be any real-valued function whose domain includes $v_1, v_2, \ldots$. Then

$$\mathbf{E}[\varphi(Z)] = \sum_{i=1}^{\infty} \varphi(v_i) \mathbf{P}(D_i), \tag{14.22}$$

in the sense that $\mathbf{E}[\varphi(Z)]$ exists if and only the series on the right converges absolutely, and in this case equality holds.

14.8 Calculating expectation using the tail of the distribution

Definition 14.12 (The tail of a distribution). For any real-valued random variable X, probabilities of the form $\mathbf{P}(X > t)$ are sometimes called *tail* probabilities, particularly when one is studying the behavior of $\mathbf{P}(X > t)$ as $t \to \infty$. As a function of t, $\mathbf{P}(X > t)$ referred to as the tail of the distribution. (Similar terminology applies to $\mathbf{P}(X < -t)$.)

If we know the tail of the distribution, then we can calculate everything else about the distribution, with a little work. In particular, there is a nice recipe for calculating expected values of nonnegative random variables. The next lemma gives the most common case.

Lemma 14.13 (A tail expectation formula). Let X be a nonnnegative random variable, and let $a > 0$ be such that the range of X is contained in the set of numbers na, $n = 0, 1, \ldots$. Then

$$\mathbf{E}[X] = a \sum_{k=1}^{\infty} \mathbf{P}(X \geq ka) \tag{14.23}$$

Proof.

$$\begin{aligned}
&a \sum_{k=1}^{\infty} \mathbf{P}(X \geq ka) \\
&= a \sum_{k=1}^{\infty} (\mathbf{P}(X = ka) + \mathbf{P}(X = (k+1)a) + \mathbf{P}(X = (k+2)a) + \ldots) \\
&= a \sum_{k=1}^{\infty} \sum_{\ell \geq k}^{\infty} \mathbf{P}(X = \ell a) = a \sum_{\ell=1}^{\infty} \mathbf{P}(X = \ell a) \sum_{k=1}^{\ell} 1 \\
&= \sum_{\ell=0}^{\infty} \ell a \mathbf{P}(X = \ell a) = \mathbf{E}[X].
\end{aligned}$$

□

Exercise 14.6. Use equation (14.23) to calculate $\mathbf{E}[T]$, where T is the waiting time defined in section 13.2.

Note that this is the same calculation used in Example 13.6.

[Solution]

We have not yet defined expected values for general random variables, but expectation can be defined in general. Lemma 14.13 is a special case of the following general theorem, which holds for *every* random variable.

Theorem 14.14 (Expectation using the tail integral formula). Let X be a nonnegative random variable. Then

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}(X > t)\, dt. \tag{14.24}$$

Equation (14.24) is completely general, in the sense that if either side of this equation exists then both sides exist and are equal.

Exercise 14.7. Suppose that X satisfies the assumptions of Lemma 14.13. Show that equation (14.24) implies equation (14.23).

[Solution]

14.9 Solutions for Chapter 14

Solution (Exercise 14.1). Since $p > 0$, we know that $\mathbf{P}(T = \infty) = 0$.

So we want to show by calculation that

$$\mathbf{P}(T_n = n) = \sum_{k=n}^{\infty} \mathbf{P}(T = k).$$

Using equation (13.14) and equation (13.5), we want to show that:

$$q^{n-1} = \sum_{k=n}^{\infty} q^{k-1} p.$$

Let $j = k - n$. As k runs from n to ∞, j runs from 0 to ∞. Thus

$$\sum_{k=n}^{\infty} q^{n-1}p = pq^{n-1}\sum_{j=0}^{\infty} q^j = pq^{n-1}\frac{1}{1-q} = pq^{n-1}\frac{1}{p} = q^{n-1},$$

as claimed.

Solution (Exercise 14.2).

$$\{T \text{ even}\} = \{T = 2\} \cup \{T = 4\} \cup \{T = 6\} \cup \ldots.$$

Using countable additivity,

$$\mathbf{P}(T \text{ even}) = \mathbf{P}(T = 2) + \mathbf{P}(T = 4) + \mathbf{P}(T = 6) + \ldots = \sum_{\ell=1}^{\infty} \mathbf{P}(T = 2\ell) = \sum_{\ell=1}^{\infty} q^{2\ell-1}p$$

$$= p\sum_{j=0}^{\infty} q^{2j+1} = pq\sum_{j=0}^{\infty}(q^2)^j = \frac{pq}{1-q^2} = \frac{pq}{(1-q)(1+q)} = \frac{q}{1+q}.$$

Solution (Exercise 14.3).

(i) A mathematical random variable has to be defined on a sample space. So let's try $\Omega = \{1, 2, \ldots\}$, and define $X(n) = n$. Then at least this X will have positive integer values.

We are supposed to have

$$\mathbf{P}(X = n) = \frac{c}{n^2}$$

for some constant c.

With our definition of X, $\mathbf{P}(X = n) = \mathbf{P}(\{n\})$. So we want to have

$$\mathbf{P}(\{n\}) = \frac{c}{n^2}.$$

This definition will give us a genuine distribution provided that

$$\sum_{n=1}^{\infty} \mathbf{P}(\{n\}) = 1,$$

i.e.

$$\sum_{n=1}^{\infty} \frac{c}{n^2} = 1.$$

So finally we see the essential requirement: is it true that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} < \infty?$$

Well, yes, it is true, by the Integral Test from calculus. So now we just define

$$c = \frac{1}{\sum_{n=1}^{\infty} \frac{1}{n^2}},$$

and we have a genuine distribution, such that X does have the stated properties.

(ii) From our definitions, if $\mathbf{E}[X]$ exists then

$$\mathbf{E}[X] = \sum_{n=1}^{\infty} n\mathbf{P}(X = n) = \sum_{n=1}^{\infty} \frac{c}{n}.$$

But it is a well-known calculus fact is that $\sum_{n=1}^{\infty} \frac{1}{n}$ does not converge. So $\mathbf{E}[X]$ does not exist.

Solution (Exercise 14.4).

(i) By definition,

$$\{X = b_{j+1} - b_j\} = [b_j, b_{j+1}).$$

Thus

$$\mathbf{P}(X = b_{j+1} - b_j) = b_{j+1} - b_j.$$

Every point t in the sample space is a member of exactly one of the sets $[b_j, b_{j+1})$. Hence the range of X consists of the points $b_{j+1} - b_j$. By definition,

$$\mathbf{E}[X] = \sum_{j=0}^{\infty} (b_{j+1} - b_j)\mathbf{P}(X = b_{j+1} - b_j) = \sum_{j=0}^{\infty} (b_{j+1} - b_j)^2.$$

(ii) In this case,

$$\begin{aligned}\mathbf{E}[X] &= \sum_{j=0}^{\infty} \left(\frac{1}{2^j} - \frac{1}{2^{j+1}}\right)^2 = \sum_{j=0} \left(\frac{1}{2^{j+1}}\right)^2 = \sum_{j=0}^{\infty} \left(\frac{1}{4}\right)^{j+1} \\ &= \frac{1}{4} \sum_{j=0}^{\infty} \left(\frac{1}{4}\right)^j = \frac{1}{4} \frac{1}{1 - \frac{1}{4}} = \frac{1}{3}.\end{aligned}$$

Solution (Exercise 14.5). Every point t in the sample space is a member of exactly one of the sets $[b_j, b_{j+1})$. Hence the range of Y consists of the points

$$\frac{\sin(b_{j+1}) - \sin(b_j)}{b_{j+1} - b_j}.$$

By definition,

$$\begin{aligned}\mathbf{E}[Y] &= \sum_{j=0}^{\infty} \left(\frac{\sin(b_{j+1}) - \sin(b_j)}{b_{j+1} - b_j}\right) \mathbf{P}\left(Y = \frac{\sin(b_{j+1}) - \sin(b_j)}{b_{j+1} - b_j}\right) \\ &= \sum_{j=0}^{\infty} \left(\frac{\sin(b_{j+1}) - \sin(b_j)}{b_{j+1} - b_j}\right) (b_{j+1} - b_j) = \sum_{j=0}^{\infty} (\sin(b_{j+1}) - \sin(b_j)).\end{aligned}$$

The final series telescopes:

$$\begin{aligned}&\sum_{j=0}^{n} (\sin(b_{j+1}) - \sin(b_j)) \\ &= (\sin(b_1) - \sin(b_0)) + (\sin(b_2) - \sin(b_1)) + \ldots + (\sin(b_n) - \sin(b_{n-1})) \\ &\qquad = \sin(b_n) - \sin(b_0) = \sin(b_n).\end{aligned}$$

Letting $n \to \infty$, we see that

$$\sum_{j=0}^{\infty} (\sin(b_{j+1}) - \sin(b_j)) = \sin 1,$$

so $\mathbf{E}[Y] = \sin 1$.

Solution (Exercise 14.6). By equation (13.13),

$$\mathbf{P}(T \geq k) = \mathbf{P}(T > (k-1))q^{k-1}.$$

By equation (14.23),

$$\mathbf{E}[T] = \sum_{k=1}^{\infty} q^{k-1} = \sum_{n=0}^{\infty} = q^n = \frac{1}{1-q} = \frac{1}{p},$$

in agreement with equation (13.19).

Solution (Exercise 14.7). By assumption, the range of X is contained in the set of numbers na, $n = 0, 1, \ldots$. Thus for $na < t \leq (n+1)a$, $\{X > t\} = \{X > na\}$.

Hence for $na < t \leq (n+1)a$, $\mathbf{P}(X > t) = \mathbf{P}(X > na)$. Thus

$$\int_{na}^{(n+1)a} \mathbf{P}(X > t)\, dt = a\mathbf{P}(X > na). \tag{14.25}$$

By equation (14.24),

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}(X > t)\, dt = \sum_{n=0}^{\infty} \int_{na}^{(n+1)a} \mathbf{P}(X > t)\, dt = \sum_{n=0}^{\infty} a\mathbf{P}(X > na).$$

Let $k = n+1$ in the summation. Then k runs from 1 to ∞, and $\{X > na\} = \{X \geq ka\}$. This gives equation (14.23).

Chapter 15

Exponential waiting times and general random variables

An exponential waiting time is the continuous-time analog of the coin-tossing waiting time that was introduced in Section 13.2. The range of the exponential waiting time is not a countable set. On the contary, it is the whole interval $[0, \infty)$.

15.1 The exponential distribution

Definition 15.1 (The exponential distribution). For each $\lambda > 0$, let h_λ be the function on $\mathbb{R}$ defined by

$$h_\lambda(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{15.1}$$

It was shown in Exercise 3.11 that $\int_{-\infty}^{\infty} h_\lambda = 1$, so that h_λ is a probability density on $\mathbb{R}$.

The function h_λ is referred to as the exponential density with parameter λ. The distribution with probability density h_λ is called the *exponential distribution* with parameter λ.

Any random variable having this probability distribution will be referred to as an exponential waiting time.

We'll discuss examples of random variables with exponential distributions in Section 15.4. A typical experimental example involves recording events which occur randomly, at an "average rate" λ per unit time. Let X denote the length of time until the first event is recorded. It often turns out that X is a random variable whose distribution is exponential with parameter λ. For $t \geq 0$, we then have

$$\mathbf{P}(X > t) = \int_{\{X>t\}} h_\lambda = \int_t^\infty \lambda\, e^{-\lambda u}\, du = -e^{-\lambda u}\Big|_t^\infty = e^{-\lambda t}. \tag{15.2}$$

Equation (15.2) characterizes the exponential distribution.

In order to establish properties of exponential waiting times, we are going to have to calculate some expected values. The next two sections say how to do that in general.

15.2 Facts about general expectations

In this section we'll outline some facts that are true for all expectations.

We use expected values freely in this book, but we will not state a rigorous definition of expectation for general random variables, although such a definition can be given fairly easily (see Appendix N for an outline). The general definition is consistent with the definitions given previously for random variables with a finite range or a countable range. The properties which hold for expectation are summarized in the next theorem.

Theorem 15.2 (Properties of expectations of general random variables).

(i) For every bounded random variable X, $\mathbf{E}[X]$ exists. (Bounded functions are defined in Definition 10.18.)

If X is an unbounded random variable then $\mathbf{E}[X]$ exists if X is not too large as a function on the sample space.

(ii) The value of $\mathbf{E}[X]$ is determined by the distribution of X, i.e. random variables with the same distribution have the same expected value. (And since random variables with the same distribution must have the same expectation, we sometimes speak of "the expectation of the distribution", rather than the expectation of the random variable.)

(iii) The four key properties of expectation stated in Theorem 14.9 hold for the general definition of expected value: linearity, monotonicity, the

formula for the expectation of an indicator function (equation (11.4)), and the comparison principle.

In the case of a finite-range random variable, the expected value has a frequency interpretation, as stated in Probability Fact 10.1. What can we say more generally?

Remark 15.3 (Bounded random variables and experiments). Suppose that a bounded random variable X is intended to model a measured value in an experiment, and X does not have finite range. (Perhaps the experiment involves measuring a random distance along a road, or the weight of a random lump of butter, so it would be unnatural to use a finite-range random variable.) In this case the $\mathbf{E}[X]$ has the same physical interpretation described in Section 10.3. It is the long-run average of the measured value of X.

See the discussion in Remark N.5 of Appendix N.

What about the interpretation of an unbounded random variable? If X is unbounded, even if $\mathbf{E}[X]$ exists we won't try to state a *direct* physical interpretation for $\mathbf{E}[X]$. In our work we will think of unbounded random variables simply as mathematical tools, which help us to understand bounded random variables.

Example 15.4 (Using the rules). We didn't give a rigorous definition for the expectation of a general random variable, but Theorem 15.2 says that the four key properties of expectation which were stated earlier (in Theorem 14.9) continue to hold for the general definition of expected value. That is usually all you need.

For example, we studied the Markov inequality in Section 12.7, and then used that inequality to understand the behavior of random walk. The derivation of that inequality only used linearity, monotonicity, and equation (11.4).

15.3 Expectations when there is a density on the sample space

We have noted two situations where it can be natural to use probability densities. In Section 3.2, the sample space is an interval of the real line, and probabilities are defined by equation (3.7). When the sample space is a region in the plane, we mentioned that probability densities can be used in a similar way, though we didn't bother to give details.

Whenever a distribution has a probability density, the probability of an event is given by integrating the density over the event.

Let's recall the concept of integration over a set (introduced after equation (3.12) in Section 3.4).

Suppose we have a sample space Ω on which integration is defined. Ω doesn't have to be an interval of the real line, as long as we know how to integrate. So Ω might be a region in the plane, or more generally a region in $\mathbb{R}^n$, or even something else.

Let $\int f$ denote whatever form of integral we are using on the sample space Ω. If Ω is the real line, calculus books often write $\int f$ as $\int_{-\infty}^{\infty} f$, and if Ω is the plane, $\int f$ is often written as in calculus books as

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx$$

The integral of a function f over a set A is denoted by $\int_A f$. If A is an interval $[a, b]$ of the real line, calculus books often write $\int_A f$ as $\int_a^b f(x), dx$, and if A is a subset of the plane, $\int_A f$ is often written in calculus books as

$$\int\int_A f.$$

For thinking about the logic of a problem, it is likely clearer to just write the integral over a set A as $\int_A f$, as long as the reader understands what sort of integral you are working with.

Incidentally, a precise general definition of integrating a function f over a set A, that works on any space, is given in Definition 3.6.

Probability densities were introduced in Definition 3.4, in the setting of the real line. A general definition of a probability density is the following.

Definition 15.5 (Probability densities for probability distributions (general formulation)). In general, to say that f is a probability density simply means that f is a nonnegative function whose integral over the whole space is equal to one, i.e. $\int f = 1$. Here it is assumed that integration on the space has been defined.

To say that a distribution has a probability density f means that the probability of an event is given by integrating f over the event, i.e. for any event A,

$$\mathbf{P}(A) = \int_A f. \tag{15.3}$$

Remark 15.6 (Comparison with the previous density definition). Remark 3.7 shows that Definition 15.5 is consistent with the original definition given in Definition 3.4 for the real line case.

(In Definition 3.4, equation (15.3) is only required to hold for intervals, but Remark 3.7 states that if equation (15.3) holds for events which are intervals of the real line, then it actually holds for all events.)

Examples of densities on subsets of the real line were given in Sections 3.4 and 3.7. Appendix F has some examples of using probability densities on subsets of the plane.

As noted in Example 10.20, if probabilities are given by a probability density f on the sample space, we can also use f to find expected values. For any random variable X on the sample space,

$$\mathbf{E}\,[X] = \int Xf, \tag{15.4}$$

provided of course that the integral of Xf exists.

Here $\int Xf$ means the integral of Xf over the sample space Ω. If Ω is an interval of the real line, say $\Omega = [s, t]$, then equation (15.4) is the same statement as equation (10.33):

$$\mathbf{E}\,[X] = \int_s^t X(u)f(u)\,du.$$

But equation (15.4) holds for any sample space, for example if the sample space is a region in the plane or in $\mathbb{R}^n$. The only difference in those other cases is that the integral in the equation (15.4) may require more work.

We are ready to start computing expectations when probabilities are given by a density on the sample space. But what if we not told the definition of a random variable X on any sample space? Instead, suppose that someone simply gives us the probability distribution of X? Can we still find $\mathbf{E}[X]$?

Theorem 15.2, part (ii), says that the distribution of a random variable determines the expected value. So, yes, in principle it must be true that we can find $\mathbf{E}[X]$, if someone tells us the distribution of X.

If it happens that the probability distribution of X is given by a density function h on the real line, then there is a neat formula:

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} th(t)\,dt, \tag{15.5}$$

provided of course that the integral of $th(t)$ exists.

More generally, for any function φ on the real line,

$$\mathbf{E}[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(t)h(t)\,dt, \tag{15.6}$$

provided that the integral of $\varphi(t)h(t)$ exists.

Equation (15.6) **is the formula for the expected value of a function of a random variable, in the case that the distribution of the random variable has a density.**

Equation (15.6) is often useful. Taking φ to be the function $\varphi(t) = t$ shows that equation (15.6) implies equation (15.5).

One can derive equation (15.6) from equation (15.4). The argument is given in Appendix E.

Example 15.7 (Mean of a uniform distribution). Consider a random variable X whose distribution is uniform on $[a, b]$. (This is a short way of saying that the distribution of X on the real line is uniform on $[0, 5]$ and zero everywhere else. We talked about the probability density for the distribution of such a random variable in Example 9.13 and Remark 9.14.)

Let's find $\mathbf{E}[X]$

By Exercise 3.5, a density for the uniform distribution on $[a, b]$ is given by the probability density f which is constant and equal to $1/(b - a)$ at all

points of $[a, b]$. Using the function f in quation (15.4), we have

$$\mathbf{E}[X] = \int_a^b \frac{t}{b-a}\,dt = \frac{1}{b-a}\frac{t^2}{2}\bigg|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}. \tag{15.7}$$

Thus the mean of X is the midpoint of the interval $[a, b]$.

s

Exercise 15.1 (The Cauchy distribution). Let X be a random variable whose probability distribution has a density h given by

$$h(x) = \frac{c}{1+x^2} \tag{15.8}$$

for all $x \in \mathbb{R}$. Then X is said to have a *standard Cauchy distribution.*

1. Find c.

2. Does $\mathbf{E}[X]$ exist?
[Solution]

Exercise 15.2. Let X be a random variable whose distribution is uniform on $[0, 5]$. (This is a short way of saying that the distribution on the real line is uniform on $[0, 5]$ and zero everywhere else. We talked about the probability density for the distribution of such a random variable in Example 9.13 and Remark 9.14.)

Find $\mathbf{E}[\sin X]$.
[Solution]

15.4 Properties of the exponential distribution

Now we are ready to get back to exponential waiting times.

Exercise 15.3 (Mean of the exponential distribution). Let T be a random variable whose distribution is exponential with parameter λ (Definition 15.1). Show that

$$\mathbf{E}[T] = \int_0^\infty \lambda t e^{-\lambda t}\, dt = \frac{1}{\lambda}. \tag{15.9}$$

[Solution]

As noted in Section 14.8, for any random variable T the function $t \mapsto \mathbf{P}(T > t)$ is sometimes called the *tail* of the distribution (on the right).

One of the things we learn from the tail is the rate at which the probability of $\mathbf{P}(T > t)$ approaches zero as $t \to \infty$. It is easy to calculate the tail function for the exponential distribution.

Let T be a random variable having exponential distribution with parameter λ. The tail function is $\mathbf{P}(T > t)$. By equation (15.2), the tail function for the exponential distribution with parameter λ is given by $\mathbf{P}(T > t) = e^{-\lambda t}$ for any $t \geq 0$.

Of course, since T has an exponential distribution, $\mathbf{P}(T \geq 0) = 1$, so $\mathbf{P}(T > t) = 1$ for all $t \leq 0$.

Clearly the tail of an exponential distribution approaches zero rapidly as $t \to \infty$.

Exercise 15.4. In Theorem 14.14 we stated a useful general formula for expectation called "The Tail-Integral formula for expectation", although no proof was given. Test the tail integral formula by calculating the mean of the exponential distribution again.

[Solution]

Exercise 15.5 (Expectation of square of random variable with exponential distribution). The exponential distribution with parameter λ is defined in Exercise 15.3. Suppose that X has this distribution.

Calculate $\mathbf{E}[X^2]$.

[Solution]

Recall from equation (13.13) and Definition 13.3: when T is the waiting time for first success in coin-tossing with $p > 0$, the distribution of T is the geometric distribution, and $\mathbf{P}(T > k) = q^k$, where $q = 1 - p$ and k is a nonnegative integer.

Let $r = -\log q$, where $\log q$ denotes the logarithm of q with base e. Since $q < 1$, r is a positive number. We have $\mathbf{P}(T > k) = e^{-r\,k}$, so the tail of the geometric distribution seems quite similar to the tail of the exponential distribution.

The similarity of these two distributions suggests that a random variable T with exponential distribution is the continuous-time analogue of the waiting time for first success in coin-tossing. And in fact, a random variable T with exponential distribution is used as a model for a "lifetime" or a waiting time in many situations.

For example, suppose you are measuring the rate of decay of some radioactive material, using a detection device such as a Geiger counter. The random time until the first emission of a particle from the sample has an exponential distribution. Similarly, the time spent waiting for a telephone call at a sales desk, or a data request at a computer server, will often have a distribution which is approximately exponential.

Discussing any kind of waiting times, one presumably wants to know when the waiting *begins*. When waiting for the emission of a particle from radioactive material, the wait starts when some observer starts to record data. But radioactivity happens continually, so the starting time is not connected at all to the physical process. Thus it seems as if the choice of starting time could affect the observed statistical distribution of the waiting time.

However, for the exponential distribution, just as with coin-tossing (Exercise 13.3), the observed distribution does not depend on the choice of starting time when waiting.

Lemma 15.8 (The "memoryless" property). Let T be a waiting time having an exponential distribution with parameter λ. For some given time $s \geq 0$, let $A = \{T > s\}$. Then for any $t \geq 0$,

$$\mathbf{P}(T - s > t \,|\, A) = \mathbf{P}(T > t), \tag{15.10}$$

i.e.

$$\mathbf{P}(T > s + t \,|\, A) = \mathbf{P}(T > t). \tag{15.11}$$

In Lemma 15.8, one can think of T as the time that some observer has to wait, for a particular event to occur. If a new observer arrives at time $s \geq 0$, there are two possibilities.

- If the event has already occurred, the new observer has nothing to wait for, and does not record the result.
- If the event has not yet occurred (i.e. if we are in the situation described by A), then the new observer is waiting along with the original observer. The new observer will wait time $T - s$ until the event takes place.

  The left side of equation (15.10) describes the statistical properties of the time that the new observer records, given that the event has not already occurred when the new observer arrives.

Exercise 15.6. Prove Lemma 15.8.
[Solution]

Remark 15.9 (Memoryless really means memoryless). In the setting of Lemma 15.8, think of T as the waiting time for some physical event D.

Let $[a, b]$ be a subinterval of (s, ∞). Let $T^* = T - s$. Equation 15.10 says that

$$\mathbf{P}(T^* > t \mid \{T > s\}) = \mathbf{P}(T > t). \tag{15.12}$$

Let $\mathbf{P}^*$ denote probabilities conditioned on $\{T > s\}$. Equation 15.12 says that $\mathbf{P}^*(T^* > t) = \mathbf{P}(T > t)$ for all $t \geq 0$. Since these random variables are nonnegative, $\mathbf{P}^*(T^* > t) = 1 = \mathbf{P}(T > t)$ for all $t < 0$.

So $\mathbf{P}^*(T^* > t) = \mathbf{P}(T > t)$ for all t.

Based on that equality, one can show that the whole distribution of T^* is the same as the whole distribution of T. So all statistical properties must be the same for both random variables.

Remark 15.9 explains why we do not need to specify the starting time for the exponential distribution.

Let's think more about the parameter λ in the exponential distribution for T. We know now that $\mathbf{P}(T > t) = e^{-\lambda t}$ and that $\mathbf{E}[T] = 1/\lambda$. Both those equations tell us that as λ increases the waiting time T becomes smaller. We can make that statement more precise by considering the tail function $s(t) = \mathbf{P}(T > t)$. Thinking about T as the lifetime of a randomly selected object, we might call s the "survival function", since it gives the probability that the randomly selected object is still alive at time t. Thinking of the randomly selected object as part of a large population, the frequency interpretation says that $s(t)$ represents the fraction of the population that is still alive at time t. Notice that $s(t) = e^{-\lambda t}$ satisfies a simple differential equation on $(0, \infty)$:

$$s'(t) = -\lambda s(t), \tag{15.13}$$

Suppose that the initial size of the population is N, where N is some large number. We would expect that at time t the surviving population would have size approximately equal to $Ns(t)$.

At time $t + \Delta t$, the size of the population is approximately $Ns(t + \Delta t)$. The number of objects that have died during the time interval $[t, t + \Delta t]$ is approximately $Ns(t) - Ns(t + \Delta t)$. Using equation (15.13),

$$s(t) - s(t + \Delta t) \approx -s'(t)\Delta t = \lambda s(t)\Delta t.$$

Thus the number of objects that have died during the time interval $[t, t+\Delta t]$ is approximately $N\lambda s(t)\Delta t$.

The number of living objects at time t is approximately $Ns(t)$. Thus the fraction of the current population which dies during $[t, t+\Delta]$ is approximately

$$\frac{N\lambda s(t)\Delta t}{Ns(t)} = \lambda\Delta.$$

Dividing by Δt shows that the average death rate per object per unit time is λ.

With this interpretation in mind, one might call λ the "death rate", or more briefly the *rate* for the exponential distribution which has parameter λ.

Equation (15.13) is a differential equation for the survival function s. We assume $s(0) = 1$, since there has not been time for any deaths. Hence we are interested in a solution s of equation (15.13) which satisfies the initial condition $s(0) = 1$.

Exercise 15.7 (Uniqueness for the solution of equation (15.13) **).** Let s be a solution of equation (15.13). Show that $s(t) = s(0)e^{-\lambda t}$ for all t. (This is often shown in calculus courses.) Hint: Define $f(t) = s(t)e^{\lambda t}$. Calculate $f'(t)$.

[Solution]

15.5 Solutions for Chapter 15

Solution (Exercise 15.1). (i) Since h is a probability density we must have $\int h = 1$.

$$\int h = \int_{-\infty}^{\infty} \frac{c}{1+x^2}\,dx = \lim_{b\to\infty} \int_{-b}^{b} = \lim_{b\to\infty} c \left(\arctan x \Big|_{-b}^{b} \right)$$
$$= c \lim_{b\to\infty} (\arctan b - \arctan(-b)) = 2c \lim_{b\to\infty} \arctan b = 2c\frac{\pi}{2} = c\ \pi.$$

Hence $c = 1/\pi$.

(ii) By equation (15.5),

$$\mathbf{E}\,[X] = \int_{-\infty}^{\infty} t\,h(t)\,dt.$$

However, we have to be careful in evaluating this integral, because the integrand has both a positive and a negative part. The integral of the positive part is

$$\int_{0}^{\infty} t\,h(t)\,dt = \frac{1}{\pi} \lim_{b\to\infty} \int_{0}^{b} \frac{t}{1+t^2}\,dt.$$

Using the fact that $t^2 \geq 1$ on $[1, \infty)$, we have

$$\int_{0}^{\infty} t\,h(t)\,dt \geq \frac{1}{\pi} \lim_{b\to\infty} \int_{1}^{b} \frac{t}{2t^2}\,dt = \frac{1}{2\pi} \lim_{b\to\infty} \int_{1}^{b} \frac{1}{t}\,dt = \frac{1}{2\pi} \lim_{b\to\infty} (\log b - \log 2) = \infty.$$

Thus $\mathbf{E}\,[X]$ does not exist.

(In this calculation, we use log to denote logarithms to the base e.)

Solution (Exercise 15.2). Let f be the function on $[0,5]$ defined by $f(x) = 1/5$ for all $x \in [0,5]$. Then f is a probability density for the uniform distribution on $[0,5]$.

As in Example 9.13, we obtain a probability density h for the distribution of X by extending f. We extend it to be equal to zero on the complement of $[0,5]$, so h is given as in equation (9.18):

$$h(x) = \begin{cases} \frac{1}{5} & \text{if } x \in [0,5], \\ 0 & \text{otherwise.} \end{cases}$$

By equation (15.6),

$$\begin{aligned} \mathbf{E}[\sin X] = \int_{-\infty}^{\infty} \sin(x) h(s)\, dx = \int_0^5 \sin(x) \frac{1}{5}\, dx = -\frac{1}{5} \cos x \Big|_0^5 \\ = -\frac{1}{5}(\cos 5 - \cos 0) = \frac{1}{5}(1 - \cos 5). \end{aligned}$$

Solution (Exercise 15.3). By equation (15.5),

$$\mathbf{E}[T] = \int_{-\infty}^{\infty} t\, \lambda e^{-\lambda t}\, dt = \int_0^{\infty} t\, \lambda e^{-\lambda t}\, dt = \lim_{b \to \infty} \int_0^b t\, \lambda e^{-\lambda t}.$$

Using integration by parts,

$$\mathbf{E}[T] = \lim_{b \to \infty} \left(t e^{-\lambda t} \Big|_0^b + \int_0^{\infty} e^{-\lambda t}\, dt \right) = \lim_{b \to \infty} \left(t e^{-\lambda t} \Big|_0^b - \frac{1}{\lambda} e^{-\lambda t} \Big|_0^b \right).$$

We know that $te^{-\lambda b} \to 0$ as $b \to \infty$, a consequence of L'Hôpital's Rule, and of course $e^{-\lambda b} \to 0$ as $b \to 0$. Hence

$$\mathbf{E}[T] = \frac{1}{\lambda}.$$

Solution (Exercise 15.4). Let X be a random variable having exponential distribution with parameter λ.

By Theorem 14.14,

$$\mathbf{E}[X] = \int_0^{\infty} \mathbf{P}(X > t)\, dt = \int_0^{\infty} e^{-\lambda t}\, dt = -\frac{e^{-\lambda t}}{\lambda} \Big|_0^{\infty} = \frac{1}{\lambda},$$

which is consistent with equation (15.9).

Solution (Exercise 15.5). By equation (15.6),

$$\mathbf{E}\left[X^2\right] = \int_0^\infty x^2 \lambda e^{-x\lambda}\, dx = \lim_{b\to\infty} \int_0^b x^2 \lambda e^{-x\lambda}\, dx.$$

We will use integration by parts to calculate the integral, and then use the fact that $b^2 e^{-b} \to 0$ and $be^{-b} \to 0$ as $b \to \infty$. Thus

$$\begin{aligned}
\mathbf{E}\left[X^2\right] &= \lim_{b\to\infty} \left(-x^2 e^{-x\lambda}\Big|_0^b + \int_0^b 2xe^{-x\lambda}\, dx \right) \\
&= \lim_{b\to\infty} \left(-x^2 e^{-x\lambda}\Big|_0^b - \frac{2x}{\lambda} e^{-x\lambda}\Big|_0^b + \int_0^b \frac{2}{\lambda} e^{-x\lambda}\, dx \right) \\
&= \lim_{b\to\infty} \left(-x^2 e^{-x\lambda}\Big|_0^b - \frac{2x}{\lambda} e^{-x\lambda}\Big|_0^b - \frac{2}{\lambda^2} e^{-\lambda x}\Big|_0^b \right) = \frac{2}{\lambda^2} \qquad (15.14)
\end{aligned}$$

Solution (Exercise 15.6).

$$\mathbf{P}(T > s+t \mid A) = \frac{\mathbf{P}(T > s+t)}{\mathbf{P}(A)} = \frac{\mathbf{P}(T > s+t)}{\mathbf{P}(T > s)}.$$

By equation (15.10),

$$\mathbf{P}(T > s+t \mid A) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbf{P}(T > t).$$

Solution (Exercise 15.7). Let $f(t) = s(t)e^{\lambda t}$.

$$f'(t) = s'(t)e^{\lambda t} + s(t)\lambda e^{\lambda t} = -\lambda s(t) e^{\lambda t} + s(t)\lambda e^{\lambda t} = 0.$$

By the Mean Value Theorem $f(t) - f(0) = 0$ for every t, so $f(t) = f(0)$, i.e. $s(t)e^{\lambda t} = f(0) = s(0)$, and hence $s(t) = s(0)e^{-\lambda t}$.

Chapter 16

Moments and inequalities

The mean of a distribution can be thought of as an average value for the random variable which has that distribution. It can also be thought of as a "central point" of the distribution. In this chapter we introduce the concept of the variance of a distribution, which can be thought of as a measure of the "width" of the distribution.

Readers may not wish to work all the exercises in this chapter. The goal should be to develop a feeling for how the concept of variance is used.

16.1 Moments

If a random variable has a large range, then its distribution can be complicated, even if the range is finite. We need to identify simple properties that help us to understand the behavior of the random variable.

The expected value of a random variable X is usually the most important such property, but an expected value is just one number. We can learn more by calculating *moments* of the random variable.

Definition 16.1 (Moments of a random variable). For $n = 0, 1, 2, \ldots$, the n-th moment of the random variable X is $\mathbf{E}[X^n]$, provided that it exists. The n-th absolute moment is defined to be $\mathbf{E}[\,|X|^n\,]$.

The definition here is general, so we are using the fact that expected values can be defined for general random variables (see Section 15.2).

The first moment of X is the expected value $\mathbf{E}[X]$. When it exists, and when X represents some property of an experiment, we know $\mathbf{E}[X]$ is likely to be close to the average measured value in repeated experiments. The first absolute moment gives the same sort of information about the absolute value of X. Thus the first absolute moment gives the average *size* of the random variable.

By definition, the second moment of X is $\mathbf{E}[X^2]$, so it gives the average size of the square of the random variable. The significance of the second moment will become clearer as we study the concept of variance.

All moments are expected values, and we have noted that mathematical random variables can be so large that their expected values do not exist. A random variable with a Cauchy distribution (Exercise 15.8) is a typical example. And even if the expected value exists, higher moments may not.

Here are a few examples.

Exercise 16.1.

(i) Let $\Omega = (0, 1]$, with uniform distribution. Let $X(t) = 1/t$. Show that $\mathbf{E}[X]$ does not exist.

(ii) Let $\Omega = (0, 1]$, with uniform distribution. Let $X(t) = 1/\sqrt{t}$. Show that $\mathbf{E}[X]$ exists but $\mathbf{E}[X^2]$ does not exist.

(iii) Let $\Omega = \{1, 2, \ldots\}$, and let $\mathbf{P}$ be a distribution on Ω such that $\mathbf{P}(\{n\}) = c/n^5$ for some constant c. Let $X(j) = j$. Show that $\mathbf{E}[X]$, $\mathbf{E}[X^2]$ and $\mathbf{E}[X^3]$ exist but $\mathbf{E}[X^4]$ does not exist.

[Solution]

Section 16.8 has some information that you can use if you are trying to confirm that a moment exists.

16.2 Variance

Variance is a key concept in probability theory. The variance of X is simply the second moment of the centered version of X.

Definition 16.2 (Centered random variables). A random variable X will be said to be centered if $\mathbf{E}[X] = 0$. In this case X is also said to be a mean zero random variable.

Let X be a random variable such that $\mathbf{E}[X]$ exists. The centered version of X is the random variable $X - \mathbf{E}[X]$. The value of $X - \mathbf{E}[X]$ is also called the *deviation* of X from its mean.

In calculations we often write $\mathbf{E}[X]$ as μ, so that the deviation of X from its mean is written as $X - \mu$.

Definition 16.3 (Variance). Let X be a random variable whose expectation exists.

The centered second moment, $\mathbf{E}[(X - \mathbf{E}[X])^2]$, is called the *variance* of X, and is denoted by $\mathbf{Var}(X)$, when this expected value exists. It is often referred to as the *mean square deviation* of X.

The positive square root of the variance of X is called the *standard deviation* of X, and is often written as σ in calculations. In that case $\mathbf{Var}(X) = \sigma^2$.

If one must describe the properties of a probability distribution using only two numbers, the mean and the variance of a distribution are usually the most informative. The variance tells us how "spread-out" the distribution is.

We often write the expression for $\mathbf{Var}(X)$ more neatly using μ to denote $\mathbf{E}[X]$:

$$\mathbf{Var}(X) = \mathbf{E}\left[(X - \mu)^2\right]. \tag{16.1}$$

Remark 16.4 (Variance of a distribution). The distribution of X determines $\mathbf{E}[X^n]$ and $\mathbf{Var}(X)$, so we will at times speak of the "the moments of a distribution" and "the variance of a distribution", or "the moments of a density" and "the variance of a density".

Thinking about existence of the variance, note that we only speak about the variance of X in situations where $\mathbf{E}[X]$ exists.

Denote $\mathbf{E}[X]$ by μ. Since $X - \mu = X + (-\mu)$, the $n = 2$ case of Lemma 16.28 tells us that $\mathbf{E}[(X-\mu)^2]$ exists if $\mathbf{E}[X^2]$ exists.

Since $X = (X - \mu) + \mu$, the $n = 2$ case of Lemma 16.28 also says that if $\mathbf{E}[(X-\mu)^2]$ exists then $\mathbf{E}[X^2]$ exists.

Thus we have the following:

Existence Fact Whenever $\mathbf{E}[X]$ exists, $\mathbf{Var}(X)$ exists if and only if $\mathbf{E}[X^2]$ exists.

That's all we have to say about existence of the variance.

The next exercise is of major importance!

Exercise 16.2 ("Mean square minus square mean"). By expanding equation (16.1), show that

$$\mathbf{Var}(X) = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2. \tag{16.2}$$

[Solution]

Notice that equation (16.2) shows immediately that $\mathbf{Var}(X) \leq \mathbf{E}[X^2]$. Also, the definition of $\mathbf{Var}(X)$ shows that $\mathbf{Var}(X) \geq 0$, so equation (16.2) tells us that

$$(\mathbf{E}[X])^2 \leq \mathbf{E}\left[X^2\right]. \tag{16.3}$$

Incidentally, we can replace X by $|X|$ in equation (16.3), so we also have:

$$(\mathbf{E}[\,|X|\,])^2 \leq \mathbf{E}\left[X^2\right]. \tag{16.4}$$

Exercise 16.3 (Variance of a constant). Let X be a constant random variable in some probability model, so that $X = 7$ everywhere. Find $\mathbf{E}[X]$ and $\mathbf{Var}(X)$.

[Solution]

Remark 16.5 (When the variance is zero). Let X be a random variable with mean μ and variance zero. Then $\mathbf{E}[(X-\mu)^2] = 0$.

Since $(X-\mu)^2$ is a nonnegative random variable with mean zero, it is tempting to conclude that $(X-\mu)^2$ is the zero random variable. That is not quite true, but it's almost true.

The following fact holds: for any nonnegative random variable Y, if $\mathbf{E}[Y] = 0$ then $\mathbf{P}(Y \neq 0) = 0$, i.e. $\mathbf{P}(Y = 0) = 1$. (See Appendix G for a derivation.)

So when $\mathbf{Var}(X-\mu) = 0$, we know that $\mathbf{P}(\,(X-\mu)^2 = 0\,) = 1$, i.e. $\mathbf{P}(X = \mu) = 1$. So a random variable with zero variance is equal to its mean with probability one.

Exercise 16.4 (Scaling the variance). Show that

$$\mathbf{Var}(cX) = |c|^2 \mathbf{Var}(X). \tag{16.5}$$

[Solution]

Exercise 16.5 (Shifting preserves variance). Prove that for any real number c,

$$\mathbf{Var}(X+c) = \mathbf{Var}(X). \tag{16.6}$$

[Solution]

Remark 16.6 (The standard version of a random variable). Let X be a random variable with mean μ and variance $\sigma^2 > 0$.

By the linearity of expectation, $\mathbf{E}[X-\mu] = 0$. By equation (16.5),

$$\mathbf{Var}(X-\mu)\,\sigma = \frac{1}{\sigma^2}\mathbf{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1.$$

Thus the random variable

$$\frac{X-\mu}{\sigma}$$

has mean zero and variance one. We will call this random variable the *standard version of* X, or sometimes say that we "standardize" X to form this random variable.

Exercise 16.6 (Variance of a uniform distribution). Let X be the random variable on $[0, 5]$ with $X(\omega) = \omega$. Using the uniform distribution on $[0, 5]$, calculate the mean and variance of X.

Then, generalize your work. Find the mean and variance of a random variable X whose distribution is uniform on an interval $[s, t]$.

[Solution]

Example 16.7 (Variance for a coin toss). Let X represent the result of tossing a coin. $X = 1$ means a head (success) $X = 0$ means a tail. Assume $\mathbf{P}(X = 1) = p$.

Then $\mathbf{E}[X] = 1 \cdot \mathbf{P}(X = 1) + 0 \cdot \mathbf{P}(X = 0) = p$.

Since $X = X^2$ for this random variable, $\mathbf{E}[X^2] = p$ also.

By equation (16.2),

$$\mathbf{Var}(X) = p - p^2 = p(1 - p). \tag{16.7}$$

Example 16.8 (Variance of a binomial random variable). Let S_n be the number of successes in n tosses of a coin, when the coin has success probability p. Using equation (9.3), we will show that

$$\mathbf{Var}(S_n) = np(1 - p). \tag{16.8}$$

However, recall that we used additivity in section 10.5.1 to find mean values easily. This suggests that hammering away with equation (9.3) may not the easiest way to calculate $\mathbf{Var}(S_n)$.

So you may want to leave the justification of equation (16.8) until you have learned the general formula for the variance of a sum of independent random variables, which is given in equation (16.30). Exercise 16.15 asks you to derive equation (9.3) using that formula. (Incidentally, the proof of Lemma 12.14 already used the method of Exercise 16.15!)

Nevertheless, for those who are interested, here's the algebraic calculation for justifying equation (16.8), using equation (9.3).

We can learn from Exercise 10.9, where we found

$$\mathbf{E}[S_n] = np,$$

using algebraic manipulations. We will extend that method here.

By Theorem 10.8,

$$\mathbf{E}\left[S_n^2\right] = \sum_{k=0}^{n} k^2 \mathbf{P}(S_n = k).$$

Looking at the solution for Exercise 10.9, it seemed to depend on cancelling out the factor k from $k!$. But here we seem to need to remove a factor k^2 from $k!$. It's not clear how to do that.

After some meditation, we decide to find a related quantity, namely $\mathbf{E}\left[S_n\left(S_n-1\right)\right]$.

By the formula for the expectation of a function of a random variable, Theorem 10.8, we have

$$\mathbf{E}\left[S_n\left(S_n-1\right)\right] = \sum_{k=0}^{n} k(k-1)\mathbf{P}(S_n = k) = \sum_{k=0}^{n} k(k-1)p^k \frac{n!}{k!(n-k)!}.$$

Thus

$$\mathbf{E}\left[S_n\left(S_n-1\right)\right] = \sum_{k=2}^{n} k(k-1)p^k \frac{n!}{k!(n-k)!},$$

since the $k=0$ and $k=1$ terms are zero. So

$$\mathbf{E}\left[S_n\left(S_n-1\right)\right] = \sum_{k=2}^{n} p^k \frac{n!}{(k-2)!(n-k)!} = p^2 n(n-1) \sum_{k=2}^{n} p^{k-2} \frac{(n-2)!}{(k-2)!(n-k)!}.$$

We notice that $(n-k)! = (\,(n-2)-(k-2)\,)!$. This suggests replacing $k-2$ by j in the sum. Using equation (9.3) with n replaced by $n-2$, we obtain

$$\mathbf{E}\left[S_n\left(S_n-1\right)\right] = p^2 n(n-1) \sum_{j=0}^{n-2} p^j \frac{(n-2)!}{j!(\,(n-2)-j\,)!} = p^2 n(n-1) \sum_{j=0}^{n-2} \mathbf{P}(S_{n-2} = j).$$

Since the range of S_{n-2} consists of the numbers $j = 0, 1, \ldots, n-2$, we know that

$$\sum_{j=0}^{n-2} \mathbf{P}(S_{n-2} = j) = 1.$$

Hence $\mathbf{E}\left[S_n\left(S_n-1\right)\right] = p^2 n(n-1)$. That is, $\mathbf{E}\left[S_n^2\right] - \mathbf{E}\left[S_n\right] = p^2 n(n-1)$. Hence

$$\mathbf{E}\left[S_n^2\right] = \mathbf{E}\left[S_n\right] + p^2 n(n-1) = pn + p^2 n^2 - p^2 n = p(1-p)n + p^2 n^2.$$

By equation (16.2),

$$\mathbf{Var}\left(S_n\right) = \mathbf{E}\left[S_n^2\right] - \left(\mathbf{E}\left[S_n\right]\right)^2 = p(1-p)n.$$

Example 16.9 (Variance of a geometric distribution). Let T be the time of first success in ∞ Bernoulli trials, when the success probability on each trial is p.

T is defined in section 13.2. We will show that $\mathbf{Var}\left(T\right)$ is given by

$$\mathbf{Var}\left(T\right) = \frac{q}{p^2}. \tag{16.9}$$

First note that by equation 13.19,

$$\mathbf{E}\left[T\right] = \frac{1}{p}.$$

To get more information, we're going to use the trick of differentiating a series term-by-term.

We used that trick (for a finite series) in one of the derivations of the expected value of T_n (see equation (13.8)).

In Exercise 13.4 we used the differentiation trick as one method to find $\mathbf{E}\left[T\right]$.

Now we want to use the same trick here, to get $\mathbf{E}\left[T^2\right]$.

Using the formula for the sum of a geometric series, we know that for $x \in (-1, 1)$,

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + x^4 + x^5 + \ldots.$$

Differentiating term-by-term,

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + 4x^3 + 5x^4 + \ldots.$$

Differentiating term-by-term *again*, we have

$$\frac{2}{(1-x)^3} = 2 \cdot 1 + 3 \cdot 2x^1 + 4 \cdot 3x^2 + 5 \cdot 4x^3 + \ldots.$$

Setting $x = q$ we have

$$\frac{2}{p^3} = 2 \cdot 1 + 3 \cdot 2q + 4 \cdot 3q^2 + 5 \cdot 4q^3 + \ldots = \sum_{k=2}^{\infty} k(k-1)q^{k-2}.$$

Then

$$\frac{2}{p^3} = \sum_{k=1}^{\infty} k(k-1)q^{k-2},$$

because the first term of this series is zero. Multiplying by pq,

$$\frac{2q}{p^2} = \sum_{k=1}^{\infty} k(k-1)q^{k-1}p = \sum_{k=1}^{\infty} k(k-1)\mathbf{P}(T=k).$$

By the formula for the expectation of a function of a random variable, Theorem 14.11 (which is the generalization of Theorem 10.8 to the countable-range case),

$$\sum_{k=1}^{\infty} k(k-1)\mathbf{P}(T=k) = \mathbf{E}\left[T(T-s)\right].$$

Thus we have shown that

$$\frac{2q}{p^2} = \mathbf{E}\left[T(T-1)\right] = \mathbf{E}\left[T^2\right] - \mathbf{E}\left[T\right] = \mathbf{E}\left[T^2\right] - \frac{1}{p}.$$

Thus

$$\mathbf{E}\left[T^2\right] = \frac{2q}{p^2} + \frac{1}{p}.$$

By equation (16.2),

$$\mathbf{Var}\left(T\right) = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q+p-1}{p^2} = \frac{q+q+p-1}{p^2} = \frac{q}{p^2}. \tag{16.10}$$

Exercise 16.7 (Variance of an exponential waiting time). Let T be the exponential waiting time (Definition 15.1). Find $\mathbf{Var}\left(T\right)$.

[Solution]

For every nonnegative integer n, by linearity we have

$$\mathbf{E}\left[(cX)^n\right] = c^n \mathbf{E}\left[X\right], \ \mathbf{E}\left[|cX|^n\right] = |c|^n \mathbf{E}\left[|X|\right]. \tag{16.11}$$

Which is more a meaningful measure of deviation: $\mathbf{E}\left[\,|X-\mu|\,\right]$ or the variance, $\mathbf{E}\left[(X-\mu)^2\right]$?

Either one can be larger than the other. Which is more significant in a practical situation may depend on whether you consider that a few large deviations should be regarded as more important than a large number of smaller deviations.

We will see later that $\mathbf{Var}\left(X\right)$ is the most useful measure of deviation for theoretical purposes.

Example 16.10 (A case where $\mathbf{E}\left[X^2\right] = \left(\mathbf{E}\left[\,|X|\,\right]\right)^2$). Figure 16.1 and 16.2 show examples of random variables X, Y on $\Omega = [0,1]$. The probability on Ω is assumed to be uniform.

You can check that $\mathbf{E}\left[\,|X|\,\right] = \mathbf{E}\left[\,|Y|\,\right]$, $\mathbf{E}\left[X^2\right] = \left(\mathbf{E}\left[\,|X|\,\right]\right)^2$, and $\mathbf{E}\left[Y^2\right] > \left(\mathbf{E}\left[\,|Y|\,\right]\right)^2$.

X is such that equality holds in equation (16.3). Notice that $|X|$ is constant. Using equation (16.2) and Remark 16.6 you can show that this is not a coincidence.

Exercise 16.8. Prove that

$$\mathbf{Var}\left(X\right) - \mathbf{Var}\left(|X|\right) = \left(\mathbf{E}\left[\,|X|\,\right]\right)^2 - \left(\mathbf{E}\left[X\right]\right)^2. \tag{16.12}$$

[Solution]

We showed long ago, in equation (10.31), that

$$\left|\,\mathbf{E}\left[X\right]\,\right| \leq \mathbf{E}\left[\,|X|\,\right].$$

Using this fact and equation (16.12) gives us another inequality:

$$\mathbf{Var}\left(|X|\right) \leq \mathbf{Var}\left(X\right). \tag{16.13}$$

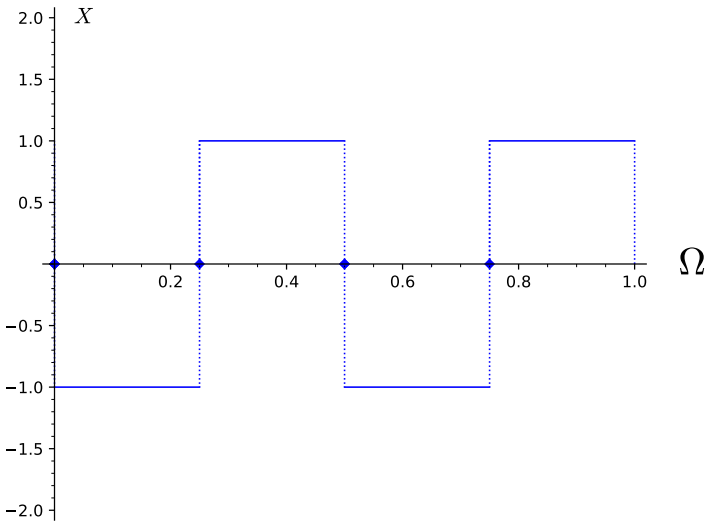


Figure 16.1: Unusual case: square of centered absolute first moment equals the variance.

Exercise 16.9 (A minimum property for the variance). Let X be any random variable such that $\mathbf{E}[X^2]$ exists, and let c be any real number. Let $\mu = \mathbf{E}[X]$. After writing $X - c$ as $(X - \mu) + (\mu - c)$, show that

$$\mathbf{E}\left[(X-c)^2\right] = \mathbf{Var}(X) + (\mu - c)^2. \tag{16.14}$$

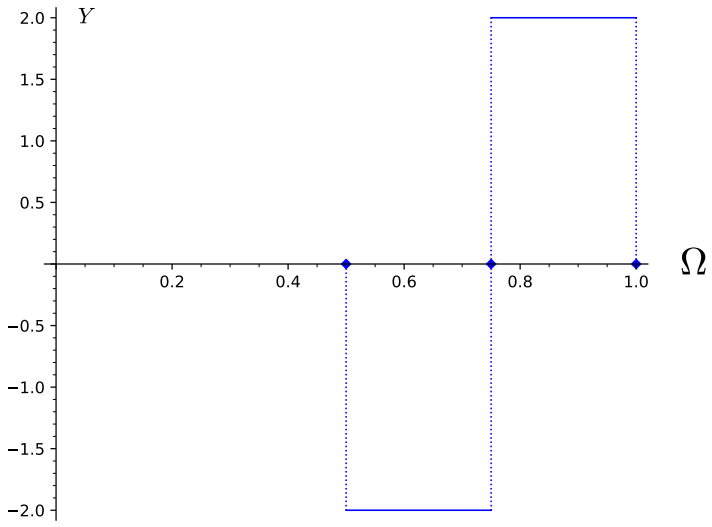
Notice that equation (16.14) tells us that the mean square deviation of X from c is smallest when $c = \mu$. If you must describe the distribution of X by a single number, this suggests that μ is the best choice.

[Solution]

Exercise 16.10. Let X_1 and X_2 be independent random variables with the same distribution. Suppose that $\mathbf{E}[X_i^2]$ exists. Prove that

$$\mathbf{E}\left[(X_1 - X_2)^2\right] = 2\mathbf{Var}(X_i). \tag{16.15}$$

[Solution]



(a) Y

Figure 16.2: Typical case: square of centered absolute first moment less than variance.

16.3 The Chebyshev Inequality

We can use the Markov inequality (Lemma 12.15) to estimate the probability that a random variable deviates from its mean, as follows.

Lemma 16.11 (Chebyshev's Inequality). Let X be a random variable such that the mean and variance of X exist. Then for any real number $a > 0$,

$$\mathbf{P}\left(|X - \mathbf{E}[X]| \geq a\right) \leq \frac{\mathbf{Var}(X)}{a^2}. \tag{16.16}$$

Proof. Let Y be the square of the deviation from the mean, i.e. $Y = (X - \mathbf{E}[X])^2$. Then $\mathbf{Var}(X) = \mathbf{E}[Y]$. By the Markov inequality,

$$a^2\mathbf{P}\left(Y \geq a^2\right) \leq \mathbf{E}[Y].$$

Since $\{Y \geq a^2\} = \{(X - \mathbf{E}[X])^2 \geq a^2\} = \{|X - \mathbf{E}[X]| \geq a\}$ and $\mathbf{E}[Y] = \mathbf{Var}(X)$, this gives equation (16.16).

□

Remark 16.12 (Chebyshev and the search for Charlie). Now that we have stated the Chebyshev inequality, we can see that Exercise 12.6 is a typical application of that inequality.

Recall that in the solution to Exercise 12.6, we applied the Markov inequality to obtain an estimate for $\mathbf{P}(|S_n| \geq 500)$.

This is the same as estimating $\mathbf{P}(S_n^2 \geq 250000)$, and using the Markov inequality we had:

$$\mathbf{P}(|S_n| \geq 500) = \mathbf{P}(S_n^2 \geq 250000) \leq \frac{\mathbf{E}\left[S_n^2\right]}{250000}. \tag{16.17}$$

Let $\mu = \mathbf{E}\left[S_n\right]$. In this problem $\mathbf{E}\left[S_n\right] = 0$, so $\mathbf{Var}\left(S_n\right) = \mathbf{E}\left[S_n^2\right]$, and

$$\mathbf{P}(|S_n| \geq 500) = \mathbf{P}(|S_n - \mu| \geq 500).$$

Thus equation (16.17) is exactly the estimate given by the Chebyshev inequality:

$$\mathbf{P}(|S_n - \mu| \geq 500) \leq \frac{\mathbf{Var}\left(S_n\right)}{250000}.$$

With the usual notation of μ for $\mathbf{E}\left[X\right]$ and σ for the standard deviation of X, σ^2 is the variance of X. One often writes Chebyshev's inequality as

$$\mathbf{P}\left(|X - \mu| \geq a\right) \leq \frac{\sigma^2}{a^2}. \tag{16.18}$$

Equation (16.18) suggests that it might be useful to measure deviation from the mean in units of the standard deviation σ. Thus we can rephrase the Chebyshev inequality as:

$$\mathbf{P}\left(|X - \mu| \geq c\sigma\right) \leq \frac{1}{c^2}. \tag{16.19}$$

Note that the estimate in this equation does not depend on the value of σ.

Exercise 16.11. Let W be a random variable such that $\mathbf{P}(W = 3) = 1/18$, $\mathbf{P}(W = -3) = 1/18$, and $\mathbf{P}(W = 0) = 8/9$. Find the mean and standard deviation of W. Find the probability that W deviates from its mean by at least three standard deviations.

Note that for this particular random variable, the probability you found is not very small.

Compare your answer with the estimate obtained using the Chebyshev inequality.

Now repeat these steps for the probability that W deviates from its mean by at least 3.1 standard deviations

[Solution]

Exercise 16.12. Let Y be a random variable on $[-2, 2]$ defined by $Y(t) = t$. With $\mathbf{P}$ equal to the uniform distribution on $[-2, 2]$, find the mean, variance and standard deviation of Y.

Also find the probability that Y deviates from its mean by at least three standard deviations.

[Solution]

16.4 Covariance of two real-valued random variables

We observed in Section 16.1 that calculating the moments of a random variable can help us to understand its distribution. When dealing with two random variables X, Y, we can of course calculate the means, moments and centered moments of X and Y separately. But it is also useful to have quantities which tell us about the relation between X and Y. One such quantity is $\mathbf{E}[XY]$. For example, if $\mathbf{E}[XY] \neq \mathbf{E}[X]\,\mathbf{E}[Y]$ we at least know that X and Y are not independent.

We can learn more by calculating the mean of the product of the centered random variables, which is referred to as the *covariance* of X, Y.

Definition 16.13 (Covariance of two random variables). For any random variables X, Y, the covariance of X, Y is denoted by $\mathbf{Cov}(X, Y)$, and is defined by

$$\mathbf{Cov}(X, Y) = \mathbf{E}\left[(X - \mathbf{E}[X])\,(Y - \mathbf{E}[Y])\right], \tag{16.20}$$

provided that the expected values exist.

As usual, if $\mathbf{E}[X] = \mu$ and $\mathbf{E}[Y] = \nu$, we can write

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mu)(Y - \nu)], \tag{16.21}$$

Remark 16.14 (Existence of $\mathbf{Cov}(X, Y)$). The comparison principle can be used to show that $\mathbf{Cov}(X, Y)$ exists if $\mathbf{E}[X]$, $\mathbf{E}[Y]$ and $\mathbf{E}[XY]$ exist.

By Lemma 16.29, $\mathbf{E}[XY]$ exists if $\mathbf{E}[X^2]$ exists and $\mathbf{E}[Y^2]$ exists.

Assuming $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ exist, $\mathbf{E}[X^2]$ exists if and only if $\mathbf{E}[(X - \mu)^2]$ exists. Thus $\mathbf{Cov}(X, Y)$ exists if $\mathbf{Var}(X)$ exists and $\mathbf{Var}(Y)$ exist.

Since

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mu)(Y - \nu)] = \mathbf{E}[(X(Y - \nu)] - \mu\mathbf{E}[Y - \nu],$$

and $\mathbf{E}[Y - \nu] = 0$, we have

$$\mathbf{Cov}(X, Y) = \mathbf{E}[X(Y - \nu)]. \tag{16.22}$$

Similarly

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mu)Y]. \tag{16.23}$$

Thus it is only necessary to center one of the two random variables when calculating the covariance.

From the definition of covariance,

$$\mathbf{Cov}(X, X) = \mathbf{Var}(X). \tag{16.24}$$

Much as in Exercise 16.5, shifting random variables by constants has no effect on covariance:

$$\mathbf{Cov}(X - a, Y - b) = \mathbf{Cov}(X, Y). \tag{16.25}$$

Writing variance as mean square minus square mean is often useful. Covariance has a similar property.

Exercise 16.13 (Mean product minus product mean). Generalize Exercise 16.2, by proving that

$$\mathbf{Cov}\left(X,Y\right)=\mathbf{E}\left[XY\right]-\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]. \tag{16.26}$$

[Solution]

Lemma 16.15 (Expanding the variance of a sum). Let X and Y be any random variables such that $\mathbf{E}\left[X\right],\mathbf{E}\left[Y\right]$ exist. If $\mathbf{Var}\left(X\right),\mathbf{Var}\left(Y\right)$ exist, then $\mathbf{Var}\left(X+Y\right)$ exists. Furthermore $\mathbf{Cov}\left(X,Y\right)$ exists, and

$$\mathbf{Var}\left(X+Y\right)=\mathbf{Var}\left(X\right)+2\mathbf{Cov}\left(X,Y\right)+\mathbf{Var}\left(Y\right). \tag{16.27}$$

Thus $2\mathbf{Cov}\left(X,Y\right)$ is the "cross term" in the expansion of $\mathbf{Var}\left(X+Y\right)$.

Proof. We have already discussed existence.

The algebra is routine:

$$\begin{aligned}\mathbf{Var}\left(X+Y\right)&=\mathbf{E}\left[\left(X+Y-(\mu+\nu)\right)^2\right]=\mathbf{E}\left[\left(\left(X-\mu\right)+\left(Y-\nu\right)\right)^2\right]\\&=\mathbf{E}\left[\left(X-\mu\right)^2\right]+2\mathbf{E}\left[\left(X-\mu\right)(Y-\nu)\right]+\mathbf{E}\left[\left(Y-\nu\right)^2\right]\\&=\mathbf{Var}\left(X\right)+2\mathbf{E}\left[\left(X-\mu\right)(Y-\nu)\right]+\mathbf{Var}\left(Y\right)\\&=\mathbf{Var}\left(X\right)+2\mathbf{Cov}\left(X,Y\right)+\mathbf{Var}\left(Y\right)\end{aligned} \tag{16.28}$$

□

The random variables X,Y in equation (16.27) could be independent. In that case the next lemma says that their covariance is zero.

Lemma 16.16 (Independence implies zero covariance). Let X,Y be any independent random variables such that $\mathbf{E}\left[X\right]$ exists and $\mathbf{E}\left[Y\right]$ exists. Then $\mathbf{Cov}\left(X,Y\right)$ exists, and $\mathbf{Cov}\left(X,Y\right)=0$.

Proof. By additivity, $\mathbf{E}[X - \mu]$ exists and $\mathbf{E}[Y - \nu]$ exists, and these expectations are zero.

By Lemma 12.10, $X - \mu$ and $Y - \nu$ are independent.

By Theorem 12.11, $\mathbf{E}[(X - \mu)(Y - \nu)]$ exists, and $\mathbf{E}[(X - \mu)(Y - \nu)] = \mathbf{E}[X - \mu]\,\mathbf{E}[Y - \nu] = 0 \cdot 0 = 0$.

And by definition $\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mu)(Y - \nu)]$.

□

Corollary 16.17 (Additivity of variance for independent). Let X and Y be independent random variables whose variances exist. Then the variance of $X + Y$ exists, and $\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$.

Proof. Apply Lemma 16.16 to equation (16.27).

□

The next lemma extends Lemma 16.15, with the same proof.

Lemma 16.18 (Expanding the variance of a sum of n random variables). Let $X_1, \ldots, X_n$ be real-valued random variables. Assume that the mean μ_i of each X_i exists, and the variance $\mathbf{Var}(X_i)$ of each X_i exists. Let $S_n = X_1 + \ldots + X_n$.

Then $\mathbf{Var}(S_n)$ exists, and

$$\mathbf{Var}(S_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{Cov}(X_i, X_j). \tag{16.29}$$

Of course $\mathbf{Cov}(X_i, X_i) = \mathbf{Var}(X_i)$ for each i (equation (16.24)).

If it happens that $X_1, \ldots, X_n$ is an independent sequence, then $\mathbf{Cov}(X_i, X_j) = 0$ for all $i \neq j$, and

$$\mathbf{Var}(S_n) = \mathbf{Var}(X_1) + \ldots + \mathbf{Var}(X_n). \tag{16.30}$$

Exercise 16.14. Write out the proof of Lemma 16.18.

[Solution]

Exercise 16.15 (Variance of a binomial random variable revisited). Use equation (16.30) to prove equation (16.8) in an efficient manner.

As in section 10.5.1, start by writing $S_n = X_1 + \ldots + X_n$.

Your goal is to show that

$$\mathbf{Var}\,(S_n) = n(1-p)p. \tag{16.31}$$

[Solution]

Remark 16.19 (Random walk again). Section 12.6 introduced random walk. In simple symmetric random walk, the "walker" takes independent steps X_i, where $\mathbf{P}(X_i = 1) = 1/2 = \mathbf{P}(X_i = -1)$. In this case $\mathbf{E}\,[X_i] = 0$ and $\mathbf{E}\,[X_i^2] = \mathbf{E}\,[1] = 1$. From the definition, $\mathbf{Var}\,(X_i) = 1$.

As in equation (12.16), let

$$S_n = X_1 + ... + X_n \text{ for each } n = 1, 2, ...$$

Then $\mathbf{E}\,[S_n] = 0$, so from the definition $\mathbf{E}\,[S_n^2] = \mathbf{Var}\,(S_n^2)$.

Applying equation (16.30),

$$\mathbf{Var}\,(S_n) = \mathbf{Var}\,(X_1) + \ldots + \mathbf{Var}\,(X_n) = 1 + \ldots + 1 = n.$$

Thus we have proved Lemma 12.14. Of course this is not a new proof. We have taken the idea of Lemma 12.14 and generalized it to obtain a powerful tool, stated in equation (16.30).

Life would be simpler if the converse to Corollary 16.17 were true. However, it ain't.

Example 16.20. Consider tossing a fair coin twice, and then rolling a fair die.

Let $G = 1$ if the first toss gives success, and $G = -1$ otherwise.

Let $H = 1$ if the second toss gives success, and $H = -1$ otherwise.

Let K be the result of rolling the fair die, so $\mathbf{P}(K = i) = 1/6$ for $i = 1, \ldots, 6$.

It is clear physically that G, H, K is an independent sequence of random variables.

Notice that

$$\mathbf{P}(GK = 6) = \mathbf{P}(G = 1)\mathbf{P}(K = 6) = \frac{1}{12}.$$

Similarly

$$\mathbf{P}(HK = 1) = \mathbf{P}(H = 1)\mathbf{P}(K = 6) = \frac{1}{12}.$$

However,

$$\mathbf{P}(HK = 1 \,|\, GK = 6) = 0,$$

since the result of the roll of the die cannot be both equal to 6 and equal to 1.

Thus GK and HK are *not* independent.

On the other hand, $\mathbf{E}[GK] = \mathbf{E}[G]\,\mathbf{E}[K] = 0$, and $\mathbf{E}[HK] = \mathbf{E}[H]\,\mathbf{E}[K] = 0$, so

$$\mathbf{Cov}(GK, HK) = \mathbf{E}\left[GHK^2\right] = \mathbf{E}[GH]\,\mathbf{E}\left[K^2\right] = \mathbf{E}[G]\,\mathbf{E}[H]\,\mathbf{E}\left[H^2\right] = 0.$$

Although covariance zero does not imply independence, we often think of covariance as a rough measure of the degree of dependence between two random variables.

Here's some terminology.

Definition 16.21 (Uncorrelated random variables). If $\mathbf{Cov}(X, Y)$ exists and $\mathbf{Cov}(X, Y) = 0$ then X, Y are said to be uncorrelated.

The word "uncorrelated" is also used in ordinary language, with a less precise meaning. As usual, one must judge what is meant from the context.

Exercise 16.16. Let X and Y be mean zero random variables such that $|X| = |Y| = 1$ everywhere. Suppose that $\mathbf{Cov}(X, Y) < 0$. Show that

$$\mathbf{P}(X = Y) < \mathbf{P}(X \neq Y).$$

[Solution]

16.5 The Weak Law of Large Numbers

We mentioned earlier that the frequency interpretation of probability does not tell us how many repetitions of an experiment are likely to be needed in order to reliably estimate a probability value using an average value. More generally, the frequency interpretation of expected value (Fact 10.1) has the same deficiency. Theorem 16.22 in this section allows us to make these frequency statements a little more precise.

Let X be a mathematical random variable which represents some property of an experiment. Let $X_1, \ldots, X_n$ be independent random variables with the same distribution as X, and let $S_n = X_1 + \ldots + X_n$. Then S_n/n represents the average measured value for the property in n repetitions of the experiment. We would like to know whether $\mathbf{E}[X]$ is a reliable estimate for the average measured value of the property. In other words, how likely is it that S_n/n deviates significantly from $\mathbf{E}[X]$?

The next theorem attempts to answer this question, with "significantly" interpreted as "by more than ε" and "likely" expressed as a probability.

Theorem 16.22 (The Weak Law of Large Numbers). Let $X_1, \ldots, X_n$ be independent random variables on some sample space, such that each X_i has the same mean μ and the same standard deviation σ. Let $S_n = X_1 + \ldots + X_n$. Then

$$\mathbf{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}. \tag{16.32}$$

Proof. For each i, $\mathbf{Var}(X_i) = \sigma^2$.

By equation equation (16.30), $\mathbf{Var}(S_n) = n\sigma^2$.

Thus $\mathbf{Var}(S_n/n) = (1/n^2)\mathbf{Var}(S_n) = \sigma^2/n$ (using equation (16.5)).

Chebyshev's inequality (equation (16.16)) then gives equation (16.32). □

When using equation (16.32), it is important to remember that in σ is the standard deviation of each X_i, not the standard deviation of the random variable S_n/n.

The variance and standard deviation of S_n are given by

$$\mathbf{Var}(S_n) = n\sigma^2, \ \sqrt{\mathbf{Var}(S_n)} = \sqrt{n}\sigma, \tag{16.33}$$

so

$$\mathbf{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n},\ \sqrt{\mathbf{Var}\left(\frac{S_n}{n}\right)} = \frac{\sigma}{\sqrt{n}}. \tag{16.34}$$

The Weak Law of Large Numbers is important theoretically, since it removes some of the vagueness from the frequency interpretation of expected value. For practical purposes, one usually finds that the bound for the probability of error which is given in equation (16.32) is not very precise, i.e. the actual probability is considerably smaller.

We mentioned earlier that there is another mathematical law of large numbers, called the Strong Law of Large Numbers. From the name one might hope that the Strong Law would give a better estimate. Unfortunately, although the Strong Law is indeed stronger for theoretical purposes, it does nothing to improve the probability estimate. We will need another approach, such as the Central Limit Theorem ([10]), which is studied in Chapter 18.

16.6 Covariance is bilinear

If you wish to calculate the covariance of random variables which are given as algebraic expressions in terms of other random variables, you can always use the definition of covariance in terms of products. However, it may be simpler to use the algebraic properties of covariance directly.

The general concept of a *linear operation* was defined in Definition 10.12. An operation is linear if one can take sums "through" the operation, and one can also take multiplication by a constant through the operation. Now we introduce the general concept of a *bilinear operation*.

Definition 16.23 (Bilinear operations). A bilinear operation is an operation on two elements $\mathbf{x}$ and $\mathbf{y}$ which depends linearly on $\mathbf{x}$ when $\mathbf{y}$ is fixed, and depends linearly on $\mathbf{y}$ when $\mathbf{x}$ is fixed.

Covariance is defined in terms of expectations of products, and it is bilinear in the sense of Definition 16.23. That is, $\mathbf{Cov}(X, Y)$ is a linear function of X when Y is held fixed, and $\mathbf{Cov}(X, Y)$ is a linear function of Y when X is held fixed. We state this formally in the next lemma.

Lemma 16.24 (Covariance is a bilinear function). Covariance is bilinear, so that for any random variables X, Y, Z, and any numbers c_1, c_2,

$$\begin{aligned} \mathbf{Cov}\left(c_1X + c_2Y, Z\right) &= c_1\mathbf{Cov}\left(X, Z\right) + c_2\mathbf{Cov}\left(Y, Z\right), \text{ and} \\ \mathbf{Cov}\left(Z, c_1X + c_2Y\right) &= c_1\mathbf{Cov}\left(Z, X\right) + c_2\mathbf{Cov}\left(Z, Y\right). \end{aligned} \tag{16.35}$$

Of course the second equality in equation (16.35) is redundant, since covariance is clearly a *symmetric* operation:

$$\mathbf{Cov}\left(X, Y\right) = \mathbf{Cov}\left(Y, X\right) \tag{16.36}$$

Exercise 16.17. Prove the first equality in equation (16.35).
[Solution]

Multiplication of two numbers is a simple example of a bilinear operation. The "bilinear" property in this case is just another way of describing the distributive law. Our familiarity with the algebra of numbers makes it easy for us to use the bilinear property for other operations, such as covariance of random variables, or the dot product of vectors.

Example 16.25. The algebra needed to derive equation (16.29) was given in the solution to Exercise 16.14. If we want to make use of the bilinear property of covariance in the same proof, we would write the same manipulations a bit differently.

$$\mathbf{Var}\left(S_n\right) = \mathbf{Cov}\left(S_n, S_n\right) = \mathbf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right).$$

Then, using bilinearity as much as possible, again we arrive at equation (16.29):

$$\mathbf{Var}\left(S_n\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{Cov}\left(X_i, X_j\right).$$

Using bilinearity is bit shorter, but not dramatically so. Still, you will likely find that de-cluttering equations helps to clarify your work.

Exercise 16.18. Let X, Y, Z be independent random variables, such that $\mathbf{Var}(X) = 1$, $\mathbf{Var}(Y) = 2$ and $\mathbf{Var}(Z) = 3$.

Calculate $\mathbf{Cov}(5X - Y + Z, X + 3Y - Z)$.

[Solution]

16.7 Variance of a hypergeometric random variable

Let $L_{N,K,n}$ be the random variable defined in Definition 9.10, so that $L_{N,K,n}$ has a hypergeometric distribution with parameters N, K, n.

The experiment consists of randomly selecting n objects for a set S of N objects, where a certain target set T of K objects has been specified. $L_{N,K,n}$ is the number of selected objects which lie in the target set T.

We have already dealt with $\mathbf{E}[L_{N,K,n}]$ in Section 10.5.2, using Method 1 of that section.

We found that

$$\mathbf{E}[L_{N,K,n}] = n\frac{K}{N}.$$

Consider the case that $n = 1$ in this equation. Since $\mathbf{E}[L_{N,K,1}]$ is either zero or one, $\frac{K}{N}$ is also the probability that when a single element is randomly selected, it will lie in the target set T. Let us denote this probability by p. We will say it is the *success probability* when choosing a *single* element.

In the present example we wish to calculate $\mathbf{Var}(L_{N,K,n})$ for all n. The properties of the covariance function will allow us to do that fairly efficiently.

For each member σ of S, let X_σ be equal to 1 if σ is in the selected subset of n elements, and $X_\sigma = 0$ otherwise. By symmetry, $\mathbf{E}[X_\sigma]$ is the same for all σ and $\mathbf{Var}(X_\sigma)$ is the same for all σ. Let $\mathbf{E}[X_\sigma] = \mu$ and let $\mathbf{Var}(X_\sigma) = v$. Of course $\mathbf{Cov}(X_\sigma, X_\sigma) = \mathbf{Var}(X_\sigma) = v$.

Let

$$Z = \sum_{\sigma \in S} X_\sigma.$$

From the description of the experiment, Z is constant and $Z = n$. Hence $\mathbf{E}[Z] = n$ (and $\mathbf{Var}(Z) = 0$).

It follows that $\mathbf{E}[Z] = N\mu$, so $N\mu = n$, so

$$\mu = \frac{n}{N}. \tag{16.37}$$

Since the possible values of X_σ are 0 and 1, we see that $\mathbf{P}(X_\sigma = 1) = \mu$ for each σ.

Since Z is constant, $\mathbf{Var}(Z) = 0$, i.e. $\mathbf{Cov}(Z, Z) = 0$:

$$\mathbf{Cov}\left(\sum_{\sigma\in S} X_\sigma, \sum_{\tau\in S} X_\tau\right) = 0.$$

Expanding using bilinearity, this gives

$$\sum_{\sigma,\tau\in S} \mathbf{Cov}(X_\sigma, X_\tau) = 0.$$

Grouping like terms,

$$\sum_{\sigma\in S} \mathbf{Cov}(X_\sigma, X_\sigma) \; + \sum_{\sigma,\tau\in S,\ \sigma\neq\tau} \mathbf{Cov}(X_\sigma, X_\tau) = 0. \tag{16.38}$$

Since $\mathbf{E}[X_\sigma] = \mu$, we know that $\mathbf{P}(X_\sigma = 1) = \mu$, so

$$v = \mathbf{Var}(X_\sigma) = \mathbf{E}\left[X_\sigma^2\right] - (\mathbf{E}[X_\sigma])^2 = \mu - \mu^2 = \mu(1-\mu),$$

as usual. Thus

$$\sum_{\sigma\in S} \mathbf{Cov}(X_\sigma, X_\sigma) = N\mu(1-\mu).$$

By symmetry, $\mathbf{Cov}(X_\sigma, X_\tau)$ is the same for any σ, τ with $\sigma \neq \tau$. Call this number c. There are $N(N-1)$ choices for σ, τ with $\sigma \neq \tau$ (N ways to choose σ, and then, for that σ, $N-1$ ways to choose τ). Hence we have

$$\sum_{\sigma,\tau\in S,\ \sigma\neq\tau} \mathbf{Cov}(X_\sigma, X_\tau) = N(N-1)c.$$

Substituting in equation (16.38),

$$N\mu(1-\mu) + N(N-1)c = 0,$$

and so

$$c = -\frac{\mu(1-\mu)}{N-1}. \tag{16.39}$$

From the definitions,

$$L_{N,K,n} = \sum_{\sigma \in T} X_\sigma.$$

Since $\mathbf{Var}\,(L_{N,K,n}) = \mathbf{Cov}\,(L_{N,K,n}, L_{N,K,n})$, by expanding and grouping like terms we have

$$\mathbf{Var}\,(L_{N,K,n}) = \sum_{\sigma \in T} \mathbf{Cov}\,(X_\sigma, X_\sigma) + \sum_{\sigma,\tau \in T, \tau \neq \sigma} \mathbf{Cov}\,(X_\sigma, X_\tau)\,.$$

Thus

$$\mathbf{Var}\,(L_{N,K,n}) = K\mu(1-\mu) - \frac{K(K-1)\mu(1-\mu)}{N-1} = K\mu(1-\mu)\left(1 - \frac{K-1}{N-1}\right)$$
$$= K\frac{n(N-n)}{N^2}\left(\frac{N-K}{N-1}\right) = \frac{K}{N}\frac{N-K}{N}n\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right),$$

so we have obtained the following formula.

The variance of a hypergeometric random variable:

$$\mathbf{Var}\,(L_{N,K,n}) = np(1-p)\left(\frac{N-n}{N-1}\right) \tag{16.40}$$

Remark 16.26 (Smaller variance than S_n).

For $n > 1$, clearly $(N-n)/(N-1) < 1$. Thus equation (16.40) shows that for $n > 1$,

$$\mathbf{Var}\,(L_{N,K,n}) < np(1-p). \tag{16.41}$$

Notice that $np(1-p) = \mathbf{Var}\,(S_n)$, where S_n is the number of successes in n coin-tosses, when the coin has success probability p, Thus equation (16.41) says that $\mathbf{Var}\,(L_{N,K,n}) < \mathbf{Var}\,(S_n)$. This gives an answer to the question posed at the end of the solution for Exercise 8.9, which asks why the graph of the hypergeometric distribution should look narrower than the graph of corresponding binomial distribution.

16.8 Estimates for moments

In this section we discuss existence of moments, a topic that was raised Example 16.1. This leads to consideration of inequalities, an important topic in its own right.

Lemma 16.27 (Existence of lower moments). If $\mathbf{E}[X^n]$ exists then $\mathbf{E}\left[X^k\right]$ exists for all $k \leq n$.

Proof. Here's a handy fact: for any nonnegative integer $k \leq n$,

$$|x|^k \leq 1 + |x|^n , \tag{16.42}$$

for all x.

To check equation (16.42), consider two cases: $|x| \leq 1$ and $|x| > 1$. In the first case, $|x|^k \leq 1$, In the second case, $|x|^k \leq |x|^n$.

Since $|X|^k \leq 1 + |X|^n$, the statement of the lemma follows by the comparison principle for expected values (part (iv) of Theorem 14.9).

□

Lemma 16.28 (Moments of a sum). Let n be a positive integer. If $\mathbf{E}[|X|^n]$ exists and $\mathbf{E}[\,|Y|^n\,]$ exists, then $\mathbf{E}[\,|X+Y|^n\,]$ exists.

Proof. Claim: for any numbers x, y, and any positive integer n,

$$|x+y|^n \leq 2^n |x|^n + 2^n |y| . \tag{16.43}$$

To justify the claim, remember the *triangle inequality* (Appendix B): for any numbers x, y,

$$|x+y| \leq |x| + |y| . \tag{16.44}$$

So equation (16.43) certainly holds for $n = 1$. In fact, equation (16.43) is a cruder inequality than equation (16.44), isn't it? But hey, we ain't bein' paid to be fancy here. We just need to get an upper bound for $|x+y|^n$.

Anyway, for general n we have

$$|x+y|^n \leq (|x| + |y|)^n .$$

Now consider the case that $|x| \leq |y|$. In this case, $|x| + |y| \leq 2|y|$, and so

$$|x+y|^n \leq (2|y|)^n = 2^n |y|^n ,$$

so equation (16.43) holds.

The other possible case is that $|y| < |x|$, and of course a similar argument works there too.

This proves the claim.

By equation (16.43),

$$|X + Y|^n \leq 2^n |X|^n + 2^n |X|^n .$$

The expected value of $2^n |X|^n + 2^n |X|^n$ exists, by additivity (Theorem 14.9).

And then the comparison principle (part (iv) of Theorem 14.9) says that the expected value of $|X + Y|^n$ exists.

□

Lemma 16.28 dealt with existence of moments when we *add* random variables. Lemma 16.29 will help us deal with expected values of *products*.

Before going on to consider that lemma, please be sure to work through the next exercise. It provides us with a pleasing inequality that is often useful.

Exercise 16.19 (Inequality for a product). Let x, y be real numbers. Show that

$$2xy \leq x^2 + y^2. \tag{16.45}$$

You can begin by noting that $(x - y)^2 \geq 0$ is always true.

As long as you are showing that equation (16.45) holds, you might as well also show:

$$(x + y)^2 \leq 2x^2 + 2y^2. \tag{16.46}$$

[Solution]

Since the inequality in equation (16.45) holds for *all* x and y, we can of course replace x by $|X|$ and y by $|Y|$ in this equality, and obtain:

$$2 |X\, Y| \leq X^2 + Y^2. \tag{16.47}$$

By linearity, $\mathbf{E}\left[\frac{1}{2}(X^2 + Y^2)\right]$ exists. Hence by the the comparison principle, $\mathbf{E}\left[|X\, Y|\right]$ exists.

Using monotonicity, $\mathbf{E}\left[|X\, Y|\right] \leq \mathbf{E}\left[\frac{1}{2}(X^2 + Y^2)\right]$.

Linearity then gives

$$2\mathbf{E}\left[|X\,Y|\right] \leq \mathbf{E}\left[X^2\right] + \mathbf{E}\left[Y^2\right]. \tag{16.48}$$

Sjnce $\mathbf{E}\left[|X\,Y|\right]$ exists, by the definition of expected value we know that $\mathbf{E}\left[X\,Y\right]$ exists. By monotonicity, $\mathbf{E}\left[X\,Y\right] \leq \mathbf{E}\left[|X\,Y|\right]$.

We have proved the following.

Lemma 16.29 (Existence of the expectation of a product). If $\mathbf{E}\left[X^2\right]$ exists and $\mathbf{E}\left[Y^2\right]$ exists, then $\mathbf{E}\left[XY\right]$ and $\mathbf{E}\left[|X\,Y|\right]$ exist, and

$$2\mathbf{E}\left[XY\right] \leq 2\mathbf{E}\left[|X\,Y|\right] \leq \mathbf{E}\left[X^2\right] + \mathbf{E}\left[Y^2\right]. \tag{16.49}$$

Exercise 16.20 (First moment exists if second does). Show that if $\mathbf{E}\left[X^2\right]$ exists then $\mathbf{E}\left[X\right]$ exists. Do this in *two ways*: using Lemma 16.27 and using Lemma 16.29.

[Solution]

Inequalities play a crucial role in advanced mathematics, and they are also fun. Appendix O deals with the Schwarz inequality, which is often useful.

16.9 Solutions for Chapter 16

Solution (Exercise 16.1).

(i) By equation (15.4),

$$\mathbf{E}\left[X\right] = \int_0^1 \frac{1}{t}\,dt = \lim_{a \searrow 0} \int_a^1 \frac{1}{t}\,dt = \lim_{a \searrow 0} \log(t)\Big|_a^1 = \lim_{a \searrow 0} (\log 1 - \log a) = \infty.$$

Thus $\mathbf{E}\left[X^2\right]$ does not exist.

(Remember that in our work, log means logarithm to the base e.)

(ii)

$$\mathbf{E}[X] = \int_0^1 \frac{1}{\sqrt{t}}\,dt = \lim_{a\searrow 0}\int_a^1 \frac{1}{\sqrt{t}}\,dt = \lim_{a\searrow 0} 2\sqrt{t}\Big|_a^1 = \lim_{a\searrow 0} 2\left(\sqrt{1}-\sqrt{a}\right) = 2.$$

$$\mathbf{E}\left[X^2\right] = \int_0^1 \frac{1}{t}\,dt = \lim_{a\searrow 0}\int_a^1 \frac{1}{t}\,dt = \lim_{a\searrow 0} \log t\Big|_a^1 = \lim_{a\searrow 0} -\log a = \infty.$$

Thus $\mathbf{E}[X^2]$ does not exist.

(iii) For any positive integer k,

$$\mathbf{E}\left[X^k\right] = \sum_{n=1}^{\infty} n^k \frac{c}{n^5} = \sum_{n=1}^{\infty} \frac{c}{n^{5-k}}.$$

By the integral test, this series converges if and only if

$$\int_1^{\infty} \frac{1}{x^{5-k}}\,dx \text{ exists.}$$

Also

$$\int_1^{\infty} \frac{1}{x^{5-k}}\,dx = \lim_{b\to\infty}\int_1^b \frac{1}{x^{5-k}}\,dx.$$

For $k < 4$,

$$\lim_{b\to\infty}\int_1^b \frac{1}{x^{5-k}}\,dx = \lim_{b\to\infty} -\frac{1}{4-k}\left(\frac{1}{x^{4-k}}\right)\Big|_1^b = \lim_{b\to\infty}\frac{1}{4-k}\left(1-\frac{1}{b^{4-k}}\right) = \frac{1}{4-k}.$$

Thus $\mathbf{E}\left[X^k\right]$ exists for $k < 4$.

When $k = 4$,

$$\lim_{b\to\infty}\int_1^b \frac{1}{x^{5-k}}\,dx = \lim_{b\to\infty}\log(x)\Big|_1^b = \lim_{b\to\infty}(\log(b)-0) = \infty.$$

Thus $\mathbf{E}[X^4]$ does not exist.

Solution (Exercise 16.2).

$$\begin{aligned}\mathbf{Var}(X) = \mathbf{E}\left[(X-\mu)^2\right] = \mathbf{E}\left[X^2 - 2\mu X + \mu^2\right] &= \mathbf{E}\left[X^2\right] - 2\mu\mathbf{E}[X] + \mu^2 \\ &= \mathbf{E}\left[X^2\right] - 2\mu^2 + \mu^2 = \mathbf{E}\left[X^2\right] - \mu^2.\end{aligned}$$

Solution (Exercise 16.3).

$$\mathbf{E}\,[X] = \mathbf{E}\,[7] = 7.$$

$$\mathbf{Var}\,(X) = \mathbf{E}\left[(X-7)^2\right] = \mathbf{E}\left[(7-7)^2\right] = 0.$$

Solution (Exercise 16.4). Let $\mu = \mathbf{E}\,[X]$.

By linearity,

$$\mathbf{E}\,[cX] = c\mu,$$

and so

$$\mathbf{Var}\,(cX) = \mathbf{E}\left[\,(cX - c\mu)^2\,\right] = \mathbf{E}\left[\,c^2(X-\mu)^2\,\right] = c^2\mathbf{E}\left[\,(X-\mu)^2\,\right] = c^2\mathbf{Var}\,(X)\,.$$

Solution (Exercise 16.5). Let $\mu = \mathbf{E}\,[X]$.

$$\mathbf{E}\,[X-c] = \mathbf{E}\,[X] - \mathbf{E}\,[c] = \mu - c.$$

$$\mathbf{Var}\,(X-c) = \mathbf{E}\left[\,((X-c)-(\mu-c))^2\,\right] = \mathbf{E}\left[\,(X-\mu)^2\,\right] = \mathbf{Var}\,(X)\,.$$

Solution (Exercise 16.6).

$$\mathbf{E}\,[X] = \int_0^5 x\frac{1}{5}\,dx = \frac{1}{5}\frac{x^2}{2}\bigg|_0^5 = \frac{1}{5}\frac{1}{2}\left(5^2-0\right) = \frac{5}{2}.$$

Thus, as expected, the mean of the coordinate function is the midpoint.

$$\mathbf{E}\left[X^2\right] = \int_0^5 x^2\frac{1}{5}\,dx = \frac{1}{5}\frac{x^3}{3}\bigg|_0^5 = \frac{1}{5}\frac{1}{3}\left(5^3-0\right) = \frac{25}{3}.$$

By equation (16.2),

$$\mathbf{Var}\,(X) = \frac{25}{3} - \frac{25}{4} = \frac{25}{12}.$$

Now we generalize. Let X have a uniform distribution on $[s,t]$. An easy calculation shows that $\mathbf{E}\,[X]$ is the midpoint, i.e. $\mu = (s+t)/2$.

Also

$$\mathbf{Var}\,(X) = \mathbf{E}\left[\,(X-\mu)^2\,\right] = \int_s^t (u-\mu)^2\frac{1}{t-s}\,du = \frac{1}{t-s}\frac{1}{3}(u-\mu)^3\bigg|_s^t$$
$$= \frac{1}{t-s}\frac{1}{3}\left((t-\mu)^3 - (s-\mu)^3\right).$$

Of course $t-\mu = (t-s)/2$ and $s-\mu = -(t-s)/2$. Thus

$$\mathbf{Var}\,(X) = \frac{1}{t-s}\frac{2}{3}\frac{(t-s)^3}{8} = \frac{(t-s)^2}{12}. \tag{16.50}$$

Solution (Exercise 16.7). Let T have an exponential distribution with parameter λ.

By equation (15.9),

$$\mathbf{E}\,[T] = \frac{1}{\lambda}.$$

By the solution to Exercise 15.5,

$$\mathbf{E}\left[T^2\right] = \frac{2}{\lambda^2}.$$

Hence by equation (16.2),

$$\mathbf{Var}\,(T) = \frac{1}{\lambda^2}. \tag{16.51}$$

Solution (Exercise 16.8). By equation (16.2),

$$\mathbf{Var}\,(X) = \mathbf{E}\left[\,X^2\,\right] - (\,\mathbf{E}\,[X]\,)^2\,,$$

$$\mathbf{Var}\,(\,|X|\,) = \mathbf{E}\left[\,|X|^2\,\right] - (\,\mathbf{E}\,[\,|X|\,]\,)^2 = \mathbf{E}\left[\,X^2\,\right] - (\,\mathbf{E}\,[\,|X|\,]\,)^2\,.$$

Subtracting these equations gives equation (16.12).

Solution (Exercise 16.9).

$$\mathbf{E}\left[\,(X-c)^2\,\right] = \mathbf{E}\left[\,((X-\mu)+(\mu-c))^2\,\right] = \mathbf{E}\left[\,(X-\mu)^2 + 2(X-\mu)(\mu-c) + (\mu-c)^2\,\right].$$

Since $\mathbf{E}\,[\,2(X-\mu)(\mu-c)\,] = 2(\mu-c)\mathbf{E}\,[\,X-\mu\,] = 0$,

$$\mathbf{E}\left[\,(X-c)^2\,\right] = \mathbf{E}\left[\,(X-\mu)^2\,\right] + (\mu-c)^2 = \mathbf{Var}\,(X) + (\mu-c)^2.$$

Solution (Exercise 16.10). Let $\mathbf{E}\,[X_i] = \mu$. Using Theorem 12.11,

$$\begin{aligned}\mathbf{E}\left[(X_1-X_2)^2\right] &= \mathbf{E}\left[X_1^2 - 2X_1X_2 + X_2^2\right] \\ &= \mathbf{E}\left[X_1^2\right] - 2\mathbf{E}\,[X_1]\,\mathbf{E}\,[X_2] + \mathbf{E}\left[X_2^2\right] = 2\left(\mathbf{E}\left[X_i^2\right] - \mu^2\right).\end{aligned}$$

Apply equation (16.2).

Solution (Exercise 16.11).

$$\mathbf{E}\,[W] = \frac{1}{18}\,3 + \frac{1}{18}\,(-3) + \frac{2}{18}\,0 = 0.$$

$$\mathbf{Var}\,(W) = \mathbf{E}\left[W^2\right] = \frac{1}{18}\,9 + \frac{1}{18}\,9 = 1.$$

Thus the standard deviation σ for this random variable is one.

The probability that W deviates from its mean by at least three standard deviations is

$$\mathbf{P}(|W| \geq 3\sigma) = \mathbf{P}(|W| \geq 3) = \mathbf{P}(W = 3) + \mathbf{P}(W = -3) = \frac{1}{9}.$$

Using equation (16.19), the Chebyshev estimate is

$$\mathbf{P}(|W| \geq 3\sigma) \leq \frac{1}{9}.$$

A perfect estimate!

The probability that W deviates from its mean by at least 3.1 standard deviations is

$$\mathbf{P}(|W| \geq 3.1\sigma) = \mathbf{P}(|W| \geq 3.1) = 0.$$

Using equation (16.19), the Chebyshev estimate is

$$\mathbf{P}(|W| \geq 3.1\sigma) \leq \frac{1}{9.61}.$$

Not so good.

Solution (Exercise 16.12). The mean of uniform distribution on an interval is the midpoint, so $\mu = \mathbf{E}\,[Y] = 0$.

By equation (16.50), the variance of a uniform distribution on an interval is one-twelfth of the square of the length, so $\mathbf{Var}\,(Y) = 16/12 = 4/3$, and the standard deviation of Y is $\sigma = 2/\sqrt{3}$.

Notice that $3\sigma = 2\sqrt{3} > 2$.

Thus

$$\mathbf{P}(|Y - \mu| \geq 3\sigma) = 0.$$

Solution (Exercise 16.13).

$$\begin{aligned}\mathbf{E}\,[\,(X-\mu)(Y-\nu)\,] &= \mathbf{E}\,[XY] - \mu\mathbf{E}\,[Y] - \nu\mathbf{E}\,[X] + \mu\nu \\ &= \mathbf{E}\,[XY] - \mu\nu - \mu\nu + \mu\nu = \mathbf{E}\,[XY] - \mu\nu.\end{aligned}$$

Solution (Exercise 16.14).

Proof.

$$\begin{aligned}\mathbf{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mu_i\right)^2\right] \\ &= \mathbf{E}\left[\left(\sum_{i=1}^{n}\big(X_i-\mu_i\big)\right)^2\right]\end{aligned}$$

Using the distributive law as much as possible,

$$\left(\sum_{i=1}^{n}\big(X_i-\mu_i\big)\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\big(X_i-\mu_i\big)\big(X_j-\mu_j\big).$$

Hence

$$\begin{aligned}\mathbf{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\left[\big(X_i-\mu_i\big)\big(X_j-\mu_j\big)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{Cov}\left(X_i, X_j\right)\end{aligned}$$

□

Solution (Exercise 16.15). By equation (16.30),

$$\mathbf{Var}\left(S_n\right) = \mathbf{Var}\left(X_1\right) + \ldots + \mathbf{Var}\left(X_n\right).$$

By equation (16.7),

$$\mathbf{Var}\left(S_n\right) = np(1-p),$$

in agreement with equation (16.8).

Solution (Exercise 16.16). Since $\mathbf{E}\left[X\right] = \mathbf{E}\left[Y\right] = 0$ and $XY = \pm 1$,

$$\mathbf{Cov}\left(X, Y\right) = \mathbf{E}\left[XY\right] = \mathbf{P}(XY = 1) - \mathbf{P}(XY = -1).$$

Since $X = \pm 1$ and $Y = \pm 1$,

$$\{XY = 1\} = \{X = Y\} \text{ and } \{XY = -1\} = \{X \neq Y\}.$$

Thus $\mathbf{Cov}\left(X, Y\right) < 0$ tells us that

$$\mathbf{P}(X = Y) - \mathbf{P}(X \neq Y) < 0.$$

Solution (Exercise 16.17). We must prove that $\mathbf{Cov}(c_1X + c_2Y, Z) = c_1\mathbf{Cov}(X, Z) + c_2\mathbf{Cov}(Y, Z)$.

Let $\gamma = \mathbf{E}[Z]$. To save some writing we can use equation (16.22). By equation (16.22),

$$\begin{aligned}\mathbf{Cov}(c_1X + c_2Y, Z) &= \mathbf{E}[(c_1X + c_2Y)(Z - \gamma)] \\ &= \mathbf{E}[c_1X(Z - \gamma) + c_2Y(Z - \gamma)] \\ &= \mathbf{E}[c_1X(Z - \gamma)] + \mathbf{E}[c_2Y(Z - \gamma)] \\ &= c_1\mathbf{E}[X(Z - \gamma)] + c_2\mathbf{E}[Y(Z - \gamma)] \\ &= c_1\mathbf{Cov}(X, Z) + c_2\mathbf{Cov}(Y, Z).\end{aligned}$$

Solution (Exercise 16.18).

$$\begin{aligned}&\mathbf{Cov}(5X - Y + Z, X + 3Y - Z) \\ &= 5\mathbf{Cov}(X, X + 3Y - Z) - \mathbf{Cov}(Y, X + 3Y - Z) + \mathbf{Cov}(Z, X + 3Y - Z).\end{aligned}$$

By independence, $\mathbf{Cov}(X, Y) = 0$, $\mathbf{Cov}(X, Z) = 0$, $\mathbf{Cov}(Y, Z) = 0$. Thus

$$\mathbf{Cov}(X, X + 3Y - Z) = \mathbf{Cov}(X, X) + 0 + 0 = 1,$$

$$\mathbf{Cov}(Y, X + 3Y - Z) = 3\mathbf{Cov}(Y, Y) = 6,$$

and

$$\mathbf{Cov}(Z, X + 3Y - Z) = -\mathbf{Cov}(Z, Z) = -3.$$

Hence

$$\mathbf{Cov}(5X - Y + Z, X + 3Y - Z) = 5 - 6 - 3 = -4.$$

Solution (Exercise 16.19). Since $(x - y)^2 \geq 0$, $x^2 - 2xy + y^2 \geq 0$. Rearranging gives

$$2xy \leq x^2 + y^2.$$

Also

$$(x + y)^2 = x^2 + y^2 + 2xy \leq x^2 + y^2 + x^2 + y^2 = 2x^2 + 2y^2.$$

Solution (Exercise 16.20).

First method This is just the statement of Lemma 16.27 with $n = 2$.

Second method By assumption, $\mathbf{E}[X^2]$ exists.

Since 1 is a bounded random variable (it's even finite-range), $\mathbf{E}[1]$ exists.

By Lemma 16.29, with Y replaced by 1, $\mathbf{E}[X]$ exists.

Chapter 17

Poisson random variables

Poisson random variables are used in many applications, and have a surprisingly elegant theory.

17.1 A limit for powers

The definition and physical interpretation for a Poisson random variable is given in the next section. The following calculus fact will be useful.

$$\lim_{t\to 0}\Big(1+t\Big)^{1/t} = e. \tag{17.1}$$

The limit in equation (17.1) describes what happens as t approaches 0 but is not equal to zero. Since $1+t>0$ when t is close to 0, the expression $\big(1+t\big)^{1/t}$ in the limit makes sense.

To prove equation (17.1), we can use L'Hôpital's Formula:

$$\begin{aligned}\lim_{t\to 0}\log\bigg(\Big(1+t\Big)^{1/t}\bigg) &= \lim_{t\to 0}\frac{1}{t}\log\big(1+t\big) = \lim_{t\to 0}\frac{\log\big(1+t\big)}{t}\\ &= \lim_{t\to 0}\frac{\frac{1}{1+t}}{1} = 1.\end{aligned}$$

Here's a handy variation on equation (17.1). We will apply this lemma in the next section.

Lemma 17.1 (Exponential limit for powers). Suppose that $a_n \to \infty$, and $b_n a_n \to z$ for some number z. (z can be any real number.) Then

$$\lim_{n\to\infty} \big(1 + b_n\big)^{a_n} = e^z. \tag{17.2}$$

Proof. There are two cases.

Case 1 Suppose that $z \neq 0$.

Then for large n, $a_n b_n$ is nonzero, so b_n is nonzero. Also, since $a_n b_n \to z$, $b_n = (a_n b_n)/a_n \to 0$.

By equation (17.1),

$$\big(1 + b_n\big)^{1/b_n} \to e.$$

Then, since $a_n b_n \to z$, we have

$$\big(1 + b_n\big)^{a_n} = \Big(\big(1 + b_n\big)^{1/b_n} \Big)^{a_n b_n} \to e^z.$$

This finishes the proof for the case that $z \neq 0$.

Case 2 When $z = 0$, it is possible that $b_n = 0$ holds for infinitely many values of n. That means that throwing around expressions like $1/b_n$, as we did in Case 1, is rather obnoxious.

But we can fix that by breaking up the sequence b_n into two subsequences. Let b_{n_k} be the subsequence consisting of all elements b_n which are nonzero. If there is an infinite subsequence b_{n_k}, we handle that just as in Case 1:

$$\big(1 + b_{n_k}\big)^{1/b_{n_k}} \to e,$$

by equation (17.1), and

$$\big(1 + b_{n_k}\big)^{a_{n_k}} = \Big(\big(1 + b_{n_k}\big)^{1/b_{n_k}} \Big)^{a_{n_k} b_{n_k}} \to e^z.$$

That takes care of the subsequence b_{n_k}. The rest of the sequence b_n is just a sequence of zeros. For the elements b_n in that subsequence,

$$\big(1 + b_n\big)^{a_n} = 1^{a_n} = 1 \to 1 = e^z.$$

So equation (17.2) holds for the whole sequence.

□

In Appendix H), a different proof of Lemma 17.1 is given, using inequalities for the exponential function.

17.2 The frantic flipper and the Poisson approximation

In many experimental situations, an observer records the arrival of a message. For example, the "message" could be an event recorded by a Geiger counter, or a telephone call to a sales office, or a request to a computer server. A natural random variable in this situation is the number N of messages that arrive during a given time interval. In the present chapter we will derive a formula for the distribution of this random variable.

Physically, we are thinking of a situation in which there are many independent sources which randomly send a message to the observer. Although there are many sources, we assume that each source emits messages at a low rate, so the total number of arriving messages is not unbearably large.

In the case of particle emissions recorded by a Geiger counter, the particle emissions are caused by decay of atoms in a sample of radioactive material. Each atom has only a small chance of decaying during a given time interval, but there are many atoms in the sample. In the case of telephone calls to a sales office, there are many potential customers in the population, but for any particular potential customer, there is only a small chance that the customer will be motivated enough to call. And so on.

In order to have get an definite formula, we will consider a very familiar situation: tossing a coin. Imagine that each source of the message is tossing a coin to decide whether or not to send a message. The coin has a very low success probability, but there are many sources, and they are all tossing coins. We will use this picture in our derivation, but change it slightly.

Instead of many tossers, imagine that we have a single coin tosser, who tosses very rapidly, and sends a message every time the coin toss brings success. That seems easier to think about, and should result in the same type of formula.

Suppose the coin is tossed n times, where n is large. If the coin were *fair*, the tosser would almost certainly have an enormous number of successes, and the number of messages would be hopelessly large.

However, the coin has a very small success probability p.

We are interested in finding the distribution of the random variable N which records the number of successes.

Of course, everything depends on how large n is, and how small p is. Suppose these numbers are such that np is approximately equal to a number

λ which is of "ordinary" size. In this case we may be able to give an estimate for the probability distribution of the successes which the tosser will obtain.

Our estimate will apply when n is large, so let's analyze the situation by finding a limit as $n \to \infty$.

Lemma 17.2 (The Poisson approximation to the binomial). Consider a sequence of experiments. These experiments are *not* repetitions of the same experiment. Instead, in experiment n, the tosser records the result of n tosses, using a coin with success probability p_n.

Assume that

$$\lim_{n\to\infty} np_n = \lambda, \tag{17.3}$$

for some number λ.

Let the random variable S_n be the total number of heads obtained during the experiment with n tosses. Then:

$$\lim_{n\to\infty} \mathbf{P}(S_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{17.4}$$

As usual, $0! = 1$ in this formula. To include the special case $\lambda = 0$ in the formula, we also use the standard convention that $0^0 = 1$.

Proof. Let us think first about the result when no heads are obtained: $k = 0$. During experiment n, the probability of failure on a toss is $(1 - p_n)$. And $S_n = 0$ means all n tosses in experiment n gave failure. Using independence, $\mathbf{P}(S_n = 0) = (1 - p_n)^n$. Thus

$$\lim_{n\to\infty} \mathbf{P}(S_n = 0) = \lim_{n\to\infty} (1 - p_n)^n = e^{-\lambda}, \tag{17.5}$$

using Lemma 17.1. Note that equation (17.5) agrees with equation (17.4).

From now on we consider $k > 0$.

Using the binomial distribution,

$$\lim_{n\to\infty} \mathbf{P}(S_n = k) = \lim_{n\to\infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k}. \tag{17.6}$$

Here k is a fixed positive integer, while p_n is approaching zero in such a way that $np_n \to \lambda$.

We'll look first at one part of the expression in the limit. Note that

$$\binom{n}{k}p_n^k = \frac{1}{k!}n(n-1)\dots(n-k+1)p_n^k = \frac{1}{k!}\left(1-\frac{1}{n}\right)\dots\left(1-\frac{k-1}{n}\right)(np_n)^k.$$

Remember that k is fixed as we let $n \to \infty$. Each factor $1 - \frac{i}{n}$ converges to one, so

$$\lim_{n\to\infty}\binom{n}{k}p_n^k = \frac{\lambda^k}{k!}. \tag{17.7}$$

Now let's get back to evaluating the rest of the expression in equation (17.6), i.e finding $\lim_{n\to\infty}(1-p_n)^{n-k}$. Notice that

$$\lim_{n\to\infty} p_n(n-k) = \lim_{n\to\infty} p_n n - \lim_{n\to\infty} p_n k = \lambda - 0 = \lambda.$$

By Lemma 17.1, with $b_n = p_n$ and $a_n = n - k$,

$$(1-p_n)^{n-k} \to e^{-\lambda}. \tag{17.8}$$

Combining our facts, we have shown that equation (17.4) holds.

□

Exercise 17.1. Prove that those probability limits in equation (17.4), for $k = 0, 1, \dots$, add up to 1. That is, prove that

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda} = 1. \tag{17.9}$$

You can use the power series expansion of the exponential, namely

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}. \tag{17.10}$$

(Notice that we use the convention $0^0 = 1$ in this power series, when evaluating e^0.)

[Solution]

Exercise 17.2. Suppose that you wish to verify the power series expansion for e^x stated in equation (17.10).

First step: you can use a test for convergence of a series, for example the ratio test. Use that to prove that the series

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \tag{17.11}$$

is convergent, for every λ.

Second step: remember from calculus that a convergent power series can be differentiated term by term in the interior of its interval of convergence.

Let $f(\lambda)$ be sum of the series in equation (17.11). Differentiate that series with respect to λ, and show that $f'(\lambda) = f(\lambda)$.

Third step: Find the derivative of $f(\lambda)e^{-\lambda}$.

Fourth step: After finding $f(0)$, finish the proof of equation (17.10).

[Solution]

Definition 17.3 (Poisson random variables). Let λ be a nonnegative real number.

Let N be any random variable such that for all nonnegative integers k,

$$\mathbf{P}(N = k) = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{17.12}$$

We say that N has a Poisson distribution with parameter λ, and we also say that N is a Poisson random variable with parameter λ.

(Equation (17.9) shows that the this definition makes sense.)

The Poisson distribution is interesting in its own right, and also gives us a reliable approximation for the binomial distribution when equation (17.4) is applicable.

Exercise 17.3 (Expected value of a Poisson random variable). Let N be a Poisson random variable with parameter λ. Prove that

$$\mathbf{E}[N] = \lambda. \tag{17.13}$$

[Solution]

Exercise 17.4 (Convergence of expectations). Let S_n be the number of successes in n tosses of a coin with success probability p_n. Assume that $\lim_{n\to\infty} np_n = \lambda$.

Let N be a Poisson random variable with parameter λ. Prove that

$$\lim_{n\to\infty} \mathbf{E}\left[S_n\right] = \mathbf{E}\left[N\right]. \tag{17.14}$$

[Solution]

Equation (17.14) complements equation (17.4), and strengthens our confidence that the Poisson distribution is a reliable approximation to the binomial distribution.

Example 17.4 (Bounding e^x for $x \geq 0$). Suppose that $x \geq 0$. We will show that

$$e^x \leq 1 + x + \frac{1}{2}x^2 e^x. \tag{17.15}$$

A good approach is to use calculus. One can also use a power series.

Method 1 Let $f(x) = e^x$, $g(x) = 1 + x + \frac{1}{2}x^2 e^x$.

Then $f'(x) = e^x = f''(x)$.

$g'(x) = 1 + xe^x + \frac{1}{2}x^2 e^x$, so $g''(x) = e^x + xe^x + xe^x + \frac{1}{2}x^2 e^x$.

Thus $f(0) = 1 = g(0)$, $f'(0) = 1 = g'(0)$, and $f''(x) \leq g''(x)$ for all nonnegative x. Then for any $x \geq 0$,

$$f'(x) = 1 + \int_0^x f''(t)\,dt \leq 1 + \int_0^x g''(t)\,dt = g'(x).$$

Hence for any $x \geq 0$,

$$f(x) = 1 + \int_0^x f'(t)\,dt \leq 1 + \int_0^x g'(t)\,dt = g(x).$$

The idea of comparing the effect of the accelerations of two cars, when they start off with equal position and velocity, gives a good picture for this argument.

Method 2

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \sum_{k=2}^{\infty} \frac{x^k}{k!} = 1 + x + \sum_{j=0}^{\infty} x^2 \frac{x^j}{(j+2)!}.$$

Notice that $(j+2)! = (j+1)(j+2)(j!) \geq 2(j!)$. Hence

$$e^x \leq 1 + x + \sum_{j=0}^{\infty} x^2 \frac{x^j}{2(j!)} = 1 + x + \frac{1}{2}x^2 \sum_{j=0}^{\infty} \frac{x^j}{j!} = 1 + x + \frac{1}{2}x^2 e^x.$$

Exercise 17.5 (More than one Poisson success). Let N be a Poisson random variable with parameter λ. Prove that

$$\mathbf{P}(N > 1) \leq \frac{1}{2}\lambda^2. \tag{17.16}$$

Note that $\mathbf{P}(N > 1) = 1 - \mathbf{P}(N \leq 1)$. One might try setting $x = \lambda$ in equation (17.15).

[Solution]

If we think of N as the number of successes, Equation (17.16) tells us that when λ is small, the chance of more than one Poisson success is small, even in comparison to the small chance of one Poisson success.

Let's compare Exercise 17.5 to coin tossing.

Exercise 17.6 (More than one coin toss success). Let W be the number of successes in m independent Bernoulli trials, each trial having success probability p. Use subadditivity to justify the following bound:

$$\mathbf{P}(W) > 1) \leq \frac{m\,(m-1)\,p^2}{2}. \tag{17.17}$$

Suggestion: for any indices $i < j$, let A_{ij} be the event that both trial i and trial j give success. Note that $\{W > 1\}$ is equal to the union of all such events.

[Solution]

In Exercise 17.6, if we assume that $mp \approx \lambda$ we see that equation (17.17) is consistent with equation (17.16).

Physical settings and Poisson arrivals

We derived the Poisson distribution by thinking about the number of successes using a low-probability coin which is tossed very rapidly during a given time interval. So we might refer to N as (approximately) the number of successes.

But we mentioned at the beginning of Section 17.2 that there are many physical situations in which Poisson random variables arise. In the telephone call example, N is the number of telephone calls that arrive in an office during a fixed time interval. The office might be an office for telephone sales, or perhaps a help center. In the computer server example, N is the number of requests that arrive at a computer server which is connected to a computer network.

With these settings in mind, a Poisson random variable is often referred to as "the number of Poisson arrivals".

An impressive list of other examples for the Poisson distribution is given in Chapter VI of Feller's text [2]. The author mentions that the Poisson distribution does not just apply to random arrival times, but also applies to random points in the plane or in space, so that " Stars in space, raisins in cake, weed seeds among grass seeds, flaws in materials, animal litters in fields are distributed in accordance with the Poisson law".

17.3 Poisson approximations on all time intervals

To describe the Poisson approximation for a family of time intervals, we again imagine a sequence of experiments. We will speak about the experiments as coin-tossing, but the mathematical formulas apply to any of the other physical situations just mentioned.

As before, in experiment n a coin with success probability p_n is tossed many times. But we now imagine that the tosses are performed during a time interval named I. And for any subinterval J of that time interval, we will keep track of the number of successes during that time interval.

Let $|J|$ denote the length of J. (We've used a different notation for length in the past, but $|J|$ is convenient here.)

The rate of tossing is assumed to be *large*, and it is assumed to be constant during the time interval I. For any finite subinterval J of I, when the rate of

tossing is sufficiently large, the number of tosses during J will be proportional to the length of J.

Let $\ell_n(J)$ be the number of tosses of the coin during J. Thus $\ell_n(I)$ is the total number of tosses during the whole time interval I. In our previous discussion the total number of tosses in experiment n was equal to n.

Just as before, we assume that the probabilities p_n are such that the total number of tosses times p_n converges to a limit. So we assume that $\ell_n(I)p_n$ converges to a limit. Let's call the limit L. In our earlier discussion we called the limit λ, but now let's use λ to denote $L/|I|$. Thus λ represents an average rate of success per unit time for the whole experiment.

By definition,

$$\lambda = \lim_{n\to\infty} \frac{\ell_n(I)p_n}{|I|}. \tag{17.18}$$

Since the number of tosses during any subinterval J is assumed to be proportional to the length of the interval, it should be approximately true that $\ell_n(J)/|J| = n/|I|$. Thus we also have

$$\lambda = \lim_{n\to\infty} \frac{\ell_n(J)p_n}{|J|}. \tag{17.19}$$

We can express equation (17.19) in words by saying that λ is the limiting success rate per unit time during the time of the experiment. The whole analysis of Section 17.2 applies to the tosses during any time subinterval J. In the earlier analysis we had $np_n \to \lambda$. Now we have

$$\ell_n(J)p_n \to \lambda\,|J|\,. \tag{17.20}$$

Why is that? Well,

$$\ell_n(J)p_n = \left(\frac{\ell_n(J)p_n}{|J|}\right)|J| \to \lambda\,|J|\,.$$

Since equation (17.20) holds, and since $\ell_n(J)$ is the number of tosses of the coin during J, the following definition is appropriate.

Definition 17.5. Let $N(J)$ be a Poisson random variable with parameter $\lambda\,|J|$. Thus

$$\mathbf{P}(N(J) = k) = \frac{(\lambda\,|J|)^k}{k!}e^{-\lambda|J|}. \tag{17.21}$$

Our earlier discussion of the Poisson approximation now applies for each time interval J. When the coin is tossed rapidly, we expect that $\mathbf{P}(N(J) = k)$ is a good approximation to the probability that k successes are obtained during time interval J.

Since we sometimes say that a Poisson random variable counts "arrivals", we might say here that λ is the arrival rate per unit time.

In our mathematical model of this situation we picture all the Poisson random variables $N(J)$ as being defined on *the same sample space*. We can't define such a sample space with mathematical rigor in the present book, but it is perfectly possible, and makes sense physically.

For example, $N([0,1)) + N([1,2])$ represents the number of Poisson arrivals during the time interval $[0,2]$. $N([0,2])$ represents the same physical random quantity, so $N([0,2]) = N([0,1)) + N([1,2])$ should hold. And it does hold, in the right mathematical model.

The next lemma states an important property of Poisson random variables.

Lemma 17.6 (The sum of independent Poisson random variables is Poisson). Let N_1, N_2 be independent Poisson random variables with parameters λ_1, λ_2 respectively. Then $N_1 + N_2$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

To motivate this lemma, think about disjoint time intervals J_1, J_2 whose union is an interval. Physically,

$$N(J_1) + N(J_2) = N(J_1 \cup J_2), \tag{17.22}$$

since we are just adding up arrivals.

Coin tosses during different time intervals have no influence on each other, so the random variables $N(J_1)$ and $N(J_2)$ should be independent.

Thus we expect that the statement of the lemma applies when $N_1 = N(J_1)$ and $N_2 = N(J_2)$, for any intervals J_1, J_2.

That argument makes the lemma seem plausible. The actual proof is a short computation.

Proof. For any $k = 0, 1, \ldots$ we have

$$\begin{aligned}\mathbf{P}\left(N_1 + N_2 = k\right) &= \sum_{i=0}^{k} \mathbf{P}\left(\{N_1 = i\} \cap \{N_2 = k - i\}\right) \\ &= \sum_{i=0}^{k} \mathbf{P}\left(N_1 = i\right) \mathbf{P}\left(N_2 = k - i\right) = \sum_{i=0}^{k} \frac{\lambda_1^i}{i!} e^{-\lambda_1} \frac{\lambda_2^{k-i}}{(k-i)!} e^{-\lambda_2} \\ &= \sum_{i=0}^{k} \frac{1}{k!} \binom{k}{i} \lambda_1^i \lambda_2^{k-i} e^{-(\lambda_1+\lambda_2)} = \frac{(\lambda_1 + \lambda_2)^k}{k!} e^{-(\lambda_1+\lambda_2)}. \end{aligned} \tag{17.23}$$

□

Exercise 17.7. Justify the last two equalities in equation (17.23).
[Solution]

Exercise 17.8. By additivity, $\mathbf{E}\left[N_1 + N_2\right] = \mathbf{E}\left[N_1\right] + \mathbf{E}\left[N_2\right]$ must hold whenever the expected values exist. Check that this is true using for the random variables N_1, N_2 in Lemma 17.6, using equation (17.13).
[Solution]

Exercise 17.9 (The variance of a Poisson random variable). Let N be a Poisson random variable with parameter λ. Show that

$$\mathbf{Var}\left(N\right) = \lambda. \tag{17.24}$$

It may be helpful to calculate $\mathbf{E}\left[N(N-1)\right]$ first.
[Solution]

Let N_1, N_2 be independent Poisson random variables with parameters λ_1, λ_2 respectively. By Lemma 17.6, $N_1 + N_2$ is Poisson with parameter $\lambda_1 + \lambda_2$. Thus equation (17.24) shows that $\mathbf{Var}\left(N_1 + N_2\right) = \mathbf{Var}\left(N_1\right) + \mathbf{Var}\left(N_2\right)$, which of course is consistent with equation (16.30).

17.4 Waiting for a Poisson arrival

Back to the help center!

Suppose you are sitting at a desk in a help center, stoically waiting by your telephone for the first call of the day to arrive. Let τ be the random variable which represents how long you must wait. In the case of radiation events recorded by a Geiger counter, a similar random variable represents the time until the first event. We would like to know the distribution of τ.

For example, given a particular time t, what is $\mathbf{P}(\tau > t)$? We can answer this question surprisingly easily, by thinking about $N([0,t])$. The event that $\tau > t$ is exactly the event that $N([0,t]) = 0$. Since $N([0,t])$ is a Poisson random variable with parameter λt,

$$\mathbf{P}(\tau > t) = \mathbf{P}(N([0,t]) = 0) = e^{-\lambda t}. \tag{17.25}$$

The tail of a distribution characterizes the distribution, so equation (17.25) tells us that τ has an *exponential distribution* (recall equation 15.2).

Now let's compare equation (17.25) with coin-tossing.

Consider a sequence of experiments. In experiment n a coin with success probability p_n is tossed again and again during a long time interval I.

As in Section 17.3, we assume that there are $\ell_n(J)$ tosses during a time interval J. For each J, for large n we have by equation (17.19) that

$$\lambda \approx \frac{\ell_n(J)p_n}{|J|},$$

and so

$$|J| \approx \frac{\ell_n(J)p_n}{\lambda}.$$

Thus we convert from tosses to times by the following approximate formula:

$$\text{time} \approx \frac{\text{number of tosses } \times p_n}{\lambda}. \tag{17.26}$$

Let $W(n)$ denote the number of tosses required to obtain the first success. That is, $W(n)$ is the smallest positive integer k such that toss k produces a success. Then the distribution of $W(n)$ is the geometric distribution (Definition 13.3). Thus the distribution of $W(n)$ is given by equation the formula in equation (13.13):

$$\mathbf{P}(W(n) > k) = (1 - p_n)^k.$$

Let $T(n)$ be the *time* until the first success. Then

$$\mathbf{P}(T(n) > t) \approx \mathbf{P}\left(\frac{W(n)p_n}{\lambda} > t\right) = \mathbf{P}\left(W_n > \frac{\lambda t}{p_n}\right) \approx (1-p_n)^{\frac{\lambda t}{p_n}}.$$

By Lemma 17.2,

$$\mathbf{P}(T(n) > t) \approx e^{-\lambda t}. \tag{17.27}$$

Since $\mathbf{P}(\tau > t) = e^{-\lambda t}$, this is consistent with the Poisson approximation.

17.5 Solutions for Chapter 17

Solution (Exercise 17.1). Just set $x = \lambda$ in equation (17.10). Then multiply both sides by $e^{-\lambda}$.

Solution (Exercise 17.2).

Step 1 Ye Olde Ratio Test:

$$\lim_{n\to\infty} \frac{\frac{\lambda^{k+1}}{(k+1)!}e^{-\lambda}}{\frac{\lambda^k}{k!}e^{-\lambda}} = \lim_{n\to\infty} \frac{\lambda}{k+1} = 0 < 1.$$

So the series converges.

Step 2

$$\begin{aligned} f'(\lambda) &= \frac{d}{d\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \frac{d}{d\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k\frac{\lambda^{k-1}}{k!} \\ &= \sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = f(\lambda). \end{aligned}$$

Step 3

$$\big(f(\lambda)e^{-\lambda}\big)' = f'(\lambda)e^{-\lambda} - f(\lambda)e^{-\lambda} = 0.$$

By the Mean Value Theorem of Calculus, the expression $f(\lambda)e^{-\lambda}$ is constant, so $f(\lambda)e^{-\lambda} = f(0)$.

Step 4 $f(0) = 1$.
Since $f(\lambda)e^{-\lambda} = 1$, $f(\lambda) = e^{\lambda}$.

Solution (Exercise 17.3).

$$\begin{aligned}\mathbf{E}[N] &= \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}e^{-\lambda} \\ &= \sum_{k=1}^{\infty}\lambda\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} = \lambda\left(\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}e^{-\lambda}\right) = \lambda.\end{aligned}$$

Solution (Exercise 17.4). Let S_n be the number of heads obtained in experiment n. Then

$$\mathbf{E}[S_n] = np_n \to \lambda = \mathbf{E}[N].$$

Solution (Exercise 17.5). By equation (17.12),

$$\mathbf{P}(N \le 1) = \mathbf{P}(N=0) + \mathbf{P}(N=1) = e^{-\lambda} + \lambda e^{-\lambda} = e^{-\lambda}(1+\lambda). \quad (17.28)$$

By equation (17.15) in Example 17.4,

$$e^{\lambda} \le 1 + \lambda + \frac{1}{2}\lambda^2 e^{\lambda}.$$

Thus

$$1 + \lambda \ge e^{\lambda} - \frac{1}{2}\lambda^2 e^{\lambda}.$$

Applying this inequality to equation (17.28),

$$\mathbf{P}(N \le 1) \ge e^{-\lambda}\left(e^{\lambda} - \frac{1}{2}\lambda^2 e^{\lambda}\right) = 1 - \frac{1}{2}\lambda^2.$$

Thus

$$\mathbf{P}(N > 1) = 1 - \mathbf{P}(N \le 1) \le \frac{1}{2}\lambda^2.$$

Solution (Exercise 17.6). To say that $\{W > 1\}$ occurs is the same as saying that there are at least two tosses, say toss i and toss j, with $i < j$, such that both of those tosses give success. Thus $\{W > 1\}$ occurs when at least one of the events A_{ij} occurs. That is why

$$\{W > 1\} = \bigcup_{i<j} A_{ij}.$$

By sub-additivity (Theorem 2.25),

$$\mathbf{P}(W > 1) \leq \sum_{i<j} \mathbf{P}(A_{ij}).$$

For each $i \neq j$, $\mathbf{P}(A_{ij}) = p^2$.

The number of pairs i, j such that $i < j$ is exactly the same as the number of subsets of $\{1, \ldots, m\}$ containing two elements. Thus there are $C_2^m = m(m-1)/2$ such subsets, by Lemma 8.1.

Hence

$$\mathbf{P}(W > 1) \leq \frac{m(m-1)}{2} p^2,$$

verifying equation (17.17).

Solution (Exercise 17.7). The second last equality holds by the definition of $\binom{k}{i}$:

$$\begin{aligned}
\frac{\lambda_1^i}{i!} e^{-\lambda_1} \frac{\lambda_2^{k-i}}{(k-i)!} e^{-\lambda_2} &= \frac{1}{i!(k-i)!} \lambda_1^i \lambda_2^{k-i} e^{-\lambda_1} e^{-\lambda_2} \\
&= \frac{1}{i!(k-i)!} \lambda_1^i \lambda_2^{k-i} e^{-(\lambda_1+\lambda_2)} \\
&= \frac{1}{k!} \frac{k!}{i!(k-i)!} \lambda_1^i \lambda_2^{k-i} e^{-(\lambda_1+\lambda_2)} \\
&= \frac{1}{k!} \binom{k}{i} \lambda_1^i \lambda_2^{k-i} e^{-(\lambda_1+\lambda_2)}.
\end{aligned}$$

The final equality holds by the binomial theorem (equation (8.6)):

$$\sum_{i=0}^{k} \binom{k}{i} \lambda_1^i \lambda_2^{k-i} = (\lambda_1 + \lambda_2)^k.$$

Solution (Exercise 17.8). By equation (17.13), $\mathbf{E}[N_1] = \lambda_1$, $\mathbf{E}[N_2] = \lambda_2$, and, since $N_1 + N_2$ is Poisson with parameter $\lambda_1 + \lambda_2$, $\mathbf{E}[N_1 + N_2] = \lambda_1 + \lambda_2$.

Solution (Exercise 17.9). By Theorem 14.11,

$$\begin{aligned}
\mathbf{E}[N(N-1)] &= \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} = \lambda^2 \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} = \lambda^2.
\end{aligned}$$

Thus $\mathbf{E}[N^2 - N] = \lambda^2$, and so $\mathbf{E}[N^2] = \lambda^2 + \mathbf{E}[N] = \lambda^2 + \lambda$.

Hence $\mathbf{Var}(N) = \mathbf{E}[N^2] - (\mathbf{E}[N])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Chapter 18

Normal random variables and the Central Limit Theorem

18.1 Sums of independent random variables

Let S_n be the number of heads obtained in n tosses of a coin, where the probability of a head on any toss is p. Then S_n has a binomial distribution with parameters n, p (see Example 9.9).

We can calculate $\mathbf{P}(S_n = k)$ using equation (9.3), but this formula doesn't seem to give us a sense of the main features of the distribution, especially when n is large. We do obtain some insight by calculating $\mathbf{E}[S_n]$ and $\mathbf{Var}(S_n)$, but that is only a start. In the present chapter we will go much farther in understanding the distributions of random variables which are similar to S_n.

In the experiment of tossing a coin repeatedly, let $X_i = 1$ if toss i results in a head, and let $X_i = 0$ otherwise. Then $S_n = X_1 + \ldots + X_n$. The fact that S_n can be written as a sum of independent random variables is the key to understanding its distribution.

We saw this already in the calculation of the mean and variance of S_n. Writing S_n as a sum allowed us to use additivity of expectation, showing that $\mathbf{E}[S_n] = \mathbf{E}[X_1] + \ldots + \mathbf{E}[X_n]$. Similarly, equation (16.30) shows that $\mathbf{Var}(S_n) = \mathbf{Var}(X_1) + \ldots + \mathbf{Var}(X_n)$.

More generally, any time you have a random variable S_n such that $S_n = X_1 + \ldots + X_n$, where the random variables $X_1, \ldots, X_n$ form an independent sequence, it will be true that $\mathbf{E}[S_n] = \mathbf{E}[X_1] + \ldots + \mathbf{E}[X_n]$ and $\mathbf{Var}(S_n) =$

$\mathbf{Var}(X_1) + \ldots + \mathbf{Var}(X_n)$.

Notice that the mean and variance of S_n in this situation are completely determined by the means and variances of the X_i, and do not depend in on any other properties of the distributions of the random variables X_i.

One can say much more about the distribution of S_n in this situation. In a very wide range of cases, when S_n is the sum of independent random variables $X_1, \ldots, X_n$, and n is large, the distribution of S_n is *approximately* described by a simple analytical formula, and the formula does *not* depend on the details of the distributions of the random variables X_i. This is the content of the Central Limit Theorem (see Theorem 18.14 and Example 18.17).

To express the Central Limit Theorem, we will introduce a new probability distribution, called the normal distribution. This distribution has a smooth probability density, whose graph has a characteristic shape, sometimes called a "bell-shaped curve" (see Figure 18.7).

The Central Limit Theorem is a mathematical result, but it is consistent with experiment. Many physical random variables have distributions which are approximately normal. Because the normal distribution plays such an important role, we will develop its properties carefully. The steps are easy, and we'll try to give a full discussion of each step.

18.2 Plotting the binomial distribution

For coin-tossing, the random variable $S_n = X_1 + \ldots + X_n$ has a binomial distribution. The binomial distribution is a very special case, but we can motivate the Central Limit Theorem by plotting the values of this distribution.

In the present section we will take $p = 1/2$, so S_n represents the number of heads in n tosses of a fair coin.

In Figures 18.1, 18.2 and 18.3, we take a straightforward approach and graph all values for the probability mass function of S_n, when $n = 100$, $n = 1000$ and $n = 10000$.

There are $n + 1$ points in the range of S_n, namely $0, 1, \ldots, n$. In these figures we are plotting the points $(k, \mathbf{P}(S_n = k))$, for all k in the range of S_n.

When the points are close together, they may appear to form a continuous curve, but that is just a limitation of the picture.

The three graphs, for $n = 100$, $n = 1000$, $n = 10000$ don't look very

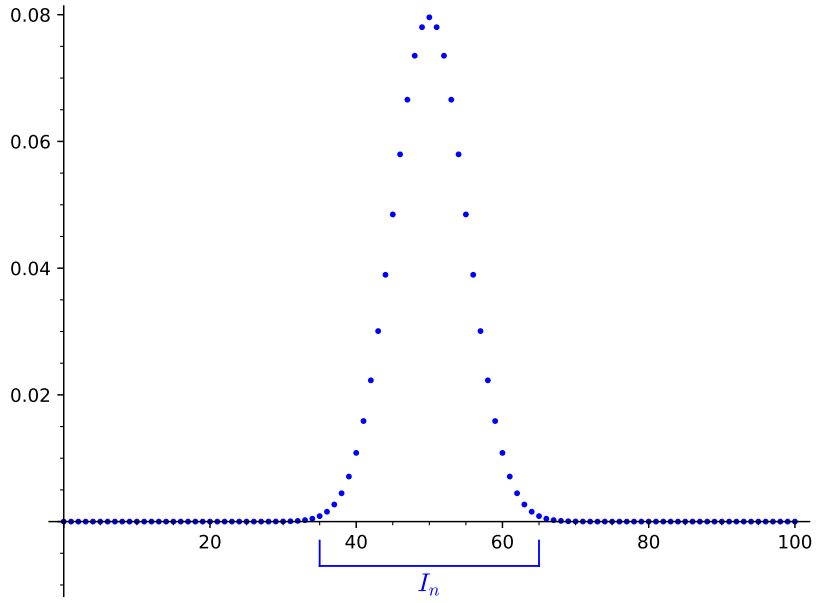


Figure 18.1: $\mathbf{P}(S_n = k)$ versus k for $n = 100$.

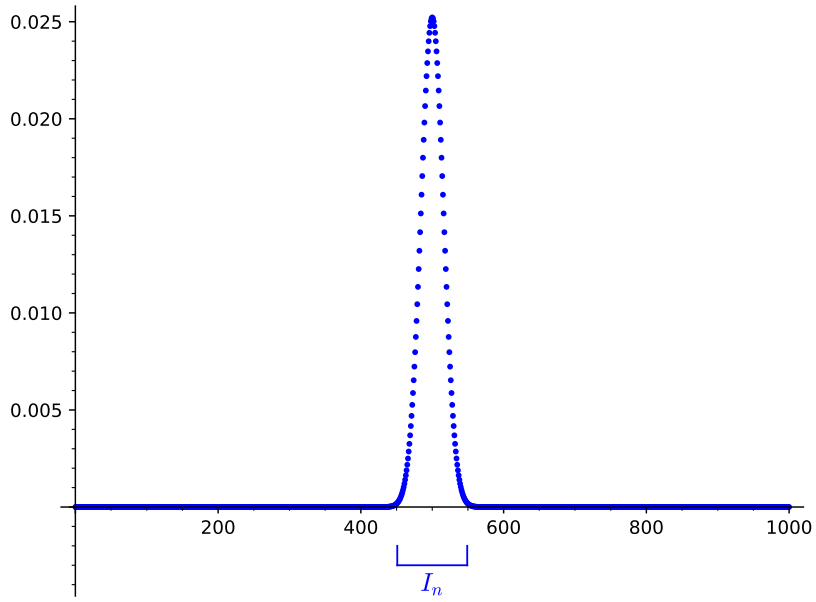


Figure 18.2: $\mathbf{P}(S_n = k)$ versus k for $n = 1000$.

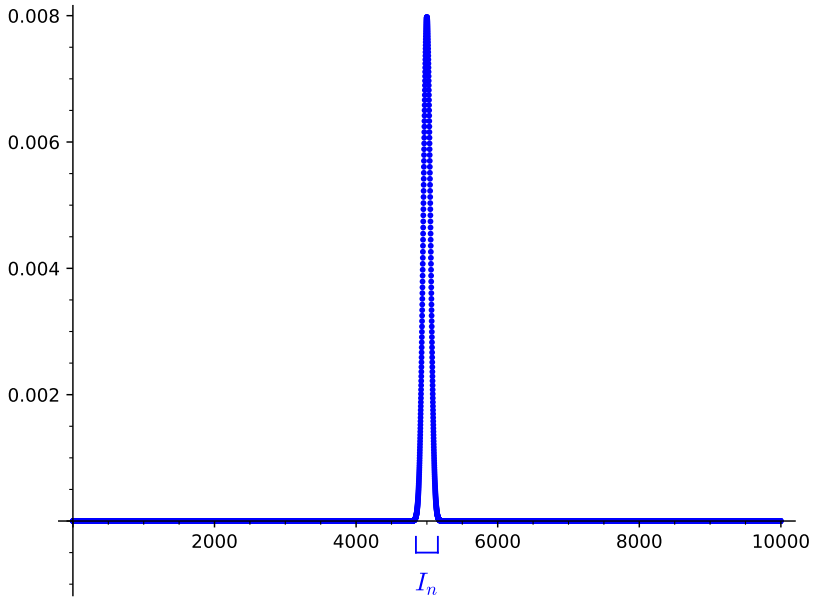


Figure 18.3: $\mathbf{P}(S_n = k)$ versus k for $n = 10000$.

similar, although in each case there is a single maximum at the center of the graph, which occurs at the mean value of S_n.

Notice that in each case the only significant probability values are found relatively near the mean value of S_n. It interesting that the interval in which the probability is concentrated is so small, relative to the whole range of S_n. This is especially true as n gets large.

Because the probability is so concentrated, it is hard to see details in the graphs when n is large. We have to do something about that problem.

For convenience, let us use the name I_n to refer to the interval where significant probability values are found in the binomial distribution. We might refer to that region verbally as the part of the graph where the "main values" of the distribution are located. The location of the interval I_n is shown in each of Figures 18.1, 18.2 and 18.3,

I_n is not a precisely defined interval but we can see roughly where it is, by looking at each graph. Let $|I_n|$ denote the number of points in I_n, As n increases, $|I_n|$ increases along with n, but it evidently increases more slowly than n.

The smallness of $|I_n|$ relative to the size of range of S_n makes the graph

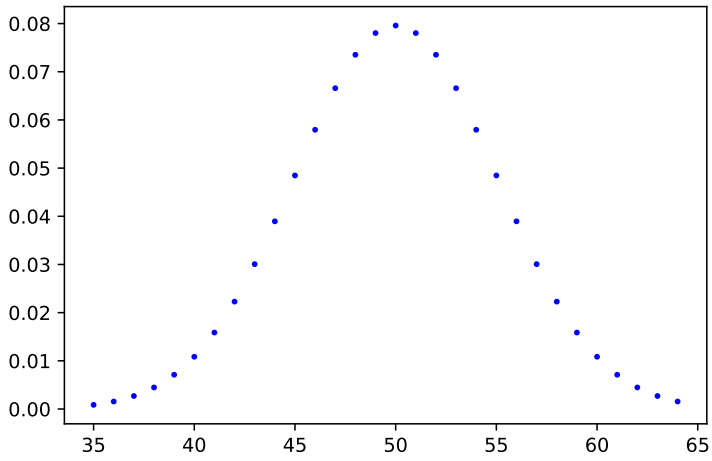


Figure 18.4: Main values of $\mathbf{P}(S_n = k)$ for $n = 100$.

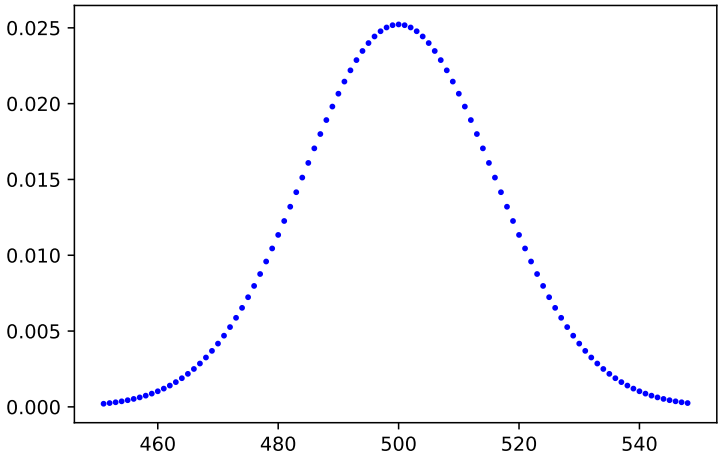


Figure 18.5: Main values of $\mathbf{P}(S_n = k)$ for $n = 1000$.

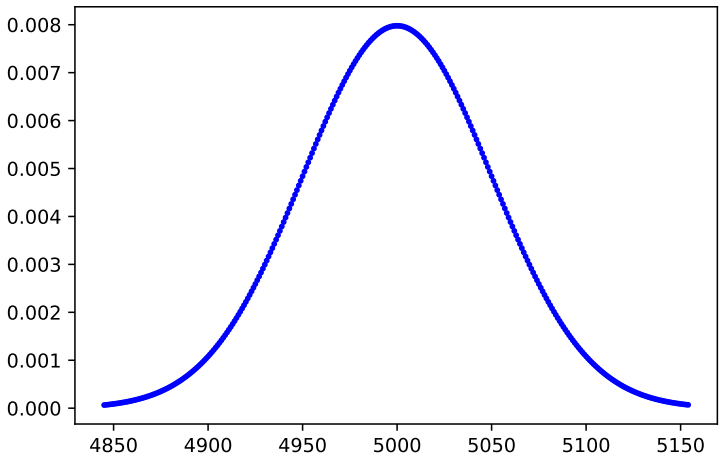


Figure 18.6: Main values of $\mathbf{P}(S_n = k)$ for $n = 10000$.

of the probability mass function appear more and more sharply peaked, as n increases.

On the other hand, since the number of points in I_n *is* growing larger, we are not surprised that the maximum value for $\mathbf{P}(S_n = k)$ becomes smaller, since the total probability is shared among more points.

It is hard to see the *precise shapes* of the graphs in Figures 18.1, 18.2 and 18.3, since the interesting part of the graph is being squeezed into a narrower and narrower spike as n gets large. In order to see more clearly what is going on, we need to focus our attention on the central region I_n.

To focus on I_n let's choose a more precise definition of I_n. We'll define a little bit of terminology, just for this discussion.

Terminology 18.1 (The "main part" of the distribution). Let's specify an interval $[\ell(n), u(n)]$ which contains most of the probability of the distribution. This interval $[\ell(n), u(n)]$ will be the interval I_n that we want to look at.

Suppose that we consider a small probability value δ, say $\delta = .001$.

We are trying to focus on the important part of the distribution. So we don't want I_n to be bigger than necessary. With that in mind, let $\ell(n)$ be the *largest* integer such that $\mathbf{P}\big(S_n < \ell(n)\big) \leq \delta$. Since δ is small, the values of S_n are unlikely to be located to the left of $\ell(n)$.

Similarly, choose $u(n)$ to be the *smallest* integer such that $\mathbf{P}\big(S_n > u(n)\big) \leq \delta$. Then $u(n)$ is such that the values of S_n are unlikely to be further to the right than $u(n)$.

From the definitions, $[\ell(n), u(n)]$ is an interval that contains at least $1-2\delta$ of the probability for this distribution. That is, $\mathbf{P}\big(\ell(n) \leq S_n \leq u(n)\big) \geq 1 - 2\delta$. If $\delta = .001$ then this interval contains 99.8% of the probability for this distribution.

Let $I_n = [\ell(n), u(n)]$. We will refer to the part of the range of S_n which lies within $[\ell(n), u(n)]$ as the "main part" of the range.

The values of $\ell(n)$ and $u(n)$ will of course change if we choose some other value for the small probability δ. In our discussion we will stick to using $\delta = .001$.

Using a computer, we find that $[\ell(n), u(n)]$ is $[35, 65]$ when $n = 100$, $[451, 549]$ when $n = 1000$, and $[4845, 5155]$ when $n = 10000$. The number of points in $[\ell(n), u(n)]$ is 31 when $n = 100$, 99 when $n = 1000$, and 311 when $n = 10000$.

Figures 18.4, 18.5 and 18.6, show the graphs of the probability mass functions only for k in $[\ell(n), u(n)]$ and $n = 100$, $n = 1000$ and $n = 10000$. So we are *zooming in* on the interval $[\ell(n), u(n)]$. This is where the action is, for these distributions.

Restricting our attention to k in $[\ell(n), u(n)]$ has made it possible to see the **shape** of the graph of the probability mass function much more clearly. It is striking how similar in shape these three graphs are now, for $n = 100, 1000, 10000$.

Figure 18.6 looks like the graph of a continuous curve, but of course it is still obtained by plotting a finite number of points. In the next section we will introduce a continuous function which has the shape shown in Figure 18.6.

18.3 A function with the right shape

We will derive some properties of the function e^{-x^2}, whose graph is shown in Figure 18.7. The **shape** of this graph is similar to Figure 18.6. The Central Limit Theorem will show that this resemblance is not an accident.

Lemma 18.2 (An important integral). The function e^{-x^2} is integrable on $\mathbb{R}$, and

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}. \tag{18.1}$$

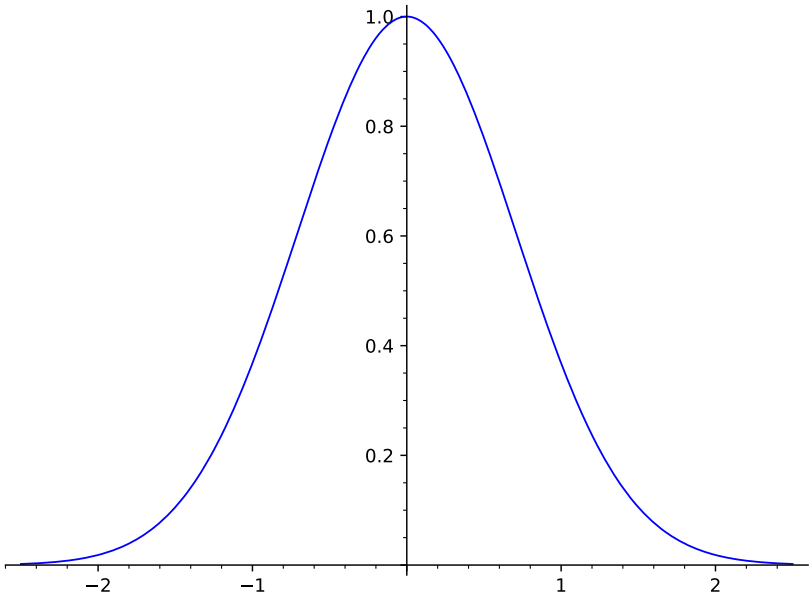


Figure 18.7: the graph of $f(x) = e^{-x^2}$, a "bell-shaped curve".

Proof. The usual way to finding $\int_{-\infty}^{\infty} e^{-x^2}\,dx$ would be to evaluate

$$\lim_{a\to\infty} \int_{-a}^{a} e^{-x^2}\,dx.$$

Unfortunately, we can't even start this procedure. The function e^{-x^2} does not have an antiderivative in terms of the calculus functions that we know and love. So we cannot calculate

$$\int_{-a}^{a} e^{-x^2}\,dx.$$

To even show that $\int_{-\infty}^{\infty} f(x)\,dx$ exists we will have to use a *comparison principle*. If we can find a nonnegative function g such that $\int_{-\infty}^{\infty} g(x)\,dx$ exists, and if $e^{-x^2} \le g(x)$ for all x, then $\int_{-\infty}^{\infty} e^{-x^2}\,dx$ exists.

In the present situation, the reader can easily check that for any x,

$$x^2 \ge |x| - 1.$$

(Consider the case $|x| \leq 1$ and the case $|x| > 1$ separately.)

Since $-(|x|-1) \geq -x^2$, and since the exponential function is an increasing function,

$$e^{1-|x|} \geq e^{-x^2}$$

holds for all x. And we can certainly use ordinary calculus methods to show that $\int_{-\infty}^{\infty} e^{1-|x|}\,dx$ exists. (Using symmetry, this integral is the same as $2\int_0^{\infty} e^{1-|x|}\,dx = 2e\int_0^{\infty} e^{-x}\,dx$, etc.)

Hence, by the comparison principle for integrals, we know that $\int_{-\infty}^{\infty} e^{-x^2}\,dx$ exists.

But we want to know the value of this integral, not just that it exists. To evaluate the integral, we use a trick!

First, we move the problem to $\mathbb{R}^2$, by noticing the following convenient fact.

$$\begin{aligned}\iint_{\mathbb{R}^2} e^{-(s^2+t^2)}\,ds\,dt &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(s^2+t^2)}\,ds\,dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-s^2}e^{-t^2}\,ds\,dt \\ &= \left(\int_{-\infty}^{\infty} e^{-s^2}\,ds\right)\left(\int_{-\infty}^{\infty} e^{-t^2}\,dt\right) = \left(\int_{-\infty}^{\infty} e^{-s^2}\,ds\right)^2.\end{aligned}$$

We can use polar coordinates to evaluate the integral over $\mathbb{R}^2$.

$$\iint_{\mathbb{R}^2} e^{-(s^2+t^2)}\,ds\,dt = \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r\,dr\,d\theta = 2\pi\int_0^{\infty} e^{-r^2} r\,dr = \pi\int_0^{\infty} e^{-r^2} 2r\,dr$$

Now we have an integral for which calculus methods work.

$$\int_0^{\infty} e^{-r^2} 2r\,dr = \lim_{b\to\infty}\int_0^{b} e^{-r^2} 2r\,dr = \lim_{b\to\infty} -e^{-r^2}\Big|_0^{b} = \lim_{b\to\infty}\left(1 - e^{-b^2}\right) = 1. \tag{18.2}$$

Thus

$$\left(\int_{-\infty}^{\infty} e^{-s^2}\,ds\right)^2 = \pi.$$

□

Exercise 18.1 (A random variable whose distribution density has the right shape). Equation (18.1) tells us that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-y^2}\,dy = 1,$$

so $\frac{1}{\sqrt{\pi}}e^{-y^2}$ is a probability density.

Let Y be a random variable whose distribution has probability density $\frac{1}{\sqrt{\pi}}e^{-y^2}$. (And such a random variable certainly exists. This is spelled out in Appendix C.)

(i) By equation (15.6) (the formula for the expected value of a function of a random variable whose distribution has a density), we have

$$\mathbf{E}\left[Y^2\right] = \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{\pi}} e^{-y^2}\, dy, \tag{18.3}$$

provided that the integral on the right exists.

Show that $\mathbf{E}\left[Y^2\right]$ exists and

$$\mathbf{E}\left[Y^2\right] = 1/2. \tag{18.4}$$

(Integration by parts is useful here.)

(ii) Show that

$$\mathbf{E}\left[Y\right] = 0. \tag{18.5}$$

(iii) Show that

$$\mathbf{E}\left[\,|Y|\,\right] = \frac{1}{\sqrt{\pi}}. \tag{18.6}$$

[Solution]

The density $\frac{1}{\sqrt{\pi}}e^{-y^2}$ used in Exercise 18.1 is an example of a *normal density*.

Definition 18.3 (Normal densities and distributions). Let κ, m be real numbers with $\kappa > 0$. Let g be the probability density given by

$$g(x) = \frac{1}{\sqrt{\kappa^2 \pi}} e^{-\frac{(x-m)^2}{\kappa^2}}. \tag{18.7}$$

Then g is said to be a normal probability density.

Any distribution with a normal density is referred to as a normal distribution, and any random variable with a normal distribution is said to be a normal random variable.

Incidentally, the word "normal" in this definition is a special usage. One should not draw the conclusion that there is something wrong with a random variable if it is not a normal random variable.

Normal densities are also referred to as Gaussian densities (in honor of the mathematician Carl Friedrich Gauss).

Let Y be a random variable like the one in Exercise 18.1, meaning that the distribution of Y has probability density $\frac{1}{\sqrt{\pi}}e^{-y^2}$. Exercise 18.2 will show that a probability density g for the distribution of $\kappa Y+m$ is given by equation (18.7).

The most common way of writing a normal density is actually the one given below in equation (18.13), which differs slightly from equation (18.7). It is useful to be able to recognize normal densities which are written in various forms (see Lemma 18.28).

18.4 Rescaling and shifting random variables and distributions

Before discussing properties of normal distributions, we need to discuss transformations of distributions. This will give some motivation for the definition of a normal density, and is relevant for another reason: in applications we often have to transform one normal distribution into another.

In Exercise 18.1 of Section 18.3 we introduced a random variable Y whose distribution has a probability density $\frac{1}{\sqrt{\pi}}e^{-y^2}$. We said that the shape of the graph of this density resembles the shape of the distribution of a binomial random variable S_n.

Of course, we haven't defined exactly what is meant by "shape". But in this section we note some transformations which preserve the kind of shape we are interested in.

You've likely seen such transformations when sketching graphs of functions in calculus. For example, after you've seen the graph of $y = x^2$, you can easily draw a rough sketch of the graph of $y = 57x^2$, without plotting any points.

Just to have some terminology, we will speak of *shifting* a graph (moving the graph horizontally or vertically), *rescaling* a graph (stretching the graph horizontally or vertically), and *reflecting* a graph (vertically or horizontally).

These transformations are carried out in the following ways.

- The graph of a function f is moved to the right by m to obtain the graph of $f(x-m)$. (If m is negative this actually means that the graph is moved to the left.) We refer to this movement as shifting the graph. (The graph could be moved *up* by b to obtain the graph of $f(x)+b$, but we won't have occasion to do this for graphs of probability densities.)

  Is it obvious that the graph of $f(x-m)$ is obtained by moving the graph of $f(x)$ to the right by m? Here's an argument. Think of travelling to the right along the x-axis. Suppose that for some x you have $f(x) = .3$. How much further to the right do you have to go, to find the same value for $f(x-m)$?

- The graph of a function f is stretched horizontally by $\kappa > 0$ to obtain the graph of $f(x/\kappa)$. And when $\kappa < 0$ the graph of f is stretched by $|\kappa|$ and reflected horizontally to obtain the graph of $f(x/\kappa)$.

- The graph of a function f is stretched vertically by $\kappa > 0$ to obtain the graph of $\kappa f(x)$. And of course when $\kappa < 0$ the graph of f is stretched vertically by $|\kappa|$ and reflected vertically to obtain the graph of $\kappa f(x)$.

Remark 18.4 (Stretching factors). The graph of $f(x/100)$ is 100 times wider than the graph of $f(x)$. Can you see why?

The reason is that everything happens 100 times more slowly when the function is $f(x/100)$. Consider moving from x to $x+\Delta$. This causes a change $f(x+\Delta) - f(x)$ in the value of f. When the function is $f(x/100)$, you will need to move from $100x$ to $100x + 100\Delta$ to obtain the same change in the value of the function. So you have to move 100 times as far to cause the same change.

When f is a probability density, the function $(1/|\kappa|)f((x-m)/\kappa)$ is again a probability density, as you will show in Exercise 18.2.

Exercise 18.2 (Rescaling and shifting probability densities). Let X be a random variable whose distribution has a probability density f. Using Definition 3.4 and Remark 9.12, check the following.

(i) Let m be any real number. Let $W = X + m$.

Show that $f(w-m)$ is a probability density for the distribution of W.

(ii) Let κ be any nonzero real number. Let $V = \kappa X$.

Show that $(1/|\kappa|)f(v/\kappa)$ is a probability density for the distribution of V.

(iii) Let κ, m be any real numbers with $\kappa \neq 0$. Let $U = \kappa X + m$.

Show that $(1/|\kappa|)f((u-m)/\kappa)$ is a probability density for the distribution of U.

[Solution]

When f is a probability density, we can sometimes understand the properties of f more clearly by considering a transformed version of f, such as $(1/|\kappa|)f((x-m)/\kappa)$, for suitable values of m and κ.

Lemma 18.5 (General normal by rescaling and shifting). Let Y be the random variable defined in Exercise 18.1, so that a density for the distribution of Y is the function f defined by

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

Let W be a random variable whose distribution has density g, where g is given by equation (18.7) in the definition of a normal density, so that

$$g(x) = \frac{1}{\sqrt{\kappa^2 \pi}} e^{-\frac{(x-m)^2}{\kappa^2}}.$$

Then

$$g(x) = \frac{1}{|\kappa|} \frac{1}{\sqrt{\pi}} e^{-\left(\frac{x-m}{\kappa}\right)^2}, \tag{18.8}$$

and W and $\kappa Y + m$ have the same distribution.

Proof. Equation (18.8) is clearly equivalent to equation (18.7).

Equation (18.8) and Exercise 18.2 tell us that g is a density for the distribution of $\kappa Y + m$.

□

Lemma 18.6 (Variance and mean for a rescaled and shifted density). Let f be the density of the distribution of a random variable Y. Suppose that $\mathbf{E}[Y]$ and $\mathbf{Var}(Y)$ exist.

Let κ, m be numbers with $\kappa \neq 0$. Let h be the function defined by

$$h(x) = \frac{1}{|\kappa|} f\left(\frac{1}{\kappa}(x-m)\right). \tag{18.9}$$

Let W be a random variable whose distribution has density h.

Then W and $\kappa Y + m$ have the same distribution, and

$$\mathbf{Var}(W) = \kappa^2 \mathbf{Var}(Y), \ \mathbf{E}[W] = \kappa \mathbf{E}[Y] + m. \tag{18.10}$$

Proof. Exercise 18.2 says that W and $\kappa Y + m$ have the same distribution.

Equations (16.6) and (16.5) tell us that

$$\mathbf{Var}(\kappa Y + m) = \kappa^2 \mathbf{Var}(Y),$$

and by linearity we have

$$\mathbf{E}[\kappa Y + m] = \kappa \mathbf{E}[Y] + m.$$

□

Corollary 18.7 (Variance and mean for a normal density). Let W be a random variable whose distribution has density g, where g is defined by

$$g(x) = \frac{1}{|\kappa|} \frac{1}{\sqrt{\pi}} e^{-\left(\frac{x-m}{\kappa}\right)^2} = \frac{1}{\sqrt{\kappa^2 \pi}} e^{-\frac{(x-m)^2}{\kappa^2}}. \tag{18.11}$$

Then

$$\mathbf{Var}(W) = \frac{\kappa^2}{2}, \ \mathbf{E}[W] = m. \tag{18.12}$$

Proof. Let Y be the random variable defined in Exercise 18.1.

By Lemma 18.5, W and $\kappa Y + m$ have the same distribution.

We showed in Exercise 18.1 that $\mathbf{E}[Y^2] = 1/2$ and $\mathbf{E}[Y] = 0$. Thus $\mathbf{Var}(Y) = 1/2$.

The conclusion follows from equation (18.10).

□

Remark 18.8 (Any rescaled and shifted normal is normal!). Lemma 18.5 shows that a random variable X is a normal random variable if and only if for some real numbers κ, m with $\kappa \neq 0$, we have $X = \kappa Y + m$, where Y is the random variable defined in Exercise 18.1.

When $X = \kappa Y + m$, suppose we now form a new random variable $W = \tau X + v$, where τ, v are real numbers and $\tau \neq 0$.

Note that

$$W = \tau(\,\kappa Y + m\,) + v = \tau\kappa Y + (\tau m + v).$$

Hence W is also normal.

18.5 Properties of normal densities

In this section we'll state and derive the main facts that are used when calculating with normal densities. We can think of these facts as a "toolbox". In the next section we will finally put normal densities to work, when we state the Central Limit Theorem.

Because any normal density involves a function similar to e^{-x^2}, and such function do not have elementary antiderivatives, one might suspect that computations with normal densities must be hard. But the calculations performed in this section are easy, even though dealing with e^{-x^2} is inherently messier than dealing with e^{-x}.

We have deliberately written several different-looking formulas for normal densities, because normal densities may be encountered in such forms. But there is a "best" way to write any normal density, as follows.

Lemma 18.9 (Mean and variance: the best form of a normal density). Suppose that the distribution of X has a normal probability density g. Then g can be written as

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{18.13}$$

where

$$\mathbf{Var}\,(X) = \sigma^2,\ \mathbf{E}\,[X] = \mu. \tag{18.14}$$

Any normal density is completely determined by its mean and variance.

If W is a normal random variable such that $\mathbf{Var}(W) = t\mathbf{Var}(X)$ and $\mathbf{E}[W] = \sqrt{t}\mathbf{E}[X] + v$, then

$$W \text{ and } \sqrt{t}X + v \text{ have identical distributions.} \tag{18.15}$$

Proof. By Corollary 18.7, with $\kappa^2 = 2\sigma$, we have $\mathbf{E}[X] = \mu$ and

$$\mathbf{Var}(X) = \frac{2\sigma^2}{2} = \sigma^2.$$

These facts give equation (18.14).

Equation (18.14) determines σ in terms of $\mathbf{Var}(X)$ and μ in terms of $\mathbf{E}[X]$. So to show that X and $\sqrt{t}X + v$ have identical distributions, it is sufficient to check that these two random variables have the same variance and the same mean.

□

It follows from Remark 18.8 that all normal random variables can be obtained from any one normal random variable by rescaling and shifting. So there is no reason to think of the distribution of any normal random variable as being special! However, we will pick one normal distribution to be "standard".

Definition 18.10 (The standard normal). Let Z be a random variable with probability density η given by

$$\eta(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}. \tag{18.16}$$

The mean and variance of Z are particularly simple: $\mathbf{E}[Z] = 0$, $\mathbf{Var}(Z) = 1$. For this reason, Z is given a special title, and is said to be a *standard normal random variable*. The density η is called the standard normal density.

The graphs of all normal densities have a similar shape. For what it's worth, a graph of the standard normal density is shown in Figure 18.8.

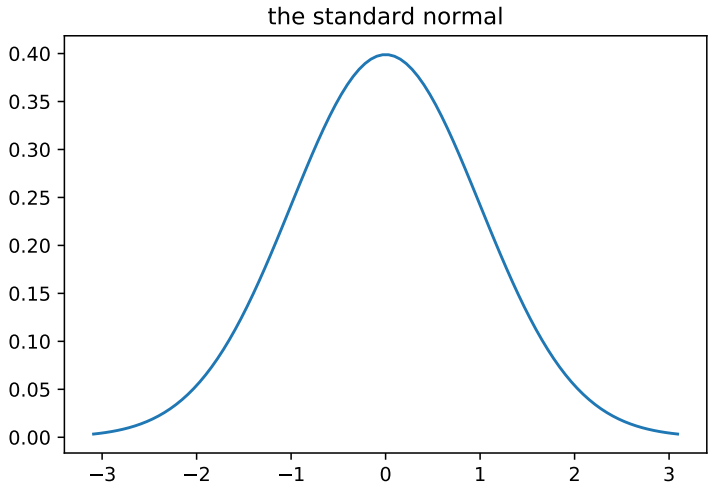


Figure 18.8: The standard normal density η.

Notice that all derivatives of η exist, so the smooth appearance of the graph in Figure 18.8 is not an illusion.

The next exercise repeats the information in Remark 16.6.

Exercise 18.3 ("Standardizing" a random variable). Let X be a random variable, with $\mathbf{Var}(X) = \sigma^2$ and $\mathbf{E}[X] = \mu$.

Let

$$Z = \frac{X - \mu}{\sigma}. \tag{18.17}$$

Check that $\mathbf{Var}(Z) = 1$ and $\mathbf{E}[Z] = 0$. Thus whenever X is normal, $(X - \mu)/\sigma$ is standard normal!

[Solution]

Equation (18.17) says that

$$Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma}, \tag{18.18}$$

so Z is obtained from X by rescaling and shifting.

Of course, whenever equation (18.17) holds, we have

$$X = \sigma Z + \mu, \tag{18.19}$$

which can be convenient.

Example 18.11 (Scaled probabilities for normal deviations). Let X be any normal random variable, with mean μ and variance σ^2. Then X and $\sigma Z + \mu$ have the same distribution. Hence for any interval $[a, b]$,

$$\begin{aligned} \mathbf{P}(X \in [a,b]) &= \mathbf{P}(\sigma Z + \mu \in [a,b]) \\ &= \mathbf{P}(a \leq \sigma Z + \mu \leq b) \\ &= \mathbf{P}\left(Z \in [\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}]\right). \end{aligned} \tag{18.20}$$

(The final equality in equation (18.20) uses the fact that $\sigma > 0$.)

When $\mu = 0$, $a = -c\sigma$ and $b = c\sigma$, we obtain using equation (18.20) that

$$\mathbf{P}(|X| \leq c\sigma) = \mathbf{P}(X \in [-c\sigma, c\sigma]) = \mathbf{P}(Z \in [-c, c]) = \mathbf{P}(|Z| \leq c). \tag{18.21}$$

If X is any normal random variable with mean μ and variance σ^2, we know that $X - \mu$ has mean zero and variance σ^2, so equation (18.21) tells us that

$$\mathbf{P}(|X - \mu| > c\sigma) = \mathbf{P}(|Z| > c). \tag{18.22}$$

Thus for any normal random variable X with mean μ and variance σ^2, we can conveniently calculate the probability of a deviation by X from its mean by *measuring the size of the deviation of X from its mean in units of σ.*

For example,

$$\begin{aligned} \mathbf{P}(|X - \mu| > \sigma) &\approx 0.31731050786291415, \\ \mathbf{P}(|X - \mu| > 2\sigma) &\approx 0.04550026389635839, \\ \mathbf{P}(|X - \mu| > 3\sigma) &\approx 0.0026997960632601866. \end{aligned} \tag{18.23}$$

Thus a deviation by one standard deviation is not uncommon, while a deviation by three standard deviations is rare.

Example 18.12. Let Z be standard normal. We will derive the following facts.

(i)

$$\mathbf{E}\left[|Z|\right] = \frac{2}{\sqrt{2\pi}}, \tag{18.24}$$

(ii) For each nonnegative integer n, if $\mathbf{E}\left[|Z|^n\right]$ exists, then $\mathbf{E}\left[|Z|^{n+2}\right]$ exists, and

$$\mathbf{E}\left[|Z|^{n+2}\right] = (n+1)\mathbf{E}\left[|Z|^n\right]. \tag{18.25}$$

(Since $\mathbf{E}\left[|Z|^0\right] = \mathbf{E}\left[1\right] = 1$, equation (18.25) tells us again that the variance of a standard normal random variable is equal to 1.)

(iii) $\mathbf{E}\left[|Z|^n\right]$ exists for each nonnegative integer n.

Proof of (i) Let Y be the random variable defined in Exercise 18.1. By that exercise, $\mathbf{Var}\left(Y\right) = 1/2$, $\mathbf{E}\left[Y\right] = 0$, $\mathbf{E}\left[\,|Y|\,\right]$ exists and $\mathbf{E}\left[\,|Y|\,\right] = \frac{1}{\sqrt{\pi}}$.

By Exercise 18.3, $\sqrt{2}Y$ is a standard normal random variable. Hence $\mathbf{E}\left[\,|Z|\,\right] = \frac{\sqrt{2}}{\sqrt{\pi}}$.

Proof of (ii) For any nonnegative integer n, assume that $\mathbf{E}\left[|Z|^n\right]$ exists. We will show that $\mathbf{E}\left[|Z|^{n+2}\right]$ exists, and equation (18.25) holds.

Note that

$$\mathbf{E}\left[|Z|^n\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |t|^n\, e^{-t^2/2}\, dt = 2\frac{1}{\sqrt{2\pi}}\int_0^{\infty} t^n e^{-t^2/2}\, dt.$$

Similarly

$$\mathbf{E}\left[|Z|^{n+2}\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |t|^{n+2}\, e^{-t^2/2}\, dt = 2\frac{1}{\sqrt{2\pi}}\int_0^{\infty} t^{n+2} e^{-t^2/2}\, dt,$$

in the sense that the expected value exists if the integral exists, and they are equal.

For any $b > 0$, using integration by parts we have

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\int_0^b t^{n+2}e^{-t^2/2}\, dt &= -\frac{1}{\sqrt{2\pi}}t^{n+1}e^{-t^2/2}\Big|_0^b + \frac{n+1}{\sqrt{2\pi}}\int_0^b t^n e^{-t^2/2}\, dt \\ &= -\frac{1}{\sqrt{2\pi}}b^{n+1}e^{-b^2/2} + (n+1)\frac{1}{\sqrt{2\pi}}\int_0^b t^n e^{-t^2/2}\, dt. \end{aligned} \tag{18.26}$$

By L'Hôpital's rule, $\lim_{b\to\infty} b^{n+1}e^{-b^2/2} = 0$. Also,

$$\lim_{b\to\infty} \int_0^b t^n e^{-t^2/2}\,dt = \int_0^\infty t^n e^{-t^2/2}\,dt.$$

Letting $b \to \infty$ in equation (18.26),

$$\lim_{b\to\infty} \frac{1}{\sqrt{2\pi}} \int_0^b t^{n+2} e^{-t^2/2}\,dt = (n+1)\frac{1}{\sqrt{2\pi}} \int_0^\infty t^n e^{-t^2/2}\,dt,$$

which gives equation (18.25).

Proof of (iii) Since $\mathbf{E}\left[\,|Z|^0\,\right] = \mathbf{E}\,[1] = 1$ exists, using equation (18.25) repeatedly shows that $\mathbf{E}\left[\,|Z|^n\,\right]$ exists for every *even* nonnegative integer n.

Since $\mathbf{E}\,[\,|Z|\,] = \frac{\sqrt{2}}{\sqrt{\pi}}$, using equation (18.25) repeatedly shows that $\mathbf{E}\left[\,|Z|^n\,\right]$ exists for every *odd* nonnegative integer n.

Exercise 18.4. Let Z be standard normal. Show that for any odd positive integer, $\mathbf{E}\,[Z^n] = 0$.

[Solution]

Exercise 18.5 (Testing Chebyshev with a normal random variable). Let X be a normal random variable with mean μ and standard deviation σ.

(i) Find the probability density for the random variable Z defined by

$$Z = \frac{X-\mu}{\sigma}.$$

(ii) Show that for any $c \geq 1$,

$$\mathbf{P}\left(|X-\mu| \geq c\sigma\right) \leq \frac{2}{\sqrt{2\pi}} e^{-\frac{c^2}{2}}. \tag{18.27}$$

(Hint: you can convert equation (18.27) into an inequality involving an integral of the density of Z, namely

$$\mathbf{P}(|Z| \geq c) = 2\mathbf{P}(Z \geq c) = 2\int_c^\infty \eta(z)\,dz = 2\int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz. \tag{18.28}$$

Then notice that for $z \geq 1$,

$$e^{-\frac{1}{2}z^2} \leq ze^{-\frac{1}{2}z^2}. \tag{18.29}$$

This trivial fact can be used to get something that we know how to integrate.)

(iii) Show that equation (18.27) gives a much sharper inequality than Chebyshev (equation (16.19)), for large c.

[Solution]

18.6 The Central Limit Theorem

Suppose that a random variable S_n is equal to the sum of n independent random variables, where n is a large number. The Central Limit Theorem says that under the right circumstances, a normal distribution can be used as a good approximation to the distribution of S_n. This type of approximation is suggested by the plots we made of various binomial distributions, in Section 18.2. The actual statement of the Central Limit Theorem makes a giant leap in generality from those binomial examples. We'll state this theorem carefully and then show more examples.

In this chapter we will try to give a thorough discussion of the Central Limit Theorem. This includes stating the theorem in more than one way. But a reader who understands the statement and the examples in the present section will have the main idea. After that the most important remaining point is the idea of stating the Central Limit Theorem for "standardized" random variables. That is discussed in Corollary 18.20 to Theorem 18.19.

First we need a simple definition.

Definition 18.13 (Identically distributed random variables).
Let $X_1, \ldots, X_n$ be random variables defined for some probability model. If all the random variables X_i have the same distribution, we say that the random variables in the sequence are identically distributed.

A sequence of random variables which is both independent and identically distributed is said to be an *IID sequence of random variables*.

The random variables X_i in Section 18.1, which give the results of a sequence of coin tosses, are a typical example of an IID sequence.

The idea of the Central Limit Theorem is sometime expressed by saying that any physical random variable whose value is the sum of "many small independent effects" should have a distribution which is similar to a normal distribution.

A mathematical special case of "many independent effects" is an IID sequence $X_1, \ldots, X_n$. We assume that $\mathbf{E}[X_i]$ and $\mathbf{Var}(X_i)$ exist. The basic form of the Central Limit Theorem says that for such a sequence, if n is large enough then the random variable $S_n = X_1 + \ldots + X_n$ has a distribution which is similar to a normal distribution. That is, if n is large enough, and if W_n is a normal random variable with the **same mean** and the **same variance** as S_n, then for every interval J we have

$$\mathbf{P}(S_n \in J) \approx \mathbf{P}(W_n \in J). \tag{18.30}$$

The formal statement is as follows.

Theorem 18.14 (The Central Limit Theorem). Let $X_1, \ldots, X_n$ be an independent, identically distributed sequence of random variables. Let $S_n = X_1 + \ldots + X_n$.

Suppose that the mean of each X_i exists and that the variance of each X_i exists and is nonzero.

(Since the X_i have identical distributions the mean of each X_i is the same, and the variance of each X_i is the same. Let the mean of X_i be μ and let the variance of X_i be $\sigma^2 > 0$. By additivity of expectation, the mean of S_n is $n\,\mu$, and equation (16.30) shows that $\mathbf{Var}(S_n) = n\sigma^2$.)

Let W_n be a normal random variable such that the mean and variance of W_n are the same as the mean and variance of S_n.

For any $\varepsilon > 0$, there exists n_0, such that for all $n \geq n_0$,

$$|\, \mathbf{P}(S_n \in J) - \mathbf{P}(W_n \in J) \,| < \varepsilon, \tag{18.31}$$

for *all* intervals J (including both finite intervals J and infinite intervals J).

As in Definition 3.1, in equation (18.31) we include intervals J which are one-point sets.

When applying the Central Limit Theorem we often deal with expressions like $\mathbf{P}\big(S_n \in J\big)$ more explicitly. Thus

$$\mathbf{P}\big(S_n \in \big(a, b\big)\big) = \mathbf{P}\big(a < S_n < b\big), \tag{18.32}$$

$$\mathbf{P}\big(S_n \in \big(-\infty, b\big)\big) = \mathbf{P}\big(S_n < b\big), \tag{18.33}$$

and so on.

A proof will not be given for the Central Limit Theorem. Figures 18.9 and J.8 illustrate the theorem, when S_n is defined by tossing a fair coin.

The Central Limit Theorem is sometimes referred to briefly as the CLT. It should be mentioned that the CLT can be expressed in various ways, and readers may find different-looking statements in other books. We'll also restate the theorem later ourselves, in Theorem 18.19 and Corollary 18.20, as well as Theorem J.8.

Motivation A skeptical person might wonder if Theorem 18.14 is worthwhile. After all, we replace the problem of finding $\mathbf{P}\left(S_n \in J\right)$ with the problem of finding $\mathbf{P}\left(W_n \in J\right)$, and the new problem may be just as hard as the old problem. But it isn't! We have a nice formula for the density of a normal random variable. Also, remember that *all* normal random variables have *the same shape*. So if we understand the distribution of the single standard normal random variable Z defined in Definition 18.10, then we understand *all* normal random variables. And the density of Z has a simple and beautiful shape, so we can hope to understand the theoretical behavior of normal random variables more easily than general ones.

Remark 18.15 (Approximation, but no rate of convergence). We know how to compute numerical probabilities using normal densities, so Theorem 18.14 shows us how to compute an approximation for $\mathbf{P}(S_n \in J)$ which is close to within an error bound of ε, for every interval J.

The approximation holds for every interval J for the same n, if n is large enough. But the statement of this theorem does not tell us how large n must be.

So, despite its formality, equation (18.31) is not more specific than equation (18.30).

From the standpoint of applications, we tend to expect such limitations. General mathematical theorems tell us what sort of behavior to look for, but precise error estimates may not be easy to come by.

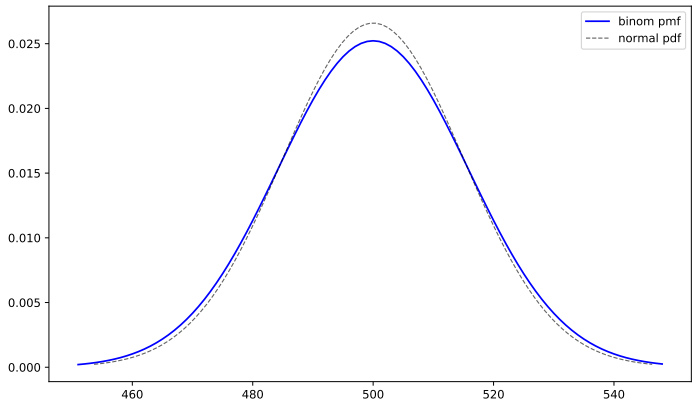


Figure 18.9: Comparing the binomial distribution ($p = .5$, $n = 1000$) with the normal density having the same mean and variance (mean $= np$, variance $= np(1-p)$).

One theoretical estimate is given by the Berry–Esseen Theorem, which says that (under slightly more restrictive conditions than Theorem 18.14) there exists a numerical constant C such that if W_n is normal and has the same mean and variance as S_n, then

$$|\mathbf{P}(W_n \in J) - \mathbf{P}(S_n \in J)| \leq C \frac{\mathbf{E}\left[|X_1|^3\right]}{\sigma^3} \frac{1}{\sqrt{n}} \tag{18.34}$$

for every interval J. Here $S_n = X_1 + \ldots + X_n$ and σ^2 is the variance of X_1. But we won't take time to study theorems of this sort.

Remark 18.16 (Do we need *identically distributed* random variables in the CLT?). In Theorem 18.14 it is assumed that the random variables $X_1, X_2, \ldots$ all have the same distribution. This is convenient as a simple mathematical assumption, but it is not necessary.

Remember that it was suggested earlier that any physical random variable whose value is the sum of many small independent effects should have a distribution which is similar to a normal distribution. In a real-world situ-

ation, it doesn't seem natural that many independent effects would all have the same probability distribution.

There are more general forms of Theorem 18.14, in which the random variables $X_1, X_2, \ldots$ are independent but do not all have the same distribution. In this situation one must impose an extra mathematical condition to obtain the result of the Central Limit Theorem. Very roughly, the idea is that the sum for S_n should be made up of terms which are comparable in size.

We won't pursue this topic mathematically, but such theorems make us more confident that the approximation described in the Central Limit Theorem is valid in many situations.

Example 18.17 (A basic approximation example). Let $X_1, \ldots, X_{10000}$ be a sequence of independent random variables.

Suppose that the distribution of each X_i has a probability density f, where $f(x) = cx^2$ on the interval $[-1, 2]$, and f is zero everywhere on the complement of $[-1, 2]$.

The constant c is of course determined by the fact that $\int f = 1$.

Let $n = 10000$. Our goal in this example to find the approximate value of $\mathbf{P}(S_n < 12600)$.

The Central Limit Theorem suggests a way to do this, by means of a normal random variable W_n with the same mean and variance as S_n.

So we need to find the mean and variance of S_n.

Since $\int f = 1$,

$$1 = \int_{-1}^{2} cx^2 \, dx = \frac{cx^3}{3}\bigg|_{-1}^{2} = \frac{8c}{3} + \frac{c}{3} = 3c.$$

Thus $c = 1/3$.

Then

$$\mathbf{E}[X_i] = \int_{-1}^{2} xcx^2 \, dx = \frac{cx^4}{4}\bigg|_{-1}^{2} = \frac{16c}{4} - \frac{c}{4} = \frac{15c}{4} = \frac{5}{4}.$$

Also

$$\mathbf{E}\left[X_i^2\right] = \int_{-1}^{2} x^2 cx^2 \, dx = \frac{cx^5}{5}\bigg|_{-1}^{2} = \frac{32c}{5} + \frac{c}{5} = \frac{33c}{5} = \frac{11}{5}.$$

Thus

$$\mathbf{Var}\,(X_i) = \frac{11}{5} - \left(\frac{5}{4}\right)^2 = \frac{11}{5} - \frac{25}{16} = \frac{176-125}{80} = \frac{51}{80}.$$

Thus, for any n, $\mathbf{Var}\,(S_n) = 51n/80$, while $\mathbf{E}\,[S_n] = 5n/4$.

Now we are ready to use the Central Limit Theorem.

For any n, let W_n have a normal distribution, with $\mathbf{Var}\,(W_n) = 51n/80$ and $\mathbf{E}\,[W_n] = 5n/4$.

Because of the Central Limit Theorem, we know that as n becomes large, the distributions of W_n and S_n become similar.

In particular, when n is large, $\mathbf{P}(W_n < a)$ and $\mathbf{P}(S_n < a)$ are close, for all a.

We will try to use this approximation when $a = 12600$ and $n = 10000$. We hope that this n is large enough so that

$$\mathbf{P}(S_n < 12600) \approx \mathbf{P}(W_n < 12600).$$

We can calculate $\mathbf{P}(W_n < 12600)$ using a computer program.

Some programs have a predefined function for working with normal distributions, but we won't assume that we have such a program. Instead, we'll work with the normal density ourselves, and then ask a computer to perform a routine numerical integration.

Let h be a probability density for the distribution of W_n, with $n = 10000$. Then

$$\mathbf{P}(W_n < 12600) = \int_{-\infty}^{12600} h.$$

What is the formula for h?

Let $m = \mathbf{E}\,[W_n] = \mathbf{E}\,[S_n] = 12500$, and let $v = \mathbf{Var}\,(W_n) = \mathbf{Var}\,(S_n) = 10000 \cdot (51/80) = 125 \cdot 51$. By Lemma 18.13, a valid density h is given by

$$h(u) = \frac{1}{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-(u-m)^2(/2v)}.$$

Using a computer program to perform the numerical integration, we find

$$\int_{-\infty}^{12600} h = \int_{-\infty}^{12600} \frac{1}{\sqrt{125 \cdot 51}} \frac{1}{\sqrt{2\pi}} e^{-(u-12500)^2/(2 \cdot 125 \cdot 51)}\, du = 0.8947967735713137. \tag{18.35}$$

So that's our approximation:

$$\mathbf{P}(S_n < 12600) \approx 0.8947967735713137 \tag{18.36}$$

See Figure 18.10. $\mathbf{P}(W_n < 12600)$ is the area of the shaded region under the graph. Note that this picture is uses a *very different scale* for vertical and horizontal distances. If we used the same scale on the vertical and horizontal axes, then the graph would be extremely low in comparison to its width!

The various numbers in equation (18.35) seem large, don't they? Computers are apparently smart enough to compute efficiently with such formulas, perhaps by a change of variable. But we know how to do that too.

Notice that one can apply a simple change of variable, $t = (u-12500)/\sqrt{125 \cdot 51}$ to equation (18.35), and it turns into an integral of the standard normal density. We'll feed this integral into the computer too.

$$\int_{-\infty}^{100/\sqrt{125\cdot 51}} h = \int_{-\infty}^{100/\sqrt{125\cdot 51}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} du \approx 0.8947967734523156 \quad (18.37)$$

Rescaling and shifting the variable of integration to get smaller numbers would make the numerical integration easier, if humans had to do it.

We can use also perform this same rescaling and shifting at a higher conceptual level, by rescaling and shifting the random variable itself. The integral in equation (18.37) then follows automatically. That is what is going on in Corollary 18.20 to Theorem 18.19.

See Remark 18.21 for further discussion of rescaling and shifting normal densities.

We have focused here on a numerical example as a way of understanding the statement of the Central Limit Theorem. But it should be noted that the Central Limit Theorem is not just a tool for obtaining approximations. It also gives us insight into the general behavior of sums of random variables. In modern probability theory that is likely the most important application of this theorem.

Exercise 18.6. Let $X_1, \ldots, X_{10000}$ be a sequence of independent random variables. For each i, $\mathbf{P}(X = -1) = \frac{2}{3}$, $\mathbf{P}(X = 2) = \frac{1}{6}$, $\mathbf{P}(X = 5) = \frac{1}{6}$.

Find a reasonable approximation to the value of $\mathbf{P}(S_{10000} < 5200)$.

Your final answer can be in the form of the integral of an explicitly given function.

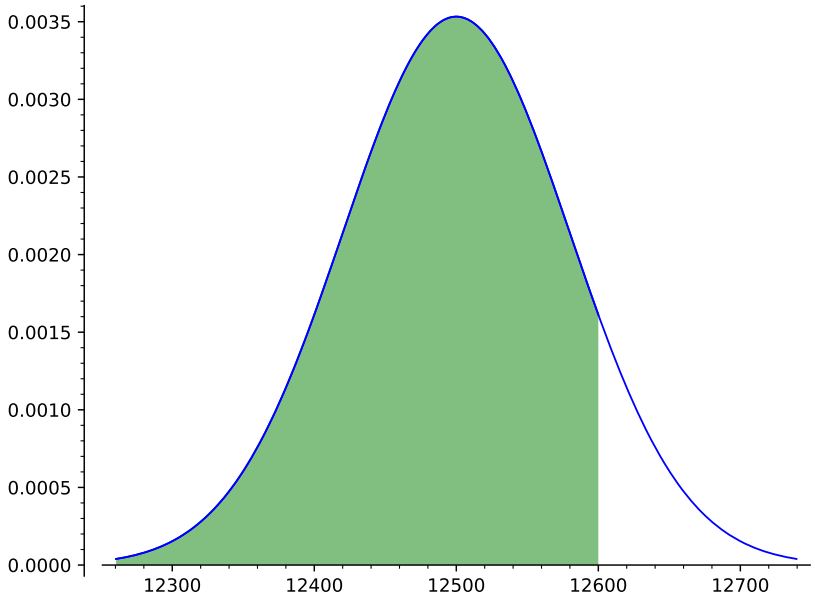


Figure 18.10: Graph of the density of W_n, showing $\mathbf{P}(W_n < 12600)$

[Solution]

Exercise 18.7. Use the Central Limit Theorem to estimate the value of the sum in equation (9.8).

Your final answer can be in the form of the integral of an explicitly given function.

[Solution]

Remark 18.18 (The case of one-point intervals). It was mentioned in the statement of Theorem 18.14 that the statement of the theorem holds for one-point intervals. Let's think about that. For any number b, by Definition 3.1, $[b, b]$ is the one-point interval containing b.

Since any normal distribution has a density, we know that for the normal random variables W_n appearing in Theorem 18.14, we always have $\mathbf{P}(W_n = b) = 0$, for all real numbers b. In other words, $\mathbf{P}(W_n \in [b, b]) = 0$ for all b.

That sort of statement is certainly not always true for S_n, as we see by thinking about coin-tossing! But the Central Limit Theorem tells us that when n is large, $\mathbf{P}(S_n \in [b,b])$ must be small for all b. And $S_n \in [b,b]$ is just another way of saying that $S_n = b$. So the Central Limit Theorem tells us that $\mathbf{P}\big(S_n = b\big)$ is small for all b, when n is sufficiently large. This is a useful fact.

But suppose for a moment that we had forgotten to include the case of one-point intervals in our statement of the Central Limit Theorem. In this remark we will note that the one-point case automatically follows when the statement holds for intervals with positive length. To see that, notice that for any interval J, since a density for W_n is given by

$$\frac{1}{\sqrt{n}\sigma} e^{-\frac{z^2}{\sqrt{n}\sigma}},$$

we have

$$\mathbf{P}\big(W_n \in J\big) = \int_J \frac{1}{\sqrt{n}\sigma} e^{-\frac{z^2}{\sqrt{n}\sigma}}\, dz \leq \frac{1}{\sqrt{n}\sigma}\mathbf{length}(J) \leq \frac{1}{\sigma}\mathbf{length}(J). \quad (18.38)$$

Thus for any $\varepsilon > 0$, for sufficiently large n we have by equation (18.31) that

$$\mathbf{P}(S_n \in J) \leq \varepsilon + \frac{1}{\sigma}\mathbf{length}(J)$$

for all intervals J.

For any point b, consider an interval J containing b with positive length less than ε. Then for sufficiently large n, since $S_n = b$ implies $S_n \in J$, we have

$$\mathbf{P}(S_n = b) \leq \mathbf{P}(S_n \in J) \leq \varepsilon + \frac{1}{\sigma}\varepsilon.$$

Since ε can be arbitrarily small, this shows that for sufficiently large n, $\mathbf{P}(S_n = b)$ is small for all b, just as Theorem 18.14 claims.

18.7 Checking the answer in Example 18.17

How accurate is the approximation in equation (18.36)? We are not going to discuss theoretical upper bounds for the error. But we're going to carry out a numerical check.

Using a computer, we will simulate the independent sequence $X_1, \ldots, X_n$, with $n = 10000$, getting a sequence of values $v_1, \ldots, v_n$. The sum $v_1 + \ldots + v_n$ is a sample value for S_n.

Call this sample value s_n. If $s_n < 12600$, we will say that the event $\{S_n < 12600\}$ occurred in the simulation.

We'll ask the computer to perform that whole procedure 1000000 times!

The fraction of the those times that $\{S_n < 12600\}$ occurs will be an estimate for $\mathbf{P}(S_n < 12600)$, based on the frequency interpretation of probability.

We can compare this estimate with the one given in equation (18.36).

Doing these simulations sounds like a lot of work. But the work is done by the computer, not us. It does take some time.

Incidentally, Section J.6 discusses an interesting transformation that makes the task easier for the computer.

In any case this simulation procedure is slower than using the Central Limit Theorem. There seems to be a trade-off. We have to think harder about concepts in order to use the Central Limit Theorem, but there is less computational work.

Doing the simulation 1000000 times, and calculating the frequency, gives the following estimate:

$$\mathbf{P}(S_n < 12600) \approx 0.895182.$$

Is that better or worse than the estimate in equation (18.36)? Your author doesn't know for sure. But at least the two estimates seem consistent.

18.8 Formulating the CLT using convergence of sequences

Stating the Central Limit Theorem in various ways can be useful. The statement of the CLT in Theorem 18.14 talked about the error between probabilities for S_n and probabilities for the corresponding normal distribution, and stated that this error is small when n is large. The next theorem expresses the same fact using familiar language about convergence of sequences.

Theorem 18.19 (The Central Limit Theorem using sequential convergence). Let $X_1, \ldots, X_n$ be an independent, identically distributed sequence of random variables. Let $S_n = X_1 + \ldots + X_n$.

Suppose that the mean of each X_i exists and that the variance of each X_i exists and is nonzero.

Since the X_i have identical distributions the mean of each X_i is the same, and the variance of each X_i is the same. Let the mean of X_i be μ and let the variance of X_i be $\sigma^2 > 0$. By additivity of expectation, the mean of S_n is $n\,\mu$, and equation (16.30) shows that $\mathbf{Var}\,(S_n) = n\sigma^2$.

Let W_n be a normal random variable such that the mean and variance of W_n are the same as the mean and variance of S_n.

Let J_n be any sequence of intervals.

Then

$$\lim_{n\to\infty} \Big(\mathbf{P}\big(S_n \in J_n\big) - \mathbf{P}\big(W_n \in J_n\big) \Big) = 0. \tag{18.39}$$

As in Definition 3.1, in equation (18.31) we include intervals J which are one-point sets.

Taking J_n to be the one-point interval $\big[b_n\big]$ in equation (18.39), we see that for any sequence of numbers b_n,

$$\lim_{n\to\infty} \mathbf{P}\big(S_n = b_n\big) = 0. \tag{18.40}$$

Proof. The usual definition of sequential convergence shows that Theorem 18.14 implies that equation (18.39) of Theorem 18.19 holds for any sequence of intervals J_n. The converse implication is not hard to prove, but requires a small argument, one which readers might have seen when studying uniform convergence in calculus. The details are omitted. □

In applications one can of course use whichever formulation of the Central Limit Theorem seems most convenient. The following corollary uses "standardized" versions of random variables (see Exercise 18.3) and also writes out $\mathbf{P}(S_n \in J_n)$ explicitly using inequalities, as we saw in equations (18.32) and (18.33).

Corollary 18.20 (Standardizing the statement of the Central Limit Theorem). Let $X_1, \ldots, X_n$ satisfy the assumptions of Theorem 18.19.

Note that

$$\begin{aligned} \mathbf{E}\left[\frac{S_n - n\mu}{\sqrt{n}\sigma}\right] = 0 \text{ and } \mathbf{Var}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma}\right) = 1 \\ \mathbf{E}\left[\frac{W_n - n\mu}{\sqrt{n}\sigma}\right] = 0 \text{ and } \mathbf{Var}\left(\frac{W_n - n\mu}{\sqrt{n}\sigma}\right) = 1. \end{aligned} \tag{18.41}$$

Let Z be a standard normal random variable (See Definition 18.10).

(i)

For any sequence of numbers a_n, b_n with $a_n \leq b_n$,

$$\lim_{n\to\infty}\left(\mathbf{P}\left(\frac{a_n-n\mu}{\sqrt{n}\sigma} \leq \frac{S_n-n\mu}{\sqrt{n}\sigma} \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right)-\mathbf{P}\left(\frac{a_n-n\mu}{\sqrt{n}\sigma} \leq Z \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right)\right)=0. \tag{18.42}$$

(ii) For any sequence of numbers b_n,

$$\lim_{n\to\infty}\left(\mathbf{P}\left(\frac{S_n-n\mu}{\sqrt{n}\sigma} \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right)-\mathbf{P}\left(Z \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right)\right)=0. \tag{18.43}$$

Similar statements to (i), (ii) hold for any sequence of intervals.

Proof. For any sequence of intervals $\big(a_n, b_n\big]$,

$$\begin{aligned}\mathbf{P}\big(a_n \leq S_n \leq b_n\big) &= \mathbf{P}\big(a_n-n\mu \leq S_n-n\mu \leq b_n-n\mu\big)\\ &= \mathbf{P}\left(\frac{a_n-n\mu}{\sqrt{n}\sigma} \leq \frac{S_n-n\mu}{\sqrt{n}\sigma} \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right),\end{aligned} \tag{18.44}$$

$$\begin{aligned}\mathbf{P}\big(a_n \leq W_n \leq b_n\big) &= \mathbf{P}\big(a_n-n\mu \leq W_n-n\mu \leq b_n-n\mu\big)\\ &= \mathbf{P}\left(\frac{a_n-n\mu}{\sqrt{n}\sigma} \leq \frac{W_n-n\mu}{\sqrt{n}\sigma} \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right).\end{aligned} \tag{18.45}$$

Applying these facts to equation (18.39) gives equation (18.42).

Using the same arguments again, for any sequence of intervals $\big(-\infty, b_n\big]$,

$$\begin{aligned}\mathbf{P}\big(S_n \leq b_n\big) &= \mathbf{P}\big(S_n-n\mu \leq b_n-n\mu\big)\\ &= \mathbf{P}\left(\frac{S_n-n\mu}{\sqrt{n}\sigma} \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right),\end{aligned} \tag{18.46}$$

$$\begin{aligned}\mathbf{P}\big(W_n \leq b_n\big) &= \mathbf{P}\big(W_n-n\mu \leq b_n-\mu\big)\\ &= \mathbf{P}\left(\frac{W_n-n\mu}{\sqrt{n}\sigma} \leq \frac{b_n-n\mu}{\sqrt{n}\sigma}\right).\end{aligned} \tag{18.47}$$

Applying these facts to equation (18.39) gives equation (18.43).

□

Remark 18.21 (Natural scaling). Equations (18.42) and (18.43) seem natural if we are trying to picture the distribution of S_n, as $n \to \infty$.

For example, by replacing S_n by the centered random variable $S_n - n\mu$, we are trying ensure that the distribution stays in the picture. By subtracting the mean we keep the distribution from drifting off to ∞ or $-\infty$ as $n \to \infty$.

However, the distribution of $S_n - n\mu$ still gets wider and wider as n grows, so it is still escaping from our view. To prevent this widening we replace $S_n - n\mu$, by $(S_n - n\mu)/(\sqrt{n}\sigma)$, which has variance equal to one.

Example 18.22. In the setting of Theorem 18.19, suppose that the random variables X_i have mean zero and variance σ^2.

What can we say about the value of $\mathbf{P}\big(S_n \le 100\sqrt{n}\big)$, when n is large?

Let $b_n = 100\sqrt{n}$.

By equation (18.43),

$$\lim_{n\to\infty} \left(\mathbf{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le \frac{b_n - n\mu}{\sqrt{n}} \right) - \mathbf{P}\left(Z \le \frac{b_n - n\mu}{\sqrt{n}\sigma} \right) \right) = 0.$$

In this equation μ is the mean of X_i, so μ is zero.

Applying equation (18.46) to this equation gives

$$\lim_{n\to\infty} \left(\mathbf{P}\big(S_n \le b_n\big) - \mathbf{P}\left(Z \le \frac{b_n}{\sqrt{n}\sigma} \right) \right) = 0. \tag{18.48}$$

Since $b_n = 100\sqrt{n}$, this says that

$$\lim_{n\to\infty} \left(\mathbf{P}\big(S_n \le 100\sqrt{n}\big) - \mathbf{P}\left(Z \le \frac{100}{\sigma} \right) \right) = 0. \tag{18.49}$$

Since $100/\sigma$ does not depend on n, this equation says that $\mathbf{P}\big(S_n \le 100\sqrt{n}\big) \to \mathbf{P}\left(Z \le \frac{100}{\sigma} \right)$, i.e.

$$\lim_{n\to\infty} \mathbf{P}\big(S_n \le 100\sqrt{n}\big) = \mathbf{P}\left(Z \le \frac{100}{\sigma} \right). \tag{18.50}$$

Using the density for the standard normal distribution,

$$\lim_{n\to\infty} \mathbf{P}\big(S_n \le 100\sqrt{n}\big) = \int_{-\infty}^{100/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz. \tag{18.51}$$

The integral on the right side of this equation can be computed numerically.

Exercise 18.8. In the setting of Theorem 18.19, suppose that the random variables X_i have mean zero and variance σ^2.

(i) What can you say about $\mathbf{P}(S_n \leq 7)$ when n is large?

(ii) Let $a_n = 2\sqrt{n}$, and let $b_n = 11\sqrt{n}$. Find

$$\lim_{n\to\infty} \mathbf{P}(S_n \in [a_n, b_n]).$$

Your answer may involve an integral.

[Solution]

Example 18.23. In the setting of Theorem 18.19, let a_n and b_n any sequences of numbers with $a_n \leq b_n$. We would like to know whether

$$\mathbf{P}(a_n \leq S_n \leq b_n) \approx 0$$

when n is large.

In other words, we would like to know whether or not

$$\lim_{n\to\infty} \mathbf{P}(a_n \leq S_n \leq b_n) = 0. \tag{18.52}$$

Let

$$u_n = \frac{a_n - n\mu}{\sqrt{n}\sigma}, \; v_n = \frac{b_n - n\mu}{\sqrt{n}\sigma}.$$

By equations (18.42) and (18.44),

$$\lim_{n\to\infty} \Big(\mathbf{P}(a_n \leq S_n \leq b_n) - \mathbf{P}(u_n \leq Z \leq v_n) \Big) = 0.$$

Thus $\mathbf{P}(a_n \leq S_n \leq b_n) \to 0$ if $\mathbf{P}(u_n \leq Z \leq v_n) \to 0$.

Using the density for the standard normal distribution, we have

$$\mathbf{P}(u_n \leq Z \leq v_n) = \int_{u_n}^{v_n} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz.$$

So it's sufficient to show that

$$\int_{u_n}^{v_n} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \to 0. \tag{18.53}$$

Equation (18.53) holds when $u_n - v_n \to 0$, because

$$\left| \int_{u_n}^{v_n} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \right| \leq \frac{1}{\sqrt{2\pi}} \left| v_n - u_n \right|.$$

As concrete example of this situation, suppose that $b_n = a_n + 10000000 n^{1/4}$, so that $b_n - a_n \to \infty$. It might not be clear whether or not equation (18.52) holds. However, it is easy to see that $u_n - v_n \to 0$, and so equation (18.52) does indeed hold.

Remark 18.24 (Another case). In Example 18.23 didn't cover every situation in which $\mathbf{P}(a_n \leq S_n \leq b_n) \to 0$. A similar argument shows that the same convergence holds when $u_n \to \infty$ or when $v_n \to -\infty$.

That is because

$$\lim_{n\to\infty} \int_{u_n}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = 0$$

and

$$\lim_{n\to\infty} \int_{-\infty}^{v_n} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = 0,$$

using the definition of an improper integral.

Exercise 18.9 (Expressing the Central Limit Theorem again as a standardized limit). In the setting of Theorem J.8, prove that for any real number a,

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq a \right) = \mathbf{P}(Z \leq a), \tag{18.54}$$

where Z is standard normal.

You will sometimes see this statement in textbooks. It looks weaker than Theorem 18.19, since it only involves a single sequence. However, as usual, one can show that it is equivalent.

[Solution]

18.9 Checking the Central Limit Theorem for another binomial distribution

Recall that we started our discussion of the Central Limit Theorem by noting the shape of the graphs in Figures 18.4, 18.5 and 18.6. These graphs are only a tiny example, since they only deal with binomials for which $p = .5$. Theorem 18.14 of course applies to every binomial distribution, and to an enormous zoo of other distributions.

Although we won't do much testing of the Central Limit Theorem, we can at least try another binomial distribution. Let's graph some binomial distributions for $p = .99$. This value of p favors success in an extreme way, so it certainly changes the shape of the binomial distributions. The Central Limit Theorem asserts that these binomial distributions will still become approximately normal as n grows large.

Recall the interval $[\ell(n), u(n)]$ defined in Terminology 18.1, which contains most of the probability in the distribution. With $p = .99$, we compute that $\mu = 99$ and $[\ell(n), u(n)] = [95, 100]$ when $n = 100$, $\mu = 990$ and $[\ell(n), u(n)] = [979, 998]$ when $n = 1000$, $\mu = 9900$ and $[\ell(n), u(n)] = [9868, 9929]$ when $n = 10000$.

Graphs showing the main parts of the binomial distribution, for $p = .99$ and $n = 100, 1000, 10000$, are given in Figures 18.11, 18.12 and 18.13.

Things look bad when $n = 100$! However, we can see that the shape of the graph seems to be somewhat similar to the shape of a normal density when $n = 1000$, and is more similar when $n = 10000$, in rough agreement with the Central Limit Theorem.

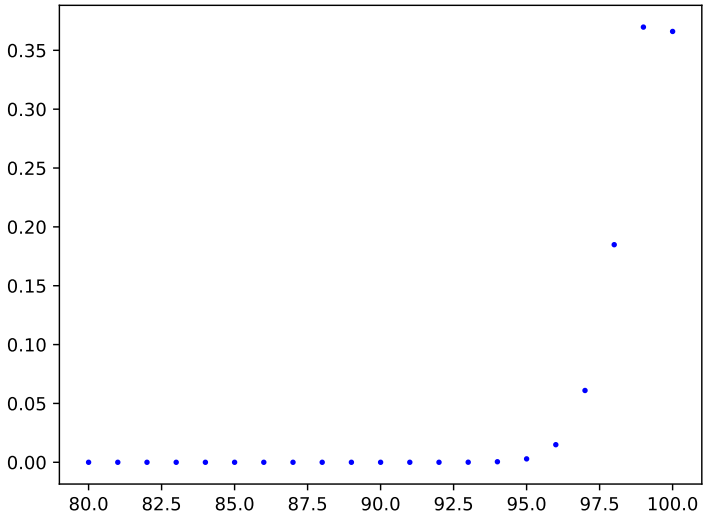


Figure 18.11: Main values of $\mathbf{P}(S_n = k)$ for $n = 100$, $p = .99$.

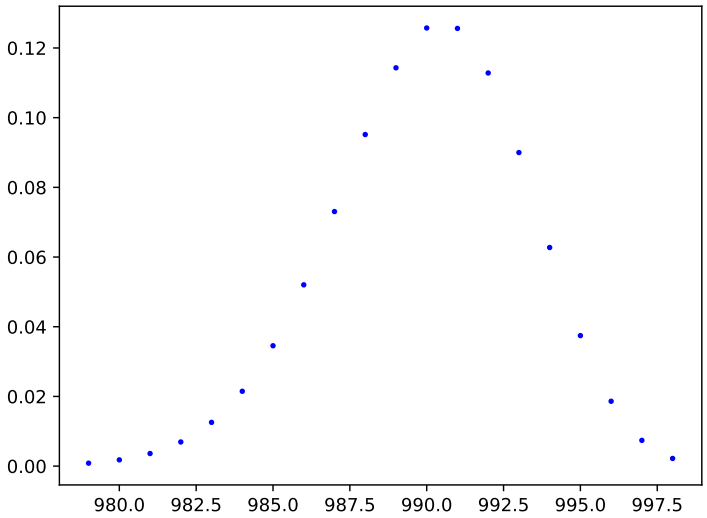


Figure 18.12: Main values of $\mathbf{P}(S_n = k)$ for $n = 1000$, $p = .99$

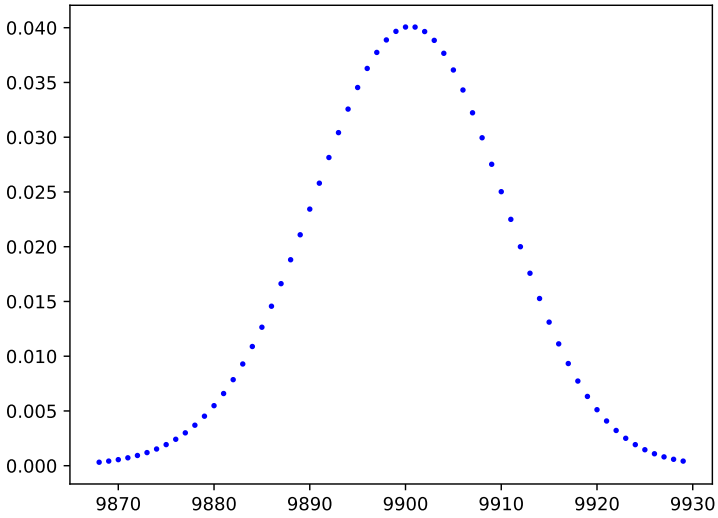


Figure 18.13: Main values of $\mathbf{P}(S_n = k)$ for $n = 10000$, $p = .99$

18.10 Sums of independent normals

In this section we consider the density of a random variable which is formed by adding up independent normal random variables. Theorem 18.27 states very useful fact: it is a normal density too! That is the main point of the section, together with Section 18.11, which shows how the Central Limit Theorem helps us to think about this theorem.

The proof of Theorem 18.27 assumes knowledge of the convolution operation for functions, defined in Section K.5. By equation (K.13), for any integrable functions f, g on the real line,

$$f * g(t) = \int_{-\infty}^{\infty} f(x)g(t-x)\,dx. \tag{18.55}$$

The proof of Theorem 18.27 also uses an alternate form of the normal density. Such alternate forms are discussed in Section 18.12.

Definition 18.25 (The mean zero normal densities). In equation (18.16) we defined the density η, the standard normal density with mean zero and variance one.

For any $t > 0$, define η_t by

$$\eta_t(x) = \frac{1}{\sqrt{t}}\eta(\tfrac{x}{\sqrt{t}}) = \frac{1}{\sqrt{2\pi t}}e^{-\frac{x^2}{2t}}. \tag{18.56}$$

By Corollary 18.7, η_t is a density for a random variable with mean zero and variance t.

Of course, $\eta_1 = \eta$.

Lemma 18.26. Let Z_a be a random variable with mean zero and variance a, and let Z_b be a random variable with mean zero and variance b. Suppose that Z_a and Z_b are independent.

Then $Z_a + Z_b$ is normal, with mean zero and variance $a + b$. Also

$$\eta_a * \eta_b = \eta_{a+b}. \tag{18.57}$$

Here $\eta_a * \eta_b$ is the convolution of the functions η_a, η_b, and is given by

$$\eta_a * \eta_b(t) = \int_{-\infty}^{\infty} \eta_a(x)\eta_b(t-x)\,dx. \tag{18.58}$$

Proof. As noted in Definition 18.25, η_a is a density for Z_a and η_b is a density for Z_b.

Therefore $\eta_a * \eta_b$ is a density for $Z_a + Z_b$, by Section K.5.

We will perform a calculation to show that $\eta_a * \eta_b$ is a normal density. This will show that $Z_a + Z_b$ is normal, and the rest will follow.

We have

$$\eta_a * \eta_b(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi a}} e^{-\frac{t^2}{2a}} \frac{1}{\sqrt{2\pi b}} e^{-\frac{(z-t)^2}{2b}}\, dt.$$

Hence

$$\begin{aligned} \eta_a * \eta_b(z) &= \frac{1}{\sqrt{2\pi a}} \frac{1}{\sqrt{2\pi b}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2a} - \frac{(z-t)^2}{2b}}\, dt \\ &= \frac{1}{\sqrt{2\pi a}} \frac{1}{\sqrt{2\pi b}} \int_{-\infty}^{\infty} e^{-v}\, dt, \end{aligned} \tag{18.59}$$

where

$$v = \frac{t^2}{2a} + \frac{(z-t)^2}{2b} = \frac{t^2}{2a} + \frac{1}{2b}\left(z^2 - 2zt + t^2\right). \tag{18.60}$$

Thus

$$v = t^2\left(\frac{1}{2a} + \frac{1}{2b}\right) - \frac{z}{b}t + \text{ stuff},$$

where the "stuff" does not involve t, and is a quadratic polynomial in z.

After completing the square for the variable t, we see that

$$v = \alpha_1(t - \alpha_2 z)^2 + \text{ stuff},$$

where α_1, α_2 are constants, $\alpha_1 > 0$ and the "stuff" does not involve t, and is a quadratic polynomial in z. Thus

$$\int e^{-v}\, dt = e^{-\text{ stuff}} \int e^{-\alpha_1(t-\alpha_2 z)^2}\, dt.$$

Letting $s = t - \alpha_2 z$, we can see that

$$\int e^{-\alpha_1(t-\alpha_2 z)^2}\, dt = \int e^{-\alpha_1 s^2}\, ds = \alpha_3,$$

where α_3 is some constant which does not depend on z.

So we know that

$$\eta_a * \eta_b = de^{-k_2 z^2 - k_1 z - k_0},$$

for some constants k_2, k_1, k_0, d. We also know that the convolution of probability densities is again a probability density, by Section K.5. And so d cannot be zero, since the integral of a density must be equal to one.

If k_2 were not a positive number, $\int_{-\infty}^{\infty} de^{-k_2 z^2 - k_1 z - k_0}\, dz$ could not exist. But it does, and so we conclude that $k_2 > 0$.

By Lemma 18.28, $de^{-k_2 z^2 - k_1 z - k_0}$ is a normal density, i.e. $\eta_a * \eta_b$ is a normal density.

Hence $Z_a + Z_b$ is a normal random variable.

Since Z_a and Z_b are mean zero random variables, $Z_a + Z_b$ has mean zero.

Also, by Corollary 16.17, $\mathbf{Var}(Z_a + Z_b) = \mathbf{Var}(Z_a) + \mathbf{Var}(Z_b) = a + b$. Since we know that the mean and the variance of $Z_a + Z_b$ match the mean and variance of η_{a+b}, we know that the distribution of $Z_a + Z_b$ has density η_{a+b}.

□

Lemma 18.26 implies:

Theorem 18.27 (A sum of independent normals is normal). Let X and Y be normal random variables. If X and Y are independent then $X + Y$ is normal.

Proof. Lemma 18.26 takes care of the mean zero case.

Thus $X - \mathbf{E}[X] + Y - \mathbf{E}[Y]$ is normal.

The random variable $X + Y$ is obtained from $X - \mathbf{E}[X] + Y - \mathbf{E}[Y]$ by shifting (adding $\mathbf{E}[X] + \mathbf{E}[Y]$). Since a shifted normal distribution is normal (Remark 18.8), $X + Y$ is normal.

□

18.11 Why should we have expected that Theorem 18.27 holds?

Much as in the case of the Poisson distribution (Lemma 17.6), our picture of the normal distribution as an approximation to a sum of independent random

variables suggests that $X + Y$ must be normal, even without a proof. One can use the following argument.

We can think of X as statistically similar to a sum of many small independent random variables $U_1, \ldots, U_n$.

Similarly we can think of Y as statistically similar to a sum of small independent random variables $V_1, \ldots, V_m$.

We can imagine that the whole sequence $U_1, \ldots, U_n, V_1, \ldots, V_m$ is independent. This is because we can think of measuring values of X for repetitions of an experiment, and measuring values for Y for repetitions of a completely different experiment.

The physical motivation for the Central Limit Theorem tells us that the distribution of $U_1 + \ldots + U_n + V_1 + \ldots + V_m$ should be approximately a normal distribution. Thus the statistics for $X+Y$ should be that of a normal distribution.

18.12 Manipulating normal densities

The next lemma gives us a way to recognize a normal density if its formula is written in a nonstandard way. Despite its messy appearance, it is not hard.

Lemma 18.28 (Equivalent forms of normal densities). The following statements are equivalent.

(i) g is a normal density, i.e. for some κ, m with $\kappa \neq 0$,

$$g(x) = \frac{1}{\sqrt{\kappa^2 \pi}} e^{-\frac{(x-m)^2}{\kappa^2}}. \tag{18.61}$$

(ii) g is a probability density such that

$$g(x) = d e^{-(k_2 x^2 + k_1 x + k_0)}, \tag{18.62}$$

where k_2, k_1, k_0, d are constants and $k_2 > 0$.

Equations (18.61) and (18.62) hold for the same probability density g when:

$$\begin{aligned} k_2 &= \tfrac{1}{\kappa^2}, \\ k_1 &= -\tfrac{2m}{\kappa^2} \end{aligned} \tag{18.63}$$

and also

$$d = \frac{1}{\sqrt{\kappa^2 \pi}} e^{k_0 - \frac{m^2}{\kappa^2}}. \tag{18.64}$$

Proof. **Going from Equation** (18.61) **to Equation** (18.62)**:**

Suppose that g given by equation (18.61).

Expanding the square in the exponent, we can rewrite $\frac{(x-m)^2}{\kappa^2}$ as

$$k_2 x^2 + k_1 x + k_0,$$

where equation (18.63) holds and also

$$k_0 = \frac{m^2}{\kappa^2}. \tag{18.65}$$

Thus equation (18.62) holds with $d = \frac{1}{\sqrt{\kappa^2 \pi}}$, and equation (18.64) holds.

Going from Equation (18.62) **to Equation** (18.61)**:**

If g is given by equation (18.62), we will obtain an equation for g which is similar to equation (18.61), by "completing the square". (Readers have likely seen such manipulations before. Appendix I recalls that procedure.)

Let

$$\kappa = \frac{1}{\sqrt{k_2}}$$

and let

$$m = -\frac{\kappa^2 k_1}{2} = -\frac{k_1}{2k_2},$$

Then expanding $(x-m)^2$ shows that

$$\frac{(x-m)^2}{\kappa^2} = k_2 x^2 + k_1 x + \frac{m^2}{\kappa^2} = k_2 x^2 + k_1 x + \frac{k_1^2}{4k_2} = k_2 x^2 + k_1 x + k_0 + \left(\frac{k_1^2}{4k_2} - k_0 \right).$$

Hence

$$d e^{-\left(k_2 x^2 + k_1 x + k_0\right)} e^{-\frac{k_1^2}{4k_2} + k_0} = d e^{-\frac{(x-m)^2}{\kappa^2}},$$

and so

$$d e^{-\left(k_2 x^2 + k_1 x + k_0\right)} = d e^{\frac{k_1^2}{4k_2} - k_0} e^{-\frac{(x-m)^2}{\kappa^2}}. \tag{18.66}$$

Since g is a density,

$$\int de^{-\left(k_2x^2+k_1x+k_0\right)}\,dx = 1.$$

Since $\frac{1}{\sqrt{\kappa^2\pi}}e^{-\frac{(x-m)^2}{\kappa^2}}$ is a density,

$$\int \frac{1}{\sqrt{\kappa^2\pi}}e^{-\frac{(x-m)^2}{\kappa^2}} = 1, \text{ i.e. } \int e^{-\frac{(x-m)^2}{\kappa^2}} = \sqrt{\kappa^2\pi}.$$

Thus integrating equation (18.66) shows that equation (18.64) holds. Since

$$de^{\frac{k_1^2}{4k_2}-k_0} = \frac{1}{\sqrt{\kappa^2\pi}},$$

equation (18.66) shows that g satisfies equation (18.61).

□

Example 18.29 (Identifying densities using the variable parts). Let h be a probability density given by $h(x) = de^{-\left(ax^2+bx+c\right)}$, for some constants d, a, b, c.

Let h_1 be probability density given by $h_1(x) = d_1e^{-\left(a_1x^2+b_1x+c_1\right)}$, for some constants $d_1, a_1, b_1, 1$.

Suppose that $a_1 = a$ and $b_1 = b$. We will show that then

$$d_1e^{-c_1} = de^{-c}, \text{ and so } h_1 = h. \tag{18.67}$$

Indeed, since $\int h = 1 = \int h_1$,

$$\int de^{-c}e^{-\left(ax^2+bx\right)} = \int d_1e^{-c_1}e^{-\left(a_1x^2+b_1x\right)}.$$

That is,

$$de^{-c_1}\int e^{-\left(ax^2+bx\right)} = d_1e^{-c_1}\int e^{-\left(a_1x^2+b_1x\right)},$$

and the two integrals in this equation are identical, so equation (18.67) holds.

Remark 18.30 (Absorbing a constant). Suppose that a probability density h is written as:

$$h(x) = de^{-\left(ax^2+bx+c\right)}, \tag{18.68}$$

where $d, a, b, c.$ are constants. As we have noticed, we can write

$$g(x) = re^{-(ax^2+bx)},$$

where $r = de^{-c}$. In this situation one sometimes says that we have absorbed the constant c into the constant r.

18.13 Solutions for Chapter 18

Solution (Exercise 18.1).

(i)

$$\mathbf{E}\left[Y^2\right] = \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{\pi}} e^{-y^2}\, dy = \frac{2}{\sqrt{\pi}} \int_0^{\infty} y^2 e^{-y^2}\, dy,$$

provided that the integral exists. We show the integral exists and evaluate it at the same time:

$$\begin{aligned}\int_0^{\infty} y^2 e^{-y^2}\, dy &= \lim_{b\to\infty} \int_0^b y^2 e^{-y^2}\, dy = \frac{1}{2} \lim_{b\to\infty} \int_0^b y \cdot 2ye^{-y^2}\, dy \\ &= \lim_{b\to\infty} \frac{1}{2} \left(-ye^{-y^2}\Big|_0^b + \int_0^b e^{-y^2}\, dy \right) \\ &= 0 + \frac{1}{2}\int_0^{\infty} e^{-y^2}\, dy = \frac{1}{4}\int_{-\infty}^{\infty} e^{-y^2}\, dy = \frac{\sqrt{\pi}}{4}.\end{aligned}$$

We used integration by parts to obtain the second line of the equation.

Combining our facts, $\mathbf{Var}\,(Y) = \frac{2}{\sqrt{\pi}}\frac{\sqrt{\pi}}{4} = 1/2$.

(ii) By equation (15.6),

$$\mathbf{E}\,[Y] = \int_{-\infty}^{\infty} y \frac{1}{\sqrt{\pi}} e^{-y^2}\, dy = \int_{-\infty}^{0} y \frac{1}{\sqrt{\pi}} e^{-y^2}\, dy + \int_0^{\infty} y \frac{1}{\sqrt{\pi}} e^{-y^2}\, dy.$$

A trivial change of variable shows that

$$\int_{-\infty}^{0} y\frac{1}{\sqrt{\pi}}e^{-y^2}\,dy = -\int_{0}^{\infty} y\frac{1}{\sqrt{\pi}}e^{-y^2}\,dy,$$

so cancellation takes place, and equation (18.5) holds.

(iii)
By equation (15.6),

$$\begin{aligned}\mathbf{E}\left[|Y|\right] &= \int_{-\infty}^{\infty} |y|\,\frac{1}{\sqrt{\pi}}e^{-y^2}\,dy = \frac{1}{\sqrt{\pi}}\int_{0}^{\infty} 2ye^{-y^2}\,dy = \frac{1}{\sqrt{\pi}}\lim_{b\to\infty}\int_{0}^{b} 2ye^{-y^2}\,dy \\ &= \lim_{b\to\infty}\frac{1}{\sqrt{\pi}}\left(-e^{-y^2}\Big|_0^b\right) = \lim_{b\to\infty}\frac{1}{\sqrt{\pi}}\left(1-e^{-b^2}\right) = \frac{1}{\sqrt{\pi}}.\end{aligned}$$

Solution (Exercise 18.2).

(i) For an interval $[a, b]$,

$$\begin{aligned}\mathbf{P}(W\in[a,b]) &= \mathbf{P}(X+m\in[a,b]) = \mathbf{P}(a\le X+m\le b) \\ &= \mathbf{P}(a-m\le X\le b-m) = \mathbf{P}(X\in[a-m,b-m]) \\ &= \int_{a-m}^{b-m} f(x)\,dx = \int_{a}^{b} f(w-m)\,dw.\end{aligned}$$

By Remark 9.12, $f(w-m)$ is a probability density for the distribution of W.

(ii) Suppose that $\kappa > 0$. For an interval $[a, b]$,

$$\begin{aligned}\mathbf{P}(V\in[a,b]) &= \mathbf{P}(\kappa X\in[a,b]) = \mathbf{P}(a\le\kappa X\le b) \\ &= \mathbf{P}\left(\frac{a}{\kappa}\le X\le\frac{b}{\kappa}\right) = \mathbf{P}\left(X\in\left[\frac{a}{\kappa},\frac{b}{\kappa}\right]\right) \\ &= \int_{a/\kappa}^{b/\kappa} f(x)\,dx = \frac{1}{\kappa}\int_{a}^{b} f(v/\kappa)\,dv.\end{aligned}$$

By Remark 9.12, $(1/\kappa)f(v/\kappa)$ is a probability density for the distribution of V.

Suppose that $\kappa < 0$. Then $|\kappa| = -\kappa$. For an interval $[a, b]$,

$$\begin{aligned}\mathbf{P}(V \in [a,b]) &= \mathbf{P}(\kappa X \in [a,b]) = \mathbf{P}(a \le \kappa X \le b) \\ &= \mathbf{P}\left(\frac{b}{\kappa} \le X \le \frac{a}{\kappa}\right) = \mathbf{P}\left(X \in \left[\frac{b}{\kappa}, \frac{a}{\kappa}\right]\right) \\ &= \int_{b/\kappa}^{a/\kappa} f(x)\,dx = \frac{1}{\kappa}\int_b^a f(v/\kappa)\,dv = \frac{1}{|\kappa|}\int_a^b f(v/\kappa)\,dv.\end{aligned}$$

By Remark 9.12, $(1/|\kappa|)f(v/\kappa)$ is a probability density for the distribution of V.

(iii) Let $g(u) = (1/|\kappa|)f(u/\kappa)$.

By part (ii), g is a probability density for the distribution of κX.

Then by part (i), $g(u-m)$ is a probability density for the distribution of $\kappa X + m$.

And $g(u-m) = (1/|\kappa|)f(\,(u-m)/\kappa\,)$.

Solution (Exercise 18.3).

$$\mathbf{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathbf{Var}\,(X-\mu) = \frac{1}{\sigma^2}\mathbf{Var}\,(X) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

$$\mathbf{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}\mathbf{E}\,[X-\mu] = \frac{1}{\sigma}\,(\mu-\mu) = 0.$$

Solution (Exercise 18.4). Let X be standard normal and let n be an odd nonnegative integer. We must show that $\mathbf{E}\,[X^n] = 0$.

By equation (15.6),

$$\mathbf{E}\,[X^n] = \int_{-\infty}^{\infty} x^n \frac{1}{\sqrt{\pi}} e^{-x^2}\,dx = \int_{-\infty}^{0} x^n \frac{1}{\sqrt{\pi}} e^{-x^2}\,dx + \int_0^{\infty} x^n \frac{1}{\sqrt{\pi}} e^{-x^2}\,dx$$

A trivial change of variable shows that

$$\int_{-\infty}^{0} x^n \frac{1}{\sqrt{\pi}} e^{-x^2}\,dx = -\int_0^{\infty} x^n \frac{1}{\sqrt{\pi}} e^{-x^2}\,dx,$$

so $\mathbf{E}\,[X^n] = 0$.

Solution (Exercise 18.5).

(i) By Exercise 18.3, Z is standard normal. Hence a density for the distribution of Z is η, where

$$\eta(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

(ii)

$$\mathbf{P}(|X-\mu| \geq c\sigma) = \mathbf{P}\left(\left|\frac{X-\mu}{\sigma}\right| \geq c\right) = \mathbf{P}(|Z| \geq c).$$

Since $e^{-\frac{z^2}{2}} \leq z e^{-\frac{z^2}{2}}$, equation (18.28) implies that

$$\mathbf{P}(|Z| \geq c) \leq 2\int_c^\infty \frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}}\, dz = \frac{\sqrt{2}}{\sqrt{\pi}}(-1)e^{-\frac{z^2}{2}}\Big|_c^\infty = \frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{c^2}{2}}.$$

(iii) Equation(16.19) says that

$$\mathbf{P}\left(|X-\mu| \geq c\sigma\right) \leq \frac{1}{c^2}.$$

So we are asking which is a better bound when c is large:

$$\frac{1}{c^2} \text{ or } \frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{c^2}{2}}?$$

As $c \to \infty$,

$$\frac{\frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{c^2}{2}}}{\frac{1}{c^2}} \to 0,$$

(and does so very rapidly), so $\frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{c^2}{2}}$ is a much sharper bound when c is large.

Solution (Exercise 18.6).

$$\mathbf{E}\left[X_i\right] = -\frac{2}{3} + \frac{2}{6} + \frac{5}{6} = \frac{1}{2}.$$

$$\mathbf{E}\left[X_i^2\right] = \frac{2}{3} + \frac{4}{6} + \frac{25}{6} = \frac{33}{6} = \frac{11}{2}.$$

$$\mathbf{Var}\,(X_i) = \frac{11}{2} - \left(\frac{1}{2}\right)^2 = \frac{21}{4}.$$

Each X_i has standard deviation σ given by

$$\sigma = \frac{\sqrt{21}}{2}.$$

We will try to find an approximation for $\mathbf{P}(S_n < a)$, for any a.
By additivity, $\mathbf{E}\,[S_n] = \frac{n}{2}$.
By equation (16.30), $\mathbf{Var}\,(S_n) = \frac{21n}{4}$.
Let W_n have a normal distribution, with $\mathbf{Var}\,(W_n) = \frac{21n}{4}$ and $\mathbf{E}\,[W_n] = \frac{n}{2}$.
Because of the Central Limit Theorem, we know that for large n we have

$$\mathbf{P}(S_n < 5200) \approx \mathbf{P}(W_n < 5200).$$

We hope that $n = 10000$ will give a reasonable approximation.

Let h be the probability density for the distribution of W_n, with $n = 10000$. Then

$$\mathbf{P}(W_n < 5200) = \int_{-\infty}^{5200} h. \tag{18.69}$$

To finish the problem, we need to know the formula for h.

Let $m = \mathbf{E}\,[W_n] = \mathbf{E}\,[S_n] = 5000$, and let $v = \mathbf{Var}\,(W_n) = \mathbf{Var}\,(S_n) = 2500 \cdot 21$. By Lemma 18.13,

$$h(u) = \frac{1}{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-(u-m)^2/2v}.$$

This completes the solution. To obtain the actual numerical value of $\mathbf{P}(W_n < 5200)$ one can use a computer program to perform the numerical integration. This gives

$$\int_{-\infty}^{5200} h = \int_{-\infty}^{5200} \frac{1}{\sqrt{2500 \cdot 21}} \frac{1}{\sqrt{2\pi}} e^{-(u-5000)^2/(2 \cdot 2500 \cdot 21)}\, du = 0.8086334555573861$$

This integral can be simplified numerically by a simple change of variable $t = (u - 5000)/\sqrt{2500 \cdot 21}$, although computer algebra programs for finding integrals can handle it as is. See the comment in Example 18.17 about equation (18.37), relating such transformations to the idea of rescaling and shifting random variables.

Solution (Exercise 18.7). Let $S_n = X_1 + \ldots + X_n$, where the X_i are independent, identically distributed random variables with $\mathbf{P}(X_i = 1) = \frac{1}{2}$, $\mathbf{P}(X_i = 0) = \frac{1}{2}$.

We will try to find an approximation for S_n.

Clearly $\mathbf{E}[X_i)] = \frac{1}{2}$, $\mathbf{E}[X_i^2] = \frac{1}{2}$, $\mathbf{Var}(X_i) = \mathbf{E}[X_i^2] - \big(\mathbf{E}[X_i]\big)^2 = \frac{1}{2} - \big(\frac{1}{2}\big)^2 = \frac{1}{4}$.

By additivity, $\mathbf{E}[S_n] = \frac{n}{2}$.

By equation (16.30), $\mathbf{Var}(S_n) = \frac{n}{4}$.

Let W_n have a normal distribution, with $\mathbf{Var}(W_n) = \frac{n}{4}$ and $\mathbf{E}[W_n] = \frac{n}{2}$.

Because of the Central Limit Theorem, we know that for large n we have

$$\mathbf{P}(S_n < 499500) \approx \mathbf{P}(W_n < 499500).$$

We hope that $n = 1000000$ will give a reasonable approximation for probabilities of the form $\mathbf{P}(S_n < a)$.

Let h be the probability density for the distribution of W_n, with $n = 1000000$. Then

$$\mathbf{P}(W_n < 499500) = \int_{-\infty}^{499500} h. \tag{18.70}$$

To finish the problem, we need to know the formula for h.

Let $m = \mathbf{E}[W_n] = \mathbf{E}[S_n] = 500000$, and let $v = \mathbf{Var}(W_n) = \mathbf{Var}(S_n) = 250000$. By Lemma 18.13,

$$h(u) = \frac{1}{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-(u-m)^2/2v}.$$

This completes the solution. To obtain the actual numerical value of $\mathbf{P}(W_n < 499500)$ one can use a computer program to perform the numerical integration. This gives

$$\int_{-\infty}^{499500} h = \int_{-\infty}^{499500} \frac{1}{\sqrt{250000}} \frac{1}{\sqrt{2\pi}} e^{-(u-500000)^2/(2\cdot 250000)}\, du \approx 0.1586552539314554 \tag{18.71}$$

And as usual, the integral in equation (18.71) could be simplified numerically by the change of variable $t = (u - 500000)/\sqrt{250000}$.

Solution (Exercise 18.8). (i) Equation (18.46) says that

$$\mathbf{P}\big(S_n \le b_n\big) = \mathbf{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le \frac{b_n - n\mu}{\sqrt{n}\sigma}\right).$$

Applying this equation to equation (18.43), with $\mu = 0$, gives

$$\lim_{n\to\infty}\left(\mathbf{P}\big(S_n \le b_n\big) - \mathbf{P}\left(Z \le \frac{b_n}{\sqrt{n}\sigma}\right)\right) = 0.$$

That is,

$$\lim_{n\to\infty}\left(\mathbf{P}\big(S_n \le b_n\big) - \int_{-\infty}^{\frac{b_n}{\sqrt{n}\sigma}} e^{-z^2/2}\,dz\right) = 0.$$

Since $b_n/(\sqrt{n}\sigma) \to 0$,

$$\lim_{n\to\infty}\left(\mathbf{P}\big(S_n \le b_n\big) - \int_{-\infty}^{0} e^{-z^2/2}\,dz\right) = 0,$$

and so

$$\lim_{n\to\infty}\mathbf{P}\big(S_n \le 7\big) = \int_{-\infty}^{0} e^{-z^2/2}\,dz.$$

But

$$\int_{-\infty}^{0} e^{-z^2/2}\,dz = \frac{1}{2},$$

since the function $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-z^2/2}$ is symmetric and

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz = 1.$$

Thus

$$\lim_{n\to\infty}\mathbf{P}\big(S_n \le 7\big) = \frac{1}{2}.$$

(ii)

For any sequence of intervals $\big[a_n, b_n\big]$, by equation (18.44) we have

$$\mathbf{P}\big(a_n \ \le S_n \le b_n\big) = \mathbf{P}\left(\frac{a_n - n\mu}{\sqrt{n}\sigma} \le \frac{S_n - n\mu}{\sqrt{n}\sigma} \le \frac{b_n - n\mu}{\sqrt{n}\sigma}\right).$$

Applying this fact to equation (18.42) gives

$$\lim_{n\to\infty}\left(\mathbf{P}\big(a_n \ \le S_n \le b_n\big) - \mathbf{P}\left(\frac{a_n - \mu}{\sqrt{n}\sigma} \le Z \le \frac{b_n - \mu}{\sqrt{n}\sigma}\right)\right) = 0.$$

Since $a_n = 2\sqrt{n}$, and $b_n = 11\sqrt{n}$ and $\mu = 0$,

$$\lim_{n\to\infty} \left(\mathbf{P}\big(2\sqrt{n} \le S_n \le 11\sqrt{n}\big) - \mathbf{P}\left(\frac{2}{\sigma} \le Z \le \frac{11}{\sigma} \right) \right) = 0.$$

Thus

$$\lim_{n\to\infty} \mathbf{P}\big(2\sqrt{n} \ \le S_n \le 11\sqrt{n}\big) = \mathbf{P}\left(\frac{2}{\sigma} \le Z \le \frac{11}{\sigma} \right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{2}{\sigma}}^{\frac{11}{\sigma}} e^{-\frac{z^2}{2\pi}}\, dz.$$

Solution (Exercise 18.9).

Let

$$b_n = \sqrt{n}\sigma a + n\mu.$$

Then

$$a = \frac{b_n - n\mu}{\sqrt{n}\sigma}. \tag{18.72}$$

By equation (18.43),

$$\lim_{n\to\infty} \left(\mathbf{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le \frac{b_n - n\mu}{\sqrt{n}} \right) - \mathbf{P}\left(Z \le \frac{b_n - n\mu}{\sqrt{n}\sigma} \right) \right) = 0.$$

By equation (18.72), this says that

$$\lim_{n\to\infty} \left(\mathbf{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le a \right) - \mathbf{P}(Z \le a) \right) = 0,$$

which is equivalent to equation (18.54).

APPENDICES

These appendices contain additional details about topics discussed earlier. Appendices J, K, L, M and O also introduce subjects that are not covered in this book, but which readers may encounter later.

APPENDICES

Appendix A

Some practice with averages

Averages occur throughout probability theory. This section reviews the general concepts (Definition A.1, Definition A.2) and gives some exercises to illustrate the properties of averages. Readers might benefit by sampling the exercises and testing the statements against their own intuitions.

Definition A.1 (Linear combinations). Suppose that numbers $v_1, \ldots, v_n$ and $a_1, \ldots, a_n$ are given. The expression

$$a_1 v_1 + \ldots + a_n v_n \tag{A.1}$$

is said to be a *linear combination* of $v_1, \ldots, v_n$, using *coefficients* $a_1, \ldots, a_n$.

In this section, we consider the special case in which the coefficients $a_1, \ldots, a_n$ in equation (A.1) are nonnegative. Nonnegative coefficients will be referred to here as *weights*.

Definition A.2 (Weighted sums and averages). Let $v_1, \ldots, v_n$ be numbers which we will call the values, and let $w_1, \ldots, w_n$ be nonnegative numbers which we will call the weights. The "weighted sum" of the values v_i, using the weights w_i, is

$$\sum_{i=1}^{n} w_i v_i. \tag{A.2}$$

When the numbers w_i add up to one, we say that the weights are *normalized*, and in this case the weighted sum in equation (A.2) is called the "weighted average" of the values v_i.

Any average is also called a *mean*. A weighted average in which all the weights are equal is called the *arithmetic mean*. Since the weights must add to one, in this case each weight w_i must be equal to $1/n$, where n is the number of weights. Thus the arithmetic mean can be calculated by dividing the sum of the values by the number of values. This is what is usually meant by the word "average" in ordinary speech!

For brevity, we sometimes use an overline to denote an average value. Thus given some values $v_1, \ldots, v_n$ and weights $w_1, \ldots, w_n$, the weighted average of $v_1, \ldots, v_n$ might be denoted by $\bar{v}$.

By definition, weighted averages always use normalized weights, but it will be convenient to extend the terminology about weighted averages slightly.

Suppose we are given weights $w_1, \ldots, w_n$ which are not normalized. Let $W = w_1 + \ldots + w_n$. Then $w_1/W, \ldots, w_n/W$ are normalized weights which are proportional to $w_1, \ldots, w_n$. To save words, if we are given values $v_1, \ldots, v_n$ and unnormalized weights $w_1, \ldots, w_n$, we will say that

$$\bar{v} = \frac{w_1}{W} v_1 + \ldots + \frac{w_n}{W} v_n \tag{A.3}$$

is "the weighted average with weights $w_1, \ldots, w_n$".

Thus, in a problem, if a weighted average is requested, and the given weights are not normalized, it is understood that the weights should be normalized before calculating the average.

Exercise A.1. A sequence $\mathbf{x} = (x_1, \ldots, x_5)$ of values and a sequence $\mathbf{w} = (w_1, \ldots, w_5)$ of weights is given, such that $x_1 = \frac{1}{2}$, $x_2 = \frac{2}{3}$, $x_3 = 1$, $x_4 = \frac{1}{4}$, $x_5 = \frac{1}{2}$ and $w_1 = \frac{1}{3}$, $w_2 = \frac{3}{2}$, $w_3 = 3$, $w_4 = 2$, $w_5 = 2$. Find the weighted average of the values in the sequence $\mathbf{x}$.

[Solution]

Example A.3 (Mean value equals center of mass). Anyone who has seen a center of mass calculation in a physics course has encountered a weighted average.

Suppose we have seven point masses on a line. The position coordinates of the masses are $v_1 = -1.0$, $v_2 = -.4$, $v_3 = .1$, $v_4 = 1.5$, $v_5 = 2.2$, $v_6 = 3.2$,

$v_7 = 3.7$, while the weights of the masses are $w_1 = 1$, $w_2 = 3$, $w_3 = 3$, $w_4 = 2$, $w_5 = 1$, $w_6 = 1$, $w_7 = 3$.

The sum of weights is $W = 14$, and the center of mass coordinate is defined to be the usual weighted average $\bar{v}$, which is given by

$$\bar{v} = \frac{w_1}{W}v_1 + \frac{w_2}{W}v_2 + \frac{w_3}{W}v_3 + \frac{w_4}{W}v_4 + \frac{w_5}{W}v_5 + \frac{w_6}{W}v_6 + \frac{w_7}{W}v_7 = \frac{17.6}{14} \approx 1.25714. \tag{A.4}$$

See Figure A.1. We can think that each weight w_i is attached to a rigid bar at v_i. If the bar is free to turn about a pivot located at $\bar{v}$, it will balance, and remain stationary.

When X is a random variable with values $v_1, \ldots, v_7$, and $w_i = \mathbf{P}(X = x_i)$, then $W = w_1 + \ldots + w_7 = 1$. Then equation (A.4) says that

$$\bar{v} = w_1 v_1 + \ldots w_7 v_7,$$

and this sum is $\mathbf{E}[X]$, by definition. In the same way, for any finite-range random variable X, $\mathbf{E}[X]$ is equal to the center of mass of the distribution of X, provided that we represent the distribution by putting a lump of probability mass equal to $\mathbf{P}(X = x_i)$ at each point x_i.

Exercise A.2 (Average of a constant). Let $v_1, \ldots, v_n$ be values, such that each v_i is equal to the same number c. Let $w_1, \ldots, w_n$ be any sequence of weights. Show that the weighted average $v_1, \ldots, v_n$ is equal to c.

This exercise is *not* hard, but it is a useful observation.

[Solution]

Another simple but useful property of averages is the following.

Lemma A.4 (Scaling and bounds for averages). If all the values in a sequence are multiplied by the same number c, then any weighted average is also multiplied by c.

Any weighed average of a sequence of values always lies between the smallest value and the largest value. That is, if $v_1, \ldots, v_n$ are the values and $w_1, \ldots, w_n$ are any weights, then

$$\min(v_1, \ldots, v_n) \leq \bar{v} \leq \max(v_1, \ldots, v_n). \tag{A.5}$$

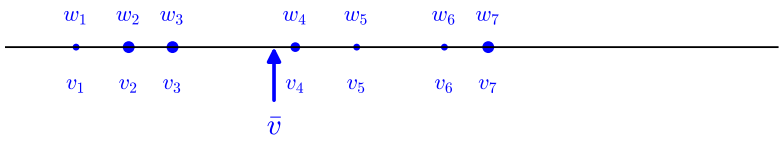


Figure A.1: $\bar{v}$ is the center of mass for the seven masses

Proof. Let $W = w_1 + \ldots + w_n$. The first fact in the statement of the lemma just says that

$$\frac{1}{W}\sum_{i=1}^{n} w_i c\, v_i = c\,\frac{1}{W}\sum_{i=1}^{n} w_i v_i.$$

To derive the second fact, let $m = \min(v_1, \ldots, v_n)$ and let $M = \max(v_1, \ldots, v_n)$. Then

$$\bar{v} = \frac{1}{W}\sum_{i=1}^{n} w_i v_i \leq \frac{1}{W}\sum_{i=1}^{n} w_i M = \frac{1}{W} W M = M.$$

Similarly $\bar{v} \geq m$.

□

If it happens that all the values $v_1, \ldots, v_n$ are equal to the same number c, then the maximum and minimum of the sequence are both equal to c. Thus Lemma A.4 implies the result of Exercise A.2.

Exercise A.3 (Replacing values in a weighted sum by a constant). Let $v_1, \ldots, v_n$ be real numbers and let $w_1, \ldots, w_n$ be weights. Let

$$M = w_1 v_1 + \ldots + w_n v_n,$$

and let

$$\bar{v} = \frac{w_1}{W} v_1 + \ldots + \frac{w_n}{W} v_n,$$

where W is the sum of the weights.

Show that

$$M = w_1 \bar{v} + \ldots + w_n \bar{v}. \tag{A.6}$$

Using the terminology of weighted sums and averages, this exercise tells us another useful fact:

"The value of a weighted sum is unchanged if all the values are replaced by their weighted average."

Note carefully that equation (A.6) holds even if the weights $w_1, \ldots, w_n$ are *not* normalized, but the weighted average $\bar{v}$ is of course always calculated using normalized weights obtained from $w_1, \ldots, w_n$.

[Solution]

Exercise A.4 (Replacing some of the values by the average of those values). Let $v_1, \ldots, v_{m+n}$ be real numbers and let $w_1, \ldots, w_{m+n}$ be weights. Let s be the weighted sum of $v_1, \ldots, v_{m+n}$, using the weights $w_1, \ldots, w_{m+n}$.

Let z be the weighted average of $v_1, \ldots, v_m$, using the weights $w_1, \ldots, w_m$. Note carefully that the definition of z does not involve any of the values other than $v_1, \ldots, v_m$.

Let $x_i = z$ for $i = 1, \ldots, m$, $x_i = v_i$ for $i = m + 1, \ldots, m + n$.

Prove that the weighted sum of $x_1, \ldots, x_{m+n}$ is equal to s.

[Solution]

Exercise A.5 (Replacing some values by the wrong average). In the setting of Exercise A.3, where we are given values $v_1, \ldots, v_n$ and weights $w_1, \ldots, w_n$, again let the weighted average be $\bar{v}$. Suppose we replace v_1 by $\bar{v}$ but leave $v_2, \ldots, v_n$ unchanged. Show by an example that the value of the weighted sum may change as a result of this substitution.

[Solution]

Exercise A.6 (Averaging pooled data). This problem contains some important techniques if you have to work with averages.

A sequence $\mathbf{v}$ of data has length 4700. Denote the arithmetic mean of $\mathbf{v}$ by $\bar{v}$.

Let

$$\begin{aligned}\mathbf{x} &= v_1, \ldots, v_{1500},\\ \mathbf{y} &= v_{1500+1}, \ldots, v_{3000},\\ \mathbf{z} &= v_{3000+1}, \ldots, v_{4700}.\end{aligned}$$

We might refer to $\mathbf{x}, \mathbf{y}, \mathbf{z}$ as "blocks of data".

Denote the arithmetic means of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ by $\bar{x}$, $\bar{y}$ and $\bar{z}$ respectively.

Suppose you are given $\bar{x}$, $\bar{y}$ and $\bar{z}$.

Derive a formula for $\bar{v}$ in terms of $\bar{x}$, $\bar{y}$ and $\bar{z}$.

One sometimes says that the data sequence $\mathbf{v}$ is obtained by "pooling" the data in sequences $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. The answer to this problem says that the "pooled average" of the data can be calculated as a *weighted* average of the averages of the three blocks of data $\mathbf{x}, \mathbf{y}, \mathbf{z}$.

This illustrates a small theorem, which you are finding in this exercise. It comes with a slogan:

> "The average of the averages is the average."

[Solution]

Exercise A.7 (A weighted average for pooled data). Consider the same data sequences $\mathbf{v}, \mathbf{x}, \mathbf{y}, \mathbf{z}$ studied in Exercise A.6. The rule given in that exercise generalizes to weighted averages, as you will now show.

In the present situation suppose that you wish to find the *weighted* average of $\mathbf{v}$, using weights $w_1, \ldots, w_{4700}$. You are not told the weights w_i, but you are told numbers r, s, t, where $r = w_1 + \ldots w_{1500}$, $s = w_{1500+1} + \ldots + w_{3000}$, and $t = w_{3000+1} + \ldots + w_{4700}$.

Using the weights w_i, let $\bar{v}$ be the *weighted* average of $\mathbf{v}$, and let $\bar{x}$, $\bar{y}$ and $\bar{z}$ be the *weighted* averages of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, respectively.

Find a formula for $\bar{v}$ in terms of $\bar{x}, \bar{y}, \bar{z}, \bar{r}, \bar{s}, \bar{t}$. As always, justify your answer.

[Solution]

Exercise A.8 (Weighted sum of a sum of sequences). Let $w_1, \ldots, w_n$ be a sequence of weights. Let $x_1, \ldots, x_n$ be a sequence of values, and let $y_1, \ldots, y_n$ be a sequence of values.

Let s be the weighted sum of $x_1, \ldots, x_n$, using the weights $w_1, \ldots, w_n$. Let t be the weighted sum of $y_1, \ldots, y_n$, using the *same* weights $w_1, \ldots, w_n$.

Prove that the weighted sum of $x_1 + y_1, \ldots, x_n + y_n$, using those weights $w_1, \ldots, w_n$, is equal to $s + t$.

Of course, since a weighted average is simply a weighted sum using normalized weights, you have also shown the following. If $\bar{x}$ is the weighted average of the x_i and $\bar{y}$ is the weighted average of the y_i, then the weighted average of the numbers $x_i + y_i$ is equal to $\bar{x} + \bar{y}$.

One might write this rule as $\overline{x + y} = \bar{x} + \bar{y}$.

And one might say: "The average of a sum is the sum of the averages."

[Solution]

Remark A.5 (A danger when comparing averages). Here's an example of a problem. We'll phrase it in terms of batting averages in baseball, but it can come up in other situations, for example in testing medical procedures.

A baseball player's batting average, over a given time period, is defined as the fraction of the times at bat which actually result in a hit.

Suppose you are comparing two players, named A and B. You don't know the details of their records, but you feel, quite reasonably, that their batting averages will give you a good idea of their abilities as hitters.

But suppose you don't learn their batting averages over the entire season, but instead you are given their averages over the first half of the season, and then separately are given their averages over the second half of the season.

Suppose that player A has a better batting average than B over the first half of the season, and also has a better average than B over the second half of the season.

Can we conclude that A has a better average than B over the whole season? One might jump to that conclusion after reading Exercise A.6, which

gives a formula for combining averages from different sets of data. However, in this case, there is an important extra piece of information: different players may have different numbers of times at bat.

Suppose, for example, that neither player A nor player B did particularly well during the first half of the season, and both had very similar records, with A slightly better. And suppose that player A had a great batting average during the second half of the season, but had very few times at bat during that period. In the second half of the season, player B had a batting average that was very good, and had a large number of times at bat.

When combining the averages for these players, player B's very good average during the second half of the season will correctly receive much more weight than player A's great average. As a result, player B may have the best overall average.

A.1 Solutions for Appendix A

Solution (Exercise A.1). Let

$$w = w_1 + w_2 + w_3 + w_4 + w_5.$$

Then

$$\bar{x} = (x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 + x_4 \cdot w_4 + x_5 \cdot w_5)/w \approx 0.641509433962264.$$

Solution (Exercise A.2). Let

$$w = w_1 + \ldots + w_n.$$

Then

$$\bar{v} = \frac{w_1\, v_1 + \ldots + w_n\, v_n}{w} = \frac{w_1\, c + \ldots + w_n\, c}{w} = \frac{w\, c}{w} = c.$$

Solution (Exercise A.3). By definition,

$$\bar{v} = \frac{w_1}{W} v_1 + \ldots + \frac{w_n}{W} v_n = \frac{1}{W}\left(w_1 v_1 + \ldots + w_n v_n\right) = \frac{M}{W},$$

Thus $W\bar{v} = M$.

Hence

$$w_1\bar{v} + \ldots + w_n\bar{v} = \left(w_1 + \ldots + w_n\right)\bar{v} = W\bar{v} = M.$$

Solution (Exercise A.4). Let $K = w_1 + \ldots + w_m$. Then

$$z = \frac{w_1}{K} v_1 + \ldots + \frac{w_m}{K} v_m,$$

so $Kz = w_1 v_1 + \ldots + w_m v_m$.

Then

$$\begin{aligned} w_1 x_1 + \ldots + w_{m+n} x_{m+n} &= (w_1 z + \ldots + w_m z) + (w_{m+1} v_{m+1} + \ldots + w_{m+n} v_{m+n}) \\ &= Kz + (w_{m+1} v_{m+1} + \ldots + w_{m+n} v_{m+n}) \\ &= (w_1 v_1 + \ldots + w_m v_m) + (w_{m+1} v_{m+1} + \ldots + w_{m+n} v_{m+n}) \,. \end{aligned}$$

Solution (Exercise A.5). Let $w_1 = w_2 = 1/2$, and let $v_1 = 1$, $v_2 = -1$.

Then $\bar{v} = 0$ and $w_1 v_1 + w_2 v_2 = 0$.

Replacing v_1 by $\bar{v}$, the weighted sum becomes

$$w_1\, 0 + w_2 v_2 = -\frac{1}{2}.$$

Solution (Exercise A.6).

$$\bar{x} = \frac{v_1 + \ldots + v_{1500}}{1500}.$$

$$\bar{y} = \frac{v_{1501} + \ldots + v_{3000}}{1500}.$$

$$\bar{z} = \frac{v_{3001} + \ldots + v_{4700}}{1700}.$$

Then

$$\bar{v} = \frac{v_1 + \ldots + v_{4700}}{4700} = \frac{1500\bar{x} + 1500\bar{y} + 1700\bar{z}}{4700} = \frac{1500}{4700}\bar{x} + \frac{1500}{4700}\bar{y} + \frac{1700}{4700}\bar{z}$$

Solution (Exercise A.7). By definition,

$$\bar{x} = \frac{w_1 v_1 + \ldots + w_{1500} v_{1500}}{r}.$$

$$\bar{y} = \frac{w_{1501} v_{1501} + \ldots + w_{1501} v_{3000}}{s}.$$

$$\bar{z} = \frac{w_{3001} v_{3001} + \ldots + w_{4700} v_{4700}}{t}.$$

Hence

$$w_1 v_1 + \ldots + w_{1500} v_{1500} = r\bar{x},$$
$$w_{1501} v_{1501} + \ldots + w_{1501} v_{3000} = s\bar{y},$$
$$w_{3001} v_{3001} + \ldots + w_{4700} v_{4700} = t\bar{z}.$$

Then

$$\bar{v} = \frac{w_1 v_1 + \ldots + w_{4700} v_{4700}}{w_1 + \ldots + w_{4700}}$$
$$= \frac{(w_1 v_1 + \ldots w_{1500} v_{1500}) + (w_{1501} v_{1501} + \ldots + w_{3000} v_{3000}) + (w_{3001} v_{3001} + \ldots + w_{4700} v_{4700})}{r+s+t}$$
$$= \frac{r\bar{x} + s\bar{y} + t\bar{z}}{r+s+t} = \frac{r}{r+s+t}\bar{x} + \frac{s}{r+s+t}\bar{y} + \frac{t}{r+s+t}\bar{z}.$$

Solution (Exercise A.8).

$$s = \sum_{i=1}^{n} w_i x_i,$$

$$t = \sum_{i=1}^{n} w_i y_i.$$

Adding these equations,

$$s + t = \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i \left(x_i + y_i\right).$$

Appendix B

The triangle inequality

Proving the triangle inequality The triangle inequality says that for any numbers x, y,

$$|x + y| \leq |x| + |y| \,. \tag{B.1}$$

Proof. By definition $|x| = x$ if $x \geq 0$, and $|x| = -x$ if $x < 0$. Thus either $|x + y| = x + y$ or $|x + y| = -x - y$.

The definition of $|x|$ tells us that

$$x \leq |x| \text{ and } -x \leq |x| \,. \tag{B.2}$$

Adding the inequalities $x \leq |x|$ and $y \leq |y|$ gives $x + y \leq |x| + |y|$.

Adding the inequalities $-x \leq |x|$ and $-y \leq |y|$ gives $-x - y \leq |x| + |y|$.

Since $|x + y| = x + y$ or $|x + y| = -x - y$, in every possible case we have $|x + y| \leq |x| + |y|$.

□

Equation (B.1) deals with a sum, but the triangle inequality also applies to differences:

$$|x - y| = |x + (-y)| \leq |x| + |-y| = |x| + |y| \,. \tag{B.3}$$

Why is the triangle inequality called "the triangle inequality"? See Figure B.1. This picture shows that the length of any side of a triangle is less than or equal to the sum of the lengths of the other two sides. In vector language: $\|\vec{\mathbf{a}} + \vec{\mathbf{b}}\| \leq \|\vec{\mathbf{a}}\| + \|\vec{\mathbf{b}}\|$.

It's interesting that the triangle inequality works for vectors, not just numbers. The proof in that case is more complicated.

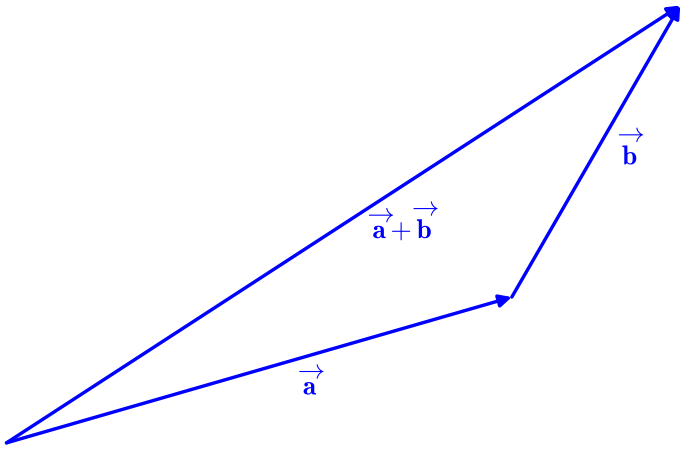


Figure B.1: The sum of two geometric vectors.

Exercise B.1. For readers who would like more practice with absolute values.

In equation (B.1), replace x by $z - w$ and replace y by w. Use the result to show that

$$|z| - |w| \leq |z - w| \,.$$

Do this again, with the roles of z, w reversed.

Then obtain:

$$\big|\, |z| - |w| \,\big| \leq |z - w| \,. \tag{B.4}$$

This inequality is sometimes useful. If you think about $z - w$ as a *change* in some quantity, this inequality says: "The change in the absolute value is no larger than the absolute value of the change."

You would expect that to be true, but it's nice to have a guarantee.

[Solution]

As usual, once we have the triangle inequality for the sum of two numbers, we can get a similar inequality for any sum of numbers:

$$|x_1 + \ldots + x_k| \leq |x_1| + \ldots + |x_k| \,. \tag{B.5}$$

To establish equation (B.5), one can use the Old Induction Trick (Exercise 2.23). Another approach is to reorder the terms, in order to separate

the positive and negative numbers in the sum. So write $x_1 + \ldots + x_k$ as $y_1 + \ldots + y_m - (z_1 + \ldots + z_n)$, where all the numbers $y_1, \ldots, y_m, z_1, \ldots, z_n$ are nonnegative. Then use the triangle inequality for the sum of two numbers:

$$y_1 + \ldots + y_m - (z_1 + \ldots + z_n) \leq |y_1 + \ldots + y_m| + |z_1 + \ldots + z_n| \,.$$

And since now we have nonnegative numbers, $|y_1 + \ldots + y_m| = y_1 + \ldots + y_m$, $|z_1 + \ldots + z_n| = z_1 + \ldots + z_n$.

B.1 Solutions for Appendix B

Solution (Exercise B.1). The requested substitution produces:

$$|(z - w) + w| \leq |z - w| + |w| \,.$$

Thus

$$|z| \leq |z - w| + |w| \,,$$

and so

$$|z| - |w| \leq |z - w| \,.$$

Exchanging z and w gives

$$|w| - |z| \leq |z - w| \,.$$

And one of the numbers $|z| - |w|$, $|w| - |z|$ is equal to $\big|\, |z| - |w| \,\big|$, so we have obtained equation (B.4).

If you like, you can also check directly that equation (B.4) always holds, as follows.

It is easy to see that equality holds in equation (B.4) if either of z and w is zero, or if both have the same sign.

Suppose that $z > 0, w < 0$. Then $z - w = z + |w| = |z| + |w|$, while $\big|\, |z| - |w| \,\big| = \pm\,(|z| - |w|)$. Then

$$|z - w| - \big|\, |z| - |w| \,\big| = 2\,|w| \text{ or } 2\,|z| \,.$$

Thus equation (B.4) holds. A similar argument works if $z < 0$, $w > 0$.

Appendix C

Defining Z with a given distribution density on the real line

Suppose that we are given a probability density function h on $\mathbb{R}$. We would like to construct a random variable Z whose distribution is given by h. And we would prefer to make Z as simple a possible.

Here's how to do that.

Let $\Omega = \mathbb{R}$ and let the probability distribution $\mathbf{P}$ on Ω be given by the density function h. Define the function Z on $\mathbb{R}$ by

$$Z(u) = u. \tag{C.1}$$

We claim that Z is an example of a random variable with probability density h.

To check that, note that from the definition of Z,

$$\{Z \in S\} = \{u : \ Z(u) \in S\} = \{u : \ u \in S\} = S. \tag{C.2}$$

Thus $\mathbf{P}(Z \in S) = \mathbf{P}(S)$. Since $\mathbf{P}$ is given by h,

$$\mathbf{P}(Z \in S) = \int_S h.$$

By Definition 9.11, h is a density for the distribution of Z, as desired.

Appendix D

Distribution of a function of a random variable

Let X and Y be real-valued physical random variables. The experiment for which X is defined need not be the same as the experiment for which Y is defined. Suppose that φ is a mathematical function on the real line. We can think about applying φ to the output of X. This defines a new physical random variable $\varphi(X)$. Similarly we can think about applying φ to the output of Y, to obtain another physical random variable $\varphi(Y)$.

Now suppose that X and Y have the same distribution. This means that for any set S of numbers, the probability that the output of X lies in S is equal to the probability that the output of Y lies in S.

We claim that then $\varphi(X)$ and $\varphi(Y)$ also have the same distribution. To see that, note that the statement that the output of $\varphi(X)$ lies in S is a statement about the output of X, since the output of X certainly determines the output of $\varphi(X)$. Similarly the statement that the output of $\varphi(Y)$ lies in S is a statement about the output of Y, and indeed it is *the same* statement about the output.

Since any statement about the output of X has the same probability as the same statement about the output of Y, we conclude that the probability that the output of $\varphi(X)$ lies in S must be equal to the probability that the output of $\varphi(Y)$ lies in S. This is what it means to say that $\varphi(X)$ and $\varphi(Y)$ have the same distribution.

The next lemma establishes the corresponding fact for mathematical random variables X and Y.

Lemma D.1. Let X and Y be random variables having the same distribution. These random variables do not have to be defined on the same sample space.

Suppose that φ is any function on the real line. Then $\varphi(X)$ and $\varphi(Y)$ have the same distribution.

Using the notation in Definition 9.7, we are saying that

$$X \sim Y \implies \varphi(X) \sim \varphi(Y). \tag{D.1}$$

Proof. To show that, let S be a subset of $\mathbb{R}$. We have to show that the probability of the event $\{\varphi(X) \in S\}$, on the sample space of X, is exactly the same as the probability of the event $\{\varphi(Y) \in S\}$, on the sample space of Y.

So let $T = \{u : \varphi(u) \in S\}$. Both S and T are subsets of the real line.

From the definitions,

$$\{\varphi(X) \in S\} = \{X \in T\}.$$

Also from the definitions,

$$\{\varphi(Y) \in S\} = \{Y \in T\}.$$

Since X and Y have the same distribution, the probability of the event $\{X \in T\}$ is exactly equal to the probability of the event $\{Y \in T\}$.

But then the probability of the event $\{\varphi(X) \in S\}$ is exactly equal to the probability of the event $\{\varphi(Y) \in S\}$, as claimed. □

Appendix E

A density formula for the expected value of a function of X

The expected value of $\varphi(X)$ using the density of the distribution of X

Our goal here is to derive equation (15.6) from equation (15.4).

Since you don't know the sample space that X is defined on, go ahead and define your own sample space Ω. Let it be the real line, with probabilities given by the density function h. Then define a new random variable Z on the real line, given by $Z(t) = t$.

Appendix C tells us that X and Z have exactly the same distribution.

Appendix D then tells us that $\varphi(X)$ and $\varphi(Z)$ have exactly the same distribution.

And we know how to find $\mathbf{E}[\varphi(Z)]$. By equation (15.4) (with X in that equation replaced by $\varphi(Z)$ on both sides, and f replaced by h), we have

$$\mathbf{E}[\varphi(Z)] = \int \varphi(Z)h = \int_{-\infty}^{\infty} \varphi(t)h(t)\,dt.$$

Remember that by part (ii) of Theorem 15.2, the expected value of a random variable is determined by its distribution, so $\varphi(X)$ and $\varphi(Z)$ have the same expected value,

Thus we have obtained equation (15.6).

Appendix F

Practice using densities

There are no new ideas in this appendix, just some practice to get a feeling for the way densities work.

Example F.1 (Probability of missing the central region - density case). Someone is throwing darts at a target represented by a disc of radius 5, centered at the origin of $\mathbb{R}^2$. See Figure 3.4.

The point of impact (x, y) is random, but the thrower is trying hard to hit a point near the center of the target. Thus the probability density for the distribution of impact points is not uniform: it is described by a probability density f on the target which is large near the center.

In fact, we will use a model with f defined by:

$$f((x, y)) = \frac{c}{\sqrt{x^2 + y^2}} = \frac{c}{r},$$

where r is the distance of the point (x, y) from the center of the target.

Let A be the set of points (x, y) in the target such that $\sqrt{x^2 + y^2} > 2$. A represents the physical event that the dart lands more than two units of distance from the center. In Figure 3.4, A is the shaded ring.

We will use equation (15.3) to find $\mathbf{P}(A)$. Fortunately, in this example we can calculate integrals using polar coordinates.

The first step is to find c.

$$1 = \int f = \int_0^{2\pi} \int_0^5 \frac{c}{r} r \, dr \, d\theta = 10\pi.$$

Hence $c = \frac{1}{10\pi}$. Thus

$$\mathbf{P}(A) = \int_A f = \int_0^{2\pi} \int_2^5 \frac{1}{10\pi} \frac{1}{r}\, r\, dr\, d\theta = \frac{1}{10\pi}\,(2\pi \cdot 3) = \frac{3}{5}.$$

Exercise F.1. In the dart-throwing experiment, let $h(x, y)$ be the probability density for the random location where the dart lands. If A is a region of the target board, then the probability of hitting A is given by

$$\mathbf{P}(A) = \int_A h. \tag{F.1}$$

In calculus notation,

$$\mathbf{P}(A) = \int\int_A h(x, y)\, dx\, dy.$$

Consider the situation when the target region T is a circular disc with radius one centered at the origin, and assume that the thrower has a tendency to throw toward the right. More precisely, assume that $h(x, y)$ on T is proportional to $2 + x$.

(i) Find the exact formula for $h(x, y)$.

(ii) Find the probability that the dart lands in the right half of the target. That is, if $A = \{(x, y) :\ (x, y) \in T,\ x \geq 0\}$, find $\mathbf{P}(A)$.

[Solution]

Exercise F.2. In this problem we have a square target.

Let Ω be the rectangle consisting of all points x, y such that $0 \leq x \leq 1$ and $0 \leq y \leq 3$.

Let h be a probability density on Ω given by

$$h(x, y) = c\, x \sin(x\, y), \tag{F.2}$$

where c is an appropriate constant.

Let $\mathbf{P}$ be the distribution on Ω with probability density h.

Consider choosing a random point in Ω using this distribution. Let A be the event that the chosen point (x, y) is such that $y < 1$.

Find $\mathbf{P}(A)$.

[Solution]

Exercise F.3. Consider the probability model with sample space $[0, \pi/4]$ and probability density $f(t) = \sqrt{2}\cos t$.

Let X be the random variable on $[0, \pi/4]$ defined by $X(t) = \sin t$. Find $\mathbf{E}[X]$.

[Solution]

Remark F.2 (Are unbounded densities ok?). By definition, any non-negative function f with $\int f = 1$ is a probability density. So a probability density f is not necessarily a bounded function. Theorem 15.2 does tell us that $\mathbf{E}[X]$ is always defined if X is a bounded random variable. But what about $\int Xf$? Does that integral always exist when X is bounded? Sufficiently paranoid people (like us) might worry: if f is unbounded, could that spoil $\int Xf$, and contradict equation (15.4)?

Fortunately not, because of the Comparison Principle for integrals. Since $\int f$ is defined, $\int cf$ is also defined for any constant c. And if X is a bounded random variable, by definition this means that for some value of c we have $|X| \leq c$. But that implies that $|Xf| \leq cf$. And so the comparison principle for integrals guarantees that $\int Xf$ exists.

Exercise F.4. Let $\Omega = [0, 4]$.

Let f be the probability density on the interval $[0, 4]$, such that $f(t) = c/\sqrt{t}$. (An unbounded density)

Let $\mathbf{P}$ be the probability set-function on $[0, 4]$ with probability density f.

Let X be the random variable on $[0, 4]$ defined by $X(t) = t^{-1/4}$.

Find c, and then find $\mathbf{E}[X]$.

[Solution]

Exercise F.5 (Finding the expected value of the first coordinate). Consider the setting of Exercise F.1.

The sample space Ω is the unit disc T. Let X be the mathematical random variable defined on T by the equation

$$X(\,(x,y)\,) = x.$$

The physical interpretation of X is that it represents the x-coordinate of the spot where a dart strikes the target.

By assumption, a density for the probability function $\mathbf{P}$ on T is given by h, where

$$h(x,y) = c(2+x)$$

for an appropriate value of c. From the solution to Exercise F.1, we know that $c = 1/(2\pi)$.

Find $\mathbf{E}\,[X]$.

[Solution]

Exercise F.6. Let Ω be the disc with radius 5 centered at the origin. Let H be the function on Ω defined by $H((x,y)) = e^r = e^{\sqrt{x^2+y^2}}$.

(i) Consider the probability model with sample space Ω and uniform probability distribution $\mathbf{P}$ on Ω. Find $\mathbf{E}\,[H]$.

(ii) Consider the probability model with sample space Ω and probability distribution $\mathbf{P}$ as in Example F.1, so that $\mathbf{P}$ has a density given by $\frac{1}{10\pi r}$. Find $\mathbf{E}\,[H]$.

[Solution]

F.1 Solutions for Appendix F

Solution (Exercise F.1).

(i) We are told that $h(x, y) = c(2 + x)$ for some proportionality constant c.

We know that h must have integral equal to one, since it is a probability density. Hence

$$\int\int_T c(2 + x)\, dy\, dx = 1.$$

We note that

$$\int\int_T dy\, dx = \pi,$$

since this integral is the area of the unit circle.

Also

$$\int\int_T x\, dy\, dx = 0$$

by symmetry! Applying these facts gives $2c\pi + 0 = 1$, so $c = 1/(2\pi)$, and $h(x, y) = \frac{1}{2\pi}(2 + x)$.

(ii) Using the formula for h, and equation (F.1),

$$\mathbf{P}(A) = \frac{1}{2\pi}\int\int_A (2 + x)\, dy\, dx.$$

Since A is the right half of the unit circle,

$$\int\int_A 2\, dy\, dx = \pi.$$

Also

$$\int\int_A x\, dy\, dx = \int_0^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} x\, dy\, dx = \int_0^1 2x\sqrt{1 - x^2}\, dx = -\frac{1}{3}\left(1 - x^2\right)^{3/2}\Big|_0^1 = \frac{2}{3}.$$

Hence

$$\mathbf{P}(A) = \frac{1}{2\pi}\left(\pi + \frac{2}{3}\right).$$

Solution (Exercise F.2). Since $\int_\Omega h = 1$,

$$1 = \int_0^1 \int_0^3 c\,x \sin(x\,y)\,dy\,dx = \int_0^1 c\left(-\cos(x\,y)\Big|_0^3\right)\,dx = c\int_0^1 (1-\cos(3\,x))\,dx$$
$$= c\left(1-\left(\frac{1}{3}\sin(3\,x)\Big|_0^1\right)\right) = c\left(1-\frac{\sin(3)}{3}\right).$$

Thus

$$c = \frac{1}{1-\frac{\sin(3)}{3}}.$$

$$\mathbf{P}(A) = \int_0^1 \int_0^1 c\,x \sin(x\,y)\,dy\,dx = \int_0^1 c\left(-\cos(x\,y)\Big|_0^1\right)\,dx = c\int_0^1 (1-\cos(x))\,dx$$
$$= c\left(1-\sin(x)\Big|_0^1\right) = c\,(1-\sin(1)) = \frac{1-\sin(1)}{1-\frac{\sin(3)}{3}}.$$

Solution (Exercise F.3). By equation (15.4),

$$\mathbf{E}\,[X] = \int Xf = \int_0^{\pi/4} (\sin t)(\sqrt{2}\cos t)\,dt = \frac{1}{\sqrt{2}}\sin^2 t\Big|_0^{\pi/4} = \frac{1}{\sqrt{2}}\sin(\pi/4)^2 - 0$$
$$= \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}}\right)^2 = \frac{1}{2\sqrt{2}}.$$

Solution (Exercise F.4).

$$1 = \int_0^4 \frac{c}{\sqrt{x}}\,dx = 2c\sqrt{x}\Big|_0^4 = 2c\,(2-0) = 4c,$$

so $c = 1/4$.

By equation (15.4),

$$\mathbf{E}\,[X] = \int_0^4 t^{-1/4}\frac{1}{4}\frac{1}{\sqrt{t}}\,dt = \frac{1}{4}\int_0^4 t^{-3/4}\,dt = \frac{1}{4}4t^{1/4}\Big|_0^4 = 4^{1/4} - 0 = \sqrt{2}.$$

Solution (Exercise F.5). By equation (15.4),

$$\mathbf{E}\,[X] = \frac{1}{2\pi}\int\!\!\int_T X(\,(x,y)\,)(2+x)\,dy\,dx = \frac{1}{2\pi}\int_{-1}^{1}\left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} X(\,(x,y)\,)(2+x)\,dy\right)dx$$

$$= \frac{1}{2\pi}\int_{-1}^{1}\left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} x\,(2+x)\,dy\right)dx = \frac{1}{2\pi}\int_{-1}^{1} x\,\left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}(2+x)\,dy\right)dx.$$

Hence

$$\mathbf{E}\,[X] = \frac{1}{2\pi}\int_{-1}^{1} 2x(2+x)\sqrt{1-x^2}\,dx.$$

Nowadays one can use a computer algebra system to evaluate the integral. However, a reasonable manual evaluation is the following.

By symmetry,

$$\int_{-1}^{1} x\sqrt{1-x^2}\,dx = 0 \text{ and } \int_{-1}^{1} x^2\sqrt{1-x^2}\,dx = 2\int_{0}^{1} x^2\sqrt{1-x^2}\,dx.$$

Thus

$$\mathbf{E}\,[X] = \frac{2}{\pi}\int_{0}^{1} x^2\sqrt{1-x^2}\,dx.$$

Let $x = \sin\theta$. Then $dx = \cos\theta\,d\theta$. When $\theta = 0$, $x = 0$. When $\theta = \pi/2$, $x = 1$. Hence

$$\mathbf{E}\,[X] = \frac{2}{\pi}\int_{0}^{\pi/2} \sin^2\theta\cos^2\theta\,d\theta = \frac{1}{2\pi}\int_{0}^{\pi/2} \sin^2 2\theta\,d\theta.$$

Letting $\theta = \frac{1}{2}\varphi$, $d\theta = \frac{1}{2}d\varphi$, gives

$$\mathbf{E}\,[X] = \frac{1}{2\pi}\int_{0}^{\pi} \left(\sin^2\varphi\right)\frac{1}{2}\,d\varphi = \frac{1}{4\pi}\int_{0}^{\pi} \sin^2\varphi\,d\varphi,$$

The fastest way to evaluate this definite integral is to note that $\sin^2 + \cos^2 = 1$, and $\sin^2$ and $\cos^2$ have equal integrals over the interval $[0,\pi]$. Hence $\int_0^\pi \sin^2 = (1/2)\pi$, and so

$$\mathbf{E}\,[X] = \frac{1}{4\pi}\frac{1}{2}\pi = \frac{1}{8}.$$

Solution (Exercise F.6).

(i) By equation (15.4),

$$\begin{aligned}\mathbf{E}[H] &= \int_0^{2\pi}\int_0^5 e^r \frac{1}{25\pi} r\,dr\,d\theta = \frac{2}{25}\int_0^5 e^r\,r\,dr \\ &= \frac{2}{25}\left(re^r - e^r\right)\Big|_0^5 = \frac{2}{25}\left(5e^5 - e^5 + 1\right) = \frac{2}{25}\left(4e^5 + 1\right).\end{aligned}$$

We use integration by parts to calculate the integral.

(ii) By equation (15.4),

$$\mathbf{E}[H] = \int_0^{2\pi}\int_0^5 e^r \frac{1}{10\pi}\frac{1}{r} r\,dr\,d\theta = \frac{1}{5}\int_0^5 e^r\,dr\,d\theta = \frac{1}{5}e^r\Big|_0^5 = \frac{1}{5}\left(e^5 - 1\right).$$

Appendix G

Nonnegative random variables with zero expectation

Let Y be a nonnegative random variable such that $\mathbf{E}[Y] = 0$. We'll give an argument to show that $\mathbf{P}(Y > 0) = 0$.

If Y has finite range, a direct proof is not hard, based on definition of expected values for finite-range random variables. But there is an easy proof for general random variables. We start by using the Markov inequality (equation (12.19) of Lemma 12.15, with $\alpha = 1/n$):

$$\frac{1}{n}\mathbf{P}\left(Y > \frac{1}{n}\right) \leq \mathbf{E}[Y]$$

for every positive integer n. Since $\mathbf{E}[Y] = 0$, this inequality tells us that

$$\mathbf{P}\left(Y > \frac{1}{n}\right) = 0 \tag{G.1}$$

for every positive integer n.

Since $1/n$ can be made as small as we like by taking large values of n, equation (G.1) says that, for example, $\mathbf{P}(Y > .0000000000001) = 0$, and so on, for any number of decimal places! Surely that is enough to guarantee that $\mathbf{P}(Y > 0) = 0$. Isn't it?

Well, we're being fussy here, but if you want to be completely rigorous, a little more discussion is still needed, as follows.

Notice that $\{Y > 0\}$ is the union of the following sets:

$$\{Y > 1\}, \left\{1 \geq Y > \frac{1}{2}\right\}, \left\{\frac{1}{2} \geq Y > \frac{1}{3}\right\}, \left\{\frac{1}{3} \geq Y > \frac{1}{4}\right\}, \ldots$$

Remember that we always assume that we have *countable additivity* for our models (Section 14.3). Thus

$$\mathbf{P}(Y > 0) = \mathbf{P}(Y > 1) + \mathbf{P}\left(1 \geq Y > \frac{1}{2}\right) + \mathbf{P}\left(\frac{1}{2} \geq Y > \frac{1}{3}\right) + \mathbf{P}\left(\frac{1}{3} \geq Y > \frac{1}{4}\right) + \ldots,$$

and that sum is indeed zero.

Appendix H

Inequalities for log and exponential

As usual in mathematics, $\log x$ will denote the logarithm of x using base e.

Lemma H.1 (Basic inequalities for log and exponential).

$$\log(1+x) \le x \text{ for every } x \in (-1, \infty), \tag{H.1}$$

and

$$1 + x \le e^x \text{ for every } x \in (-\infty, \infty). \tag{H.2}$$

Proof. We'll start by deriving equation (H.1).

Let

$$f(x) = x - \log(1+x).$$

We need to show that $f(x) \ge 0$ for all $x \in (-1, \infty)$.

Note that $f(0) = 0$. Also,

$$f'(x) = 1 - \frac{1}{1+x}.$$

Clearly,

$$1 + x > 1 \text{ for } x > 0;\ 1 + x < 1 \text{ for } -1 < x < 0.$$

Hence

$$f'(x) > 0 \text{ for } x > 0;\ f'(x) < 0 \text{ for } -1 < x < 0.$$

Since f is decreasing on $(-1, 0)$ and increasing on $(0, \infty)$, the minimum of f on $(-1, \infty)$ occurs at $x = 0$, and so the minimum value of f on $(-1, \infty)$ is $f(0) = 0$.

Thus for $x \in (-1, \infty)$ we have $f(x) \geq 0$, i.e.

$$x - \log(1 + x) \geq 0, \text{ i.e. } x \geq \log(1 + x).$$

This proves equation (H.1).

The exponential function is an increasing function. Hence if we take exponential of both sides of an inequality we get another true inequality. Taking the exponential of both sides of equation (H.1) gives the inequality in equation (H.2), for $x > -1$. Since $e^x > 0$ is always true, the inequality in equation (H.2) also holds for $x \leq -1$.

□

See Figure H.1 for the picture of the inequality in equation (H.1), and Figure H.2 for the inequality in equation (H.2).

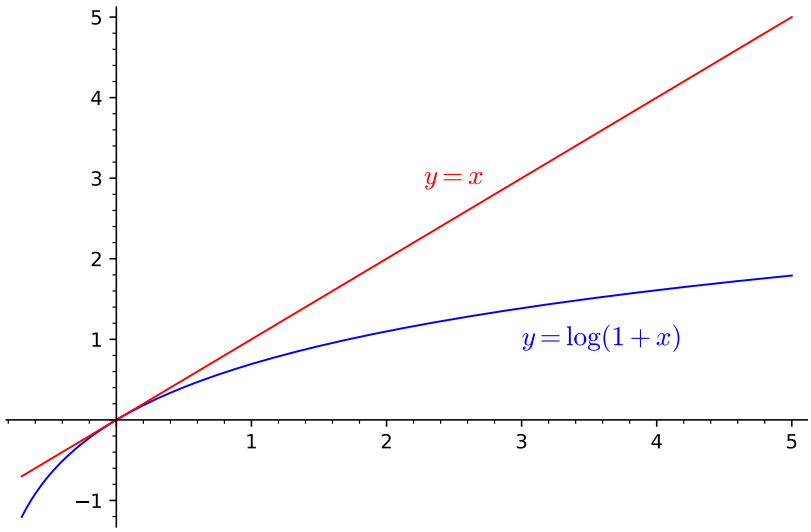


Figure H.1: x is above $\log(1 + x)$. The curves are tangent at $x = 0$.

The function $\log(1 + x)$ grows slowly, so it seems natural that we have a simple upper bound for $\log(1 + x)$. Similarly it seems natural that we have a simple lower bound for e^x. But one sometimes needs inequalities in the other direction. There is probably no "neatest" way to state these. Here's one example of a bound in the other direction.

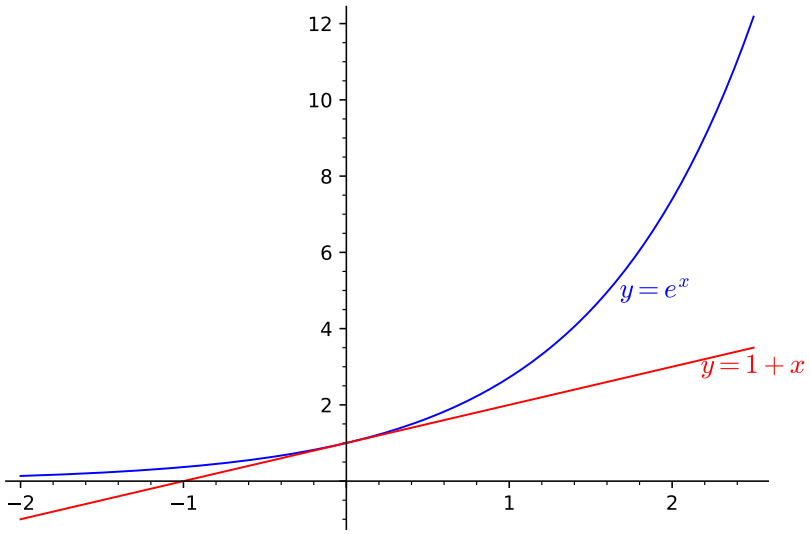


Figure H.2: $1 + x$ is below e^x. The curves are tangent at $x = 0$.

Lemma H.2 (Reversed bounds). For any real number x with $-\frac{1}{2} < x$,

$$x - x^2 \leq \log(1 + x), \tag{H.3}$$

and

$$e^{x-x^2} \leq 1 + x. \tag{H.4}$$

See Figure H.3 for a picture of the inequality in equation (H.3).

Note that equations (H.3) and (H.4) are uninteresting when x is large.

Proof. Since the exponential is an increasing function, the inequality in equation (H.4) is equivalent to equation (H.3), so we only need to prove that inequality.

Define f on $(-\frac{1}{2}, \infty)$ by

$$f(x) = \log(1 + x) - x + x^2.$$

We'll be finished as soon as we prove that $f(x) \geq 0$ for any real number x with $-\frac{1}{2} < x$.

So let's find the minimum value of $f(x)$ on this interval, and see if it's greater than or equal to zero.

We have

$$f'(x) = \frac{1}{1+x} - 1 + 2x = \frac{1 - 1 - x + 2x + 2x^2}{1+x} = \frac{x + 2x^2}{1+x} = \frac{x(1+2x)}{1+x}.$$

Clearly $f'(x) > 0$ for $x \in (0, \infty)$. Thus for $x > 0$ we have $f(x) > f(0) = 0$.

Also, $f'(x)$ has the same sign as $x(1+2x)$ for $x < 0$. Since $1 + 2x > 0$ for $x > -\frac{1}{2}$, we have $x(1+2x) < 0$ for $-\frac{1}{2} < x < 0$. Thus $f'(x) < 0$ for $-\frac{1}{2} < x < 0$.

Hence the minimum value of f on $(-\frac{1}{2}, \infty)$ is $f(0)$, and $f(0) = 0$.

□

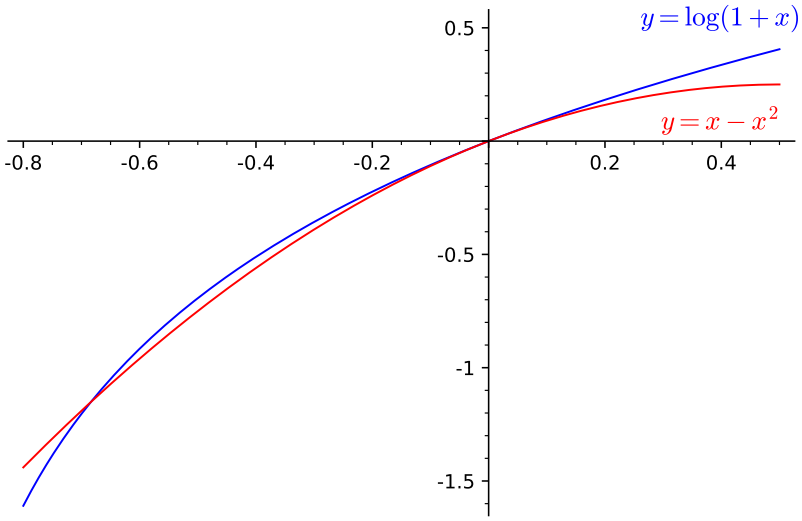


Figure H.3: A lower bound for $\log(1+x)$.

H.1 Proving equation (17.1) again

Proof. Since $a_n \to \infty$,

$$\lim_{n\to\infty} b_n = \lim_{n\to\infty} (a_n b_n)\left(\frac{1}{a_n}\right) = z \cdot 0 = 0.$$

Hence $b_n^2 a_n = b_n \cdot b_n a_n \to 0 \cdot z = 0$.

For n large enough that $b_n > -1$, we can use equation (H.4) and equation (H.2), giving:

$$\left(e^{b_n - b_n^2}\right)^{a_n} \leq (1 + b_n)^{a_n} \leq \left(e^{b_n}\right)^{a_n},$$

i.e.

$$e^{b_n a_n - b_n^2 a_n} \leq (1 + b_n)^{a_n} \leq e^{b_n a_n}.$$

Since $b_n a_n \to z$ and $b_n^2 a_n \to 0$,

$$\lim_{n\to\infty} e^{b_n a_n - b_n^2 a_n} = e^z \text{ and also } \lim_{n\to\infty} e^{b_n a_n} = e^z.$$

Thus $(1 + b_n)^{a_n}$ is trapped between two quantities that both converge to e^z, and so we also have

$$\lim_{n\to\infty} (1 + b_n)^{a_n} = e^z.$$

□

Appendix I

Completing the square

This appendix is for those who have not previously used the procedure of completing the square.

Let $q(x) = ax^2 + bx + c$, with a nonzero. We will prove that q can be written as $a(x + r)^2 + \textbf{stuff}$ for some number r, where **stuff** is a constant expression.

We begin by writing

$$q = a\left(x^2 + (b/a)x + (c/a)\right).$$

If we get $x^2 + (b/a)x + (c/a)$ into the form we want, then mutiplying by a should be no problem. So from now on let's just work on $x^2 + (b/a)x + (c/a)$.

We would like r to be such that

$$x^2 + \frac{b}{a}x + \frac{c}{a} = (x + r)^2 + \textbf{stuff},$$

where **stuff** is a constant expression.

Then

$$\begin{aligned}\textbf{stuff} &= x^2 + \frac{b}{a}x + \frac{c}{a} - (x + r)^2 \\ &= x^2 + \frac{b}{a}x + \frac{c}{a} - \left(x^2 + 2rx + r^2\right) \\ &= \frac{b}{a}x - 2rx + \frac{c}{a} - r^2. \end{aligned} \tag{I.1}$$

To ensure that **stuff** is constant, we need $b/a = 2r$, i.e. $r = b/(2a)$.

Since this choice makes the variable terms cancel out in equation (I.1), we have

$$\textbf{stuff} = \frac{c}{a} - r^2 = \frac{c}{a} - \left(\frac{b}{2a}\right)^2. \tag{I.2}$$

Hence

$$x^2 + \frac{b}{a}x + \frac{c}{a} = \left(x + \frac{b}{2a}\right)^2 + \frac{c}{a} - \left(\frac{b}{2a}\right)^2. \tag{I.3}$$

The original expression was expression $x^2 + \frac{b}{a}x + \frac{c}{a}$. Now we've written it in the form $\textbf{blob}^2 + \textbf{stuff}$, where

$$\textbf{blob} = x + \frac{b}{2a},\ \textbf{stuff} = \frac{c}{a} - \left(\frac{b}{2a}\right)^2. \tag{I.4}$$

Why is this good? Well, we can manipulate **blob** in just the same way that we manipulated x, and **stuff** is just a number, so we have reduced the complexity of the expression.

Exercise I.1 (Completing the square to solve quadratics). The usual "quadratic formula" for solving a quadratic equation is derived by completing the square.

Illustrate this approach by solving the equation $x^2+x-1 = 0$, by completing the square for the given quadratic polynomial. Do not use the quadratic formula.

[Solution]

I.1 Solutions for Appendix I

Solution (Exercise I.1). We want to solve $x^2 + x - 1 = 0$.

"Completing the square" for this polynomial means writing

$$x^2 + x - 1 = (x + r)^2 + \textbf{stuff}, \tag{I.5}$$

where **stuff** is a constant expression.

We don't have to remember any formulas here. Just rearrange equation (I.5), giving:

$$\textbf{stuff} = x^2+x-1-(x+r)^2 = x^2+x-1-\left(x^2+2rx+r^2\right) = x-2rx-1-r^2. \tag{I.6}$$

To ensue that **stuff** is constant expression, we need to have $2rx = x$, i.e. $r = 1/2$. Then equation (I.6) says that

$$\textbf{stuff} = -1 - r^2 = -1 - \frac{1}{4} = -\frac{5}{4}.$$

Thus completing the square in equation (I.5) gives us:

$$x^2 + x - 1 = \left(x + \frac{1}{2}\right)^2 - \frac{5}{4}.$$

Of course we could have used equation (I.4) to get to this point. But one forgets formulas. As long as we remember the ideas we can figure out the correct values.

Anyway, now we have simplified the expression in the equation $x^2+x-1 = 0$, and written it as $\textbf{blob}^2 + \textbf{stuff} = 0$. Solving the original equation is the same as solving

$$\left(x + \frac{1}{2}\right)^2 - \frac{5}{4} = 0,$$

i.e.

$$\left(x + \frac{1}{2}\right)^2 = \frac{5}{4},$$

i.e.

$$x + \frac{1}{2} = \pm\frac{\sqrt{5}}{2}.$$

So the solution is given by

$$x = -\frac{1}{2} \pm \frac{\sqrt{5}}{2}.$$

Appendix J

Cumulative distribution functions

J.1 Cumulative distribution functions

In this section we introduce the cumulative distribution function of a random variable. Cumulative distribution functions are useful for many purposess, and the terminology is standard throughout probability theory.

Definition J.1 (The cumulative distribution function of a random variable). Let X be a random variable. The cumulative distribution function F_X for X is the function on the real line defined by

$$F_X(a) = \mathbf{P}(X \leq a). \tag{J.1}$$

Often we refer to a cumulative distribution function simply as a "distribution function", or use the acronym **CDF**.

Recall that the tail of the distribution of X is defined as $\mathbf{P}(X > t)$, considered as a function of t (Definition 14.12). Thus the tail function is equal to $1 - F_X$, and contains exactly the same information as F_X.

Example J.2.

Let X be the random variable whose distribution is uniform on $[a, b]$ and zero on the rest of the real line.

For any t, $F_X(t) = \mathbf{P}(X \leq t)$. Thus

$$F_X(t) = \begin{cases} 0 & \text{if } t < a, \\ \frac{t-a}{b-a} & \text{if } a \leq t \leq b, \\ 1 & \text{if } t > b. \end{cases}$$

Figure J.1 shows the graph of F_X for the case $a = 2$, $b = 7$.

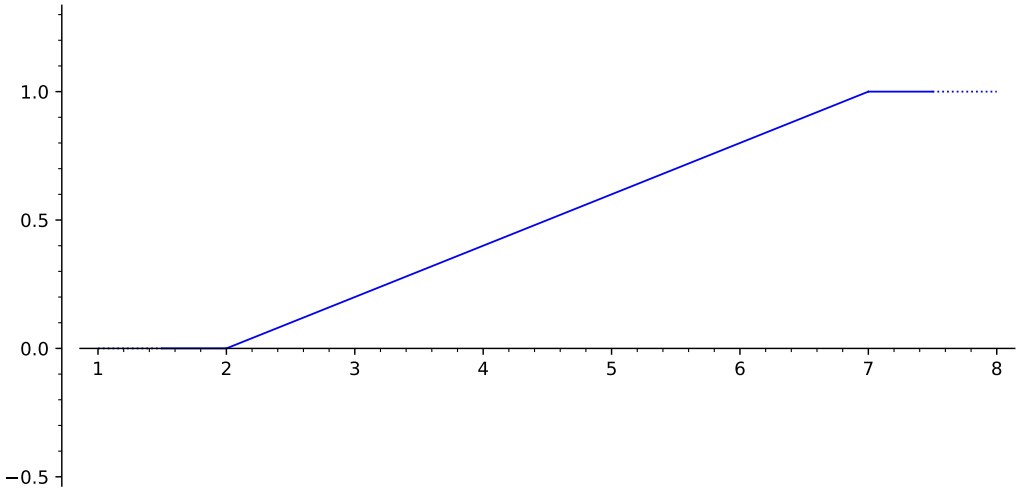


Figure J.1: CDF for X when the distribution of X is uniform on $[2, 7]$

Example J.3.

Let Z be a standard normal random variable (see Definition 18.10). Then

$$F_Z(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz.$$

F_Z increases over the whole real line, with $\lim_{t\to-\infty} F_Z(t) = 0$ and $\lim_{t\to\infty} F_Z(t) = 1$. See Figure J.2.

Incidentally, the derivative of F_Z is exactly equal to the density of the distribution. That is a general fact about distributions. See Section J.4 in Appendix J for more about that.

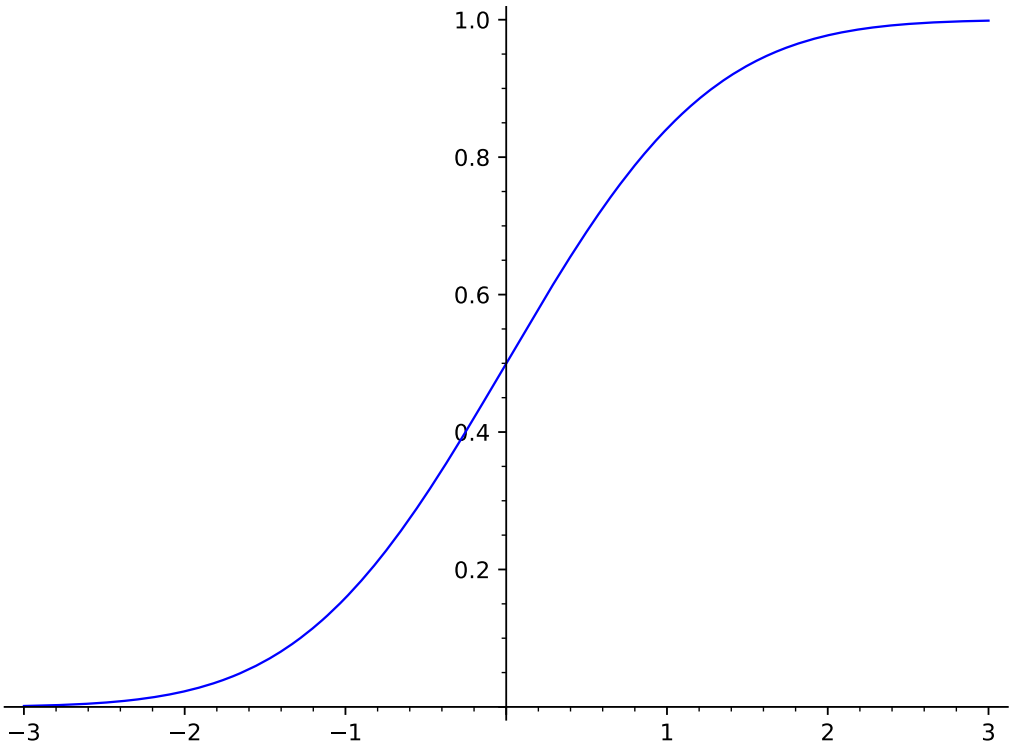


Figure J.2: CDF for the standard normal

Suppose that (in the days before computers) you wanted prepare a to table listing the CDF of a random variable X. You can't list the value of $F_X(a)$ for every real number a, but at least you could list a representative set of values, so that users could find approximations to what they need. That's the way mathematical tables work.

But could you prepare a table which approximately listed the whole distribution of X, in the same way? Your list would have to show the approximate value of $\mathbf{P}(X \in S)$ for every subset S of $\mathbb{R}$. This seems hopelessly difficult.

Even preparing a table listing $\mathbf{P}(X \in (a, b])$ for all intervals $(a, b]$ seems unbearable. However, the next exercise shows that you can at least avoid that task.

Exercise J.1 (Interval probabilities from the CDF). Show that the following statement follows from the definition of a CDF.

Let X be a random variable. For any points a, b in $\mathbb{R}$ with $a \leq b$,

$$\mathbf{P}(a < X \leq b) = F_X(b) - F_X(a). \tag{J.2}$$

[Solution]

It is an interesting fact that the CDF of a random variable determines the whole distribution. That is stated in the next lemma, but the hard part of the proof is omitted.

Lemma J.4 (Interval probabilities characterize distributions). The following statements are equivalent.

(i) X and Y have the same probability distribution. By Definition 9.7, this says that

$$\mathbf{P}(X \in S) = \mathbf{P}(Y \in S) \tag{J.3}$$

for every set S you would ever want to think about.

(ii)

$$\mathbf{P}(X \in (a, b]) = \mathbf{P}(Y \in (a, b]) \tag{J.4}$$

for every interval $(a, b]$.

(iii) $F_X = F_Y$. By Definition J.1 this says that

$$\mathbf{P}(X \in (-\infty, b]) = \mathbf{P}(Y \in (-\infty, b]) \tag{J.5}$$

for every $b \in \mathbb{R}$.

Proof. As usual, we will write $\implies$ to mean "implies".

(i) $\implies$ **(ii)** Take $S = (a, b]$ in equation (J.3).

(i) $\implies$ **(iii)** Take $S = (-\infty, b]$ in equation (J.3).

(ii) $\implies$ **(i)** This part of the proof is omitted! To see why this step is not obvious, we note that equation (J.4) is the special case of equation (J.3), where S is an interval $(a, b]$. There are lots of sets which are not intervals, so

equation (J.3) seems stronger! On the other hand, there are lots of intervals, so equation (J.4) is pretty strong too. So an argument is needed.

The ideas in the proof of this part are not deep but require technicalities from real analysis, which we won't cover.

(iii) $\implies$ (ii) This follows from equation (J.2).

□

J.2 More about cumulative distribution functions

The definition of the cumulative distribution function (CDF) for a random variable is given in Definition J.1. In this section we continue the discussion of CDF's that was begun in Section J.1.

Exercise J.2 (The simplest CDF). Let X be the constant random variable defined by $X(\omega) = c$ for all ω. Sketch the graph of F_X.

[Solution]

Exercise J.3 (Monotonicity). Using the Definition J.1, please check that every distribution function is *monotone increasing*.

[Solution]

Exercise J.4. Let X be a random variable such that $c \leq X \leq d$ always holds.

Show that $F_X(t) = 0$ for all $t < c$, and $F_X(t) = 1$ for all $t \geq d$.

[Solution]

Include the whole domain for a CDF When you are requested to find the formula for a CDF, please state the formula for all points on the real line. This may be unnecessary in many obvious cases, as exercise (J.4) illustrates, but it is a good practice.

Remark J.5 (Useful limits for CDFs). There are many probabilities which can be expressed in terms of F_X. We'll just mention two limits:

$$\lim_{a\to-\infty} F_X(a) = 0, \ \lim_{b\to\infty} F_X(b) = 1. \tag{J.6}$$

By Exercise J.4, equation (J.6) is obvious when X is bounded.

Example J.6. In the case of a simple random variable with finite range, the cumulative distribution function is likely more complicated than it's worth. But the CDF is still defined. To practice with the definition, we'll consider two examples.

(i) Consider the random variable X described in Example 9.2. Take $p = 3/5$, so that $\mathbf{P}(X = 1) = 3/5$ and $\mathbf{P}(X = 0) = 2/5$.

Figure J.3 shows the graph of F_X.

The definition of F_X implies that $F_X(t) = 2/5$ for all t with $0 \le t < 1$, while $F_X(t) = 0$ for $t < 0$ and $F_X(t) = 1$ for $t \ge 1$. The graph shows this.

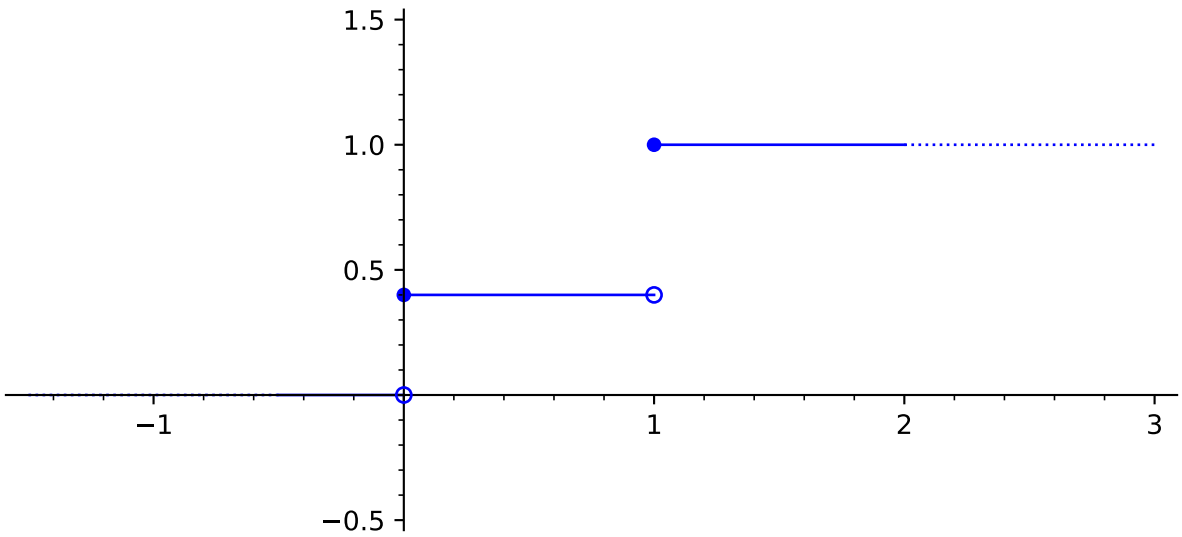


Figure J.3: CDF for result of one coin toss, $p = 3/5$.

(ii) Consider tossing a fair coin 4 times. Let X be the number of heads which are obtained.

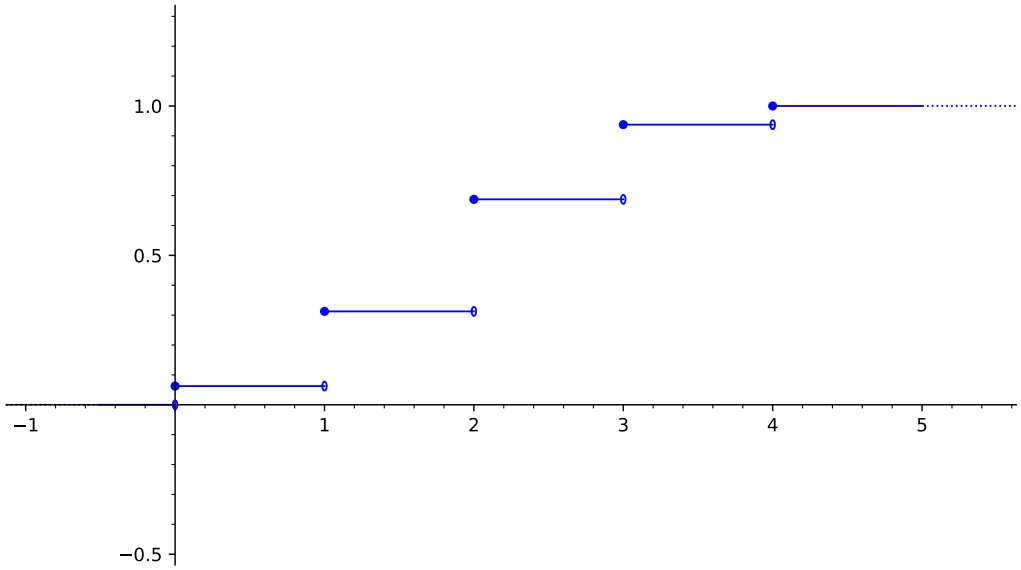


Figure J.4: CDF for result of four fair coin tosses

Figure J.4 shows the graph of F_X.

Since the possible values of T are $0, 1, 2, 3, 4$, when $k \leq t < k+1$, we have $\mathbf{P}(X \leq t) = \mathbf{P}(X \leq k)$. Thus the definition of F_X implies that when

$$k \leq t < k + 1,$$

$F_X(t) = \mathbf{P}(T \leq k) = \mathbf{P}(T = 0) + \ldots + \mathbf{P}(T = k)$. The graph shows this, with

$$\mathbf{P}(T = i) = \binom{4}{i} \left(\frac{1}{2}\right)^4.$$

In the present appendix we will mainly use CDFs for random variables whose distributions have probability densities.

Example J.7.

Let sample space for the probability model be the interval $[0, 4]$, with uniform distribution $\mathbf{P}$. See Figure J.5. Let $X(t) = t^3$.

For $0 \leq t \leq 64$, $F_X(t) = (1/4) * \mathbf{length}(\left[0, t^{1/3}\right]) = t^{1/3}/4$. By Exercise J.4, for every $t < 0$ we have $F_X(t) = 0$ and for every $t > 64$ we have $F_X(t) = 1$. See Figure J.6.

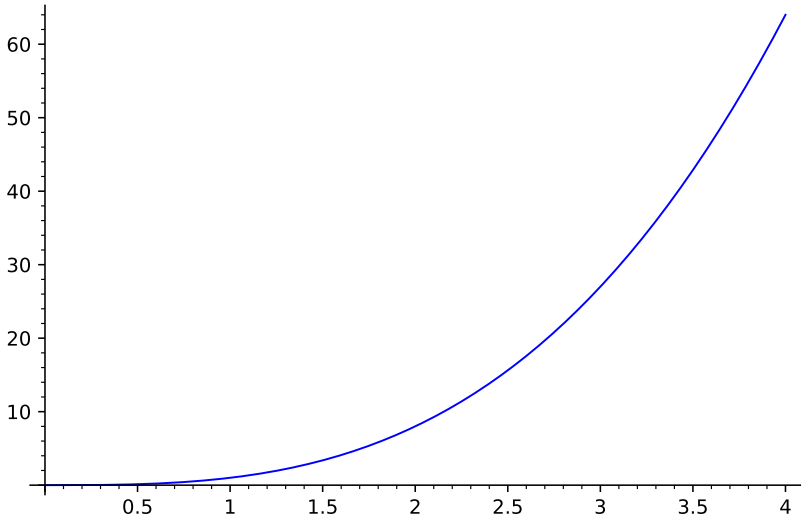


Figure J.5: $X(t) = t^3$ on the sample space $[0, 4]$.

Exercise J.5. Consider the probability model with sample space $[0, \pi/2]$ and probability density $f(u) = \sin u$. Let X be the random variable on $[0, \pi/2]$ defined by $X(u) = e^u$. Find the CDF of X.

[Solution]

The next exercise deals with a situation where it takes more work to find the CDF of the random variable. But it's doable.

Exercise J.6 (A non-monotonic random variable). Consider the probability model with sample space Ω equal to the interval $[0, 3]$ and uniform distribution. Define X on $[0, 3]$ by $X(t) = (1 - t)^2$. See Figure J.7. Find the CDF for X.

Since this random variable is not a monotonic function on its domain, a little extra care is needed in determining $\{X \leq t\}$.

[Solution]

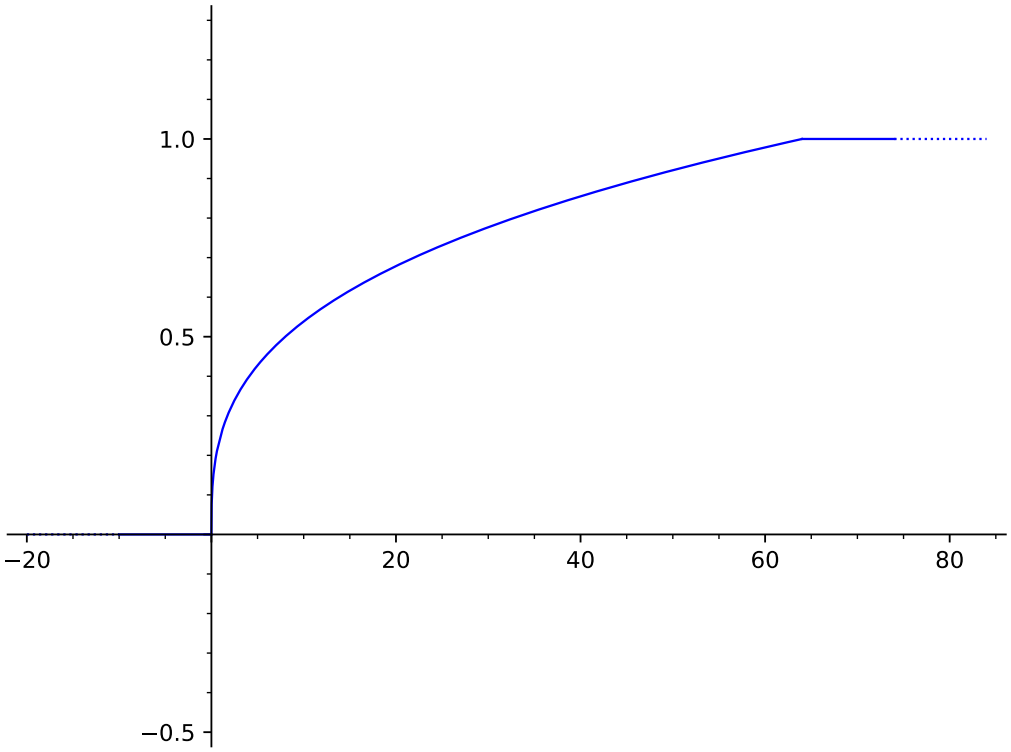


Figure J.6: $F_X(t) = 1/4\, t^{1/3}$

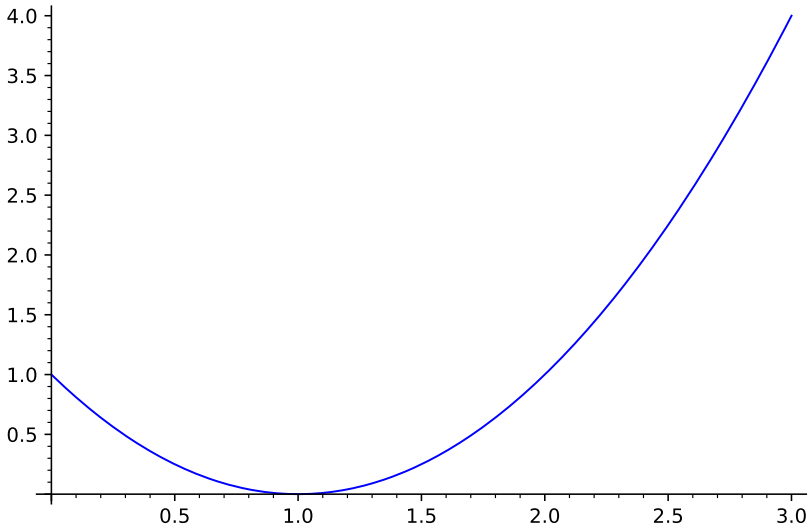


Figure J.7: $X(t) = (1-t)^2$ on the sample space $[0, 3]$.

J.3 Rephrasing the CLT using cumulative distribution functions

This section is aimed at readers who have already learned the statement of the Central Limit Theorem (Theorem 18.14). We'll return to the general properties of cumulative distribution functions in the next section.

The Central Limit Theorem says that when n is sufficiently large,

$$\mathbf{P}(S_n \in J) \approx \mathbf{P}(W_n \in J) \text{ for every interval } J, \tag{J.7}$$

where W_n is a normal random variable with $\mathbf{E}[W_n] = \mathbf{E}[S_n]$ and $\mathbf{Var}(W_n) = \mathbf{Var}(S_n)$.

We claim that equation (J.7) will hold for all J provided that

$$F_{S_n}(b) \approx F_{W_n}(b) \text{ for every } b, \tag{J.8}$$

when n is sufficiently large. Here F_{S_n} denotes the cumulative distribution function (CDF) for S_n and F_{W_n} denotes the CDF for W_n.

By the definition of the CDF, equation (J.8) is simply the statement that

$$\mathbf{P}(S_n \in (-\infty, b]) \approx \mathbf{P}(W_n \in (-\infty, b]) \tag{J.9}$$

when n is sufficiently large.

Clearly equation (J.9) is a special case of equation (J.7). We are claiming here that this special case implies all the other cases! Why is that?

The easiest other case to consider is $J = (a, b]$. We know by equation (J.2) that

$$\begin{aligned} \mathbf{P}(S_n \in (a, b]) &= F_{S_n}(b) - F_{S_n}(a), \\ \mathbf{P}(W_n \in (a, b]) &= F_{W_n}(b) - F_{W_n}(a), \end{aligned} \tag{J.10}$$

Thus

$$\mathbf{P}(S_n \in (a, b]) = F_{S_n}(b) - F_{S_n}(a) \approx F_{W_n}(b) - F_{W_n}(a) = \mathbf{P}(W_n \in (a, b]). \tag{J.11}$$

This is equation (J.7) with $J = (a, b]$.

When $J = (-\infty, b)$ we have to get more technical. The argument is not hard, but we won't take time for that.

Once we know that equation (J.7) holds for the two special cases $J = (-\infty, b]$ and $J = (-\infty, b)$, it's actually easy to show that equation (J.7)

holds for every possible interval J, by considering unions and differences of these two cases. We used that sort of argument already to obtain equation (J.11).

Our claim implies that the following form of the Central Limit Theorem holds is equivalent to the earlier statements of this theorem.

Theorem J.8 (Central Limit Theorem using CDFs). Let $X_1, \ldots, X_n$ be an IID sequence of random variables. Let $S_n = X_1 + \ldots + X_n$.

Suppose that each X_i has mean μ and has variance $\sigma^2 > 0$.

Let W_n be a normal random variable with the same mean and variance as S_n. For any $\varepsilon > 0$, there exists n_0, such that for all $n \geq n_0$,

$$\mid F_{S_n}(b) - F_{W_n}(b) \mid < \varepsilon, \tag{J.12}$$

for *all* $b \in \mathbb{R}$.

Applying equation (18.45) to equation (J.12), we also have the following form.

Let Z be a standard normal random variable. For any $\varepsilon > 0$, there exists n_0, such that for all $n \geq n_0$,

$$\left| F_{S_n}(b) - F_Z\left(\frac{b - n\mu}{\sqrt{n}\sigma}\right) \right| < \varepsilon, \tag{J.13}$$

for *all* $b \in \mathbb{R}$.

Equivalently, for any sequence b_n,

$$\lim_{n\to\infty} \left(F_{S_n}(b_n) - F_Z\left(\frac{b_n - n\mu}{\sqrt{n}\sigma}\right)\right) = 0. \tag{J.14}$$

Equation (J.12) may seem a little easier to think about than equation (J.7), because it focuses on a special case. We can plot the relevant cumulative distribution functions to see whether the stated approximation holds. For example, see Figure 18.9, which shows that $F_{S_n}(b) \approx F_{W_n}(b)$ when tossing a fair coin with $n = 1000$. (In Figure J.8, to avoid cluttering up the graph we only plot the values of F_{S_n} at integer points. The function F_{S_n} is constant on the intervals between those points.)

Statements of the Central Limit Theorem in other textbooks may be more similar to Theorem J.8 than to Theorem 18.14 or Theorem 18.19.

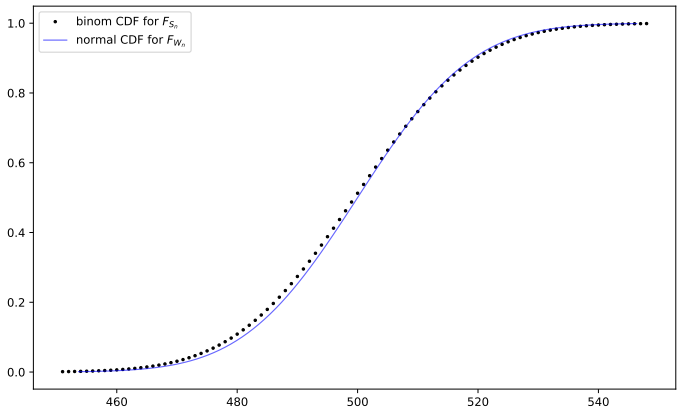


Figure J.8: Comparing the CDF of binomial distribution ($p = .5$, $n = 1000$) with the CDF of the normal distribution having the same mean and variance.

J.4 Finding a density from the CDF of a distribution

Lemma J.9 (Differentiating a CDF). Let X be a random variable whose distribution has density f.

For every real number a,

$$F_X(a) = \int_{-\infty}^{a} f(t)\,dt. \tag{J.15}$$

At any point a where f is continuous,

$$F'_X(a) = f(a). \tag{J.16}$$

Proof. Let $A = \{X \leq a\}$. By definition, $\mathbf{P}(A) = \int_A f$. This is equation (J.15).

The first statement of the Fundamental Theorem of Calculus tells us that when the integrand of an integral is continuous at the upper limit of integration, we can differentiate the integral with respect to its upper limit,

and the result of the differentiation is the value of the integrand at the upper limit. Here the integrand is f and the upper limit of integration is a. This gives equation (J.16).

□

Lemma J.9 shows that if there exists a continuous density f for the distribution of X, then $f = F'_X$. This is useful, but often we encounter random variables whose ranges have a few "bad" points, where F'_X cannot be continuous. So a careful person needs a more general statement, such as the one which is given below by Lemma J.10.

Lemma J.10 (Density from CDF derivative). Let X be a random variable.

Suppose that F_X is a continuous function.

Suppose also $F'_X(t)$ exists and is continuous at all points of $\mathbb{R}$, except possibly at finitely many points $s_1, \ldots, s_n$.

Then F'_X is a density for the distribution of X.

Proof. Since F_X is monotonic increasing, at any point a where $F'_X(a)$ exists it must be true that $F'_X(a) \geq 0$ (since the derivative is the limit of the difference quotients).

Consider an interval $[u, v]$ which does not contain any bad points. The second statement of the Fundamental Theorem of Calculus tells us that

$$\int_u^v F'_X(t)\,dt = F_X(v) - F_X(u).$$

Note that $0 \leq F_X(u) \leq F_X(v) \leq 1$.

For an interval $[u, v]$ which does not contain any bad points, if v increases to a bad point s_j, then $\int_u^v F'_X(t)\,dt$ increases to a limit. The calculus definition of an improper integral says that

$$\lim_{v \nearrow s_j} \int_u^v F'_X(t)\,dt = \int_u^{s_j} F'_X(t)\,dt,$$

and by the continuity of F_X we also have

$$\lim_{v \nearrow s_j} F_X(v) - F_X(u) = F_X(s_j) - F_X(u).$$

Thus

$$\int_u^{s_j} F_X'(t)\,dt = F_X(s_j) - F_X(u).$$

If there is a bad point s_{j-1} adjacent to s_j on the left, we can let u decrease to s_{j-1}. A similar argument to the one just given shows that

$$\int_{s_{j-1}}^{s_j} F_X'(t)\,dt = F_X(s_j) - F_X(s_{j-1}).$$

Based on these arguments, we can now say that if $[u, v]$ is any interval such that the *interior* contains no bad points, we have

$$\int_u^v F_X' = F_X(v) - F_X(u).$$

Next, think about an interval $[u, v]$ such that (u, v) contains exactly one bad point s_k. By what has already been said, we know that

$$\int_u^{s_k} F_X' = F_X(s_k) - F_X(u)$$

and

$$\int_{s_k}^v F_X' = F_X(v) - F_X(s_k).$$

Adding these two equation, we see that for any interval $[u, v]$ whose interior contains at most one bad point,

$$\int_u^v F_X' = F_X(v) - F_X(u).$$

Repeating this argument a finite number of times shows that for any interval $[u, v]$,

$$\int_u^v F_X' = F_X(v) - F_X(u).$$

Thus by Definition 3.4, F_X' is a density for the distribution of X.

This completes the proof. But readers may recall that we extended the definition of a density later, in Definition 15.5. In that definition, a density f for the distribution of X is required to satisfy

$$\mathbf{P}(X \in A) = \int_A f \tag{J.17}$$

for every event A, not just for intervals A. Do we need to check equation (J.17) for sets A which are not intervals?

Fortunately, equation (J.17) automatically holds for all sets if it holds for all intervals A (see Remark 15.6). Thus no further work is required, and we conclude that F'_X is a density for the distribution of X, in the general sense of Definition 15.5.

□

Here's a typical application of Lemma J.10.

Exercise J.7. Let X be the random variable in Exercise J.6.

Use the CDF of X to find a probability density for the distribution of X.

[Solution]

Exercise J.8. In the setting of Exercise F.3, find a probability density for the distribution of X.

[Solution]

Exercise J.9 (Checking that the distribution determines the expected value). In Exercise F.3, you found $\mathbf{E}[X]$. In Exercise J.8 you found the distribution of X, so you can find $\mathbf{E}[X]$ by a different calculation. Check that you obtain the same answer.

[Solution]

J.5 Change of variable

Let X be a random variable whose distribution has a density f on the real line.

Let φ be a continuous and strictly increasing function on an interval J of the real line, and suppose that J contains the range of X.

Does the distribution of $\varphi(X)$ necessarily have a density? And if a density exists, how do we find it?

The connection between the density and the distribution is of course based on integrating the density. For that reason, to give general answers to these questions we might want to use the theory of integration which is developed in advanced analysis courses. But we can already get useful information from calculus.

Assume that f is continuous, except possibly at a finite number of bad points. Then, at all non-bad points, Lemma J.9 tells us that $F'_X(a)$ exists and

$$F'_X(a) = f(a).$$

Suppose we know that φ' exists and is continuous and nonzero at every point of J, except possibly at a finite number of bad points. We'll give a few examples to illustrate an approach based on Lemma J.10 and the chain rule.

(i) Suppose that $\varphi(x) = e^x$. Then

$$F_{\varphi(X)}(a) = F_{e^X}(a) = \mathbf{P}(e^X \leq a)$$

The exponential function is always positive, so for $a \leq 0$, $F_{e^X}(a) = 0$.

The exponential function is an increasing one-to-one function.

Suppose that $a > 0$.

For any real number x, if $x \leq \log a$ then $e^x \leq e^{\log a} = a$.

The logarithm function is an increasing function on its domain. (Of course, it has to be an increasing function, since it is the inverse of an increasing function, but we can take its derivative to check.)

For any real number x, if $e^x \leq a$ then $\log e^x \leq \log a$, i.e. $x \leq \log a$.

We have shown that for $a > 0$ we have

$$\left\{e^X \leq a\right\} = \left\{X \leq \log a\right\}.$$

Hence

$$F_{e^X}(a) = \mathbf{P}(X \leq \log a) = F_X(\log a).$$

Thus if $F'_X(\log a)$ exists, by the chain rule we have

$$F'_{e^X}(a) = F'_X(\log a)\frac{1}{a} = \begin{cases} 0 & \text{if } a < 0, \\ \frac{1}{a} f(\log a) & \text{if } a > 0. \end{cases}$$

We note 0 may be a bad point for F'_{e^X}. There are at most finitely many other bad points. Thus Lemma J.10 says that F'_{e^X} is a density for the distribution of X, and we have the formula for this density.

(ii) Let $\varphi(x) = x + x^3$. Then φ is continuous. Since $\varphi'(x) = 1 + 3x^2$, we see that φ' exists is positive and continuous everywhere. Thus φ is strictly increasing, and in particular φ is one-to-one.

Also $\lim_{x\to\infty} \varphi(x) = \infty$ and $\lim_{x\to-\infty} \varphi(x) = -\infty$. Thus the range of φ is the whole real line.

Let θ denote the inverse function φ^{-1}.

Since φ is increasing, $x \le \theta(a)$ implies $\varphi(x) \le \varphi(\theta(a)) = a$. Since θ is increasing, $\varphi(x) \le a$ implies $\theta(\varphi(x)) \le \theta(a)$, i.e. $x \le \theta(a)$.

We have shown that for $a > 0$ we have

$$\{\varphi(X) \le a\} = \{X \le \theta(a)\}.$$

Hence

$$F_{\varphi(X)}(a) = \mathbf{P}(X \le \theta(a)) = F_X(\theta(a)). \tag{J.18}$$

The algebraic expression for θ does not seem neat, but since φ' is never zero, a calculus theorem says that θ' exists at every point. Also, since $\varphi \circ \theta(y) = y$, the chain rule says that

$$(\varphi' \circ \theta)\, \theta' = 1.$$

That is,

$$\theta' = \frac{1}{\varphi' \circ \theta}.$$

So we can find θ' if we need it.

Using equation (J.18), if $\theta(a)$ is not a bad point for F_X, then by the chain rule $F'_{\varphi(X)}(a)$ exists, and

$$F'_{\varphi(X)}(a) = F'_X(\theta(a))\theta'(a) = f(\theta(a))\theta'(a).$$

Thus Lemma J.10 says $f(\theta(a))\theta'(a)$ is a density for the distribution of $\varphi(X)$.

(iii) The map $x \mapsto x^3$ is one-to-one and onto $\mathbb{R}$. It is an increasing function, and so preserves order. Thus

$$F_{X^3}(a) = \mathbf{P}(X^3 \le a) = \mathbf{P}(X \le a^{1/3}) = F_X(a^{1/3}).$$

For $a \neq 0$, if $a^{1/3}$ is not a bad point for F_X then $F'_{X^3}(a)$ exists and

$$F'_{X^3}(a) = F'_X(a^{1/3})\frac{1}{3}a^{-2/3}.$$

Thus Lemma J.10 says $f(a^{1/3})\frac{1}{3}a^{-2/3}$ is a density for the distribution of X^3.

J.6 Converting a distribution to a uniform

Suppose that we are interested in the distribution of $S_n = X_1 + \ldots + X_n$ for an independent sequence of random variables X_i, where each X_i has the distribution described in Example 18.17. Thus the distribution of each X_i is given by the probability density f, where $f(x) = (1/3)x^2$ on the interval $[-1, 2]$, and f is zero on the rest of the real line.

Example 18.17 suggests that we can *simulate* $X_1, \ldots, X_n$ on a computer, in order to check results we obtained in that example.

Simulating $X_1, \ldots, X_n$ means running a computer program which produces a sequence of values $v_1, \ldots, v_n$ that are statistically similar to a typical sequence of values obtained from $X_1, \ldots, X_n$.

As usual, we won't discuss how to write such a computer program, but we will take note of the fact that it seems to be easier for the computer to simulate a random sequence $Y_1, \ldots, Y_n$ where each Y_i has a *uniform* distribution. So to simplify the computer program we would like to express $X_1, \ldots, X_n$ using a uniform sequence $Y_1, \ldots, Y_n$.

This section shows how to do that.

We'll start by finding F_{X_i}, which is the CDF of X_i. (All the F_{X_i} are the same, since the random variables X_i all have the same distribution.)

Since $F_{X_i}(t) = \mathbf{P}(X_i \leq t)$, for $t \in [-1, 2]$ we compute:

$$F_{X_i}(t) = \int_{-\infty}^{t} f(z)\,dz = \int_{-1}^{t} \frac{1}{3}z^2\,dz = \left(\frac{1}{9}\right)^3\Big|_{-1}^{t} = \frac{1}{9}(t^3 + 1).$$

Of course $F_{X_i}(t) = 0$ if $t < -1$ and $F_{X_i}(t) = 1$ if $t > 2$, but we will concentrate our attention on $[-1, 2]$.

Let F denote the restriction of F_{X_i} to $[-1, 2]$.

It is easy to check that F is continuous and strictly increasing, and maps $[-1, 2]$ onto $[0, 1]$ in a one-to-one fashion.

Let φ denote the inverse of F. The map φ is defined on $[0, 1]$.

Finding a formula for φ is not hard. Just solve $u = (1/9)(t^3 + 1)$ for t. We find that $\varphi(u) = (9u - 1)^{1/3}$, for $u \in [0, 1]$.

Consider $[0, 1]$ as a sample space with uniform distribution. Let Y be a random variable on $[0, 1]$ defined by $Y(u) = u$. We are going to show that $\varphi(Y)$ has the same distribution as X_i!

To do that, we will show that $F_{\varphi(Y)} = F_{X_i}$, and apply Lemma J.4.

As a first step, notice that for any number $b \in [0, 1]$,

$$\mathbf{P}(Y \leq b) = \mathbf{length}([0, b]) = b. \tag{J.19}$$

Let $a \in [-1, 2]$. We claim that:

$$\mathbf{P}(\varphi(Y) \leq a) = \mathbf{P}(Y \leq F(a)). \tag{J.20}$$

Indeed, since F is an increasing function, $\varphi(Y) \leq a$ holds if and only if $F(\varphi(Y)) \leq F(a)$, i.e. if and only if $Y \leq F(a)$. Thus equation (J.20) holds.

But Y has a uniform distribution on $[0, 1]$, so $\mathbf{P}(Y \leq F(a)) = \mathbf{length}([0, F(a)]) = F(a)$.

We conclude that for $a \in [-1, 2]$,

$$\mathbf{P}(\varphi(Y) \leq a) = F(a).$$

That is, for $a \in [-1, 2]$,

$$F_{\varphi(Y)}(a) = F(a) = F_{X_i}(a).$$

We need to check this equation for other values of a.

Since φ maps into $[-1, 2]$, if $a > 2$ it is always true that $\varphi(Y) \leq a$, and for $a < -1$ it is never true that $\varphi(Y) \leq a$.

Hence $F_{\varphi(Y)}(a) = 0$ for $a < -1$ and $F_{\varphi(Y)}(a) = 1$ for $a > 2$.

We've checked that $F_{\varphi(X)}(a) = F_{X_i}(a)$ in all possible cases for a, so

$$F_{\varphi(X)} = F_{X_i}.$$

By Lemma J.4, $\varphi(Y)$ and X_i have identical distributions.

It follows that for an independent sequence $Y_1, \ldots, Y_n$, where each Y_i has the same distribution as Y, the sequence $\varphi(Y_1), \ldots, \varphi(Y_n)$ will have the same statistical properties as $X_1, \ldots, X_n$.

And this tells us that, in order to simulate $X_1, \ldots, X_n$ on a computer, just simulate $Y_1, \ldots, Y_n$, and apply the function φ to each value in the output.

Incidentally, this trick works for any random variable X which is such that we can find the inverse of F_X.

J.7 Solutions for Appendix J

Solution (Exercise J.1). Since

$$\{X \leq a\} \cup \{a < X \leq b\} = \{X \leq b\},$$

and the union is disjoint,

$$\mathbf{P}(X \leq a) + \mathbf{P}(a < X \leq b) = \mathbf{P}(X \leq b).$$

Equation (J.2) follows.

Solution (Exercise J.2). For $t < c$, clearly $\{\omega : \ X(\omega) \leq t\}$ is empty, so $\mathbf{P}(X \leq t) = 0$, i.e $F_X(t) = 0$.

For $t \geq c$, clearly $\{\omega : \ X(\omega) \leq t\} = \Omega$, so $\mathbf{P}(X \leq c) = 1$.

See Figure J.9.

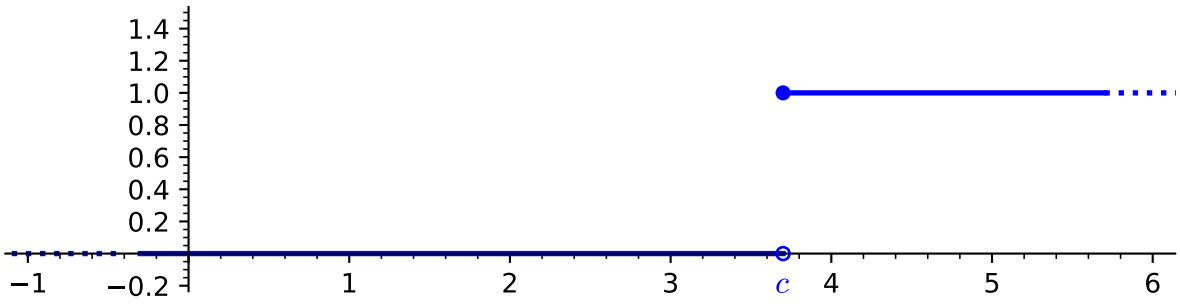


Figure J.9: CDF for a constant random variable equal to c.

Solution (Exercise J.3). Suppose that $a \leq b$. Then $X \leq a \implies X \leq b$, so $\{X \leq a\} \subset \{X \leq b\}$. Hence $\mathbf{P}(X \leq a) \leq \mathbf{P}(X \leq b)$.

Solution (Exercise J.4). For any $t < c$, $\{X \leq t\}$ is the empty set. Hence $\mathbf{P}(X \leq t) = 0$, i.e. $F_X(t) = 0$.

For any $t \geq d$, $\{X \leq t\} = \Omega$. Hence $\mathbf{P}(X \leq t) = 1$, i.e. $F_X(t) = 1$.

Solution (Exercise J.5). It is easy to check that

$$\int_0^{\pi/2} \sin t \, dt = 1,$$

so that f really is a probability density. So the problem makes sense.

The range of X is $[1, e^{\pi/2}]$.

For $t \in [1, e^{\pi/2}]$, $\{X \le t\} = \{x : \ 1 \le e^x \le t\} = [0, \log t]$. Thus for $t \in [1, e^{\pi/2}]$,

$$F_X(t) = \int_0^{\log t} \sin u \, du = -\cos u \Big|_0^{\log t} = 1 - \cos(\log t).$$

By Exercise J.4, for every $t < 1$ we have $F_X(t) = 0$ and for every $t > e^{\pi/2}$ we have $F_X(t) = 1$.

Solution (Exercise J.6). It is helpful to refer to Figure J.7 while solving this problem.

We notice that X is decreasing on $[0, 1]$ and increasing on $[1, 3]$.

The range of X is the interval $[0, 4]$.

By Exercise J.4, we know that $F_X(t) = 0$ for $t < 0$ and $F_X(t) = 1$ for $t \ge 4$.

For $u \in [0, 1]$, the values of X lie in $[0, 1]$. For $u \in [1, 2]$, the values of X also lie in $[0, 1]$. For $u \in [2, 3]$, the values of X lie in $[1, 4]$.

The solutions of $(u-1)^2 = t$ are $u = 1 - \sqrt{t}$, $u = 1 + \sqrt{t}$.

For $t \le 1$,

$$\{u : \ X(u) \le t\} = \left\{u : \ (u-1)^2 \le t\right\} = \left\{u : \ 1 - \sqrt{t} \le u \le 1 + \sqrt{t}\right\}.$$

Thus for $t \le 1$,

$$\mathbf{P}(X \le t) = \frac{1}{3}\left(2\sqrt{t}\right).$$

For $t > 1$,

$$\{u : \ X(u) \le t\} = \left\{u : \ u \le 1 + \sqrt{t}\right\}.$$

Thus for $t > 1$,

$$\mathbf{P}(X \le t) = \frac{1}{3}\left(1 + \sqrt{t}\right).$$

The graph of F_X is shown in Figure J.10.

Solution (Exercise J.7). By Exercise J.6,

$$F(t) = \begin{cases} 0 & \text{if } t < 0, \\ \frac{1}{3}\left(2\sqrt{t}\right) & \text{if } 0 \le t \le 1, \\ \frac{1}{3}\left(1 + \sqrt{t}\right) & \text{if } 1 < t \le 4, \\ 1 & \text{if } t > 4. \end{cases}$$

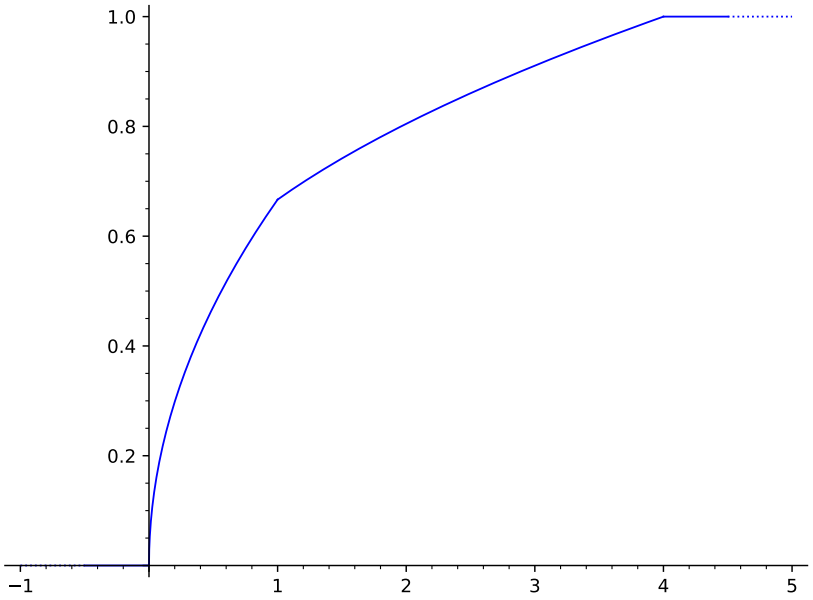


Figure J.10: CDF for $X(t) = (1-t)^2$ on the sample space $[0, 3]$.

Then

$$F'(t) = \begin{cases} 0 & \text{if } t < 0, \\ \frac{1}{3}\frac{1}{\sqrt{t}} & \text{if } 0 < t < 1, \\ \frac{1}{6}\frac{1}{\sqrt{t}} & \text{if } 1 < t < 4, \\ 0 & \text{if } t > 4. \end{cases}$$

Note that $F'(t)$ exists for all t except $t = 0, 1, 4$, and F' is continuous at every point except $0, 1, 4$.

By Lemma J.10, F' is a density for the distribution of X.

The graph of F' is shown in Figure J.11.

Solution (Exercise J.8). The function X is increasing on $[0, \pi/4]$, and has range $[0, 1/\sqrt{2}]$. Thus for $t \in [0, 1/\sqrt{2}]$,

$$\begin{aligned} F_X(t) = \mathbf{P}(X \le t) = \mathbf{P}(\{u : \ u \in [0, \pi/4], \sin u \le t\}) \\ = \{u : \ 0 \le u \le \pi/4, \ u \le \arcsin t\}. \end{aligned}$$

Hence

$$F_X(t) = \{u : \ 0 \le u \le \arcsin t\} = \mathbf{P}([0, \arcsin t]).$$

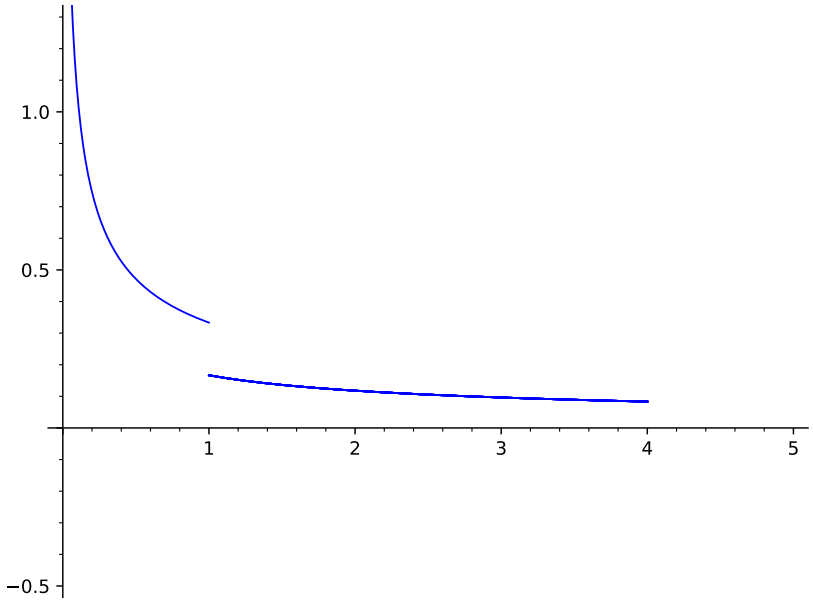


Figure J.11: distribution density for $X(t) = (1-t)^2$ on the sample space $[0,3]$.

By assumption, **P** is given by the density $\sqrt{2}\cos u$, so

$$F_X(t) = \int_0^{\arcsin t} \sqrt{2}\cos u\, du = \sqrt{2}\sin u\Big|_0^{\arcsin t} = \sqrt{2}\ t.$$

And of course, by good old Exercise J.4, we know that $F_X(t) = 0$ for $t < 0$ and $F_X(t) = 1$ for $t \geq 1/\sqrt{2}$.

Thus

$$F_X'(t) = \begin{cases} 0 & \text{if } t < 0, \\ \sqrt{2} & \text{if } 0 \leq t \leq 1/\sqrt{2}, \\ 0 & \text{if } t > 1/\sqrt{t}. \end{cases}$$

By Lemma J.10, the distribution for X has density F_X'.

Solution (Exercise J.9). In Exercise J.8 we found that the distribution of X is uniform on $[0, 1/\sqrt{2}]$.

This distribution has a probability density g given by

$$g(t) = \begin{cases} \sqrt{2} & \text{if } 0 \leq t \leq 1/\sqrt{2}, \\ 0 & \text{otherwise.} \end{cases}$$

By equation (15.6),

$$\mathbf{E}\,[X] = \int tg(t)\,dt = \int_0^{1/\sqrt{2}} t\sqrt{2}\,dt = \frac{t^2}{\sqrt{2}}\bigg|_0^{1/\sqrt{2}} = \frac{1}{2\sqrt{2}}.$$

This agrees with the result of Exercise F.3.

Appendix K

Joint distributions and densities

K.1 Random vectors and joint distributions

Suppose that two physical random variables, X and Y, are associated with some experiment. In a probability model for this experiment there will be two corresponding mathematical random variables, which we will also call X and Y.

Suppose that we know the probability distribution of X. If someone asks us a probability question about the behavior of X, we are ready to answer that question. Similarly, we can answer any probability question about Y if we know the probability distribution of Y.

Now suppose that we need the answer to a more complicated question, involving the behavior of *both* X and Y. For example, suppose we need to find $\mathbf{P}(X < Y)$. To find that probability we are going have to know something about the relationship between X and Y.

Thinking about two or more variables at once can be complicated. One sees that already when studying calculus. To deal with the complexity it is helpful to use some systematic terminology, as in the next definition.

Definition K.1 (Cartesian products). Let C and D be any sets. The set of all pairs (x, y), where $x \in C$ and $y \in D$, is called the Cartesian product of C and D, and is denoted by $C \times D$.

We can picture the Cartesian product of two intervals, $[a, b] \times [c, d]$, as the *rectangle* whose sides are $[a, b]$ and $[c, d]$. See Figure K.1.

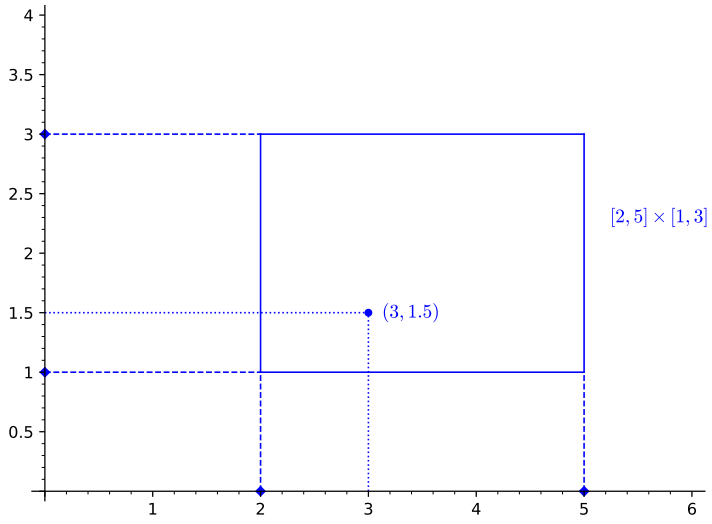


Figure K.1: $(3, 1.5)$ is a point in $[2, 5] \times [1, 3]$.

Readers who are not familiar with Cartesian product terminology should note the next example.

Example K.2. The statement

$$a \leq x \leq b \text{ and } c \leq y \leq d$$

is exactly equivalent to the statement

$$(x, y) \in [a, b] \times [c, d].$$

Thus

$$\{(x, y) : \ a \leq x \leq b \text{ and } c \leq y \leq d\} = [a, b] \times [c, d].$$

Incidentally, notice that from the definition of Cartesian product, $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, which fits our notation for $\mathbb{R}^2$.

The general definition of a random variable (Definition 9.1), states that the physical meaning of a random variable for an experiment is a quantity

whose value depends on the outcome of the experiment, and a mathematical random variable is a map from a sample space to an appropriate set of values. Up to this point we have concentrated on real-valued random variables, but it is often convenient to use vector-valued random variables.

Definition K.3 (Random vectors taking values in $\mathbb{R}^2$). Suppose that real-valued random variables X, Y are defined on the same sample space. Let F be the map from the sample space to $\mathbb{R}^2$, defined by $F(\omega) = (X(\omega), Y(\omega))$ for each sample point ω. Then F is an $\mathbb{R}^2$-valued random variable.

We refer to F as a random vector. We tend to denote a random vector by one of the usual letters we employ for random variables. So we might say we have the random vector Z defined by $Z = (X, Y)$.

Actually one often refers to the random vector using sequence notation to list the vector, so that one just says (X, Y) rather than Z.

Just as in the case of real-valued random variables, we write the set $\{\omega : \ (X(\omega), Y(\omega)) \in S\}$ more briefly as $\{(X, Y) \in S\}$.

By Example K.2, for any random vector (X, Y) taking values in $\mathbb{R}^2$, and any intervals $[a, b]$ and $[c, d]$,

$$\{a \leq X \leq b\} \cap \{c \leq Y \leq d\} = \{(X, Y) \in [a, b] \times [c, d]\}\,. \tag{K.1}$$

More generally, for any subsets A, B of the real line,

$$\{X \in A\} \cap \{Y \in B\} = \{(X, Y) \in A \times B\}\,. \tag{K.2}$$

Of course we can re-express equation (K.2) as:

$$\{X \in A \textbf{ and } \ Y \in B\} = \{(X, Y) \in A \times B\}\,. \tag{K.3}$$

Definition K.4 (Distribution of a random vector). Let $Z = (X, Y)$ be a random vector. The probability distribution of Z is the rule that specifies $\mathbf{P}(Z \in S)$, for every subset S of $\mathbb{R}^2$.

The probability distribution of Z is thus a probability set-function $\mathbf{Q}$, defined for subsets S of $\mathbb{R}^2$ as

$$\mathbf{Q}(S) = \mathbf{P}(Z \in S). \tag{K.4}$$

This definition is essentially the same as the definition of the distribution of a real-valued random variable, given in Definition 9.7. A slightly different terminology is often used:

Definition K.5 (Joint distribution terminology). For any real-valued random variables X and Y, defined mathematically on the sample space Ω of a probability model, the "joint probability distribution" of X and Y is another name for the distribution of the random vector (X, Y).

The use of the word "joint" for the distribution of the vectors is common. It emphasizes the fact that one is dealing with two real-valued random variables at the same time.

It can be shown that for a real-valued random variable X, the distribution of X is uniquely determined, once we know the value of $\mathbf{P}(a \leq X \leq b)$ for all intervals $[a, b]$. Similarly, it can be shown that the distribution of a random vector $Z = (X, Y)$ is uniquely determined, once we know $\mathbf{P}(\, Z \in R)$ for all rectangles R. We state this fact next.

Lemma K.6 (Characterizing a distribution on $\mathbb{R}^2$). Let Z, W be random vectors taking values in $\mathbb{R}^2$, such that $\mathbf{P}(\, Z \in R\,) = \mathbf{P}(\, W \in R\,)$ for every rectangle R.

Then Z and W have the same distribution.

The proof is not hard but requires technicalities, and is omitted.

Exercise K.1. Let X and Y be random variables for some probability model. Show:

$$\mathbf{P}(X \in A) = \mathbf{P}((X, Y) \in A \times \mathbb{R}) \text{ and } \mathbf{P}(Y \in B) = \mathbf{P}((X, Y) \in \mathbb{R} \times B). \tag{K.5}$$

[Solution]

Equation (K.5) tells us that whenever we know the joint distribution of X and Y, we certainly know the distributions for X and Y separately.

K.2 Marginal distributions

Definition K.7 (Marginal distributions). The separate distributions for X and Y are referred to as the *marginal* distributions associated with the joint distribution of X, Y.

The adjective "marginal" is presumably used because the word "margin" can mean "edge", and the values of the distributions of X and Y can be conveniently collected at the edges of a two-dimensional table of probabilities of the form $\mathbf{P}(\,(X,Y) = (x_i, y_j)\,)$,

Suppose that the range of X consists of the distinct values $x_1, \ldots, x_k$, and the range of Y consists of the distinct values $y_1, \ldots, y_\ell$. Then the range of (X,Y) must be included in the set of points (x_i, y_j), although not every pair (x_i, y_j) need be an actual value of (X,Y).

If we know the distribution of (X,Y), then we know $\mathbf{P}(\,(X,Y) = (x,y)\,)$ for every $(x,y) \in \mathbb{R}^2$. In particular we know $\mathbf{P}(\,(X,Y) = (x_i, y_j)\,)$ for every i, j.

Since Y always has some value,

$$\{X = x_i\} = \bigcup_{j=1}^{\ell} \{X = x_i \text{ and } Y = y_j\}\,.$$

Thus

$$\mathbf{P}(X = x_i) = \sum_{j=1}^{\ell} \mathbf{P}(X = x_i \text{ and } Y = y_j). \tag{K.6}$$

Similarly

$$\mathbf{P}(Y = y_j) = \sum_{i=1}^{k} \mathbf{P}(X = x_i \text{ and } Y = y_j). \tag{K.7}$$

In the finite range case, equations (K.6) and (K.7) show that it is easy to calculate the marginal distribution if you know the joint distribution.

For general random variables, Exercise K.1 tells us that

$$\mathbf{P}(X \in S) = \mathbf{P}((X,Y) \in S \times \mathbb{R}),\ \mathbf{P}(Y \in T) = \mathbf{P}((X,Y) \in \mathbb{R} \times T), \tag{K.8}$$

even though this does not necessarily give us a convenient formula.

Exercise K.2. Let X and Y be random variables such that $\mathbf{P}(X = 1) = \mathbf{P}(Y = 1) = 1/2$ and $\mathbf{P}(X = 2) = \mathbf{P}(Y = 2) = 1/2$.

A possible joint distribution $\mathbf{p}$ for X, Y is given by four numbers: $p_{11}, p_{12}, p_{21}, p_{22}$, where $p_{11} = \mathbf{P}(X = 1, Y = 1)$, $p_{12} = \mathbf{P}(X = 1, Y = 2)$, $p_{21} = \mathbf{P}(X = 2, Y = 1)$, $p_{22} = \mathbf{P}(X = 2, Y = 2)$. These numbers must be such that X and Y have *the correct marginal distributions*.

(i) Find three different possible joint distributions for X, Y: **a**, **b** and **c**. Distribution **a** should be such that X and Y are independent.

Display any distribution $\mathbf{p}$ as follows:

$\boldsymbol{p}$	Y	
X	p_{11}	p_{12}
	p_{21}	p_{22}

(ii) Let $\mathbf{p}$ and $\mathbf{q}$ be possible joint distributions for X, Y. Let t be a number in $[0, 1]$. Prove that $t\mathbf{p} + (1-t)\mathbf{q}$ is also a possible joint distribution for X, Y.

[Solution]

K.3 Joint and marginal densities

Equation (15.3) gives the general definition for the density of a probability distribution, on any space.

Definition 9.11 says that a function f is a probability density for the distribution of a real-valued random variable X if $\mathbf{P}(X \in S) = \int_S f$ for subsets S of the real line.

Now we consider a random vector (X, Y).

Definition K.8 (Density for a joint distribution). Let X and Y be real-valued random variables for some probability model. Suppose that there exists a probability density function h on $\mathbb{R}^2$, such that

$$\mathbf{P}\left((X, Y) \in S\right) = \int_S h(x, y)\, dy\, dx \tag{K.9}$$

for subsets S of $\mathbb{R}^2$.

This equation uses the modern notation for integration over a set, given Definition 3.6. We could also write equation (K.9) in calculus notation as

$$\mathbf{P}\left((X,Y)\in S\right)=\int\int_{S}h(x,y)\,dy\,dx. \tag{K.10}$$

When equation (K.9) holds for all S, we say that h is a probability density for the distribution of the random vector (X,Y), and we write this briefly as $(X,Y)\sim h$.

We also say that h is a probability density for the joint distribution of X and Y.

Just as in the case of the real line, if it happens that all the values of (X,Y) lie in some subset T of $\mathbb{R}^2$, then the density h can be assumed to be zero at all points in the complement of T.

Lemma K.6 can be used to justify the next fact.

Lemma K.9 (Characterizing a distribution density on $\mathbb{R}^2$). Let X,Y be real-valued random variables, and let h be a probability density function on $\mathbb{R}^2$, such that

$$\mathbf{P}(\,(X,Y)\in R\,)=\int_R h$$

for every rectangle R.

Then

$$\mathbf{P}(\,(X,Y)\in S\,)=\int_S h$$

for all sets S, so h is a density for the probability distribution of (X,Y).

Lemma K.10 (Marginal density formula from joint density). Let X and Y be real-valued random variables whose distribution has a joint probability density h defined on $\mathbb{R}^2$. Then the distribution of X has a density f on $\mathbb{R}$ and the distribution of Y has a density g on $\mathbb{R}$, given by

$$\begin{aligned} f(x)&=\int_{-\infty}^{\infty}h(x,y)\,dy,\\ g(y)&=\int_{-\infty}^{\infty}h(x,y)\,dx. \end{aligned} \tag{K.11}$$

We can express equation (K.11) by saying that we obtain the probability density for one coordinate of (X, Y) by *integrating out* the variable corresponding to the other coordinate.

Proof. Let A be any interval of $\mathbb{R}$. By equation (K.5),

$$\mathbf{P}(X \in A) = \mathbf{P}((X, Y) \in A \times \mathbb{R}) = \int\int_{A \times \mathbb{R}} h(x, y)\, dy\, dx$$
$$= \int_A \left(\int_{-\infty}^{\infty} h(x, y)\, dy \right) dx = \int_A f(x)\, dx,$$

where f is defined as in equation (K.11).

Since

$$\mathbf{P}(X \in A) = \int_A f(x)\, dx,$$

for every interval A, f is a density for the distribution of X.

The proof for Y is similar.

□

Exercise K.3. In the setting of Exercise F.1 find the density f for the distribution of X.

[Solution]

K.4 Joint density for independent random variables

Suppose that real-valued random variables X, Y are independent, and we know the probability distribution of X and the probability distribution of Y. Can we find the joint distribution for X, Y?

When X and Y have finite or countable range, the answer is easy:

$$\mathbf{P}(X = x_i \text{ and } Y = y_j) = \mathbf{P}(X = x_i)\, \mathbf{P}(Y = y_j) \tag{K.12}$$

by independence. For a subset S of $\mathbb{R}^2$, we can find $\mathbf{P}(\,(X,Y) \in S\,)$ by adding up $\mathbf{P}(X = x_i \text{ and } Y = y_j)$ for all $(x_i, y_j) \in S$.

The other easy case is the situation in which the distribution of each random variable has its own density. That's the subject of the present section. We don't assume ahead of time that there is a density for the joint distribution of X, Y, but it turns out that there is one, and the formula is similar to equation (K.12).

Lemma K.11 (Density when X, Y are independent). Let X, Y be real-valued random variables which are independent.

Suppose that f is a density for the distribution of X and g is a density for the distribution of Y. Let $h(x, y) = f(x)g(y)$. Then h is a density for the distribution of (X, Y).

Proof. Let J_1 and J_2 be intervals. Then

$$\begin{aligned}\mathbf{P}(\,(X,Y) \in J_1 \times J_2\,) = \mathbf{P}(X \in J_1)\mathbf{P}(X \in J_2) &= \left(\int_{J_1} f\right)\left(\int_{J_2} g\right) \\ &= \int\int_{J_1 \times J_2} f(x)g(y)\,dx\,dy = \int_{J_1 \times J_2} h.\end{aligned}$$

We have shown that for every rectangle R,

$$\mathbf{P}(\,(X,Y) \in R\,) = \int_R h.$$

By Lemma K.9, h is a density for the distribution of (X, Y).

□

K.5 Convolutions: finding the density for the sum of two independent random variables

Let X, Y be independent real-valued random variables whose distributions are given by densities f, g, respectively. Then a joint density h for (X, Y) is given by $h(x, y) = f(x)g(y)$.

Our goal in this section is to find a density for the distribution of $X+Y$. Let A be an interval of the real line. Let $B = \{(x,y) : \ x+y \in A\}$.

$$\mathbf{P}(X+Y \in A) = \mathbf{P}((X,Y) \in B).$$

It is easy to check from the definitions that $\mathbf{1}_B(\,(x,y)\,)$ is equal to 1 exactly when $\mathbf{1}_A(x+y)$ is equal to 1, i.e. $\mathbf{1}_B(\,(x,y)\,) = \mathbf{1}_A(x+y)$.

$$\begin{aligned}\mathbf{P}((X,Y) \in B) \int_B f(x)g(y)\,dy\,dx &= \int \mathbf{1}_B(\,(x,y)\,)f(x)g(y)\,dy\,dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_A(x+y)f(x)g(y)\,dy\,dx.\end{aligned}$$

For fixed x, change the variable in the inner integral from y to $t-x$. Then $x+y=t$, and $dy = dt$, and the double integral becomes

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_A(t)f(x)g(t-x)\,dt\,dx &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_A(t)f(x)g(t-x)\,dx\,dt \\ &= \int_{-\infty}^{\infty} \mathbf{1}_A(t)\psi(t)\,dt,\end{aligned}$$

where $\mathbf{1}_A$ is the indicator function for A (Definition 11.1) and

$$\psi(t) = \int_{-\infty}^{\infty} f(x)g(t-x)\,dx. \tag{K.13}$$

The function denoted by ψ in equation (K.13) is known as the *convolution* of f and g, and is denoted by $f * g$.

We have shown that for any interval A of the real line,

$$\mathbf{P}(X+Y \in A) = \int_A f*g(t)\,dt.$$

Thus by Definition 3.4, $f*g$ is a density for the distribution of $X+Y$. Finding this density was our goal, so we are finished.

Remark K.12 (Two physical interpretations for convolutions).
We would like to interpret the ψ in equation (K.13). This is $f*g$.

(1) Total influence field Suppose that a function g on the real line is such that $g(t)$ measures the intensity of some influence at a point t, where

the influence, whatever its nature, is produced by a particle of unit "weight" located at the origin. One might think of $g(t)$ as representing a physical field produced by a unit source located at the origin.

Assuming that the laws governing this influence are the same throughout the real line, any point can be treated as the origin of the influence, so a particle located at the point x should produce an influence with intensity of $g(t-x)$ at the point t. If the particle has "weight" c, then the intensity produced is $cg(t-x)$.

Now consider a density f of such particles. Then the weight of particles in a short interval of length Δx near the point x is given by $f(x)\Delta x$. These particles contribute $f(x)\Delta x g(t-x)$ to the influence which is felt at t. Adding up these influences requires an integral, so the total influence felt at t is then:

$$\int f(x)g(t-x)\,dx = f * g(t).$$

(2) Moving average Now suppose that g is a probability density, and the f is some function whose values we wish to study. Define h by $h(x) = g(-x)$. Then h is also a probability density, which might be called a *reflected* version of g.

Consider $g(t-x)$ as a function of t. This is equal to $h(x-t)$. As a function of x, $h(x-t)$ is a *shifted* version of h. We can think of sliding the graph of h a signed distance t along the horizontal axis.

Looking at equation (K.13) we see that the function $f * g(t)$ is an average of values of f, using the shifted probability density $h(x-t)$.

This average is calculated using the shifted probability density $h(x-t)$. We can image sliding $h(x-t)$ along the horizontal axis and calculating the average value $f * g(t)$ at each point t.

K.6 Solutions for Appendix K

Solution (Exercise K.1). For a real-valued random variable Y, to say that $Y \in \mathbb{R}$ places no restriction on the value of Y.

Thus when discussing real-valued random variables, the statements "$X \in A$ and $Y \in \mathbb{R}$" and "$X \in A$" provide the same information.

In set language,

$$\{X \in A \text{ and } Y \in \mathbb{R}\} = \{X \in A\}\,.$$

Solution (Exercise K.2).

(i) The independent case always works, and that will be our choice for $\boldsymbol{a}$.

$\boldsymbol{a}$	Y	
X	$\frac{1}{4}$	$\frac{1}{4}$
	$\frac{1}{4}$	$\frac{1}{4}$

If we move probability mass vertically in a representation like the one for $\boldsymbol{a}$, it has no effect on the marginal distribution of Y. The movement does affect the marginal distribution of X, but, since the distribution of X is obtained by summing *rows*, it doesn't matter in which column the movement takes place.

So let's obtain $\boldsymbol{b}$ from $\boldsymbol{a}$ by moving mass 1/8 upward in column one, and compensating by moving mass 1/8 downward in column two.

$\boldsymbol{b}$	Y	
X	$\frac{3}{8}$	$\frac{1}{8}$
	$\frac{1}{8}$	$\frac{3}{8}$

We'll obtain a different distribution $\boldsymbol{c}$ from $\boldsymbol{a}$ by moving mass 1/8 downward in column one, and compensating by moving mass 1/8 upward in column two.

$\boldsymbol{c}$	Y	
X	$\frac{1}{8}$	$\frac{3}{8}$
	$\frac{3}{8}$	$\frac{1}{8}$

(ii) Let $\mathbf{r} = t\mathbf{p} + (1-t)\mathbf{q}$.

$$r_{ij} = tp_{ij} + (1-t)q_{ij}.$$

Then the marginal distribution for X using $\mathbf{r}$ is given by

$$\begin{aligned}\mathbf{P}(X=i) &= r_{i1} + r_{i2} = tp_{i1} + (1-t)q_{i1} + tp_{i2} + (1-t)q_{i2} \\ &= t\,(p_{i1} + p_{i2}) + (1-t)\,(q_{i1} + q_{i2}) = t\frac{1}{2} + (1-t)\frac{1}{2} = \frac{1}{2}.\end{aligned}$$

A similar computation works for the marginal distribution for Y using $\mathbf{r}$.

Solution (Exercise K.3). We obtain the formula for f using equation (K.11). Thus

$$f(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{2\pi}(2+x)\,dy = \frac{1}{\pi}(2+x)\sqrt{1-x^2}.$$

Appendix L

More about joint distributions

L.1 Checking independence using joint distributions

A model often makes significant use of more than one random variable. It is important to be able to tell when the random variables are independent.

The definition of independence for random variables X, Y (Definition 12.2) says that real-valued random variables X, Y are independent if for any subsets S and T of $\mathbb{R}$, $\{X \in S\}$ and $\{Y \in T\}$ are independent, i.e.

$$\mathbf{P}(\{X \in S\} \cap \{Y \in T\}) = \mathbf{P}(\{X \in S\})\mathbf{P}(\{Y \in T\}).$$

In other words, X, Y are independent if for all S, T,

$$\mathbf{P}(\,(X, Y) \in S \times T\,) = \mathbf{P}(X \in S)\mathbf{P}(Y \in T). \tag{L.1}$$

This equation makes it clear that independence for random variables is a property of the *joint distribution* of the random variables.

Example L.1 (A non-independence example). In the setting of Exercise F.1 we can prove that the random variables X and Y are *not* independent.

It seems obvious physically that X and Y cannot be independent, since information about the value of X can give you information about the value of Y. For example, knowing that X is near one tells you that Y is near zero, and knowing the exact value of X tells you that Y takes one of at most two possible values.

For an argument using the mathematical definition, let $A = (1/\sqrt{2}, 1) = B$. Then $A \times B$ is a rectangle entirely outside the unit circle.

Since $\mathbf{P}(X^2 + Y^2 \leq 1) = 1$, $\mathbf{P}(A \times B) = 0$.

We know that $\{(X, Y) \in A \times B\} = \{X \in A \text{ and } Y \in B\}$, so $\mathbf{P}(X \in A \text{ and } Y \in B) = 0$. But both X and Y have densities that are positive everywhere on $(-1, 1)$ so by integrating these densities we know that $\mathbf{P}(X \in A) > 0$ and $\mathbf{P}(Y \in B) > 0$.

Thus $\mathbf{P}(X \in A \text{ and } Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B)$ is false. Hence $\{X \in A\}$ and $\{Y \in B\}$ are not independent, and so X and Y are not independent.

Lemma 12.3 tells us how to check efficiently for independence when the ranges of X and Y are finite. Now we will derive a somewhat similar criterion for independence when the distributions of X and Y are given by densities.

What we want to do now is turn Lemma K.11 around, and prove a converse statement. Given a density $h(x, y)$ for the random vector (X, Y), the idea is that X, Y will be independent if we can factor h into a product of a function of x times a function of y.

If we can factor h in this way, say $h(x, y) = f(x)g(y)$, then the factors f, g will give us the marginal densities for X and Y. But a tiny bit of extra work is necessary, because it may not be true that $\int f = 1$ and $\int g = 1$. For example, we could always multiply f by a million and divide g by a million, and obtain another perfectly correct factorization.

So what is actually true is that the factors $f(x)$ and $g(y)$ will be the marginal densities, after we *normalize* them, i.e. after we multiply each factor by an appropriate constant to ensure that its integral is equal to one. The next lemma explains all this.

Lemma L.2 (Density criterion for independence). Let X, Y be real-valued random variables in some probability model. Suppose that h is a density for the distribution of (X, Y), and and let f and g be nonnegative functions on $\mathbb{R}$, such that $h(x, y) = f(x)g(y)$ for all x, y.

Then X and Y are independent random variables. Furthermore, for some constants c_1 and c_2, $c_1 f$ is a probability density for the distribution of X and $c_2 g$ is a probability density for the distribution of Y.

Proof. By Lemma K.10, a probability density for the distribution of X is given by

$$\int_{-\infty}^{\infty} f(x)g(y)\,dy = c_1 f(x), \tag{L.2}$$

where

$$c_1 = \int_{-\infty}^{\infty} g(y)\,dy. \tag{L.3}$$

Similarly a probability density for the distribution of Y is given by

$$\int_{-\infty}^{\infty} f(x)g(y)\,dx = c_2 g(y), \tag{L.4}$$

where

$$c_2 = \int_{-\infty}^{\infty} f(x)\,dx. \tag{L.5}$$

We note that

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)g(y)\,dx\,dy$$
$$\left(\int_{-\infty}^{\infty} f(x)\,dx\right)\left(\int_{-\infty}^{\infty} g(y)\,dy\right) = c_1 c_2. \tag{L.6}$$

Let J_1, J_2 be intervals of the real line. Using the definitions,

$$\mathbf{P}(\{X_1 \in J_1\} \cap \{X_2 \in J_2\}) = \mathbf{P}(\,(X_1, X_2) \in J_1 \times J_2\,) = \int\int_{J_1\times J_2} h(x,y)\,dx\,dy$$
$$= \int\int_{J_1\times J_2} c_1 f(x) c_2 g(y)\,dx\,dy = \int_{J_1}\int_{J_2} f(x)g(y)\,dx\,dy$$
$$= \left(\int_{J_1} f(x)\,dx\right)\left(\int_{J_2} g(y)\,dy\right) = \mathbf{P}(X_1 \in J_1)\mathbf{P}(X_2 \in J_2).$$

Thus the events $\{X_1 \in J_1\}$ and $\{X_2 \in J_2\}$ are independent events. Since this is true for all intervals J_1, J_2, Lemma 12.7 shows X_1, X_2 are independent random variables.

□

When we want to prove that independence does *not* hold, it may be convenient to check densities at a single point.

Since independence is defined in terms of probabilities, not densities, we will first have to make a connection between the value of a density at one point and a probability. Continuity lets us do that.

Lemma L.3 (Density criterion for non-independence). Let X, Y be real-valued random variables for some probability model, such that the joint distribution for X, Y has a density h.

Let f and g be densities for the distributions of X and Y respectively.

Suppose for some (a, b) that f is continuous at a, that g is continuous at b, and that h is continuous at (a, b). Suppose also that $h(a, b) \neq f(a)g(b)$.

Then X and Y are *not* independent.

Proof. Assume that X and Y *are* independent. We will obtain a contradiction.

By Lemma K.11, $f(x)g(y)$ is a density for the distribution of (X, Y).

Thus $f(x)g(y)$ and $h(x, y)$ are both densities for the same distribution, and they differ at a point (a, b) where both of these densities are *continuous*. That can't happen! For example, suppose $f(a)g(b) > h(a, b)$. Let $\varepsilon = f(a)g(b) - h(a, b) > 0$.

By continuity there is a disc D around (a, b) such that

$$f(x)g(y) > h(x, y) + \frac{\varepsilon}{2}$$

holds everywhere on D. But then

$$\begin{aligned}\mathbf{P}(\,(X, Y) \in D\,) &= \int_D f(x)g(y)\,dx\,dy > \int_D h(x, y)\,dx\,dy + \frac{\varepsilon}{2}\mathbf{area}(D) \\ &= \mathbf{P}(\,(X, Y) \in D\,) + \frac{\varepsilon}{2}\mathbf{area}(D).\end{aligned}$$

That is a contradiction!

□

Note that Lemma L.3 gives us a convenient way to come to the conclusion found in Example L.1.

Remark L.4 (Existence of independent random variables). In mathematical arguments it can be useful to know that given probability densities f and g, there aways exist independent mathematical random variables X and Y such that f is a density for the distribution of X and g is a density for the distribution of Y. This point seems physically obvious (just do two separate experiments), so you may not want to worry about it. But a purely mathematical argument is easy too.

To show this, let Ω be $\mathbb{R}^2$, and let $\mathbf{P}$ be the probability distribution with density $h(x, y) = f(x)g(y)$.

Let $X(\,(x, y)\,) = x$ and let $Y(\,(x, y)\,) = y$. Using definitions one can show that h is a probability density for the distribution of the random vector $Z = (X, Y)$.

By Lemma L.2, X, Y are independent, f is a density for the distribution of X and g is a density for the distribution of Y.

L.2 Conditional densities

Let X and Y be random variables with a joint density h on $\mathbb{R}^2$. Since a probability density is a machine that produces probabilities when we integrate, we can calculate conditional probabilities involving X and Y using the standard definitions. For example, for any subsets A and B of $\mathbb{R}$,

$$\mathbf{P}(X \in A \,|\, Y \in B) = \frac{\mathbf{P}(\{X \in A\} \cap \{Y \in B\})}{\mathbf{P}(Y \in B)} = \frac{\mathbf{P}(\,(X, Y) \in A \times B)}{\mathbf{P}(Y \in B)} = \frac{\int\int_{A\times B} h(x, y)\, dx\, dy}{\int\int_{\mathbb{R}\times B} h(x, y)\, dx\, dy}. \quad \text{(L.7)}$$

See Figure L.1.

Using equation (L.7) to find conditional probabilities may require some computational work, but it does not require new ideas. However, it can at times be useful to assign a meaning to $\mathbf{P}(X \in A \,|\, Y = b)$, for some $b \in \mathbb{R}$, even when $\mathbf{P}(Y = b) = 0$. This situation is obviously not covered by Definition 4.2, since it would involve division by zero in equation (4.1). A correct definition is given below in Definition L.5.

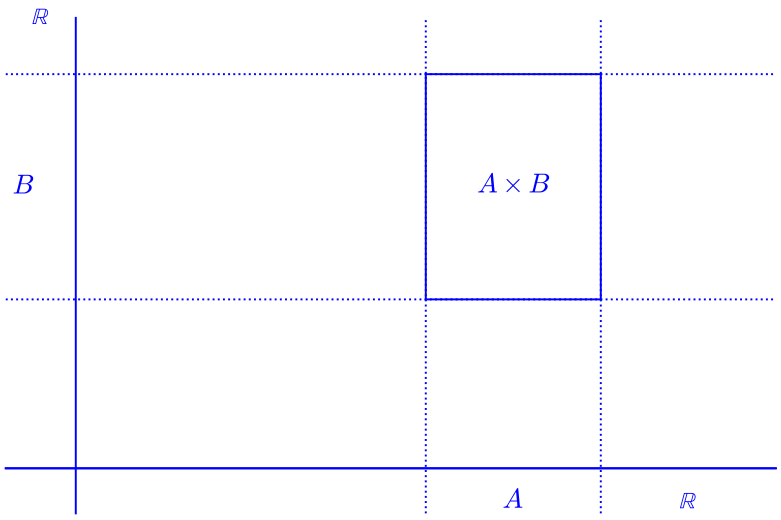


Figure L.1: Integrate h over $A \times B$ to get $\mathbf{P}(X \in A \text{ and } Y \in B)$. Integrate h over the horizontal strip $\mathbb{R} \times B$ to get $\mathbf{P}(Y \in B)$.

One can just accept that as a definition, and then see how such a conditional probability is used, in Theorem L.6. But before giving the statement of Definition L.5 it may be of interest to readers to motivate this definition by considering a limit of conditional probabilities, as in equation L.8.

Motivating Definition L.5

Physically, we can think that $\mathbf{P}(X \in A \,|\, Y = b)$ expresses the probability that $X \in A$ when given that Y has a value that is "close" to b, so that $\mathbf{P}(X \in A \,|\, Y = b)$ actually means $\mathbf{P}(X \in A \,|\, Y \approx b)$, at least in situations where $\mathbf{P}(Y \approx b) > 0$. But the event $\{Y \approx b\}$ is not precisely defined. Does it mean $\{b - .0001 < Y < b + .0001\}$, or does it mean $\{b - .0000001 < Y < b + .0000001\}$?

We have to ask this, because $\mathbf{P}(b - .0001 < Y < b + .0001)$ may be many times greater than $\mathbf{P}(b - .0000001 < Y < b + .0000001)$. We have to check mathematically that there is not a problem.

Let's look at the case that the distribution of (X, Y) has a *continuous* joint density $h(x, y)$. We wish to define a suitable value for $\mathbf{P}(X \in A \,|\, Y = b)$. In this situation let B be a small subinterval of $\mathbb{R}$, say with length 2δ, such that $b \in B$. See Figure L.2.

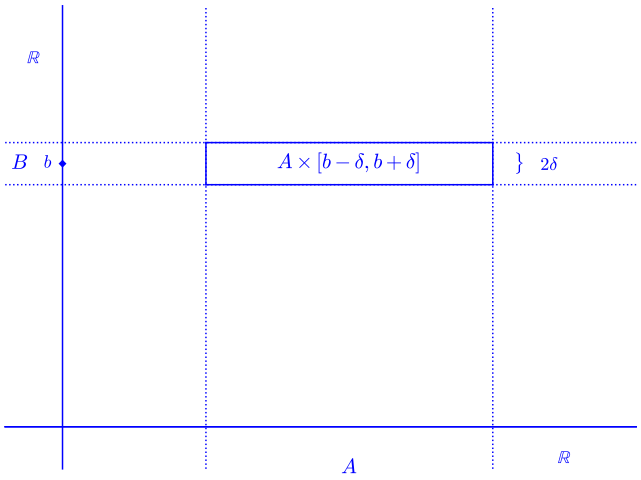


Figure L.2: Integrate h over $A \times B$ to get $\mathbf{P}(X \in A \text{ and } Y \in B)$. Integrate h over the horizontal strip $\mathbb{R} \times B$ to get $\mathbf{P}(Y \in B)$.

We think of $\mathbf{P}(X \in A \,|\, Y = b)$ as being such that

$$\mathbf{P}(X \in A \,|\, Y = b) \approx \mathbf{P}(X \in A \,|\, Y \in B). \tag{L.8}$$

To make sense of equation (L.8), this approximation should be valid for any small interval B around b!

We have chosen $B = [b - \delta, b + \delta]$. Then

$$\begin{aligned} \mathbf{P}(X \in A \,|\, Y \in B) &= \frac{\mathbf{P}(X \in A \text{ and } Y \in B)}{\mathbf{P}(Y \in B)} \\ &= \frac{\int_A \int_{b-\delta}^{b+\delta} h(x,t)\, dt\, dx}{\int_{-\infty}^{\infty} \int_{b-\delta}^{b+\delta} h(x,t)\, dt\, dx}. \end{aligned} \tag{L.9}$$

Suppose A is bounded. The continuity of h implies that if δ is small enough, for all $x \in A$ we have

$$h(x,t) \approx h(x,b) \text{ for all } t \in [b - \delta, b + \delta].$$

Applying this approximation to equation (L.9) gives

$$\mathbf{P}(X \in A \,|\, Y \in B) \approx \frac{\int_A h(x,b) 2\delta \, dx}{\int_{-\infty}^{\infty} h(x,b) 2\delta \, dx} = \frac{\int_A h(x,b) \, dx}{\int_{-\infty}^{\infty} h(x,b) \, dx}.$$

Equation (K.11) tells us that $\int_{-\infty}^{\infty} h(x,b) \, dx = g(b)$, where g is the density of Y. So

$$\mathbf{P}(X \in A \,|\, Y \in B) \approx \int_A \frac{h(x,b)}{g(b)} \, dx. \tag{L.10}$$

Notice that because the length 2δ has been cancelled out in equation (L.10), the approximate value we obtained for $\mathbf{P}(X \in A \,|\, Y \in B)$ does not depend on the choice of B. It will be a good approximation if h is continuous and B is a *small* interval containing b.

Based on equation (L.10), here is the precise definition of a conditional probability given the exact value of a random variable, when the random variable has a density.

Definition L.5 (Conditional probability given exact value). Let X, Y be random variables for some probability model. Let h be a density for the joint distribution of X, Y. Let g be a density for the distribution of Y. For any value $b \in \mathbb{R}$ such that $g(b) > 0$, and any subset A of $\mathbb{R}$, a version of the conditional probability that $X \in A$ given $Y = b$ is:

$$\mathbf{P}(X \in A \,|\, Y = b) = \int_A \frac{h(x,b)}{g(b)} \, dx. \tag{L.11}$$

If $g(y) = 0$, for convenience we will define $\mathbf{P}(X \in A \,|\, Y = y)$ to be zero. Setting $\mathbf{P}(X \in A \,|\, Y = b) = 0$ when $g(b) = 0$ has no physical meaning. We are making that definition here simply to ensure that $\mathbf{P}(X \in A \,|\, Y = b)$ is always defined mathematically.

We speak of $\mathbf{P}(X \in A \,|\, Y = b)$ as a *version* of the conditional probability since it depends on the choice of the values for $h(x,b)$ and $g(b)$, and those values (at a few points) can depend on the choice of the density h.

Remember that equation (L.5) is a new definition, not something we can derive from our previous definitions. Equation (L.10) suggests that $\mathbf{P}(X \in A \,|\, Y = b)$ ought to be a useful concept, since it is often approximately equal to the probability that $X \in A$ when Y is *close* to b.

The interpretation given by equation (L.10) was shown under a continuity assumption. However, the following theorem uses $\mathbf{P}(X \in A \,|\, Y = b)$ in an expression that *always* has a physical meaning, without any assumption about continuity.

Theorem L.6 (Total probability using exact cases). Let X, Y be random variables for some probability model. Let h be a density for the joint distribution of X, Y. Let g be a density for the distribution of Y.

Let A be a subset of $\mathbb{R}$, and let $\mathbf{P}(X \in A \,|\, Y = y)$ be defined for all y as in Definition L.5, using h and g. Then for any subset B of $\mathbb{R}$,

$$\mathbf{P}(\,(X,Y) \in A \times B) = \int_B \mathbf{P}(X \in A \,|\, Y = y) g(y)\, dy. \tag{L.12}$$

Proof.

$$\begin{aligned}\int_B \mathbf{P}(X \in A \,|\, Y = y) g(y)\, dy &= \int_B \left(\int_A \frac{h(x,y)}{g(y)}\, dx \right) g(y)\, dy \\ &= \int_{A \times B} h(x,y)\, dx\, dy = \mathbf{P}(\,(X,Y) \in A \times B).\end{aligned}$$

□

In the proof of Theorem L.6, notice how the arbitrary value we gave to $\mathbf{P}(X \in A \,|\, Y = y)$ when $g(y) = 0$ only occurs in a place where it doesn't matter!

The statement of Theorem L.6 relates $\mathbf{P}(X \in A \,|\, Y = y)$ to a probability which is physically observable in principle, namely $\mathbf{P}(\,(X,Y) \in A \times B)$. We might say that our definition of $\mathbf{P}(X \in A \,|\, Y = y)$ is valid precisely because it produces the correct value for $\mathbf{P}(\,(X,Y) \in A \times B)$.

Remark L.7 (Total probability). Theorem L.6 and Theorem 4.6 both express the law of total probability, in different situations. Equation (L.12) is the "continuous" version of equation (4.18). The role of the event C in equation (4.18) is similar to the role of $\{X \in A\}$ in equation (L.12), while the role of D in equation (4.18) is similar to the role of $\{Y \in B\}$. The sum over the events D_i is replaced by the integral over the "infinitesimal events" $\{Y = y\}$.

Equations (L.10) and (L.10) suggest one more definition.

Definition L.8 (Conditional density). Let X, Y be random variables for some probability model, and let h be a density for the joint distribution of X, Y. Let g be a density for the distribution of Y. For any value $y \in \mathbb{R}$ with $g(y) > 0$, let the conditional density $f_X(x \,|\, Y = y)$ for X given $Y = y$ be defined by

$$f_X(x \,|\, Y = y) = \frac{h(x, y)}{g(y)}. \tag{L.13}$$

For convenience, if $g(y) = 0$ let $f_X(x \,|\, Y = y)$ be defined to be zero.

We may write $f_X(x \,|\, Y = y)$ more briefly as $f(x \,|\, Y = y)$, when the random variable X is known from the context.

We have thought of a density as something that you integrate to obtain a probability. Our definition of the conditional density is consistent with this view, since by equation (L.10),

$$\mathbf{P}(X \in A \,|\, Y = y) = \int_A f(x \,|\, Y = y)\, dx. \tag{L.14}$$

Recall that any function h which produces the correct values for $\mathbf{P}(\,(X, Y) \in S\,)$ is an allowable density for the joint distribution. Thus there are many correct choices for h, and hence also for g. Is this a problem?

Not really. Notice that although our definition of $f(x \,|\, Y = y)$ depends on the choice of the densities h and g, Theorem L.6 shows that if we use $f(x \,|\, Y = y)$ to calculate an observable probability, the result will *not* depend on the choice of h and g. So the non-uniqueness of h is not a problem, as long as we stick to calculating observable probabilities.

L.3 Changing variables

Sometimes the analysis of a problem become much simpler if we define new variables, and re-express the problem in terms of the new variables. Of course, when we do that, we have to convert expressions using the old variables into equivalent expressions using the new variables. And the conversion must be done correctly!

You may encounter probability calculations which involve a two-dimensional change of variables. Just to give a sense of how that works, in this section we will briefly consider a typical case.

Suppose that X, Y are real-valued random variables, such that the range of the random vector (X, Y) is contained in some subset U of $\mathbb{R}^2$. Let h is a probability density for the distribution of the random vector (X, Y).

Let $\varphi : U \to \mathbb{R}^2$ be a map which is defined on U and has values in $\mathbb{R}^2$. Assume that φ is one-to-one. If $(x, y) \in U$ and $\varphi(x, y) = (u, v)$, we can think of (u, v) as new coordinates for the point (x, y).

Let $(U, V) = \varphi(X, Y)$. If φ is one-to-one, then we can write X, Y in terms of U, V. So we can express everything involving X and Y in terms of U, V. Doing that may require a probability density k for the distribution of (U, V). How do we find k?

Assume that φ is one-to-one and onto, has derivatives, and so on. Make it as nice as you like. We just want to get the idea of how to find k from h.

Let S be a subset of $\mathbb{R}^2$. The density k must be such that

$$\mathbf{P}(\varphi(X, Y) \in S) = \int_S k. \tag{L.15}$$

Let T be the set defined by

$$T = \{\, (x, y) : \ \varphi(x, y) \in S \}\,.$$

Then saying that $\varphi(X, Y) \in S$ is the same as saying that $(X, Y) \in T$. That is,

$$\{\varphi(X, Y) \in S\} = \{\, (X, Y) \in T \,\}\,.$$

Hence

$$\mathbf{P}(\varphi(X, Y) \in S) = \mathbf{P}(\, (X, Y) \in T \,) = \int_T h. \tag{L.16}$$

Comparing equation (L.15) with equation (L.16), we see that we need k to be such that

$$\int_S k = \int_T h.$$

This involves an old calculus topic: changing variables in an integral in the plane.

It's easier to think about functions than sets in this situation, so let's use indicators (Definition 11.1) to write everything in terms of functions. We

want to have

$$\int k\mathbf{1}_S = \int h\mathbf{1}_T. \tag{L.17}$$

Assume that the inverse map φ^{-1} exists. Call it θ. The definition of T says that $\varphi(x, y) \in S$ is equivalent to $(x, y) \in T$, so $\theta(u, v) \in T$ is equivalent to $(u, v) \in S$ and so

$$\mathbf{1}_S = \mathbf{1}_T \circ \theta.$$

Thus we want k to be such that

$$\int k\mathbf{1}_T \circ \theta = \int h\mathbf{1}_T. \tag{L.18}$$

The calculus formula for changing variables in a two-dimensional integral says that for any integrand f,

$$\int\int f = \int\int (f \circ \theta)\ |J|\,, \tag{L.19}$$

where J denotes the Jacobian determinant of the map θ.

If $\theta(u, v) = \big(\theta_1(u, v), \theta_2(u, v)\big)$, then the Jacobian determinant J is defined by

$$J = \det\left(\begin{bmatrix} \frac{\partial}{\partial u}\theta_1 & \frac{\partial}{\partial v}\theta_1 \\ \frac{\partial}{\partial u}\theta_2 & \frac{\partial}{\partial v}\theta_2 \end{bmatrix}\right). \tag{L.20}$$

In equation (L.19), the factor $|J|$ plays the role that $|\theta'|$ would play in one dimension.

Applying equation (L.19) with $f = h\mathbf{1}_T$ gives us the following general equation:

$$\int h\mathbf{1}_T = \int (h \circ \theta)(\mathbf{1}_T \circ \theta)\ |J|\,, \tag{L.21}$$

After comparing equation (L.21) with equation (L.18), we see that equation (L.18) will hold if

$$k = (h \circ \theta)\ |J|\,. \tag{L.22}$$

This is the change-of-variable formula that gives a density for the distribution of $\varphi(X, Y)$.

Sometimes there may be a few points where the change-of-coordinates map φ is undefined, or the inverse is not differentiable. (We're looking at you, polar coordinates.) Usually we can just work with φ on the rest of its domain, and still integrate the function in equation (L.22) to get probabilities.

Appendix M

Convolutions of functions on the integers

The formula for the convolution of two functions on the real line was given in equation (K.13). The convolution operation for two functions on the integers has a formula which is similar, but simpler. We can gain some insights by exploring its properties.

M.1 The general definition of convolutions of functions on the integers

Let X be an integer-valued random variable.

Then $\mathbf{P}(X = x) = 0$ if x is not an integer. Let f be the function on the integers defined by $f(n) = \mathbf{P}(X = n)$.

From the definition, f is simply the probability mass function for the distribution of X (Definition 9.8), with its domain restricted to the integers.

As in equation (14.15), f has all the information contained in the distribution of X.

We can picture a function f on the integers as a doubly infinite sequence of values:

$$\ldots, f(-3), f(-2), f(-1), f(0), f(1), f(2), f(3), \ldots.$$

For brevity, we will sometimes refer to any function on the integers as a *sequence function*. Then the function f defined by $f(n) = \mathbf{P}(X = n)$ will be called the sequence function for the distribution of X.

One of the goals of this section is to find $\mathbf{P}(X+Y=n)$ in terms of f and g.

Lemma M.1 (The sequence function for a sum of independent random variables). Suppose that X and Y are independent integer-valued random variables, whose distributions have sequence functions f and g, respectively. Then:

$$\mathbf{P}(X+Y=n) = \sum_{k=-\infty}^{\infty} \mathbf{P}(X=k)\mathbf{P}(Y=n-k) = \sum_{k=-\infty}^{\infty} f(k)g(n-k). \quad \text{(M.1)}$$

Proof. Notice that the events $\{X=k\}$, $-\infty < k < \infty$, cover all possibilities. Hence

$$\{X+Y=n\} = \bigcup_{k=-\infty}^{\infty} \{X+Y=n \text{ and } X=k\}.$$

The events in this union are obviously disjoint, since $X(\omega)$ only has one value for each ω. By countable additivity,

$$\mathbf{P}(X+Y=n) = \sum_{k=-\infty}^{\infty} \mathbf{P}(X+Y=n \text{ and } X=k).$$

Logically the statement $X+Y=n$ and $X=k$ is equivalent to the statement $X=k$ and $Y=n-k$. Hence

$$\mathbf{P}(X+Y=n) = \sum_{k=-\infty}^{\infty} \mathbf{P}(X=k \text{ and } Y=n-k).$$

Since X, Y are independent, $\mathbf{P}(X=k \text{ and } Y=n-k) = \mathbf{P}(X=k)\mathbf{P}(Y=n-k)$, and equation (M.1) follows.

□

We would like to understand the general properties of sums like the ones in equation (M.1). The next lemma tackles some analysis connected with that goal.

Lemma M.2 (The convolution sum). Let α and β be functions defined on the integers, such that

$$\sum_{n=-\infty}^{\infty} |\alpha(n)| \text{ converges and } \sum_{n=-\infty}^{\infty} |\beta(n)| \text{ converges.} \tag{M.2}$$

Then $\sum_{k=-\infty}^{\infty} |\alpha(k)\beta(n-k)|$ converges for each n.

Furthermore, the double series:

$$\sum_{n=-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} |\alpha(k)\beta(n-k)| \right) \tag{M.3}$$

is convergent.

Proof. Replacing α by $|\alpha|$ and β by $|\beta|$ does not change the value of any of the series sums in equations (M.2) and (M.3). So without loss of generality we can assume that the sequences α, β are nonnegative.

Since $\sum_{k=-\infty}^{\infty} \beta(k)$ converges, $\beta(k) \to 0$ as $k \to \pm\infty$. Hence $\beta(k)$ is bounded, i.e. there is some constant c such that $\beta(k) \le c$ for all k.

Hence $\alpha(k)\beta(n-k) \le c\alpha(k)$ for all k. Since $\sum_{k=-\infty}^{\infty} c\alpha(k)$ converges, we know that $\sum_{k=-\infty}^{\infty} \alpha(k)\beta(n-k)$ converges also, by the comparison test.

We also know that a series of nonnegative terms can be rearranged freely without altering its sum. This is also true for a doubly-indexed series. Hence

$$\sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \alpha(k)\beta(j) = \left(\sum_{k=-\infty}^{\infty} \alpha(k) \right) \left(\sum_{j=-\infty}^{\infty} \beta(j) \right),$$

showing that this double series converges.

Also

$$\sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \alpha(k)\beta(j) = \sum_{k=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} \alpha(k)\beta(j) \right).$$

Let $j = n - k$ in the inner summation, for each k. This gives

$$\sum_{k=-\infty}^{\infty} \left(\sum_{n=-\infty}^{\infty} \alpha(k)\beta(n-k) \right).$$

This shows that the double series in equation (M.3) converges.

□

Definition M.3 (The convolution operation for sequence functions). Let α and β be functions defined on the integers. These functions need not be associated with probability distributions.

Suppose that α and β are such that

$$\sum_{n=-\infty}^{\infty} |\alpha(n)| \text{ converges and } \sum_{n=-\infty}^{\infty} |\beta(n)| \text{ converges.}$$

Define a function ψ on the integers by

$$\psi(n) = \sum_{k=-\infty}^{\infty} \alpha(k)\beta(n-k). \tag{M.4}$$

Then ψ is referred to as the convolution of the sequence functions α and β, and is denoted by $\alpha * \beta$.

We defined the $*$ operation for general sequence functions. We can guess some of the properties of the $*$ operation by looking at sequence functions for distributions.

With that goal in mind, let X, Y, Z be independent integer-valued random variables, whose distributions have sequence functions α, β, γ respectively. Thus $\alpha(n) = \mathbf{P}(X = n)$, $\beta(n) = \mathbf{P}(Y = n)$, and $\gamma(n) = \mathbf{P}(Z = n)$.

By equation (M.1), $\alpha * \beta(n) = \mathbf{P}(X + Y = n)$ and $\beta * \alpha(n) = \mathbf{P}(Y + X = n)$. This shows that

$$\alpha * \beta = \beta * \alpha. \tag{M.5}$$

Similarly, $(\alpha * \beta) * \gamma(n) = \mathbf{P}((X + Y) + Z = n)$, and $\alpha * (\beta * \gamma)(n) = \mathbf{P}(X + (Y + Z) = n)$. This shows that

$$(\alpha * \beta) * \gamma = \alpha * (\beta * \gamma). \tag{M.6}$$

Equations (M.5) and (M.6) make us confident that statements (i) and (ii) of the following lemma hold.

Lemma M.4 (Commutative, associative and distributive properties).

(i) The convolution operation on general sequence functions is commutative, i.e. equation (M.5) holds for sequence functions α, β whenever $\sum_{n=-\infty}^{\infty} |\alpha(n)|$ and $\sum_{n=-\infty}^{\infty} |\beta(n)|$ converge.

(ii) The convolution operation on general sequence functions is associative, i.e. equation (M.6) holds for sequence functions α, β, γ whenever $\sum_{n=-\infty}^{\infty} |\alpha(n)|$, $\sum_{n=-\infty}^{\infty} |\beta(n)|$, and $\sum_{n=-\infty}^{\infty} |\gamma(n)|$ converge.

(iii) The distributive law holds for convolution of general sequence functions: for any sequence functions α, β, γ, whenever $\sum_{n=-\infty}^{\infty} |\alpha(n)|$, $\sum_{n=-\infty}^{\infty} |\beta(n)|$, and $\sum_{n=-\infty}^{\infty} |\gamma(n)|$ converge,

$$\alpha * (\beta + \gamma) = \alpha * \beta + \alpha * \gamma.$$

In fact, the convolution operation is bilinear (Definition 16.23):

$$\alpha * (c_1\beta + c_2\gamma) = c_1\alpha * \beta + c_2\alpha * \gamma, \tag{M.7}$$

$$(c_1\alpha + c_2\beta) * \gamma = c_1\alpha * \gamma + c_2\beta * \gamma. \tag{M.8}$$

Of course since equation (M.7) holds for all α, β, γ, and convolution is commutative, equation (M.8) is redundant here.

The proof is much like what we've already seen, and is omitted.

Lemma M.4 can be summarized briefly by saying that we can manipulate expressions involving convolution in much the same way that we manipulate expressions involving multiplication.

M.2 The δ_a function on the integers

For any integer a, let δ_a denote the sequence function defined by

$$\delta_a(n) = \begin{cases} 1 & \text{if } n = a, \\ 0 & \text{otherwise.} \end{cases} \tag{M.9}$$

Readers should be aware that the same δ_a notation is used for other mathematical objects, especially for the "Dirac delta function" located at the point a. The sequence function δ_a defined here is not the same as the Dirac delta function, but it has some similarities, so using the same notation seems appropriate.

Notice that the sequence function δ_a is the sequence function for the distribution of a constant random variable, namely the random variable which is equal to a everywhere.

Exercise M.1 (Convolution with the sequence function δ_a). Prove that for any sequence function f,

$$\delta_a * f(n) = f(n-a). \tag{M.10}$$

[Solution]

Equation (M.10) says that convolving f with δ_a shifts the values of f to the right by a.

As applications of Exercise M.1, we see that

$$\delta_0 * f = f \tag{M.11}$$

for any sequence function f, and also

$$\delta_a * \delta_b = \delta_{a+b}. \tag{M.12}$$

Let's try out these ideas on the binomial distribution with parameters n, p.

Let $X_1, \ldots, X_n$ be independent random variables, with $\mathbf{P}(X_i = 1) = p$ and $\mathbf{P}(X_i = 0) = 1 - p$ for all i. Let $S_n = X_1 + \cdots + X_n$.

The distribution of S_n is known to be binomial with parameters n, p, but suppose we are unaware of that, and wish to find the distribution of S_n.

One can start by noting that the sequence function f_i for the distribution of X_i is very simple: $f_i(1) = p$, $f_i(0) = 1 - p$, and $f_i(n) = 0$ for all other n.

In other words,

$$f_i = p\delta_1 + (1-p)\delta_0. \tag{M.13}$$

Using equation (M.1) and the associative property of convolution, we know that the sequence function g for S_n is given by

$$g = f_1 * \ldots * f_n = (p\delta_1 + (1-p)\delta_0)^{*\,n}, \tag{M.14}$$

where we use the notation $(p\delta_1 + (1-p)\delta_0)^{*\,n}$ to indicate the convolution of n identical factors.

Because the algebra of convolution is so similar to the algebra of multiplication, we can expand the convolution product in equation (M.14) using the binomial theorem. This gives

$$g = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_1^{*k} * \delta_0^{*n-k}.$$

Using equation (M.12),

$$g = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}\ \delta_k. \tag{M.15}$$

Evaluating the right side of this equation at the point j shows at once that

$$g(j) = \begin{cases} \binom{n}{j} p^j (1-p)^{n-j} & \text{for } j = 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

Thus g is the sequence function for the binomial distribution with parameters n, p.

We could also reverse this argument. Suppose we wish to study the binomial distribution without thinking about experiments. Now start with the assumption that g is the sequence function for the binomial distribution with parameters n, p.

The formula for the binomial distribution equation tells us that equation (M.15) holds. The binomial theorem then shows that equation (M.14) also holds.

M.3 Solutions for Appendix M

Solution (Exercise M.1). By definition,

$$\delta_a * f(n) = \sum_{k=-\infty}^{\infty} \delta_a(k) f(n-k).$$

By the definition of δ_a, the only surviving term in the sum on the right is the term with $k = a$. Since $\delta_a(a) = 1$, the result is $f(n-a)$.

Appendix N

Expected values for general models

N.1 Defining general expected values

The first step in defining $\mathbf{E}[X]$ is to approximate X using a random variable Y that we already understand very well. We know how to find $\mathbf{E}[Y]$, and that will give us an approximation for the value of $\mathbf{E}[X]$.

Here's how the first step works for the case of bounded random variables.

Fact N.1 (Approximation of bounded random variables with finite range random variables). For any bounded random variable X on a sample space, and any $\varepsilon > 0$, there exists a finite range random variable Y_ε that approximates X with error at most ε. That is:

$$Y_\varepsilon(\omega) - \varepsilon \leq X(\omega) \leq Y_\varepsilon(\omega) + \varepsilon \text{ for all } \omega. \tag{N.1}$$

Our experience in calculus makes Fact N.1 plausible. A formal proof of Fact N.1 is given in Section N.3 for those who are interested. But the main theoretical point to remember is that the approximations described in Fact N.1 are always available.

Remark N.2. By subtracting Y_ε throughout equation (N.1), we see that equation (N.1) is equivalent to

$$-\varepsilon \leq X(\omega) - Y_\varepsilon(\omega) \leq \varepsilon \text{ for all } \omega. \tag{N.2}$$

Equation (N.1) is equivalent to the statement that $|X(\omega) - Y_\varepsilon(\omega)| \leq \varepsilon$ for all ω.

We haven't given a mathematical definition for $\mathbf{E}[X]$ yet. To get an idea what the definition should be, assume for a moment that we have already defined $\mathbf{E}[X]$. Let's take expectations throughout equation (N.1), and see what happens.

Using monotonicity,

$$\mathbf{E}[Y_\varepsilon - \varepsilon] \leq \mathbf{E}[X] \leq \mathbf{E}[Y_\varepsilon + \varepsilon].$$

Using linearity,

$$\mathbf{E}[Y_\varepsilon] - \mathbf{E}[\varepsilon] \leq \mathbf{E}[X] \leq \mathbf{E}[Y_\varepsilon] + \mathbf{E}[\varepsilon].$$

Using equation (10.4),

$$\mathbf{E}[Y_\varepsilon] - \varepsilon \leq \mathbf{E}[X] \leq \mathbf{E}[Y_\varepsilon] + \varepsilon. \tag{N.3}$$

Equation (N.3) shows that we can find an approximation to the value of $\mathbf{E}[X]$, even if we don't have a formula to calculate $\mathbf{E}[X]$ directly. And the approximation to the value can be made to any degree of accuracy ε.

Of course, we have not yet defined $\mathbf{E}[X]$ in general. So for a general random variable X, it may seem that in equation (N.3) we are trying to approximate something that doesn't exist! But the random variable X exists, and that means that the approximation functions Y_ε for different values of ε must be *close to each other* (see Remark N.3). As a consequence, the numbers $\mathbf{E}[Y_\varepsilon]$ are close to each other, and the closeness is better and better as ε gets smaller. Standard arguments from advanced calculus then show that there must be a single well-defined number $\mathbf{E}[X]$ such that equation (N.3) holds for every $\varepsilon > 0$. And that number is our definition of $\mathbf{E}[X]$.

That's a *theoretical* definition for $\mathbf{E}[X]$, of course. For practical purposes it is unlikely to be a good way to efficiently compute $\mathbf{E}[X]$. But computation is a separate problem. Right now we can just be happy in knowing that $\mathbf{E}[X]$ is well-defined, so there really is something that deserves to be calculated.

Exercise N.1. Show that equation (N.1) is equivalent to

$$X - \varepsilon \leq Y_\varepsilon \leq X + \varepsilon. \tag{N.4}$$

[Solution]

Remark N.3. Consider two approximations using $\varepsilon = \alpha$ and $\varepsilon = \beta$. That is, suppose Y_α is such that

$$Y_\alpha(\omega) - \alpha \leq X(\omega) \leq Y_\alpha(\omega) + \alpha \text{ for all } \omega, \tag{N.5}$$

and

$$Y_\beta(\omega) - \beta \leq X(\omega) \leq Y_\beta(\omega) + \beta \text{ for all } \omega. \tag{N.6}$$

This implies that

$$Y_\beta(\omega) - (\beta + \alpha) \leq Y_\alpha(\omega) \leq Y_\beta(\omega) + (\beta + \alpha) \text{ for all } \omega, \tag{N.7}$$

and so the expected values are close:

$$\mathbf{E}[Y_\beta] - (\beta + \alpha) \leq \mathbf{E}[Y_\alpha] \leq \mathbf{E}[Y_\beta] + (\beta + \alpha). \tag{N.8}$$

Taking α and β smaller and smaller, we have better and better agreement between the approximations.

Exercise N.2. Prove equation (N.7).

When X is an unbounded random variable, we define $\mathbf{E}[X]$ similarly, except that now we must approximate $\mathbf{E}[X]$ using an infinite series, as follows.

Fact N.4 (Approximation by random variables with countable range). For any random variable X on a sample space, and any $\varepsilon > 0$, there exists a random variable Y_ε with countable range, such that

$$Y_\varepsilon(\omega) - \varepsilon \leq X(\omega) \leq Y_\varepsilon(\omega) + \varepsilon \text{ everywhere.} \tag{N.9}$$

When X is unbounded, the approximation Y_ε will also be unbounded. But assuming that $\mathbf{E}[Y_\varepsilon]$ exists, we have the same approximation for $\mathbf{E}[X]$ as before, given in equation (N.3). This leads to a similar definition for general expected values. Of course now $\mathbf{E}[Y_\varepsilon]$ may be the sum of an infinite series. Whether or not $\mathbf{E}[Y_\varepsilon]$ exists depends on the size of $|Y_\varepsilon|$.

Note that equation (N.4) continues to hold in this general case, and it shows that the size of Y_ε as a function on the sample space is essentially the same as the size of X. So if $\mathbf{E}[Y_\varepsilon]$ exists using Y_ε for *one* value of ε, then $\mathbf{E}[Y_\varepsilon]$ for *every* value of ε. It just depends on the size of $|X|$ as a function on the sample space.

Since the four key properties hold for countable range random variables (Theorem 14.9), one can show that these properties remain true for expectations of general random variables. This was asserted in Theorem 15.2.

Remark N.5 (Justifying Remark 15.3). Suppose that a bounded random variable X is intended to model a measured value in an experiment, and X does not have finite range. This suggests that we think that X represents a continuous physical quantity, or at least that there are a large number of possible values for the measurement, perhaps located throughout some interval.

Let Y be a finite range approximation for X, with an error at most ε. Does Y represent a physical random variable?

It is shown in Lemma N.10 of Section N.1 that the value of the approximation Y constructed in that lemma is determined by the value of X. So if X represents a physically measured quantity, then Y does also.

Assume that ε is smaller than the experimental error that we expect in our measurements. Then Y and X actually represent *the same* physical measurement. The difference between X and Y is just a matter of mathematical convenience!

And so $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ share the same frequency interpretation, justifying Remark 15.3.

For a general random variable, the distribution is not mentioned explicitly in the definition of expected value, but it lurks just below the surface, since for a random variable Y with finite or countable range, $\mathbf{E}[Y]$ is defined in terms of the distribution of Y (Definition 10.2 and Definition 14.6). And we

define $\mathbf{E}[X]$ by approximating X with such random variables Y. For the random variables constructed in Lemma N.10, it is not hard to show that the the distribution of Y is determined by the distribution of X. Thus $\mathbf{E}[Y]$ is determined by the distribution of X. $\mathbf{E}[Y]$ approximates $\mathbf{E}[X]$, so $\mathbf{E}[X]$ is determined too. We record this fact next.

Fact N.6 (Expectation determined by distribution). Let X be a random variable for some probability model and let Z be a random variable for some probability model. The model for Z need not be the same as the model for X.

Suppose that the distribution of X is the same as the distribution of Z.

Then X and Z have the same expected value.

Recall the notation of Definition 9.7. For random variables X and Y which have the same distribution, we can write $X \sim Y$.

With this notation, we can state Fact N.6 as

$$X \sim Z \implies \mathbf{E}[X] = \mathbf{E}[Z]. \tag{N.10}$$

This fact was also asserted in Theorem 15.2.

Since the distribution of X determines $\mathbf{E}[X]$, we may at times speak of $\mathbf{E}[X]$ as the "mean of the distribution of X".

N.2 Expected value as an integral

The general definition of expected value that was sketched in Section N.1 may not give us an efficient method of calculation. But for models that have a probability density, the machinery of integration is available.

Theorem N.7 (Expected values for models with densities). Consider a model in which the probability $\mathbf{P}(A)$ of an event A is given by equation (15.3).

For any random variable X on Ω, the expected value $\mathbf{E}[X]$ is given by

$$\mathbf{E}[X] = \int_{\Omega} Xf. \tag{N.11}$$

This equation holds in the sense that if either side of the equation is defined, then both sides are defined and they are equal.

Equation (N.11) was already stated in equation (15.4).

Example N.8 (The finite range case). Let's check equation (N.11) when X is a finite-range random variable.

Let $x_1, \ldots, x_n$ be the distinct numbers in the range of X. Let $A_i = \{X = x_i\}$. By definition,

$$\mathbf{E}[X] = x_1\mathbf{P}(A_1) + \ldots + x_n\mathbf{P}(A_n).$$

By the definition of a probability density,

$$\mathbf{P}(A_i) = \int_{A_i} f.$$

By equation (11.15),

$$X = x_1\mathbf{1}_{A_1} + \ldots + x_n\mathbf{1}_{A_n}.$$

Hence

$$\begin{aligned}\int Xf &= \int (x_1\mathbf{1}_{A_1}f + \ldots + x_n\mathbf{1}_{A_n})\, f \\ &= x_1\int \mathbf{1}_{A_1}f + \ldots + x_n\int \mathbf{1}_{A_n}f \\ &= x_1\int A_1 f + \ldots + x_n\int A_n f \\ &= x_1\mathbf{P}(A_1) + \ldots + x_n\mathbf{P}(A_n).\end{aligned}$$

This shows that equation (N.11) holds.

Example N.8 shows that equation (N.11) holds when X has finite range. When X is any bounded random variable, we can approximate X as closely as we like using a finite range random variable Y. Equation (N.11) holds with X replaced by Y. And when two random variables are close, so are their expected values **and** so are their integrals. So equation (N.11) holding for all Y easily implies equation (N.11) for all unbounded X.

The argument when X is unbounded is simiiar.

N.3 Approximation of random variables

This section gives a precise recipe for approximating general random variables by random variables with finite or countable range.

Definition N.9 (Uniform approximations). Let X and Y be functions on some set Ω (X and Y may be random variables but the definition applies to any functions.) If c is a number such that $|X - Y| \leq c$ at every point of Ω, we will say that Y approximates X uniformly on Ω to within c. The number c measure the closeness of the approximation.

Lemma N.10 (Simple approximations exist). For any real-valued random variable X and any given $\varepsilon > 0$, there exists a random variable Y with countable range that approximates X to within ε, meaning that $|X - Y| \leq \varepsilon$ holds everywhere on the sample space.

The approximating random variable Y can be chosen such that $X \leq Y$ holds everywhere, and it can also be chosen such that $Y \leq X$ holds everywhere.

The value of Y is determined by the value of X.

If X is bounded the approximating random variable Y can be chosen to have finite range.

Proof. Let X be any real-valued random variable and let $\varepsilon > 0$ be a given. For each integer k, let $A_k = \{(k-1)\varepsilon < X \leq k\varepsilon\}$, and define the random variable Y by $Y(\omega) = k\varepsilon$ for $\omega \in A_k$. See Figure N.1.

Notice that ω is a member of A_k if and only if the value of X lies in the interval $(\,(k-1)\varepsilon, k\varepsilon\,]$. Thus the value of Y is determined by the value of X.

Since the range of Y is contained in the set $\{k\varepsilon : \ k \text{ an integer}\}$, Y has a countable range.

By construction, $(k-1)\varepsilon < X \leq k\varepsilon = Y$ on A_k. Thus $Y - \varepsilon < X \leq Y$ holds everywhere.

If X is bounded, then $c \leq X \leq d$ holds for some number c, d. Then A_k will only be nonempty if $k\varepsilon \geq c$ and $(k-1)\varepsilon < d$. That is, A_k will only be nonempty if

$$\frac{c}{\varepsilon} \leq k < \frac{d}{\varepsilon} + 1.$$

This shows that Y will have a finite range if X is bounded.

□

It should be noted that we could easily have made different choices in defining the approximation Y in the proof of Lemma N.10. For example, we could have defined Y by $Y(\omega) = (k-1)\varepsilon$ for $\omega \in A_k$. In that case we would have had $Y < X \le Y + \varepsilon$ everywhere.

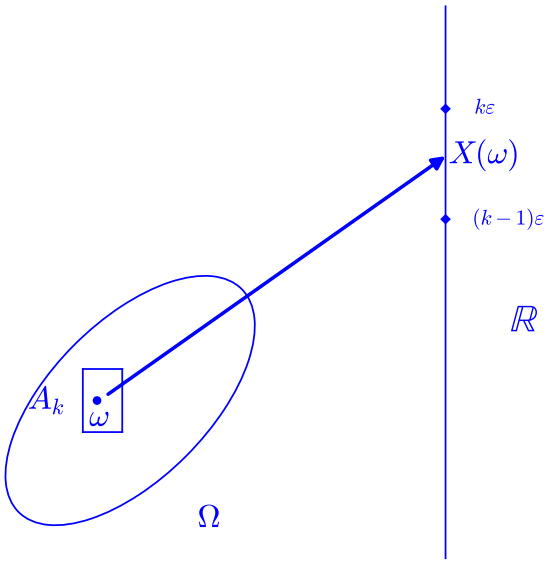


Figure N.1: $A_k = \{(k-1)\,\varepsilon < X \le k\,\varepsilon\}$.

Example N.11 (Approximating a continuous random variable). Consider a probability model with sample space Ω equal to the interval $[0, 3]$. We can take $\mathbf{P}$ to be the uniform probability distribution on $[0, 3]$, just to have a definite probability in mind, but this does not affect the construction of an approximation to a random variable.

Suppose that for some reason we want to study the random variable X on Ω, defined by $X(\omega) = (1-\omega)^2$. See Figure N.2.

Suppose $\varepsilon = .5$. As in the proof of Lemma N.10, we will divide the possible values for X into intervals $((n-1)\varepsilon, n\varepsilon]$, and $\{Y = i\,\varepsilon\}$ is the set of all ω such that $(n-1)\varepsilon < X \le n\varepsilon$.

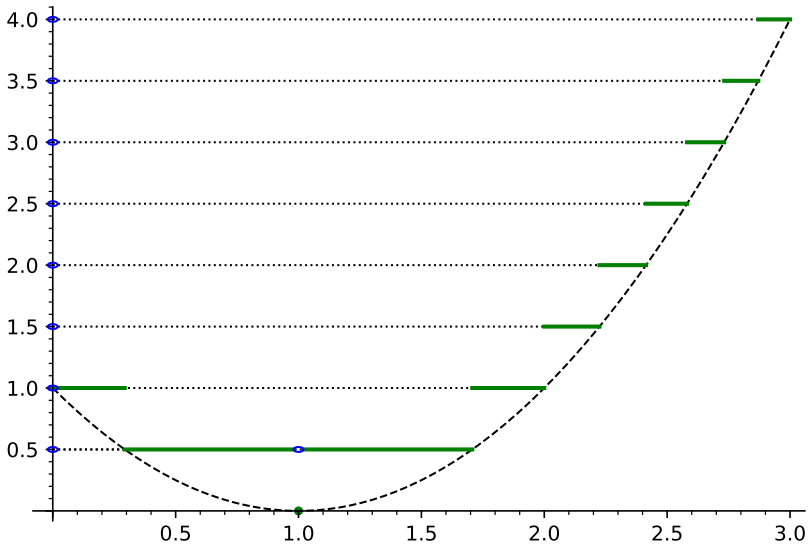


Figure N.2: The dashed graph shows $X(\omega) = (1-\omega)^2$ on the sample space $[0, 3]$. $\varepsilon = .5$. The graph of the random variable Y is shown in green.

The function Y defined in the proof of Lemma N.10 is the piecewise-constant function shown in green in Figure N.2. Notice that $X \leq Y$ holds everywhere, in accordance with the first construction for Y given in Lemma N.10.

In Figure N.2, notice that the events $\{Y = i\,\varepsilon\}$, $i = 0, 1, \ldots, 8$ are of different sizes. Also, $\{Y = .5\}$ and $\{Y = 1\}$ are each equal to the union of two disjoint intervals, while $\{Y = 0\}$ is a one-point set. In general, the sets A_n in the proof of Lemma N.10 could be much more complicated than that. The probability of each A_n is always defined, however, and that is what matters for defining expected value.

It may be instructive to compare the approximation in Example N.11 with usual calculus-style approximation by step functions. The calculus approximation to a function uses step-functions which are constructed by partitioning the domain of the function into subintervals. In contrast to that, the construction in Example N.11 partitions the *range* of the function into subintervals, rather than the domain. A graph of the calculus-style approximation is given in Figure N.3.

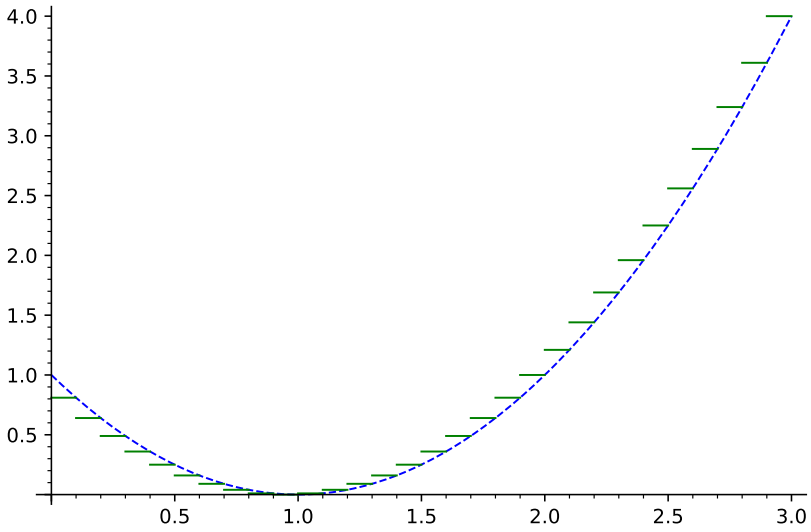


Figure N.3: $X(t) = (1-t)^2$ on the sample space $[0, 3]$. A typical step-function approximation Y in calculus is shown in green.

N.4 Solutions for Appendix N

Solution (Exercise N.1). The second inequality in equation (N.4) is equivalent to the first inequality in equation (N.9).

The first inequality in equation (N.4) is equivalent to the second inequality in equation (N.9).

Solution (Exercise N.2). Equation (N.4) with $\varepsilon. = \alpha$ tells us that

$$Y_\alpha \leq X + \alpha.$$

Equation (N.6) tells us that

$$X \leq Y_\beta + \beta.$$

Combining these inequalities,

$$Y_\alpha \leq Y_\beta + (\beta + \alpha).$$

A similar argument shows that

$$Y_\alpha \geq Y_\beta - (\beta + \alpha).$$

Appendix O

The Schwarz inequality

We have seen various inequalities for expected values, including the Markov Inequality (Lemma 12.15) and the Chebyshev Inequality (Lemma 16.11). Here we study another useful inequality, the Schwarz inequality, in the context of random variables.

Although we will call this inequality the Schwarz inequality, readers should always keep in mind that names are not a reliable guide to priority. A mathematical fact is not always named after the person who first discovered it. The names "Cauchy-Schwarz" and "Cauchy-Bunyakovsky-Schwarz" are also used for this inequality. Perhaps Schwarz benefits by having a short name, and Cauchy has too many other famous results.

O.1 The Schwarz inequality for random variables

The Schwarz inequality applies to inner products of vectors as well as to random variables. Here we will only discuss random variables, but the methods work in general.

Lemma O.1 (The Schwarz inequality for expected values). Let X and Y be real-valued random variables on some sample space. Suppose that $\mathbf{E}[X^2]$ and $\mathbf{E}[Y^2]$ are defined. Then $\mathbf{E}[XY]$ is defined, and

$$|\mathbf{E}[XY]| \leq \sqrt{\mathbf{E}[X^2]}\sqrt{\mathbf{E}[Y^2]}. \tag{O.1}$$

Proof. By equation (16.45), $|XY| \leq \frac{1}{2}(X^2 + Y^2)$. Hence the comparison principle for expected values (Fact 14.8) tells us that $\mathbf{E}[XY]$ exists.

To prove the inequality in equation (O.1) we start with a very simple fact: if a random variable is nonnegative, then its expected value has to be nonnegative also.

Hence for any real number t,

$$\mathbf{E}\left[(X - tY)^2\right] \geq 0. \tag{O.2}$$

That is,

$$\mathbf{E}\left[X^2\right] - 2t\mathbf{E}[XY] + t^2\mathbf{E}\left[Y^2\right] \geq 0. \tag{O.3}$$

If $\mathbf{E}[Y^2] = 0$ then our inequality says that

$$\mathbf{E}\left[X^2\right] \geq 2t\mathbf{E}[XY]$$

for every t. Since we are allowed to take both positive and negative values for t, this inequality could not possibly hold for all t unless $\mathbf{E}[XY] = 0$. So when $\mathbf{E}[Y^2] = 0$ it must be true that $\mathbf{E}[XY] = 0$.

And whenever $\mathbf{E}[XY] = 0$, equation (O.1) obviously holds!

So we have proved equation (O.1) for the case that $\mathbf{E}[Y^2] = 0$.

From now on we assume that $\mathbf{E}[Y^2] > 0$. One might think that this will be a harder case. But simply substituting $t = \mathbf{E}[XY]/\mathbf{E}[Y^2]$ into equation (O.3) gives an inequality which is equivalent to equation (O.1). □

How did we think of substituting $t = \mathbf{E}[XY]/\mathbf{E}[Y^2]$ to obtain Schwarz? We can motivate using this value for t by noting that it makes the inequality in equation (O.3) work as hard as possible! In other words, $t = \mathbf{E}[XY]/\mathbf{E}[Y^2]$ is the choice of t that minimizes $\mathbf{E}[X] - 2t\mathbf{E}[XY] + t^2\mathbf{E}[Y^2]$. You can check that using calculus, or by completing the square.

But we don't need to know ahead of time that $t = \mathbf{E}[XY]/\mathbf{E}[Y^2]$ is the choice of t that minimizes $\mathbf{E}[X] - 2t\mathbf{E}[XY] + t^2\mathbf{E}[Y^2]$. We could just make a wild guess, and somehow choose to substitute this particular number for t, as an experiment. If it gives us a nice inequality, we won't complain.

After all, we are allowed to substitute any value for t in equation (O.2), and it's not our fault if we have psychic powers.

Exercise O.1. By choosing a suitable random variable Y, use the Schwarz inequality to prove that $|\mathbf{E}[X]| \leq \sqrt{\mathbf{E}[X^2]}$.

This fact was already established in equation (16.3) using equation (16.2). [Solution]

Remark O.2. Notice that the right side of equation (O.1) is something that can be calculated using X and Y *separately*. The left side requires looking at X and Y together.

The version of the Schwarz inequality for geometrical vectors says that

$$\left|\vec{\mathbf{a}} \bullet \vec{\mathbf{b}}\right| \leq \|\vec{\mathbf{a}}\, \vec{\mathbf{b}}\|, \tag{O.4}$$

where $\vec{\mathbf{a}} \bullet \vec{\mathbf{b}}$ is the inner product of the geometrical vectors $\vec{\mathbf{a}}, \vec{\mathbf{b}}$, and $\|\vec{\mathbf{a}}\|, \|\vec{\mathbf{b}}\|$ are their lengths.

If we accept that $\vec{\mathbf{a}} \bullet \vec{\mathbf{b}}$ is equal to $\|\vec{\mathbf{a}}\|\, \|\vec{\mathbf{b}}\| \cos\theta$, where θ is the angle between the vectors, then the Schwarz inequality in this case simply says. that $|\cos\theta| \leq 1$.

Remark O.3 (The equality condition for Schwarz in the random variable case). We derived equation (O.1) by substituting a value for t into equation (O.3). That equation is simply the expanded form of equation (O.2). Thus the Schwarz inequality becomes an equality whenever equality holds in equation (O.2). And equality holds in equation (O.1) exactly when there is some t such that

$$\mathbf{E}\left[(X - tY)^2\right] = 0. \tag{O.5}$$

So that is the equality condition.

As noted in Appendix G, a nonnegative random variable has expectation zero if and only if the random variable is equal to zero with probability one. Thus the equality condition holds exactly when there is some t such that

$$\mathbf{P}(X \neq tY) = 0. \tag{O.6}$$

Thus equality holds when for some t, X is essentially equal to tY.

O.2 Solutions for Appendix O

Solution (Exercise O.1). In the Schwarz inequality, let $Y = 1$. Then Schwarz says that

$$|\,\mathbf{E}\,[X \cdot 1]\,| \leq \sqrt{\mathbf{E}\,[X^2]}\,\sqrt{\mathbf{E}\,[1^2]} = \sqrt{\mathbf{E}\,[X^2]},$$

as claimed.

Bibliography

[1] Cox, Richard T., *The Algebra of Probable Inference*, Johns Hopkins, Baltimore, 1961.

[2] Feller, William, *An introduction to Probability Theory and Its Applications, Vol I*, third edition, Wiley, New York 1968.

[3] Feller, William, *An introduction to Probability Theory and Its Applications, Vol II*, Wiley, New York 1966.

[4] Fowler, Michael, *Brownian Motion*, `http://galileo.phys.virginia.edu/classes/152.mf1i.spring02/BrownianMotion.htm`

[5] Jaynes, Edwin.T. *Probability Theory in Science and Engineering*, Colloquium Lectures in Pure and Applied Science, No. 4, Socony Mobil Oil Company, 1958.

[6] Lavenda, Bernard H., Brownian Motion, Scientific American **253** (1985), pp. 70-85, or `https://www.jstor.org/stable/24967570?seq=1`

[7] Matson, John, *Infinity Comes in Different Sizes* , `https://www.scientificamerican.com/article/strange-but-true-infinity-comes-in-different-sizes/`

[8] Rogers, Simon, *Data are or data is?*, `https://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular`

[9] Rosenhouse, Jason, *The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brain Teaser*, Oxford, 2009.

[10] Sanderson, Grant, *But what is the Central Limit Theorem?*, `https://www.youtube.com/watch?v=zeJD6dqJ5lo&t=143s`.

[11] Scheurer, Victor, *Convicted on Statistics?*, https://understandinguncertainty.org/node/545.

[12] Susanka, Lawrence, *The Monty Hall Problem*, https://susanka.org/Notes/montyhall.pdf

Index