Chapter 1

Combinatorics

This file contains the first three chapters (plus some appendices) of a potential book on Probability and Statistics. It does not assume knowledge of calculus. The first three chapters are titled "Combinatorics," "Probability," and "Distributions." And Appendix B gives a nice little introduction to the natural logarithm, e. Future chapters on statistics will be added in the summer of 2010.

Combinatorics is the study of how to count things. By "things" we mean the various combinations, permutations, subgroups, etc., that can be formed from a given set of objects or events. For example, how many different committees of three people can be chosen from five people? How many different full-house hands are there in poker? How many different outcomes are possible if you flip a coin four times? Knowing how to count such things is critical for an understanding of probability, because when calculating the probability of a given event, we invariably need to count the number of ways that this event can happen.

The outline of this chapter is as follows. In Section 1.1 we introduce the concept of *factorials*, which will be ubiquitous in our study of probability. In Section 1.2 we learn how to count the number of possible permutations (that is, the number of possible orderings) of a set of objects. In Section 1.3 we learn how to count the number of possible subgroups that can be formed from a set of objects. We consider both the case where the order of the objects matters, and the case where it doesn't matter. For example, the poker question posed above is one where the order of the objects (the cards) doesn't matter. In Section 1.4 we learn how to count the number of possible outcomes of a repeated experiment, where each repetition has an identical set of possible results. Examples include rolling dice or flipping coins. Finally, in Section 1.5 we look at the coin-flipping example in more detail, and we see how it relates to a set of numbers called the *binomial coefficients*.

Having learned how to count all these things, we'll then see in Chapter 2 how the results can be used to calculate probabilities. It turns out that it's generally a trivial step to obtain a probability once you've counted the relevant things, so the bulk of the work we'll need to do will be in the present chapter.

1.1 Factorials

Before getting into the discussion of actual combinatorics, we'll first need to look at a certain quantity that comes up again and again. This quantity is called the *factorial*. We'll see throughout this chapter that when dealing with a situation that involves an integer N, we often need to consider the product of the first N integers. This product is called "N

factorial," and it is denoted by the shorthand notation, "N!".[1] For the first few integers, we have:

$$\begin{aligned} 1! &= 1 \\ 2! &= 1 \cdot 2 = 2 \\ 3! &= 1 \cdot 2 \cdot 3 = 6 \\ 4! &= 1 \cdot 2 \cdot 3 \cdot 4 = 24 \\ 5! &= 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120 \\ 6! &= 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720 \end{aligned} \tag{1.1}$$

As N increases, $N!$ gets very big very fast. For example, $10! = 3,628,800$, and $20! \approx 2.43 \cdot 10^{18}$. In Chapter 3 we'll make good use of an approximate formula for $N!$, called *Stirling's formula*. This formula will make it clear what we mean by the statement, "$N!$ gets very big very fast."

We should add that 0! is defined to be 1. Of course, 0! doesn't make much sense, because when we talk about the product of the first N integers, it is understood that we start with 1. Since 0 is below this starting point, it is unclear what 0! actually means. However, there's no need to think too hard about trying to make sense out of it, because as we'll see below, if we simply define 0! to be 1, then a number of formulas turn out to be very nice.

Having defined $N!$, we can now start counting things. There are two main things we'll need to know how to count, and the results for both of these involve $N!$. These two things are (1) the *permutations* (the orderings) of N objects , and (2) the number of ways of choosing subgroups from N objects, for example, the number of different committees of three people that can be chosen from five people. Let's look at each of these in turn.

1.2 Permutations

A *permutation* of a set of objects is a way of ordering them. For example, if we have three people, Alice, Bob, and Carol, then one permutation of them is Alice, Bob, Carol. Another permutation is Carol, Alice, Bob. And another is Bob, Alice, Carol. The goal of this section is to learn how to count the number of possible permutations. We'll do this by starting off with the very simple case where we have just one object. Then we'll consider two objects, then three, and so on, until we see a pattern. As we'll find throughout this book, this is invariably a good strategy when trying to figure something out: Start with small numbers, and then gradually increase until you see a pattern.

One object

If we have only one object, then there is clearly only one way to "order" it. There is no ordering to be done. A list of one object simply consists of that one object, and that's that. If we use the notation where P_N stands for the number of permutations of N objects, then we have $P_1 = 1$.

Two objects

With two objects, things aren't completely trivial like they were in the one-object case, but they're still very simple. If we label our two objects as "1" and "2," then we can order them in two ways:

[1] I'm not sure why someone long ago picked the exclamation point for this notation. But just remember that it has nothing to do with the more common grammatical use of the exclamation point for emphasis. So try not to get too excited when you see "N!"!

1 2 or 2 1

So we have $P_2 = 2$. At this point, you might be thinking that this result, along with the previous $P_1 = 1$ result, implies that $P_N = N$ for any number N. This would imply that there should be three different ways to order three objects. Well, not so fast...

Three objects

Things get more interesting with three objects. If we call them "1," "2," and "3," then we can list out the possible orderings. (If you haven't already looked at the table below, you should cover it up with your hand and try to list out all the permutations yourself. We'll even add on this extra sentence here to make this parenthetical remark a little longer, so that you don't have any excuse for saying that you already looked at it!) The permutations are:

1 2 3	2 1 3	3 1 2
1 3 2	2 3 1	3 2 1

Table 1.1

So we have $P_3 = 6$. Note that we've grouped these six permutations into three subgroups (the three columns), according to which number comes first. It isn't necessary to group them this way, but we'll see below that this method of organization has definite advantages. It will simplify how we think about the case where the number of objects is a general number N.

REMARK: There's no need to use the numbers 1,2,3 to represent the three objects. You can use whatever symbols you want. For example, the letters A,B,C work fine, as do the letters H,Q,Z. You can even use symbols like $\otimes, \spadesuit, \heartsuit$. Or you can mix things up with $\odot$,W,7 if you want to be unconventional. The point is that the numbers/letters/symbols/whatever simply stand for three different things, and they need not have any meaningful properties except for their different appearances when you write them down on the paper.

However, there is certainly something simple about the numbers 1,2,3,..., or the letters A,B,C,..., so we'll generally work with these. In any event, it's invariably a good idea to be as economical as possible and not write down the full names, such as Alice, Bob, and Carol. Of course, with these three particular names, there's some logic in going with A,B,C. $\clubsuit$

Four objects

The pattern so far is $P_1 = 1$, $P_2 = 2$, and $P_3 = 6$. Although you might be able to guess the general rule from these three results, it will be easier to see the pattern if we look at the next case with four objects. Taking a cue from the above list of six permutations of three objects, let's organize the permutations of four object according to which number starts the list. (Again, you should cover up the following table with your hand and try to list out all the permutations yourself.) We end up with:

1 2 3 4	2 1 3 4	3 1 2 4	4 1 2 3
1 2 4 3	2 1 4 3	3 1 4 2	4 1 3 2
1 3 2 4	2 3 1 4	3 2 1 4	4 2 1 3
1 3 4 2	2 3 4 1	3 2 4 1	4 2 3 1
1 4 2 3	2 4 1 3	3 4 1 2	4 3 1 2
1 4 3 2	2 4 3 1	3 4 2 1	4 3 2 1

Table 1.2

If we look at the last column, where all the permutations start with "4," we see that if we strip off the "4," we're simply left with the six permutations of the three numbers 1,2,3 that we listed above. A similar thing happens with the column of permutations that start with "3." If we strip off the "3," we're left with the six permutations of the numbers 1,2,4. Likewise for the columns of permutations that start with "2" or "1." The 24 permutations listed above can therefore be thought of as four groups (the four columns), each consisting of six permutations.

Five objects

For five objects, you probably don't want to write down all the permutations, because it turns out that there are 120 of them. But you can *imagine* writing them all down. And for the present purposes, that's just as good as (or perhaps even better than) actually writing them down for real.

Consider the permutations of 1,2,3,4,5 that start with "1." From the above result for the $N = 4$ case, the next four numbers 2,3,4,5 can be permuted in 24 ways. So there are 24 permutations that start with "1." Likewise, there are 24 permutations that start with "2." And similarly for 3, 4, and 5. So we have five groups (columns, if you want to imagine writing them that way), each consisting of 24 permutations. The total number of permutations of five objects is therefore $5 \cdot 24 = 120$.

General case of N objects

Putting all the above results together, we have

$$P_1 = 1, \qquad P_2 = 2, \qquad P_3 = 6, \qquad P_4 = 24, \qquad P_5 = 120. \tag{1.2}$$

Do these numbers look familiar? Yes indeed, they are simply the $N!$ results from Eq. (1.1)! Does this equivalence make sense? Yes, due to the following reasoning.

- $P_1 = 1$, of course.
- $P_2 = 2$, which can be written in the suggestive form, $P_2 = 2 \cdot 1$.
- For P_3, Table 1.1 shows that $P_3 = 6$ can be thought of as three groups (characterized by which number appears first) of the $P_2 = 2$ permutations of the second and third numbers. So we have $P_3 = 3P_2 = 3 \cdot 2 \cdot 1$.
- Similarly, for P_4, Table 1.2 shows that $P_4 = 24$ can be thought of as four groups (characterized by which number appears first) of the $P_3 = 6$ permutations of the second, third, and fourth numbers. So we have $P_4 = 4P_3 = 4 \cdot 3 \cdot 2 \cdot 1$.
- And likewise, the above reasoning for $N = 5$ shows that $P_5 = 5P_4 = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. And so on and so forth. Therefore:
- At each stage, we have $P_N = N \cdot P_{N-1}$. This relation is easily seen to be satisfied by the general formula,

  $$P_N = N!. \tag{1.3}$$

  Basically, you just need to tack on a factor of N at each stage, due to the fact that the permutations can start with any of the N numbers (or whatever objects you're dealing with). The number of permutations of N objects is therefore $N!$.

The strategy of assigning seats

An equivalent way of thinking about this result is the following. For concreteness, let's say that we have four people, Alice, Bob, Carol, and Dave. And let's assume that they need to be assigned to four seats arranged in a line. The above $P_N = N!$ results tells us that there are $4! = 24$ different permutations (orderings) they can take. We'll now give an alternate derivation that shows how these 24 orderings can easily be understood by imagining the seats being filled one at a time. We'll get a lot of mileage out of this type of "seat filling" argument throughout this (and also the next) chapter.

- There are four possibilities for who is assigned to the first seat.
- For each of these four possibilities, there are three possibilities for who is assigned to the second seat (because we've already assigned one person, so there are only three people left). So there are $4 \cdot 3 = 12$ possibilities for how the inhabitants of the first two seats are chosen.
- For each of these 12 possibilities, there are two possibilities for who is assigned to the third seat (because there are only two people left). So there are $4 \cdot 3 \cdot 2 = 24$ possibilities for how the inhabitants of the first three seats are chosen.
- Finally, for each of these 24 possibilities, there is only one possibility for who is assigned to the fourth seat (because there is only one person left, so we're stuck with him/her). So there are $4 \cdot 3 \cdot 2 \cdot 1 = 24$ possibilities for how the inhabitants of all four seats are chosen. The "1" here doesn't matter, of course; it just makes the formula look nicer.

You can see how this counting works for the $N = 4$ case in Table 1.2. There are four possibilities for the first entry, which stands for the person assigned to the first seat if we label the people by 1,2,3,4. Once we pick the first entry, there are three possibilities for the second entry. And once we pick the second entry, there are two possibilities for the third entry. And finally, once we pick the third entry, there is only one possibility for the fourth entry. You can verify all these statements by looking at the table.

It should be emphasized that when dealing with situations that involve statements such as, "There are a possibilities for Event 1, and for each of these there are b possibilities for Event 2, and for each of these there are c possibilities for Event 3, and so on...," the total number of different scenarios when all the events are listed together is the *product* (not the sum!) of the different numbers of possibilities, that is, $a \cdot b \cdot c \cdots$. You should stare at Table 1.2 until you're comfortable with this.

1.3 Choosing subgroups

1.3.1 Choosing pairs

In addition to permutations, the other main thing we'll need to know how to count is the number of different subgroups of a given size, chosen from a given set of objects. For example, let's say we have five people in a room and we need to pick two of them to be on a committee. How many different pairs can we pick? (Note that the order within the pair doesn't matter.) We'll present three ways of answering this question.

First method: If we label the five people as A,B,C,D,E, we can simply list out all the possible pairs. And we can group them in the following suggestive way (remember that the order doesn't matter, so once we've listed, say, AB, we don't need to list BA):

A B			
A C	B C		
A D	B D	C D	
A E	B E	C E	D E

Table 1.3

So there are 10 possible pairs. This table also quickly tells us how many pairs there are if we have four people in all, instead of five. We simply have to remove the bottom row, since the fifth person (E) doesn't exist. We therefore have six pairs. Similarly, if we also remove the "D" row, we see that three people yield three pairs. And then two people of course yield just one pair.

We can also go in the other direction and increase the number of people. With six people in the room, we simply need to add an "F" row to the table, consisting of AF, BF, CF, DF, EF. This adds on another five pairs, so six people yield 15 pairs. In general, if we let $\binom{N}{2}$ denote the number of possible pairs (that's the "2") that can be chosen from N people,[2] then by considering the above table to be the collection of rows with increasing length (1, then 2, then 3, then 4, etc.), we find

$$\begin{aligned}
\binom{2}{2} &= 1, \\
\binom{3}{2} &= 1+2=3, \\
\binom{4}{2} &= 1+2+3=6, \\
\binom{5}{2} &= 1+2+3+4=10, \\
\binom{6}{2} &= 1+2+3+4+5=15, \\
\binom{7}{2} &= 1+2+3+4+5+6=21.
\end{aligned} \tag{1.4}$$

The number of possible pairs among N people is therefore the sum of the first $N-1$ integers. It would be nice if there were a general formula for this sum, so we wouldn't have to actually add up all the numbers. It would be a huge pain to determine $\binom{100}{2}$ this way. And indeed, there is a general formula, and it happens to be

$$\boxed{\binom{N}{2} = \frac{N(N-1)}{2}} \tag{1.5}$$

You can verify that this is consistent with the above list of $\binom{N}{2}$ values. For example, $\binom{7}{2} = 7 \cdot 6/2 = 21$.

REMARK: If you're wondering how to prove that the sum of the first $N-1$ integers equals $N(N-1)/2$, we'll end up effectively deriving this in the second and third methods below. But for now we'll just relate a story about the mathematician Carl Friedrich Gauss. One day in grade school (or so the story goes), his teacher tried to quiet the students by giving them the task of adding up the numbers 1 through 100. But to the teacher's amazement, after a few seconds Gauss

[2]This $\binom{N}{2}$ is called a *binomial coefficient*. It is read as "N choose 2." We'll talk about binomial coefficients in detail in Section 1.5.

came up with the correct answer, 5050. How did he arrive at this so quickly? Well, he wrote out the numbers in increasing order, and then below these he listed them out in decreasing order:

$$\begin{array}{ccccccc} 1 & 2 & 3 & \cdots & 98 & 99 & 100 \\ 100 & 99 & 98 & \cdots & 3 & 2 & 1 \end{array}$$

He then noted that every column of two numbers has the same sum, 101. And since there are 100 columns, the total sum is 10100. But he counted every number twice, so the sum of the numbers 1 through 100 is half of 10100, or 5050. As we saw with the triangle in Table 1.3, and as we'll see many more times, things become much clearer if you group objects in certain ways! ♣

SECOND METHOD: Given the letters A,B,C,D,E, let's write down all the possible pairs of letters, *including repetitions, and also different orderings*. There are five possibilities for the first entry, and also five possibilities for the second entry, so we end up with a 5 by 5 square of possible pairs:

AA	BA	CA	DA	EA
AB	*BB*	CB	DB	EB
AC	BC	*CC*	DC	EC
AD	BD	CD	*DD*	ED
AE	BE	CE	DE	*EE*

Table 1.4

However, the pairs with repeated letters (shown in italics) don't count, because the two people on the committee must of course be different people (no cloning allowed!). Furthermore, we aren't concerned with the ordering of the people within the pair, so AB and BA represent the same committee. Likewise for AC and CA, etc. The upper right triangle in the square therefore simply duplicates the lower left triangle, which itself is just the triangle in Table 1.3. So we end up with $\binom{5}{2} = 10$ again.

The advantage of writing down the whole square in Table 1.4 is that the resulting answer of 10 can be thought of as taking 25 (which is 5 squared) and subtracting off 5 (to eliminate the pairs with repeated letters), and then taking half of the result (due to the duplicate triangles). This way of thinking allows us to quickly write down the general result for $\binom{N}{2}$. In forming pairs from N people, we can imagine writing down an N by N square which yields N^2 pairs; and then subtracting off the N pairs with repeated letters, which leaves us with $N^2 - N$ pairs; and then taking half of the result due to the duplicate triangles (for every pair XY there is also an equivalent pair YX). So we have

$$\binom{N}{2} = \frac{1}{2}(N^2 - N) = \frac{N(N-1)}{2}\,, \tag{1.6}$$

in agreement with Eq. (1.5).

THIRD METHOD: This third method is superior to the previous two, partly because it is quicker, and partly because it can easily be generalized to subgroups involving more than two members (see Section 1.3.2 below). Our strategy will be to pick the two committee members one at a time, just as we did at the end of Section 1.2 when we assigned people to seats.

Starting again with the case of five people, we can imagine having two seats that need to be filled with the two committee members. There are 5 options for who goes in the first seat. And then for each of these possibilities there are 4 options for who goes in the second seat, since there are only 4 people left. So there are $5 \cdot 4 = 20$ ways to plop the two people down in the two seats. (This is exactly the same reasoning as with the $N!$ ways to assign

people to N seats, but we're simply stopping the assignment process after two seats. So we have only the product $5 \cdot 4$ instead of the product $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$.) However, *we double counted every pair* in this reasoning. The pair XY was counted as distinct from the pair YX. So we need to divide by 2. The number of pairs we can pick from 5 people is therefore $5 \cdot 4/2 = 10$, as we found above.

The preceding reasoning easily generalizes to the case where we pick pairs from N people. We have N options for who goes in the first seat, and then for each of these possibilities there are $N - 1$ options for who goes in the second seat. This gives $N(N-1)$ total possibilities. But since we don't care about the order, this reasoning double counted every pair. We therefore need to divide by 2, yielding the final result of

$$\binom{N}{2} = \frac{N(N-1)}{2}, \tag{1.7}$$

in agreement with the above two methods.

1.3.2 Generalizing to other subgroups

Determining $\binom{5}{3}$

What if we want to pick a committee consisting of three people? Or four? Or a general number n? (n can't be larger than the total number of people, N, of course.) For small numbers N and n, we can simply list out the possibilities. For example, if we have five people and we want to pick a committee of three (so in our above "choose" notation, we want to determine $\binom{5}{3}$), we find that there are 10 possibilities:

A B C		
A B D		
A B E		
A C D	B C D	
A C E	B C E	
A D E	B D E	C D E

Table 1.5

We've grouped these according to which letter comes first. (The order of letters doesn't matter, so we've written each triplet in increasing alphabetical order.) If you want, you can look at these columns and think of 10 as equaling $6 + 3 + 1$, or more informatively as $\binom{4}{2} + \binom{3}{2} + \binom{2}{2}$. The $\binom{4}{2}$ here comes from the fact that once we've chosen the first letter to be A, there are $\binom{4}{2} = 6$ ways to pick the other two letters from B,C,D,E. This yields the first column in the table. Likewise for the second column with $\binom{3}{2} = 3$ triplets (with two letters chosen from C,D,E), and the third column with $\binom{2}{2} = 1$ triplet (with two letters chosen from D,E). See Problem 4 for the generalization of the fact that $\binom{5}{3} = \binom{4}{2} + \binom{3}{2} + \binom{2}{2}$.

You can also think of these 10 triplets as forming a pyramid. There are six triplets (the ones that start with A) in the bottom plane, three triplets (the ones that start with B) in the middle plane, and one triplet (the one that starts with C) in the top plane. This pyramid for the triplets is the analogy of the triangle for the pairs in Table 1.3. However, the pyramid (and the exact placement of all the triplets within it) is certainly harder to visualize than the triangle, so it turns out not to be of much use. The point of listing out the possibilities in a convenient geometrical shape is so that it can help you do the counting. If the geometrical shape is a pain to visualize, you might as well not bother with it.

It's possible to explicitly list out the various possibilities (in either columns or pyramids) for a small number like 5. But this practice becomes intractable when the number of people,

N, is large. Furthermore, if you want to think in terms of pyramids, and if you're dealing with committees consisting of four people, then you'll have to think about "pyramids" in four dimensions. Not easy! Fortunately, though, the third method in Section 1.3.1 easily generalizes to triplets and larger subgroups. The reasoning is as follows.

Calculating $\binom{N}{3}$

Consider the case of triplets. Our goal is to determine the number of committees of three people that can be chosen from N people. In other words, our goal is to determine $\binom{N}{3}$.

We can imagine having three seats that need to be filled with the three committee members. There are N options for who goes in the first seat. And then for each of these possibilities there are $N-1$ options for who goes in the second seat (since there are only $N-1$ people left). And then for each of these possibilities there are $N-2$ options for who goes in the third seat (since there are only $N-2$ people left). This gives $N(N-1)(N-2)$ possibilities. However, *we counted every triplet six times* in this reasoning. All six triplets of the form XYZ, XZY, YXZ, YZX, ZXY, ZYX were counted as distinct triplets. Since they all represent the same committee (because we don't care about the order), we must therefore divide by 6. Note that this "6" is nothing other than 3!, because it is simply the number of permutations of three objects. Since we counted each permutation as distinct in the above counting procedure, the division by 3! corrects for this. (Likewise, the 2 that appeared in the denominator of $N(N-1)/2$ in Section 1.3.1 was technically 2!.) We therefore arrive at the result,

$$\binom{N}{3} = \frac{N(N-1)(N-2)}{3!}\,. \tag{1.8}$$

Plugging in $N=5$ gives $\binom{5}{3}=10$, as it should.

For future reference, note that we can write Eq. (1.8) in a more compact way. If we multiply both the numerator and denominator by $(N-3)!$, the numerator becomes $N!$, so we end up with the nice concise expression,

$$\binom{N}{3} = \frac{N!}{3!(N-3)!}\,. \tag{1.9}$$

Calculating $\binom{N}{n}$

The above reasoning with triplets quickly generalizes to committees with larger numbers of people. If we have N people and we want to pick a committee of n, then we can imagine assigning people to n seats. There are N options for who goes in the first seat. And then for each of these possibilities there are $N-1$ options for who goes in the second seat (since there are only $N-1$ people left). And then for each of these possibilities there are $N-2$ options for who goes in the third seat (since there are only $N-2$ people left). And so on, until there are $N-(n-1)$ options for who goes in the nth seat (since there are only $N-(n-1)$ people left, because $n-1$ people have already been chosen). The number of possibilities for what the inhabitants of the n seats look like is therefore

$$N(N-1)(N-2)\cdots\big(N-(n-2)\big)\big(N-(n-1)\big). \tag{1.10}$$

However, *we counted every n-tuplet $n!$ times* in this reasoning, due to the fact that there are $n!$ ways to order any group of n people, and we counted all of these permutations as distinct. Since we don't care about the order, we must divide by $n!$ to correct for this. So we arrive at

$$\binom{N}{n} = \frac{N(N-1)(N-2)\cdots\big(N-(n-2)\big)\big(N-(n-1)\big)}{n!}\,. \tag{1.11}$$

As in the $n = 3$ case, if we multiply both the numerator and denominator by $(N-n)!$, the numerator becomes $N!$, and we end up with the concise expression,

$$\boxed{\binom{N}{n} = \frac{N!}{n!(N-n)!}} \tag{1.12}$$

For example, the number of ways to pick a committee of four people from six people is

$$\binom{6}{4} = \frac{6!}{4!2!} = 15. \tag{1.13}$$

You should check this result by explicitly listing out the 15 groups of four people.

Note that because of our definition of $0! = 1$ in Section 1.1, Eq. (1.12) is valid even in the case of $n = N$, because we have $\binom{N}{N} = N!/N!0! = 1$. And indeed, there is just one way to pick N people from N people. You simply pick all of them. Another special case is the $n = 0$ one. This gives $\binom{N}{0} = N!/0!N! = 1$. It's sort of semantics to say that there is one way to pick zero people from N people (you simply don't pick any, and that's the one way). But we'll see later on, especially when dealing with the binomial theorem, that $\binom{N}{0} = 1$ makes perfect sense.

Example (Equal coefficients): We just found that $\binom{6}{4} = 6!/4!2! = 15$. But note that $\binom{6}{2} = 6!/2!4!$ also equals 15. Both $\binom{6}{4}$ and $\binom{6}{2}$ involve the product of 2! and 4! in the denominator, and since the order doesn't matter in this product, the result is the same. We also have, for example, $\binom{11}{3} = \binom{11}{8}$. Both of these equal 165. Basically, any two n's that add up to N yield the same value of $\binom{N}{n}$.

(a) Demonstrate this mathematically.

(b) Explain in words why it is true.

Solution:

(a) Let the two n values be labeled as n_1 and n_2. If they add up to N, then they must take the forms of $n_1 = a$ and $n_2 = N - a$ for some value of a. (The above example with $N = 11$ was generated by either $a = 3$ or $a = 8$.) Our goal is to show that $\binom{N}{n_1}$ equals $\binom{N}{n_2}$. And indeed,

$$\begin{aligned} \binom{N}{n_1} &= \frac{N!}{n_1!(N-n_1)!} = \frac{N!}{a!(N-a)!}\,, \\ \binom{N}{n_2} &= \frac{N!}{n_2!(N-n_2)!} = \frac{N!}{(N-a)!\big(N-(N-a)\big)!} = \frac{N!}{(N-a)!a!}\,. \end{aligned} \tag{1.14}$$

The order of the $a!$ and $(N-a)!$ in the denominators doesn't matter, so the two results are the same, as desired.[3]

(b) Imagine picking n objects from N objects and then putting them in a box. The number of ways to do this is $\binom{N}{n}$, by definition. But note that you generated *two* sets of objects in this process: there are the n objects in the box, and there are also the $N-n$ objects *outside* the box. There's nothing special about being inside the box versus being outside, so you can equivalently consider your process to be a way of picking the group of $N-n$ objects that remain outside the box. Said in another way, a perfectly reasonable way

[3] In practice, when calculating $\binom{N}{n}$, you want to cancel the larger of the factorials in the denominator. For example, you would quickly cancel the 8! in both $\binom{11}{3}$ and $\binom{11}{8}$ and write them as $(11\cdot10\cdot9)/(3\cdot2\cdot1) = 165$.

of picking a committee of n members is to pick the $N-n$ members who are *not* on the committee. There is therefore a direct correspondence between each set of n objects and the complementary (remaining) set of $N-n$ objects. The number of different sets of n objects is therefore equal to the number of different sets of $N-n$ objects, as we wanted to show.

1.3.3 Situations where the order matters

Up to this point, we have considered committees/subgroups in which the order doesn't matter. But what if the order does matter? For example, what if we want to pick a committee of three people from N people, and furthermore we want to designate one of the members as the president, another as the vice president, and the third as just a regular member? The positions are now all distinct.

As we have done previously, we can imagine assigning the people to three seats. But now the seats have the names of the various positions written on them, so *the order matters*. From the reasoning preceding Eq. (1.8), there are $N(N-1)(N-2)$ ways of assigning people to the three seats. And that's the answer for the number of possible committees in which the order matters. We're done. We *don't* need to divide by 3! to correct for multiple counting here, because we haven't done any multiple counting. The triplet XYZ is distinct from, say, XZY because although both committees have X as the president (assuming we label the first seat as the president), the first committee has Y as the vice president, whereas the second committee has Z as the vice president. They are different committees.

The above reasoning quickly generalizes to the case where we want to pick a committee of n people from N people, where all n positions are distinct. If we denote the number of possible committees (where the order matters) as C_N^n, then we find $C_N^n = N(N-1)(N-2)\cdots\big(N-(n-1)\big)$. If we multiply this by 1 in the form of $(N-n)!/(N-n)!$, we see that the number of committees of n people (where the order matters) can be written in the concise form,

$$\boxed{C_N^n = \frac{N!}{(N-n)!}} \tag{1.15}$$

This differs from the above result for $\binom{N}{n}$ only in that we don't need to divide by $n!$, because there are no issues with multiple counting.

Let's now mix things up a bit and consider a committee that consists of distinct positions, but with some of the positions being held by more than one person.

Example (Three different titles): From 10 people, how many ways can you form a committee of 7 people consisting of a president, two (equivalent) vice presidents, and four (equivalent) regular members?

SOLUTION: There are 10 (or more precisely, $\binom{10}{1}$) ways to pick the president. And then for each of these possibilities there are $\binom{9}{2}$ ways to choose the two vice presidents from the remaining 9 people (the order doesn't matter between these two people). And then for each scenario of president and vice presidents there are $\binom{7}{4}$ ways to choose the four regular members from the remaining 7 people (again, the order doesn't matter among these four people). So the total number of possible committees is

$$\binom{10}{1}\binom{9}{2}\binom{7}{4} = \frac{10}{1!}\cdot\frac{9\cdot 8}{2!}\cdot\frac{7\cdot 6\cdot 5\cdot 4}{4!} = 12,600. \tag{1.16}$$

That's the answer, but note that we also could have solved the problem in the following alternate way. There's no reason why the president has to be picked first, so let's instead pick, say, the four regular members first, then the two vice presidents, and then the president. The total number of possible committees had better still be 12,600, so let's check that this is indeed the case. There are $\binom{10}{4}$ ways to pick the four regular members, then $\binom{6}{2}$ ways to pick the two vice presidents from the remaining 6 people, then $\binom{4}{1}$ ways to pick the president from the remaining 4 people. The total number of possible committees is therefore

$$\binom{10}{4}\binom{6}{2}\binom{4}{1} = \frac{10\cdot 9\cdot 8\cdot 7}{4!}\cdot\frac{6\cdot 5}{2!}\cdot\frac{4}{1!} = 12,600, \tag{1.17}$$

as desired. Both methods yield the same result because both Eqs. (1.16) and (1.17) have the same product $10\cdot 9\cdot 8\cdot 7\cdot 6\cdot 5\cdot 4$ in the numerator (in one way or another), and they both have the same product $1!\cdot 2!\cdot 4!$ in the denominator (in one way or another). So in the end, the order in which you pick the various subparts of the committee doesn't matter. It had better not matter, of course, because the number of possible committees is a definite number and can't depend on your method of counting it (assuming your method is a valid one!).

There is a nearly endless number of subgroup-counting examples relevant to the card game of poker, one of which is the following. As in the previous example, the ordering within subgroups in this example will matter in some cases but not in others.

Example (Full houses): How many different full-house hands are possible in standard 5-card poker? A full house consists of three cards of one value plus two cards of another. An example is 9,9,9,Q,Q (the suits don't matter).[4]

SOLUTION: Our strategy will be to determine how many hands there are of a given form, say 9,9,9,Q,Q, and then multiply this result by the number of different forms.

If the hand consists of three 9's and two Queens, there are $\binom{4}{3} = 4$ ways of choosing the three 9's from the four 9's in the deck, and $\binom{4}{2} = 6$ ways of choosing the two Q's from the four Q's in the deck. So there are $4\cdot 6 = 24$ possible full houses of the form 9,9,9,Q,Q. Note that we used the "choose" notation with $\binom{4}{3}$ and $\binom{4}{2}$, because the order of the 9's and the order of the Q's in the hand doesn't matter.

How many different forms (9,9,9,Q,Q is one form; 8,8,8,3,3 is another; etc.) are there? There are 13 different values of cards in the deck, so there are 13 ways to pick the value that occurs three times, and then 12 ways to pick the value that occurs twice, from the remaining 12 values. So there are $13\cdot 12 = 156$ different forms. Note that this result is $13\cdot 12$, and *not* $\binom{13}{2} = 13\cdot 12/2$, because the order *does* matter. Having three 9's and two Q's is different from having three Q's and two 9's.[5] The total number of possible full-house hands is therefore

$$13\cdot 12\cdot\binom{4}{3}\cdot\binom{4}{2} = 3,744. \tag{1.18}$$

This should be compared with the total number of possible poker hands, which is the much larger number, $\binom{52}{5} = 2,598,960$. Many more examples of counting poker hands are given in Problem 1.

[4] A standard deck of cards consists of 52 cards, with four cards (the four suits) for each of the 13 values: 2,...,9,10,J,Q,K,A.

[5] If you want, you can think there being $\binom{13}{2} = 13\cdot 12/2$ possibilities for the two values that appear, but then you need to multiply by 2 because each pair of values represents two different forms, depending on which of the two values occurs three times. If poker hands instead consisted of only four cards, and if a full house was defined to be a hand of the form AABB, then the number of different forms *would* be $\binom{13}{2}$, because the A's and B's are equivalent; each occurs twice.

1.4 Allowing repetitions

We learned how to count permutations in Section 1.2, and then committees/subgroups in Section 1.3, first where the order didn't matter and then where it did. There is one more thing we'll need to know how to count, namely subgroups *where repetition is allowed and where the order matters*. For example, let's say we have a box containing five balls labeled A,B,C,D,E. We reach in and pull out a ball and write down the letter. Then *we put the ball back in the box*, shake it around, and pull out a second ball (which might be the same as the first ball) and write down the letter.

Equivalently, we can imagine having *two boxes* with identical sets of A,B,C,D,E balls, and we pick one ball from each box. We can think about it either way, but the point is that the act of picking a ball is identical each time. In Section 1.3, once we picked a committee member, we couldn't pick this person again. He/she was *not* put back in the room. So there were only $N-1$ possible outcomes for the second pick, and then $N-2$ for the third, and so on. In the present scenario with replacement, there are simply N possible outcomes for each pick.

How many possible different pairs of letters (where repetition is allowed and where the order matters) can we pick in this five-ball example? We actually don't need to do any work here, because we already listed out all the possibilities in Section 1.3.1. We can simply copy Table 1.4:

A A	B A	C A	D A	E A
A B	B B	C B	D B	E B
A C	B C	C C	D C	E C
A D	B D	C D	D D	E D
A E	B E	C E	D E	E E

Table 1.6

We haven't bothered writing the AA, BB, etc. pairs in italics as we did in Table 1.4, because there's nothing special about them. They're perfectly allowed, just like any of the other pairs, because we replaced the first ball we picked. Furthermore, the entire table is relevant now; we're assuming that the order matters, so we don't want to ignore the upper-right triangle of pairs in the table as we did in the reasoning following Table 1.4. So we simply end up with $5^2 = 25$ possible pairs.

In general, if we have N balls instead of 5, we obtain an N by N square of letters, so the number of possible pairs is N^2. This is a nice simple result, simpler than the $\binom{N}{2} = N(N-1)/2$ result in Section 1.3.1 for the case where the ball isn't replaced and where the order doesn't matter.

What if we pick a ball three successive times from a box containing N balls, replacing the ball after each stage? (Or equivalently we have three identical boxes of N balls.) Well, there are N possibilities for the first ball, and again N possibilities for the second ball (because we put the ball back in), and again N possibilities for the third ball (because we again put the ball back in). So there are N^3 possible outcomes for the triplet of letters (or numbers, or whatever) we write down, under the assumption that the order matters.

Extending this reasoning, we see more generally that the number of possible outcomes in the case where we pick n balls from a box containing N balls (with replacement after

each stage, and with the order mattering) is:

$$\boxed{\text{Number of possible outcomes} = N^n} \tag{1.19}$$

Note that there is no restriction on the size of n here. It is perfectly allowed for n to be larger than N.

There are two differences between this N^n result (with replacement and with the order mattering) and the $\binom{N}{n} = N(N-1)\cdots\big(N-(n-1)\big)/n!$ result in Section 1.3.2 (with no replacement and with the order not mattering). First, due to the fact that the order now matters, there is no need to include the $n!$ in the denominator; we don't need to divide by $n!$ to correct for multiple counting. And second, due to the fact that we are now replacing the objects, all the $(N-1)$, $(N-2)$, etc. factors in the $\binom{N}{n}$ formula turn into N's in the (much simpler) N^n formula; there are always N possible outcomes at each stage. Both of these differences have the effect of making the N^n result larger than the $\binom{N}{n}$ one.

If you want to compare the present N^n result with the $C_N^n = N(N-1)\cdots\big(N-(n-1)\big)$ result in Section 1.3.3, the only difference is the replacement of the $(N-1)$, $(N-2)$, etc. factors with N's. Both expressions already lack the $n!$ in the denominator, because the order matters in both cases.

There are two classic cases where this N^n type of counting comes up:

Example 1 (Rolling dice): If you roll a standard six-sided die twice (or equivalently roll two dice), how many different possible outcomes are there (where the order matters)?

SOLUTION: There are six possibilities for what the first die shows, and six for the second. So there are $6^2 = 36$ possibilities in all. If you want to list them out, they are:

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

Table 1.7

Note that a "2,5" is different from a "5,2." That is, rolling a 2 with the first die (or, say, the left die if you're rolling both at once) and then a 5 with the second die is different from rolling a 5 and then a 2. All 36 outcomes in the above table are distinct.

REMARKS: In the present scenario with the dice, we don't have to worry about replacing things, as we did in the five-ball example above that led to Table 1.6. Every roll of the die is exactly the same as every other roll. Of course, if after rolling a die you paint over the face that pointed up (so that you can't roll that number again), then the rolls would *not* be identical, and this would be analogous to picking a ball from a box and not replacing it. We would then be back in the realm of Section 1.3.3.

As a precursor to our discussion of probability in the next chapter, we can ask the question: what is the probability of obtaining a sum of 7 when rolling two dice? If we look at the above table, we see that six outcomes yield a sum of 7. They are 1,6, 2,5, 3,4, 4,3, 5,2, and 6,1. Since all 36 possibilities are equally likely (because the probability of any number showing up at any point is 1/6), and since six of them yield the desired sum of 7, the probability of rolling a sum of 7 is $6/36 = 1/6 \approx 16.7\%$. From the table, you can quickly verify that 7 is the sum that has the most outcomes corresponding to it. So 7 is the most probable sum. We'll discuss all the various nuances and subtleties about probability in the next chapter. For now, the lesson to take away from this is that the ability to count things is extremely important in calculating probabilities! ♣

Example 2 (Flipping coins): If you flip a coin four times (or equivalently flip four coins), how many different possible outcomes are there (where the order matters)?

SOLUTION: There are two possibilities (Heads or Tails) for what the first coin shows, and two for the second, and two for the third, and two for the fourth. So there are $2^4 = 16$ possibilities in all. If you want to list them out, they are:

HHHH	THHH
HHHT	THHT
HHTH	THTH
HHTT	THTT
HTHH	TTHH
HTHT	TTHT
HTTH	TTTH
HTTT	TTTT

Table 1.8

We've grouped them in two columns according to whether the first coin shows a Heads or a Tails. Each column has eight entries, because $2^3 = 8$ is the number of possible outcomes with three coins. (Just erase the first entry in each outcome, and then each column simply gives these eight possible outcomes.) Likewise, it's easy to see why five coins yield $2^5 = 32$ possible outcomes. We just need to take all 16 of the above outcomes and tack on an H at the beginning, and then take all 16 again and tack on a T at the beginning. This gives $2 \cdot 16 = 32$ possible outcomes.

REMARK: As another probability teaser, we can ask: What is the probability of obtaining exactly two Heads in four coin flips? Looking at the above table, we see that six outcomes have two Heads. They are HHTT, HTHT, HTTH, THHT, THTH, and TTHH. Since all 16 possibilities are equally likely (because the probability of either letter showing up at any point is 1/2), and since six of them yield the desired outcome of two Heads, the probability of getting two Heads is $6/16 = 3/8 = 37.5\%$. As with the sum of 7 in the previous example, you can quickly verify by looking at the table that two Heads is the most likely number of Heads that will occur. ♣

1.5 Binomial coefficients

1.5.1 Coins and Pascal's triangle

Let's look at the preceding coin-flipping example in more detail. We found that there are six different ways to obtain exactly two Heads, so we might as well also ask how many ways there are to obtain other numbers of Heads. From Table 1.8, we see that the numbers of ways of obtaining exactly zero, one, two, three, or four Heads are, respectively, 1,4,6,4,1. (These same numbers are relevant for Tails, too, of course.) The sum of these numbers equals the total number of possibilities, $2^4 = 16$, as it must.

Moving downward to three coins (the eight possibilities are obtained by taking either column in Table 1.8 and stripping off the first letter), we quickly see that the numbers of ways of obtaining exactly zero, one, two, or three Heads are 1,3,3,1. With two coins, the numbers for zero, one, or two Heads are 1,2,1. And for one coin the numbers for zero or one Heads are 1,1. Also, for zero coins, you can only obtain zero Heads, and there's just one way to do this (you simply don't list anything down, and that's that). This is somewhat a

matter of semantics, but if we use a "1" for this case, it will fit in nicely below with the rest of the results.

Note that for three coins, $1+3+3+1 = 2^3$. And for two coins, $1+2+1 = 2^2$. And for one coin, $1+1 = 2^1$. So in each case the total number of possibilities ends up being 2^N, where N is the number of coins. This must be the case, of course, because we know from Section 1.4 that 2^N is the total number of possible outcomes.[6]

We can collect the above results and list them on top of one another to form the following table. Each row lists the number of different ways to obtain the various possible numbers of Heads (these numbers range from 0 to N).

$N = 0$:				1				
$N = 1$:			1		1			
$N = 2$:		1		2		1		
$N = 3$:	1		3		3		1	
$N = 4$:	1	4		6		4		1

Table 1.9

This is known as *Pascal's triangle*. Do these numbers look familiar? A couple more rows might help. If you figure things out for the $N = 5$ and $N = 6$ coin-flipping cases by explicitly listing out the possibilities, you'll arrive at:

$N = 0$:						1					
$N = 1$:					1		1				
$N = 2$:				1		2		1			
$N = 3$:			1		3		3		1		
$N = 4$:		1		4		6		4		1	
$N = 5$:		1		5		10		10		5	1
$N = 6$:	1		6		15		20		15		6		1

Table 1.10

At this point, you might get a feeling of deja vu with the 10's and 15's, since we've seen them before at various times in this chapter. You can then make the (correct) guess that the entries in this table are nothing other than the binomial coefficients! Written out explicitly in terms of these coefficients, the table becomes:

[6] In the language of Eq. 1.19, the N there is a 2 here, because there are two possible outcomes for each coin flip. And the n there is an N here, because we're using N instead of n for the number of flips.

$N=0$:											$\binom{0}{0}$											
$N=1$:									$\binom{1}{0}$				$\binom{1}{1}$									
$N=2$:							$\binom{2}{0}$				$\binom{2}{1}$				$\binom{2}{2}$							
$N=3$:					$\binom{3}{0}$				$\binom{3}{1}$				$\binom{3}{2}$				$\binom{3}{3}$					
$N=4$:				$\binom{4}{0}$			$\binom{4}{1}$				$\binom{4}{2}$				$\binom{4}{3}$				$\binom{4}{4}$			
$N=5$:			$\binom{5}{0}$				$\binom{5}{1}$				$\binom{5}{2}$				$\binom{5}{3}$				$\binom{5}{4}$				$\binom{5}{5}$	
$N=6$:	$\binom{6}{0}$				$\binom{6}{1}$				$\binom{6}{2}$				$\binom{6}{3}$				$\binom{6}{4}$				$\binom{6}{5}$				$\binom{6}{6}$

Table 1.11

Now, observing a pattern and guessing the correct rule is most of the battle, but is there a way to prove rigorously that the entries in Pascal's triangle are the binomial coefficients? For example, can we demonstrate that the number of ways of obtaining two Heads when flipping six coins is $\binom{6}{2}$? Indeed we can. It's actually almost a matter of definition, as the following reasoning shows.

If we flip six coins, we can imagine having six blank spaces on the paper that we have to fill in with either H's or T's. If we're considering the scenarios where two Heads come up, then we need to fill in two of the blanks with H's and four of them with T's. So the question reduces to: How many different ways can we place two H's down in six possible spots? But this is *exactly* the same question as: How many different committees of two people can we form from six people? The equivalence of these two questions is made clear if we imagine six people sitting a row, and if we plop down an H on two of them, with the understanding that the two people who get tagged with an H are the two people on the committee.

In general, the $\binom{N}{n}$ ways that n Heads can come up in N flips of a coin correspond exactly to the $\binom{N}{n}$ committees of n people that can be chosen from N people. Each coin flip corresponds to a person, and the person is declared to be on the committee if the result of that coin flip is a Heads.

1.5.2 $(a+b)^N$ and Pascal's triangle

(Note: Sections 1.5.2 and 1.5.3 are a bit mathematical, so they can be skipped on a first reading.)

A quick examination of Pascal's triangle shows (as we observed above) that the sum of the numbers in a given row equals 2^N. For example, $\binom{4}{0}+\binom{4}{1}+\binom{4}{2}+\binom{4}{3}+\binom{4}{4}=2^4$, or more generally,

$$\boxed{\binom{N}{0}+\binom{N}{1}+\binom{N}{2}+\cdots+\binom{N}{N-1}+\binom{N}{N}=2^N} \tag{1.20}$$

We know that this relation must be true, because both sides represent the total number of possible outcomes for N flips of a coin (with the counting on the left side being done according to the number of Heads that show up). But is there a way to demonstrate this equality without invoking the fact that both sides are relevant to coin flips? What if someone asks you out of the blue to prove this relation? It turns out that there's a very sneaky way to do this. We'll give the proof in Section 1.5.3, but first we need some background.

Consider the quantity $(a+b)^N$. You can quickly show that $(a+b)^2 = a^2+2ab+b^2$. And then you can multiply this by $(a+b)$ to arrive at $(a+b)^3 = a^3+3a^2b+3ab^2+b^3$. And then you can multiply this by $(a+b)$ to obtain $(a+b)^4$, and so on. The results are:

$$
\begin{array}{ll}
(a+b)^1 = & a+b \\
(a+b)^2 = & a^2+2ab+b^2 \\
(a+b)^3 = & a^3+3a^2b+3ab^2+b^3 \\
(a+b)^4 = & a^4+4a^3b+6a^2b^2+4ab^3+b^4 \\
(a+b)^5 = & a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5 \\
(a+b)^6 = & a^6+6a^5b+15a^4b^2+20a^3b^3+15a^2b^4+6ab^5+b^6
\end{array}
$$

Table 1.12

The coefficients here are exactly the numbers in Pascal's triangle! And there is a very good reason for this. Consider, for example, $(a+b)^5$. This is shorthand for

$$(a+b)(a+b)(a+b)(a+b)(a+b). \tag{1.21}$$

In multiplying this out, we obtain a number of terms. 32 of them in fact, although many take the same form. There are 32 terms because in multiplying out the five factors of $(a+b)$, every term in the result will involve either the a or the b from the first $(a+b)$ factor, and similarly either the a or the b from the second $(a+b)$ factor, and so on with the third, fourth, and fifth $(a+b)$ factors. Since there are two possibilities (the a or the b) from each factor, we end up with $2^5 = 32$ different terms.

However, many of the terms are equivalent. For example, if we pick the a from the first and third terms, and the b from the second, fourth, fifth terms, then we obtain $ababb$, which equals a^2b^3. Alternatively, we can pick the a from the second and fifth terms, and the b from the first, third, and fourth terms, which gives $babba$, which also equals a^2b^3.

How many ways can we obtain a term of the form a^2b^3? Well, we have five choices (the five factors of $(a+b)$) of where to pick the three b's from (or equivalently five choices of where to pick the two a's from), so the number of ways to obtain an a^2b^3 term is $\binom{5}{3} = 10$ (or equivalently $\binom{5}{2} = 10$), in agreement with Table 1.12.

Similarly, for example, the coefficient of ab^5 in $(a+b)^6$ is $\binom{6}{5} = 6$ because there are $\binom{6}{5}$ ways to choose the five b's from the six $(a+b)$ factors (or equivalently $\binom{6}{1}$ ways to choose the one a factor). This makes it clear why the coefficients of the terms in the expansion of $(a+b)^N$ take the general form of $\binom{N}{n}$, where n is the power of b in a given term.

In general, just as with the coin flips, the $\binom{N}{n}$ ways that n b's can be chosen from the N factors of $(a+b)$ correspond exactly to the $\binom{N}{n}$ committees of n people that can be chosen from N people. Each factor of $(a+b)$ corresponds to a person, and the person is declared to be on the committee if the b is chosen from that factor.

To sum up, we've encountered three situations (committees, coins, and $(a+b)^N$) that involve the binomial coefficients, and they all involve the binomial coefficients for the same reason: they all deal with the number of ways that n things can be chosen from N things. The answer to all three of the following questions is $\binom{N}{n}$.

- How many different committees of n people can be chosen from N people?
- Flip a coin N times. How many different outcomes involve exactly n Heads?
- Expand $(a+b)^N$. What is the coefficient of $a^{N-n}b^n$ (or a^nb^{N-n})?

Historically, the name "binomial coefficient" actually comes from the third of these, even though we discussed committees and coin flips before $(a+b)^N$ in this chapter. Multiplying out $(a+b)^N$ is known as the "binomial expansion" ("bi"-nomial since there are two letters, a and b).

1.5.3 Properties of Pascal's triangle

Having established that the coefficients of the terms in the expansion of $(a+b)^N$ take the form of $\binom{N}{n}$, we can now quickly explain why Eq. (1.20) is true, without invoking anything about coins flips. We know that

$$(a+b)^N = \binom{N}{0}a^N + \binom{N}{1}a^{N-1}b + \binom{N}{2}a^{N-2}b^2 + \cdots + \binom{N}{N-1}ab^{N-1} + \binom{N}{N}b^N. \quad (1.22)$$

This holds for *any* values of a and b. So, since we are free to pick a and b to be whatever we want, so let's pick them both to be 1. Multiplication by 1 doesn't affect anything, so we can basically just erase all the a's and b's on the right side of Eq. (1.22). We then see that the right side is equal to the left side of Eq. (1.20). And the left side of Eq. (1.22) is $(1+1)^N$, which is simply 2^N, which is equal to the right side of Eq. (1.20). We have therefore demonstrated Eq. (1.20).

Another nice property of Pascal's triangle, which you can verify by looking at Table 1.10, is that each number is the sum of the two numbers above it (or just the "1" above it, if it occurs at the end of a line). For example, in the $N = 6$ line, 20 is the sum of the two 10's above it (that is, $\binom{6}{3} = \binom{5}{2} + \binom{5}{3}$), and 15 is the sum of the 5 and 10 above it (that is, $\binom{6}{2} = \binom{5}{1} + \binom{5}{2}$), etc. Written out explicitly, the rule is

$$\boxed{\binom{N}{n} = \binom{N-1}{n-1} + \binom{N-1}{n}} \quad (1.23)$$

The task of Problem 2 is to give a mathematical proof of this relation, using the explicit form of the binomial coefficients. But let's demonstrate it here in a more intuitive way by taking advantage of what the binomial coefficients mean in terms of choosing committees.

In words, Eq. (1.23) says that the number of ways to pick n people from N people equals the number of ways to pick $n-1$ people from $N-1$ people, plus the number of ways to pick n people from $N-1$ people. Does this make sense? Yes indeed, due to the following reasoning.

Let's single out one of the N people, whom we will call Alice. There are two types of committees of n people: those that contain Alice, and those that don't. How many committees of each type are there? If the committee *does* contain Alice, then the other $n-1$ members must be chosen from the remaining $N-1$ people. There are $\binom{N-1}{n-1}$ ways to do this. If the committee *doesn't* contain Alice, then all n of the members must be chosen from the remaining $N-1$ people. There are $\binom{N-1}{n}$ ways to do this. Since each of the total $\binom{N}{n}$ number of committees falls into one or the other of these two categories, we therefore arrive at Eq. (1.23), as desired.

The task of Problem 3 is to reproduce the reasoning in the preceding paragraph to demonstrate Eq. (1.23), but instead in the language of coin flips or the $(a+b)^N$ binomial expansion.

1.6 Summary

In this chapter we learned how to count things. In particular, we learned:

1. $N!$ ("N factorial") is defined to be the product of the first N integers:

$$N! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot (N-2) \cdot (N-1) \cdot N. \quad (1.24)$$

2. The number of different permutations of N objects (that is, the number of different ways of ordering them) is $N!$.

3. Given N people, the number of different ways to choose an n-person committee where the order *doesn't* matter is denoted by $\binom{N}{n}$, and it equals

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}\,. \tag{1.25}$$

4. Given N people, the number of different ways to choose an n-person committee where the order *does* matter (for example, where there are n distinct positions) equals

$$C_N^n = \frac{N!}{(N-n)!}\,. \tag{1.26}$$

5. Consider a process for which there are N possible results each time it is repeated. If it is repeated n times, then the total number of possible outcomes is given by

$$\text{Number of possible outcomes} = N^n. \tag{1.27}$$

   Examples include rolling an N-sided die n times, or picking one of N balls from a box n times, with replacement each time (so that all the trials are equivalent).

6. The binomial coefficients $\binom{N}{n}$, which can be arranged nicely in Pascal's triangle, are relevant in three situations we've discussed: (1) choosing committees, (2) flipping coins, and (3) expanding $(a+b)^N$. All three of these situations involve counting the number of ways that n things can be chosen from N things.

1.7 Problems

1. **Poker hands** $**$

   In a standard 52-card deck of cards, how many different 5-card poker hands are there of each of the following types?[7]

   (a) Full house (three cards of one value, two of another).[8]

   (b) Straight flush (five consecutive values, all of the same suit). In the spirit of being realistic, assume that aces can be either high (above kings) or low (below 2's).

   (c) Flush (five cards of the same suit), excluding straight flushes.

   (d) Straight (five consecutive values), excluding straight flushes.

   (e) One pair.

   (f) Two pairs.

   (g) Three of a kind.

   (h) Four of a kind.

   (i) None of the above.

2. **Pascal sum 1** $*$

   Using $\binom{N}{n} = N!/n!(N-n)!$, show that

   $$\binom{N}{n} = \binom{N-1}{n-1} + \binom{N-1}{n}. \tag{1.28}$$

3. **Pascal sum 2** $*$

   At the end of Section 1.5.3, we demonstrated $\binom{N}{n} = \binom{N-1}{n-1} + \binom{N-1}{n}$ by using an argument involving committees. Repeat this reasoning, but now in terms of

   (a) coin flips, and

   (b) the $(a+b)^N$ binomial expansion.

4. **Pascal diagonal sum** $*$

   In Section 1.3.2 we noted that $\binom{5}{3} = \binom{4}{2} + \binom{3}{2} + \binom{2}{2}$. You can also see from Tables 1.10 and 1.11 that, for example, $\binom{6}{3} = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2}$. More generally,

   $$\binom{N}{n} = \binom{N-1}{n-1} + \binom{N-2}{n-1} + \binom{N-3}{n-1} + \cdots + \binom{n}{n-1} + \binom{n-1}{n-1}. \tag{1.29}$$

   Or in words: A given number (for example, $\binom{6}{3}$) in Pascal's triangle equals the sum of the numbers in the diagonal string that starts with the number that is above and to the left of the given number ($\binom{5}{2}$ in this case) and proceeds upward to the right (so the string contains $\binom{5}{2}$, $\binom{4}{2}$, $\binom{3}{2}$, and $\binom{2}{2}$ in this case).

   Demonstrate this by making repeated use of Eq. (1.23), which says that each number in Pascal's triangle is the sum of the two numbers above it (or just the "1" above it, if it occurs at the end of a line). *Hint:* No math needed! You just need to draw a few pictures of Pascal's triangle after successive applications of Eq. (1.23).

Many more problems will be added...

[7] For each type, it is understood that we don't count hands that also fall into a higher category. For example, when counting the three-of-a-kind hands, we *don't* count the full-house or four-of-a-kind hands, even though they technically contain three cards of the same value.

[8] We already solved this in the second example in Section 1.3.3, but we're listing it again here so that all the results for the various hands are contained in one place.

1.8 Solutions

1. **Poker hands**

   (a) (Full house) There are 13 ways to choose the value that appears three times, and $\binom{4}{3} = 4$ ways to choose the specific three cards from the four that have this value (the four suits). And then there are 12 ways to choose the value that appears twice from the remaining 12 values, and $\binom{4}{2} = 6$ ways to choose the specific two cards from the four that have this value. The total number of full-house hands is therefore

   $$13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2} = 3,744. \tag{1.30}$$

   (b) (Straight flush) The five consecutive values can be A,2,3,4,5, or 2,3,4,5,6, or 3,4,5,6,7, and so on until 10,J,Q,K,A. There are 10 of these sequences (remember that aces can be high or low). Each sequence can occur in four possible suits, so the total number of straight-flush hands is

   $$4 \cdot 10 = 40. \tag{1.31}$$

   Of these 40 hands, four of them are the so-called Royal flushes, consisting of 10,J,Q,K,A (one for each suit).

   (c) (Flush) The number of ways to pick five cards from the 13 cards of a given suit is $\binom{13}{5}$. Since there are four suits, the total number of flush hands is $4 \cdot \binom{13}{5} = 5,148$. However, 40 of these were already counted in the straight-flush category above, so that leaves

   $$4 \cdot \binom{13}{5} - 40 = 5,108 \tag{1.32}$$

   hands that are "regular" flushes.

   (d) (Straight) The 10 sequences listed in part (b) are relevant here. But now there are four possible choices for the first card (the four suits) in a given sequence, and likewise four possible choices for each of the other four cards. So the total number of straight hands is $10 \cdot 4^5 = 10,240$. However, 40 of these were already counted in the straight-flush category above, so that leaves

   $$10 \cdot 4^5 - 40 = 10,200 \tag{1.33}$$

   hands that are "regular" straights.

   (e) (One pair) There are 13 ways to pick the value that appears twice, and $\binom{4}{2} = 6$ ways to choose the specific two cards from the four that have this value. The other three values must all be different, and they must be chosen from the remaining 12 values. There are $\binom{12}{3}$ ways to do this. But there are four possible choices (the four suits) for each of these three values, which brings in a factor of 4^3. The total number of pair hands is therefore

   $$13 \cdot \binom{4}{2} \cdot \binom{12}{3} \cdot 4^3 = 1,098,240. \tag{1.34}$$

   Alternatively, you can count this as $13 \cdot \binom{4}{2} \cdot 48 \cdot 44 \cdot 40/6 = 1,098,240$, because after picking the value for the pair, there are 48 choices for the third card (since one value is off limits), then 44 choices for the fourth card (since two values are off limits), then 40 choices for the fifth card (since three values are off limits). But we counted the 6 possible permutations of a given set of third/fourth/fifth cards as distinct. Since the order doesn't matter, we must divide by $3! = 6$, which gives the above result.

   (f) (Two pairs) There are $\binom{13}{2}$ ways to choose the two values for the two pairs. For each pair, there are $\binom{4}{2} = 6$ ways to choose the specific two cards from the four that have

this value. This brings in a factor of 6^2. And then there are 44 options for the fifth card, since two values are off limits. The total number of two-pair hands is therefore

$$\binom{13}{2} \cdot \binom{4}{2}^2 \cdot 44 = 123,552. \tag{1.35}$$

(g) (Three of a kind) There are 13 ways to pick the value that appears three times, and $\binom{4}{3} = 4$ ways to choose the specific three cards from the four that have this value. The other two values must be different, and they must be chosen from the remaining 12 values. There are $\binom{12}{2}$ to do this. But there are four possible choices (the four suits) for each of these two values, which brings in a factor of 4^2. The total number of three-of-a-kind hands is therefore

$$13 \cdot \binom{4}{3} \cdot \binom{12}{2} \cdot 4^2 = 54,912. \tag{1.36}$$

Alternatively, as in part (e), you can think of this as $13 \cdot \binom{4}{3} \cdot 48 \cdot 44/2 = 54,912$.

(h) (Four of a kind) There are 13 ways to pick the value that appears four times, and then only $\binom{4}{4} = 1$ way to choose the specific four cards from the four that have this value. There are 48 choices for the fifth card, so the total number of four-of-a-kind hands is

$$13 \cdot \binom{4}{4} \cdot 48 = 624. \tag{1.37}$$

(i) (None of the above) Since we don't want to have any pairs, we're concerned with hands where all five values are different (for example, 3,4,7,J,K). There are $\binom{13}{5}$ ways to pick these five values. However, we also don't want any straights (such as 3,4,5,6,7), so we must be careful to exclude these. As in part (d), there are 10 different sequences of straights (remembering that aces can be high or low). So the number of possible none-of-the-above sets of values is $\binom{13}{5} - 10$.

We must now account for the possibility of different suits. For each of the $\binom{13}{5} - 10$ sets of values, each value has four options for its suit, so that brings in a factor of 4^5. However, we don't want to include any flushes, so we must exclude these from this 4^5 number. There are four possible flushes (one for each suit) for each set of values, so the number of possible none-of-the-above suit combinations for each of the $\binom{13}{5} - 10$ sets of values is $4^5 - 4$. The total number of none-of-the-above hands is therefore

$$\left(\binom{13}{5} - 10\right) \cdot \left(4^5 - 4\right) = 1,302,540. \tag{1.38}$$

Alternatively, we could have calculated this by subtracting the sum of the results in parts (a) through (h) from the total number of possible poker hands, which is $\binom{52}{5} = 2,598,960$. Equivalently, let's just check that all of our results add up properly. We'll list them in order of increasing frequency:

Royal flush =	4
Straight flush (not Royal) =	36
Four of a kind =	624
Full house =	3,744
Flush (not straight flush) =	5,108
Straight (not straight flush) =	10,200
Three of a kind =	54,912
Two pairs =	123,552
One pair =	1,098,240
None of the above =	1,302,540
Total =	2,598,960

So they do indeed add up properly. Note that pairs and none-of-the-above hands account for 92% of the total number of hands.

2. **Pascal sum 1**

Using the general expression $\binom{N}{n} = N!/n!(N-n)!$, the right side of the given equation can be written as

$$\binom{N-1}{n-1} + \binom{N-1}{n} = \frac{(N-1)!}{(n-1)!(N-n)!} + \frac{(N-1)!}{n!(N-n-1)!} \tag{1.39}$$

Let's get a common denominator (which will be $n!(N-n)!$) in these fractions so we can add them. Multiplying the first by n/n and the second by $(N-n)/(N-n)$ gives

$$\begin{aligned} \binom{N-1}{n-1} + \binom{N-1}{n} &= \frac{n(N-1)!}{n!(N-n)!} + \frac{(N-n)(N-1)!}{n!(N-n)!} \\ &= \frac{N(N-1)!}{n!(N-n)!} \qquad (\textit{canceling the } \pm n(N-1)! \textit{ terms}) \\ &= \frac{N!}{n!(N-n)!} \\ &= \binom{N}{n}, \end{aligned} \tag{1.40}$$

as desired.

3. **Pascal sum 2**

(a) The binomial coefficients give the number of ways of obtaining n Heads in N coin flips. So to demonstrate the given equation, we want to show that the number of ways to get n Heads in N coin flips equals the number of ways to get $n-1$ Heads in $N-1$ coin flips, plus the number of ways to get n Heads in $N-1$ coin flips. This is true due to the following reasoning.

Let's single out the first coin flip. There are two ways to get n Heads: either we get a Heads on the first flip, or we don't. How many possibilities are there of these two types? If the first flip *is* a Heads, then the other $n-1$ Heads must come from the remaining $N-1$ flips. There are $\binom{N-1}{n-1}$ ways for this to happen. If the first flip *isn't* a Heads, then all n Heads must come from the remaining $N-1$ flips. There are $\binom{N-1}{n}$ ways to do this. Since each of the total $\binom{N}{n}$ number of ways to get n Heads falls into one or the other of these two categories, we therefore arrive at Eq. (1.23), as desired.

(b) The binomial coefficients are the coefficients of the terms in the binomial expansion of $(a+b)^N$. So to demonstrate the given equation, we want to show that the coefficient of the term involving b^n in $(a+b)^N$ equals the coefficient of the term involving b^{n-1} in $(a+b)^{N-1}$, plus the coefficient of the term involving b^n in $(a+b)^{N-1}$. This is true due to the following reasoning.

Let's write $(a+b)^N$ in the form of $(a+b)\cdot(a+b)^{N-1}$, and imagine multiplying out the $(a+b)^{N-1}$ part. The result contains many terms, but the two relevant ones are $\binom{N-1}{n-1}a^{N-n}b^{n-1}$ and $\binom{N-1}{n}a^{N-n-1}b^n$. So we have

$$(a+b)^N = (a+b)\left(\cdots + \binom{N-1}{n-1}a^{N-n}b^{n-1} + \binom{N-1}{n}a^{N-n-1}b^n + \cdots\right). \tag{1.41}$$

There are two ways to get a b^n term on the right side: either the b in the first factor gets multiplied by the $\binom{N-1}{n-1}a^{N-n}b^{n-1}$ in the second factor, or the a in the first factor gets multiplied by the $\binom{N-1}{n}a^{N-n-1}b^n$ in the second factor. The net coefficient of b^n on the right side is therefore $\binom{N-1}{n-1} + \binom{N-1}{n}$. But the coefficient of b^n on the left side is $\binom{N}{n}$, so we have demonstrated Eq. (1.23).

4. **Pascal diagonal sum**

Consider an arbitrary number in Pascal's triangle, such as the one circled in the first triangle in Fig. 1.1 (the number happens to be $\binom{5}{2}$, but this won't matter). This number equals the sum of the two numbers above it, as shown in the second triangle. At every stage from here on, we will replace the *righthand* of the two numbers (that were just circled) with the two numbers above it; this won't affect the sum. The number that just got replaced will be shown with a dotted circle. The end result is the four circled numbers in the fifth triangle, which is the desired diagonal string of numbers. Since the sum is unaffected by the replacements at each stage, the sum of the numbers in the diagonal string equals the original number in the first triangle. In this specific case, we showed that $\binom{5}{2} = \binom{4}{1} + \binom{3}{1} + \binom{2}{1} + \binom{1}{1}$, but the result holds for any starting point.

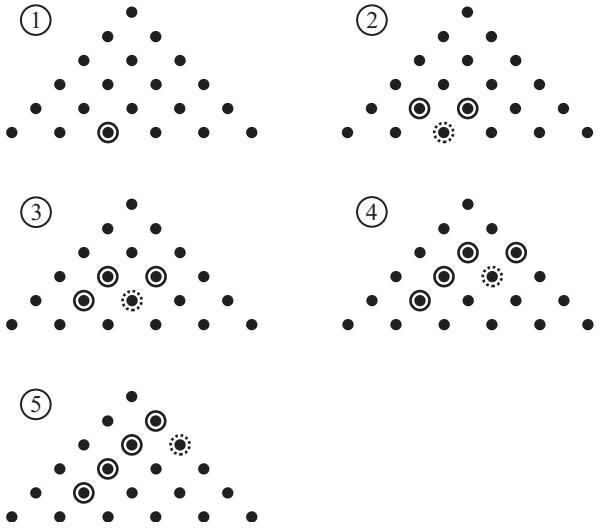


Figure 1.1

REMARK: As we saw in Table 1.5 for the case of $\binom{5}{3}$, we can interpret this result in the following general way. Let's imagine picking a committee of n people from N people, and let's label the people as 1,2,3,.... When we list out the $\binom{N}{n}$ possible committees, we can arrange them in groups according to what the lowest number in the committee is. For example, some committees have a 1, other committees don't have a 1 but have a 2, other committees don't have a 1 or a 2 but have a 3, and so on. How many committees are there of each of these types?

If the lowest number is a 1, then the other $n-1$ people on the committee must be chosen from the $N-1$ people who are 2 or higher. There are $\binom{N-1}{n-1}$ ways to do this. Similarly, if the lowest number is a 2, then the other $n-1$ people must be chosen from the $N-2$ people who are 3 or higher. There are $\binom{N-2}{n-1}$ ways to do this. Likewise, if the lowest number is a 3, then the other $n-1$ people must be chosen from the $N-3$ people who are 4 or higher. There are $\binom{N-3}{n-1}$ ways to do this. This method of counting continues until we reach the stage where there are only $n-1$ numbers higher than the lowest one (this occurs when the lowest number equals $N-(n-1)$), in which case there is just $\binom{n-1}{n-1} = 1$ way to choose the other $n-1$ people. Since the total number of possible committees is $\binom{N}{n}$, we therefore arrive at Eq. (1.29), as desired. ♣

Chapter 2

Probability

Having learned how to count things in Chapter 1, we can now talk about actual probability. We will find that in many situations it is a trivial matter to generate probabilities from our counting results, so the time and effort we spent in Chapter 1 will prove to be well worth it.

The outline of this chapter is as follows. In Section 2.1 we give the definition of probability. Although this definition is fairly easy to apply in most cases, there are a number of subtleties that come up. These are discussed in Appendix A; this is interesting material but not required for an understanding of this chapter, so feel free to skip it on a first reading. In Section 2.2 we present the various rules of probability. We show how these are applied to a few simple examples, and then we give a large number of more substantial examples in Section 2.3. In Section 2.4 we present two classic probability problems that many people find counterintuitive. In Section 2.5 we introduce the important concept of the *expectation value*, which is the expected average value of many trials of a given process. Finally, in Section 2.6 we talk about *Stirling's formula*, which gives an approximation to $N!$. We will revisit this formula in Chapter 3 when we talk about the various probability distributions.

2.1 Definition of probability

Probability gives a measure of how likely it is for something to happen. It can be defined as follows:

> *Consider a very large number of identical trials of a certain process; for example, flipping a coin, rolling a die, picking a ball from a box (with replacement), etc. If a certain event (for example, getting a Heads, rolling a 5, or picking a blue ball) happens a fraction p of the time, on average, then we say that the probability of that event occurring is p.*

Some examples are:

- The probability of getting a Heads on a coin flip is $1/2$ (or equivalently 50%), because the probabilities of getting a Heads or a Tails are equal, so they must each occur half of the time, on average.

- The probability of rolling a 5 on a standard 6-sided die is $1/6$, because the probabilities of rolling a 1, 2, 3, 4, 5, or 6 are all equal, so they must each happen one sixth of the time, on average.

- If there are three red balls and seven blue balls in a box, then the probabilities of picking a red ball or a blue ball are, respectively, 3/10 and 7/10. This follows from the fact that the probabilities of picking each of the ten balls are all equal (or let's assume they are), so they must each be picked one tenth of the time, on average. Since there are three red balls, a red ball will therefore be picked 3/10 of the time, on average. And since there are seven blue balls, a blue ball will be picked 7/10 of the time, on average.

Note the inclusion of the words "on average" in the definition and in these examples. We'll discuss this in detail in the subsection below.

Many probabilistic situations have the property that they involve a number of different possible outcomes, *all of which are equally likely*. For example, Heads and Tails are equally likely on a coin toss, the numbers 1 through 6 are equally likely on a die roll, and the ten balls in the above box are equally likely to be picked. In such a situation, the probability of a certain scenario happening is given by

$$p = \frac{\text{number of desired events}}{\text{total number of possible events}} \qquad (\textit{for equally likely events}) \tag{2.1}$$

Calculating the probability then simply reduces to a matter of counting the number of desired events, along with the total number of events. This is why we did all that counting in Chapter 1!

For example, the probability of rolling a 3 on a die is 1/6, because there is one desired event (the 3) and six total possible events (the six numbers). And the probability of rolling an even number is 1/2, because there are three desired events (2, 4, and 6) and again six total possible events (the six numbers). And the probability of picking a red ball in the above example is 3/10, because there are three desired events (picking any of the three red balls) and ten total possible events (the ten balls).

It should be stressed that Eq. (2.1) holds only under the assumption that all of the possible events are equally likely. But this isn't much of a restriction, because this assumption will usually be valid in the situations we'll be dealing with. In particular, it holds in situations dealing with permutations and subgroups, both of which we studied in detail in Chapter 1. Our ability to count these sorts of things will allow us to easily calculate probabilities via Eq. (2.1). Many examples are given in Section 2.3 below.

A word on semantics: "Chance" an "probability" mean the same thing. That is, the statement, "There is a 40% chance that the bus will be late," is equivalent to the statement, "There is a 40% probability that the bus will be late." However, the word "odds" has a different meaning; see Problem 1 for a discussion of this.

Importance of the words, "on average"

The above definition of probability includes the words "on average." These words are critical, because things wouldn't make any sense if we dropped them and instead went with the definition: "If the probability of an event occurring is p, then that event will occur in *exactly* a fraction p of the trials." This can't be a valid definition of probability, for the following reason. Consider the roll of one die, where the probability of each number occurring is 1/6. This definition would imply that on one roll of a die, we will get 1/6 of a "1," and 1/6 of a "2," and so on. But this is nonsense; you can't roll 1/6 of a "1." The number of times a "1" appears on one roll must of course be either zero or one. And in general for many rolls, the number must be an integer: 0,1,2,3,....

And there is a second problem with this definition, in addition to the problem of non integers. What if we roll a die six times? This definition would imply that we will get

exactly $(1/6) \cdot 6 = 1$ of each number. This is a little better, in that at least the proposed numbers are integers. But it still can't be right, because if you actually do the experiment and roll a die six times, you will find that you are certainly *not* guaranteed to get each of the six numbers exactly once. This scenario *might* happen (we'll calculate the probability in Section 2.3.4 below), but what is more likely is that some numbers will appear more than once, while other numbers won't appear at all.

Basically, for a small number of trials (such as six), the fractions of the time that the various events occur will most likely not look much like the various probabilities. This is where the words "very large number" in the above definition come into play. The point is that if you roll a die a huge number of times, then the fractions of the time that each of the six numbers appears will be *approximately* equal to 1/6. And the larger the number of trials gets, the closer these fractions will get to 1/6.

In Chapter 3 we will explain in detail why these fractions get closer and closer to the actual probabilities, as the number of trials gets larger and larger. For now, just take it on faith that if you, say, flip a coin 100 times, then the probability of obtaining 49, 50, or 51 Heads isn't so large. It happens to be about 24%, which tells us that there's a decent chance that the fraction of Heads you obtain will deviate at least moderately from 1/2. However, if you flip a coin, say, 100,000 times, then the probability of obtaining Heads between 49% and 51% of the time is 99.99999997%, which tells us that there's virtually no chance that the fraction of Heads you obtain will deviate much from 1/2. We'll talk in detail about such matters in Chapter 3 in Section 3.4.1.

2.2 The rules of probability

So far we've only talked about the probability of single events, for example, rolling a 3 on a die or getting Heads on a coin toss. We'll now consider two (or more) events. Reasonable questions to ask are: What is the probability that they both occur? What is the probability that either of them occurs? The rules below answer these questions. We'll present some simple examples for each rule here, but you are encouraged to reread this section after (or while) working through the examples in Section 2.3.

2.2.1 AND: The "intersection" probability, P(A and B)

Let A and B be two events. For example, let $A = \{$rolling a 2 on one die$\}$ and $B = \{$rolling a 5 on another die$\}$. Or we might have $A = \{$picking a red ball from a box$\}$ and $B = \{$picking a blue ball without replacement after the previous pick$\}$. What is the probability of both A and B occurring? In answering this question, we must consider two cases: (1) A and B are independent events, or (2) A and B are dependent events. Let's look at each of these in turn.

Independent events

Two events are said to be *independent* if they don't affect each other. Or more precisely, if the occurrence of one doesn't affect the probability that the other occurs. An example is the first situation mentioned above: rolling two dice, with $A = \{$rolling a 2 on one die$\}$ and $B = \{$rolling a 5 on the other$\}$. The probability of obtaining a 5 on the second roll is 1/6, independent of what happens on the first roll. (The events in the second situation above with the balls in the box are *not* independent; we'll talk about this below.) Another example is picking one card from a deck, with $A = \{$the card is a king$\}$ and $B = \{$the card

is a heart}. The probability of the card being a heart is 1/4, independent of whether or not it is a king.[1]

The "And" rule for independent events is:

- *If events A and B are independent, then the probability of both of them occurring equals the product of their individual probabilities:*

$$\boxed{P(A \text{ and } B) = P(A) \cdot P(B)} \tag{2.2}$$

We can quickly apply this rule to the two examples we just mentioned. The probability of rolling a 2 and then a 5 is $P(2 \text{ and } 5) = P(2) \cdot P(5) = (1/6) \cdot (1/6) = 1/36$. This agrees with the fact that one out of the 36 pairs of numbers in Table 1.7 is "2, 5." And the probability of having one card be both a king and a heart is $P(\text{king and heart}) = P(\text{king}) \cdot P(\text{heart}) = (1/13) \cdot (1/4) = 1/52$. This makes sense, because one of the 52 cards in a deck is the king of hearts.

Remarks:

1. The reasoning behind this rule is the following. Consider N different trials of a process, where N is very large. In the case of the dice, a "trial" consists of rolling both dice, so the outcome of a trial takes the form of a pair or numbers. The first number is the result of the first roll, and the second number is the result of the second roll. The fraction of these outcomes (on average) that have a 2 as the first number is $(1/6) \cdot N$. Let's now consider only these outcomes and ignore the rest. Then a fraction 1/6 *of these outcomes* have a 5 as the second number.[2] So the number of trials that have both a 2 as the first number and a 5 as the second number is 1/6 of $(1/6) \cdot N$, which equals $(1/6) \cdot (1/6) \cdot N$.

   In the case of general probabilities $P(A)$ and $P(B)$, it's easy to see that the two (1/6)'s here get replaced by $P(A)$ and $P(B)$. So the number of outcomes where both A and B occur is $P(A) \cdot P(B) \cdot N$. And since we did N trials, the fraction of outcomes where both A and B occur is therefore $P(A) \cdot P(B)$. From the definition of probability in Section 2.1, this is then the probability that both A and B occur, in agreement with Eq. (2.2).

2. A word on terminology: The words "event," "outcome," and "result" all mean essentially the same thing, so we'll use them interchangeably. They all basically mean "something that happens." But as noted in Footnote 1, you don't actually need to *do* two different things to have two different results. Even if you pick just one card from a deck, there can still be two different events/outcomes/results associated with that one card, for example, the $A = \{$the card is a king$\}$ and $B = \{$the card is a heart$\}$ events mentioned above. Or more mundanely, we could simply have $A = \{$the card is a king$\}$ and $B = \{$the card is not a king$\}$, although these events fall firmly into the "dependent" category, discussed below. ♣

If you want to think about the rule in Eq. (2.2) in terms of pictures, then consider Fig. 2.1. Without worrying about the specifics, let's assume that different points within the overall boundary represent different events. And let's assume that they're all equally likely, which means that the area of a region gives the probability that an event located in that region occurs (assuming that the area of the whole region is 1). The figure corresponds to $P(A) = 20\%$ and $P(B) = 40\%$. Events to the left of the vertical line are ones where A occurs, and events to the right of the vertical line are ones where A doesn't occur. Likewise for B and events above/below the horizontal line.

[1] Note that it is possible to have two different events even if we have only one "trial." In this example we picked only one card, but this card has two qualities (its suit and its value), and we can associate an event with each of these qualities.

[2] This is where we are invoking the independence of the events. As far as the second roll is concerned, the set of $(1/6) \cdot N$ trials that have a 2 for the first roll is no different from any other set of $(1/6) \cdot N$ trials, so the probability of rolling a 5 on the second roll is simply the standard value of 1/6.

From the figure, we see that the events where B occurs (the ones above the horizontal line) constitute 40% of the entire square. And they also constitute 40% of the vertical strip to the left of the vertical line. Since this vertical strip represents the events where A occurs, we see that B occurs 40% of the time that A occurs. In other words, B occurs 40% of the time, independent of whether or not A occurs. Basically, B couldn't care less what happens with A. Similar statements hold with A and B interchanged, so A likewise couldn't care less what happens with B.

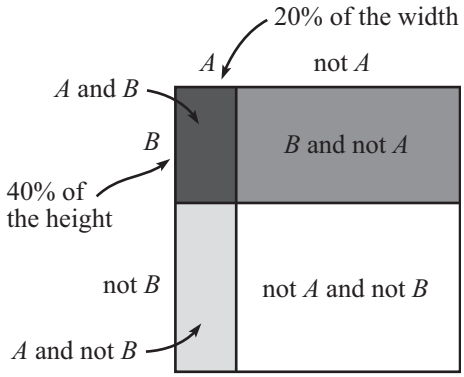


Figure 2.1

The darkly-shaded "A and B" region is the intersection of the region above the horizontal line (where B occurs) and the region to the left of the vertical line (where A occurs). Hence the word "intersection" in the title of this section. If you're wondering what the analogous figure looks like for *dependent* events, we'll draw that in Fig. 2.3 below.

Dependent events

Two events are said to be *dependent* if they *do* affect each other. Or more precisely, if the occurrence of one *does* affect the probability that the other occurs. An example is picking two balls in succession from a box containing two red balls and three blue balls (Fig. 2.2), with $A = \{$picking a red ball$\}$ and $B = \{$picking a blue ball *without replacement* after the previous pick$\}$.

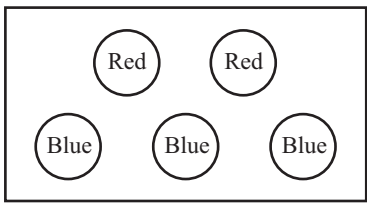


Figure 2.2

If you pick a red ball first, then the probability of picking a blue ball second is $3/4$, because there are three blue balls and one red ball left. On the other hand, if you a pick blue ball first, then the probability of picking a blue ball second is $2/4$, because there are two red balls and two blue balls left. So the occurrence of A certainly affects the probability of B.

Another example might be something like: $A = \{$it rains at 6:00$\}$ and $B = \{$you walk to the store at 6:00$\}$. People are generally less likely to go for a walk when it's raining outside, so (at least for most people) the occurrence of A affects the probability of B.

The "And" rule for dependent events is:

- *If events A and B are dependent,*[3] *then the probability of both of them occurring equals*

$$\boxed{P(A \text{ and } B) = P(A) \cdot P(B|A)} \tag{2.3}$$

where $P(B|A)$ stands for the probability of B occurring, assuming that A has occurred. It is read as "the probability of B, given A."

We can apply this rule to the above example with the balls in the box. (We won't bother trying to get quantitative about the "walking in the rain" example.) The A and B events in Eq. (2.3) are Red_1 and Blue_2, respectively, where the subscript refers to the first or second pick, to avoid any confusion. We saw above that $P(\text{Blue}_2|\text{Red}_1) = 3/4$. And we also know that $P(\text{Red}_1)$ is simply $2/5$, because there are initially two red balls and three blue balls. So Eq. (2.3) gives the probability of picking a red ball first and a blue ball second as

$$P(\text{Red}_1 \text{ and } \text{Blue}_2) = P(\text{Red}_1) \cdot P(\text{Blue}_2|\text{Red}_1) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}\,. \tag{2.4}$$

We can verify that this is correct by listing out all the possible pairs of balls that can be picked. If we label the balls as 1,2,3,4,5, and if we let 1 and 2 be the red balls, and 3, 4, 5 be the blue balls, then the possible results are (the first number stands for the first ball picked, and the second number stands for the second ball picked):

	Red first		Blue first		
Red second	—	21	31	41	51
	12	—	32	42	52
Blue second	13	23	—	43	53
	14	24	34	—	54
	15	25	35	45	—

Table 2.1

The "—" entries stand for the cases that aren't allowed; we can't pick two of the same ball, because we're not replacing the ball after the first pick. The lines are drawn for clarity; the internal vertical line separates the cases where a red or blue ball is drawn on the first pick, and the internal horizontal line separates the cases where a red or blue ball is drawn on the second pick. The six pairs in the lower left corner are the cases where a red ball (numbered 1 and 2) is drawn first and a blue ball (numbered 3, 4, and 5) is drawn second. Since there are 20 total possible outcomes, the desired probability is $6/20 = 3/10$, in agreement with Eq. (2.4).

The table also gives a verification of the $P(\text{Red}_1)$ and $P(\text{Blue}_2|\text{Red}_1)$ probabilities we wrote down in Eq. (2.4). $P(\text{Red}_1)$ equals $2/5$ because 8 of the 20 entries are to the left of the vertical line. And $P(\text{Blue}_2|\text{Red}_1)$ equals $3/4$ because of these 8 entries, 6 are below the horizontal line.

Remarks:

1. The preceding method of explicitly counting the possible outcomes shows that you don't *have* to use the rule in Eq. (2.3), and likewise Eq. (2.2), to calculate probabilities. You can often instead just count up the various possibilities and solve the problem from scratch. But the rules in Eqs. (2.2) and (2.3) allow you to take a shortcut and not go through the effort of listing all the cases out, which might be rather difficult if you're dealing with large numbers.

[3] There is actually no need for this "dependent" qualifier, as explained in the second remark below.

2. The rule in Eq. (2.2) for the "independent" case is a special case of the rule in Eq. (2.3) for the "dependent" case. This is true because if A and B are independent, then $P(B|A)$ is simply equal to $P(B)$, because the probability of B occurring is just $P(B)$, independent of whether or not A occurs. And Eq. (2.3) reduces to Eq. (2.2) when $P(B|A) = P(B)$. Therefore, there was technically no need to introduce Eq. (2.2) first. We could have started with Eq. (2.3), which covers all possible scenarios, and then showed that it reduces to Eq. (2.2) when the events are independent. But pedagogically, it's usually better to start with a special case and then work up to the more general case.

3. There's nothing special about the order of A and B in Eq. (2.3). We could just as well interchange the letters and write $P(B \text{ and } A) = P(B) \cdot P(A|B)$. But $P(B \text{ and } A) = P(A \text{ and } B)$, because it certainly doesn't matter which letter you say first when you say that two events both occur. So can also write $P(A \text{ and } B) = P(B) \cdot P(A|B)$. Comparing this with Eq. (2.3), we see that we can write $P(A \text{ and } B)$ in two different ways:

$$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B|A), \\ \text{or} \quad &= P(B) \cdot P(A|B). \end{aligned} \tag{2.5}$$

Let's verify that the second from here also works in the above example. By counting the various kinds of pairs in the above table, we find $P(\text{Blue}_2) = 12/20 = 3/5$ (by looking at all 20 pairs), and $P(\text{Red}_1|\text{Blue}_2) = 6/12 = 1/2$ (by looking at just the 12 pairs below the horizontal line). So we have

$$\begin{aligned} P(\text{Red}_1 \text{ and } \text{Blue}_2) &= P(\text{Blue}_2) \cdot P(\text{Red}_1|\text{Blue}_2) \\ &= \frac{3}{5} \cdot \frac{1}{2} = \frac{3}{10}, \end{aligned} \tag{2.6}$$

in agreement with Eq. (2.4).

4. You shouldn't take the phrase, "the probability of B, given A," to imply that A necessarily influences B. In the above "walking in the rain" example, the rain influences your likelihood of walking, but not the other way around. (It would be nice if we could control whether it rains at a given time, but we can't!) Similarly, in the "balls in the box" example, the first pick influences the second, but not the other way around. Nevertheless, it still makes sense to talk about things like, "the probability of it raining, given that you walk to the store." This probability is still well defined, even though there is no causal relation in the walking-to-raining direction. Likewise with the $P(\text{Red}_1|\text{Blue}_2)$ probability.

5. A trivial yet extreme example of two dependent events is the events A and "not A." The occurrence of A highly affect the probability of "not A" occurring: If A occurs, then "not A" occurs with probability zero. And if A doesn't occur, then "not A" occurs with probability 1. ♣

If you want to think about the rule in Eq. (2.3) in terms of pictures, then consider Fig. 2.3 (this is just a hypothetical situation, not related to the above example with the balls). This looks a lot like Fig. 2.1, but with one major difference. The one horizontal line is now two different lines. The probability of B occurring if A occurs is still 40% (this is the darkly-shaded fraction of the lefthand vertical strip), but now the probability of B occurring if A *doesn't* occur is 50% (this is the shaded fraction of the righthand vertical strip). So the occurrence of A affects the probability that B occurs.

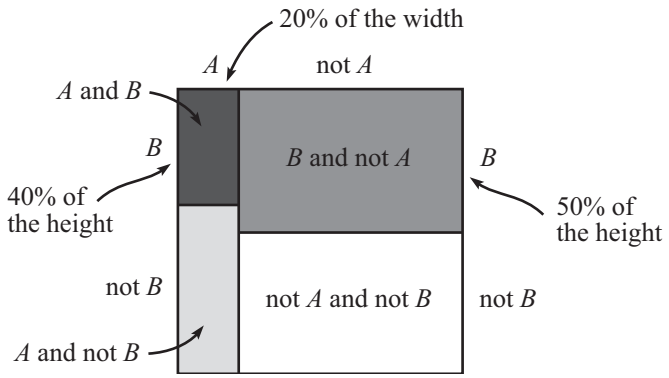


Figure 2.3

If we want to recast Table 2.1 into a form that looks like Fig. 2.3, we'll need to arrange for equal areas to give equal probabilities, so to speak. If we get rid of the "—" spaces, then all entries have equal probabilities, and the table now looks like:

12	21	31	41	51
13	23	32	42	52
14	24	34	43	53
15	25	35	45	54

Table 2.2

The upper left region corresponds to red balls on both picks. The lower left region corresponds to a red ball and then a blue ball. The upper right region corresponds to a blue ball and then a red ball. And the lower right region corresponds to blue balls on both picks. This figure makes it clear why we formed the product $(2/5) \cdot (3/4)$ in Eq. (2.4). The "2/5" gives the fraction of the outcomes that lie to the left of the vertical line (these are the ones that have a red ball first), and the "3/4" gives the fraction *of these outcomes* that lie below the horizontal line (these are the ones that have a blue ball second). The product of these fractions gives the overall fraction (namely 3/10) of the outcomes that lie in the lower left region (the ones that have red ball first and a blue ball second).

Example: For practice, let's calculate the overall probability of B occurring in the hypothetical scenario described in Fig. 2.3.

First method: The question is equivalent to finding the fraction of the total area that lies above the horizontal line segments. The upper left region is $40\% = 2/5$ of the area that lies to the left of the vertical line, which itself is $20\% = 1/5$ of the total area. And the upper right region is $50\% = 1/2$ of the area that lies to the right of the vertical line, which itself is $80\% = 4/5$ of the total area. The fraction of the total area that lies above the horizontal line segments is therefore

$$\frac{1}{5} \cdot \frac{2}{5} + \frac{4}{5} \cdot \frac{1}{2} = \frac{2}{25} + \frac{2}{5} = \frac{12}{25} = 48\%. \tag{2.7}$$

Second method: We'll use the rule in Eq. (2.3) twice. First, note that

$$P(B) = P(A \text{ and } B) + P\big((\text{not } A) \text{ and } B\big). \tag{2.8}$$

This is true because either A happens or it doesn't. We can now apply Eq. (2.3) to each of these terms to obtain

$$\begin{aligned} P(B) &= P(A) \cdot P(B|A) + P(\text{not } A) \cdot P(B|\text{not } A) \\ &= \frac{1}{5} \cdot \frac{2}{5} + \frac{4}{5} \cdot \frac{1}{2} = \frac{2}{25} + \frac{2}{5} = \frac{12}{25} = 48\%, \end{aligned} \tag{2.9}$$

in agreement with the first method. Comparing these methods makes it clear how the conditional probabilities like $P(B|A)$ are related to the fractional areas.

2.2.2 OR: The "union" probability, P(A or B)

Let A and B be two events. For example, let $A = \{$rolling a 2 on a die$\}$ and $B = \{$rolling a 5 on the *same* die$\}$. Or we might have $A = \{$rolling an even number (that is, 2, 4, or 6) on a die$\}$ and $B = \{$rolling a multiple of 3 (that is, 3 or 6) on the *same* die$\}$. What is the probability of either A or B occurring? In answering this question, we must consider two cases: (1) A and B are exclusive events, or (2) A and B are non-exclusive events. Let's look at each of these in turn.

Exclusive events

Two events are said to be *exclusive* if one precludes the other. That is, they can't both happen at the same time. An example is rolling one die, with $A = \{$rolling a 2 on the die$\}$ and $B = \{$rolling a 5 on the *same* die$\}$. These events are exclusive because it is impossible for one number to be both a 2 and a 5. (The events in the second situation above with the multiples of 2 and 3 are *not* exclusive; we'll talk about this below.) Another example is picking one card from a deck, with event $A = \{$the card is a diamond$\}$ and event $B = \{$the card is a heart$\}$. These events are exclusive because it is impossible for one card to be both a diamond and a heart.

The "Or" rule for exclusive events is:

- *If events A and B are exclusive, then the probability of either of them occurring equals the sum of their individual probabilities:*

$$\boxed{P(A \text{ or } B) = P(A) + P(B)} \tag{2.10}$$

We can quickly apply this rule to the two examples we just mentioned. The probability of rolling a 2 or a 5 on the roll of one die is $P(2 \text{ or } 5) = P(2) + P(5) = (1/6) + (1/6) = 2/6 = 1/3$. This makes sense, because two of the six numbers on a die are the 2 and 5. And the probability of a given card being either a diamond or a heart is $P(\text{diamond or heart}) = P(\text{diamond}) + P(\text{heart}) = (1/4) + (1/4) = 2/4 = 1/2$. This makes sense, because half of the 52 cards in a deck are the diamonds and hearts.

If you want to think about this rule in terms of pictures, then consider Fig. 2.4. As in Section 2.2.1, we'll assume that different points within the overall boundary represent different events. And we'll assume that they're all equally likely, which means that the area of a region gives the probability that an event located in that region happens (assuming that the area of the whole region is 1). We'll be using this figure only for its qualitative features (it's basically just a Venn diagram), so we'll simply draw the various regions as general blobs, as opposed to the specific rectangles we used for quantitative calculations in Section 2.2.1.

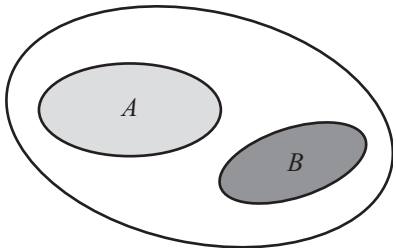


Figure 2.4

Let events A and B be signified by the regions shown. (These "events" are therefore actually the collections of many individual events, just as the $A = \{$the card is a diamond$\}$ "event" above was the collection of 13 individual events; this is perfectly fine.) The key feature of this diagram is that there is *no overlap* between the two regions, because we are assuming that A and B are exclusive.[4] The rule in Eq. (2.10) is simply the statement that the area of the union (hence the word "union" in the title of this section) of regions A and B equals the sum of the areas of A and B. There's nothing fancy going on here. This statement is no deeper than the statement that if you have two separate bowls, the total number of apples in the bowls equals the number of apples in one bowl plus the number of apples in the other bowl.

A special case of Eq. (2.10) is the "Not" rule,

$$P(A) = 1 - P(\text{not } A). \tag{2.11}$$

This is implied by Eq. (2.10) for the following reason. A and "not A" are certainly exclusive events (you can't both have something and not have it), so Eq. (2.10) gives $P(A \text{ or } (\text{not } A)) = P(A) + P(\text{not } A)$. But $P(A \text{ or } (\text{not } A)) = 1$, because every possible event can be categorized as either A or "not A" (events either happen or they don't; you can't have half of A or something like that). Therefore, we have $P(A) + P(\text{not } A) = 1$, from which Eq. (2.11) follows.

Non-exclusive events

Two events are said to be *non exclusive* if it is possible for both to happen at the same time. An example is rolling one die, with $A = \{$rolling an even number (that is, 2, 4, or 6)$\}$ and $B = \{$rolling a multiple of 3 (that is, 3 or 6) on the *same* die$\}$. If you roll a 6, then both A and B occur. Another example is picking one card from a deck, with event $A = \{$the card is a king$\}$ and event $B = \{$the card is a heart$\}$. If you pick the king of hearts, then both A and B occur.

The "Or" rule for non-exclusive events is:

- *If events A and B are non exclusive,*[5] *then the probability of either (or both) of them occurring equals:*

$$\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)} \tag{2.12}$$

[4] If there were a region that was contained in both A and B, then the events in that region would be ones for which both A and B happen at the same time, which would violate the assumption that A and B are exclusive.

[5] There is actually no need for this "non exclusive" qualifier, as explained in the third remark in the list of remarks at the end of this section.

The "or" here is the so-called "inclusive or" in the sense that we say "A or B happened" if either *or both* of the events happened.

We can quickly apply this rule to the two examples we just mentioned. In the first example, the only way to roll an even number *and* a multiple of 3 is to roll a 6, which happens with probability 1/6. So we find that the probability of rolling an even number or a multiple of 3 on the roll of one die is

$$\begin{aligned} P(\text{even or multiple of 3}) &= P(\text{even}) + P(\text{multiple of 3}) - P(\text{even and multiple of 3}) \\ &= \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{4}{6} = \frac{2}{3}\,. \end{aligned} \tag{2.13}$$

This makes sense, because four of the six numbers on a die are even numbers or multiples of 3, namely 2, 3, 4, and 6. (Remember that whenever we use "or," it means the "inclusive or.")

REMARK: The whole point of subtracting off the 1/6 in Eq. (2.13) is so that we don't double count the rolling of a 6. If we naively added up the number of ways to roll an even number (three of them: 2, 4, and 6) plus the number of ways to roll a multiple of 3 (two of them: 3 and 6), and if we then came to the conclusion that there are five ways to roll an even number or a multiple of three, then we'd end up with a probability of 5/6. But this would be wrong, because we double counted the 6. (See Fig. 2.5 for a pictorial explanation of this.) The 6 isn't "doubly good" because it satisfies both the A and B criteria. It's simply another number that satisfies the "A or B" condition, just like 2, 3, and 4. ♣

In the second example with the cards, the only way to pick a king *and* a heart is to pick the king of hearts, which happens with probability 1/52. So we find that the probability of a given card being a king or a heart is

$$\begin{aligned} P(\text{king or heart}) &= P(\text{king}) + P(\text{heart}) - P(\text{king and heart}) \\ &= \frac{1}{13} + \frac{1}{4} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}\,. \end{aligned} \tag{2.14}$$

This makes sense, because 16 of the 52 cards in a deck are kings or hearts, namely the 13 hearts, plus the kings of diamonds, spades, and clubs (we already counted the king of hearts). As in the previous example with the die, the point of subtracting off the 1/52 is so that we don't double count the king of hearts.

If you want to think about the rule in Eq. (2.12) in terms of pictures, then consider Fig. 2.5, which is a generalization of Fig. 2.4. The only difference is that we're now allowing A and B to overlap. As in Fig. 2.4, we're assuming that the area of a region gives the probability that an event located in that region happens.

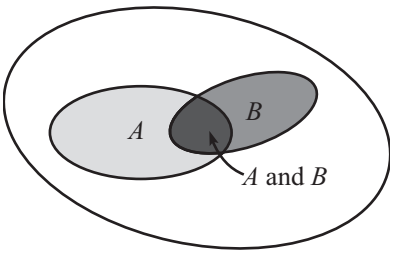


Figure 2.5

The rule in Eq. (2.12) is the statement that the area of the union of regions A and B equals the sum of the areas of A and B *minus the area of the overlap*. This subtraction is

necessary so that we don't double count the region that belongs to both A and B. (Again, this region isn't "doubly good" because it belongs to both A and B. As far as the "A or B" condition goes, the overlap region is just the same as any other part of the union of A and B.) In terms of a physical example, the rule in Eq. (2.12) is equivalent to the statement that if you have two cages that have a region of overlap, then the total number of birds in the cages equals the number of birds in one cage plus the number of birds in the other cage, minus the number of birds in the overlap region. In the situation shown in Fig. 2.6, we have $7 + 5 - 2 = 10$ birds.

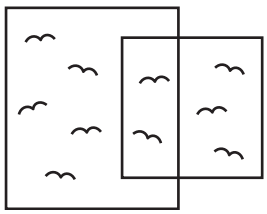


Figure 2.6

Things get a more complicated if you have three or more events, and if you want to calculate probabilities like $P(A \text{ or } B \text{ or } C)$. But in the end, the main task is to keep track of the overlaps of the various regions. See Problem 2 for a discussion of this.

Remarks:

1. If you want, you can think about the area of the union of A and B in Fig. 2.5 as the area of *only* A plus the area of *only* B, plus the area of A *and* B. (Equivalently, the number of birds in the above cages is $5 + 3 + 2 = 10$.) This is easily visualizable, because these three areas are the ones you see in the figure. However, the probabilities of *only* A and of *only* B are often a pain to deal with, so it's generally easier to think of the area of the union of A and B as the area of A plus the area of B, minus the area of the overlap. This is the way of thinking that corresponds to Eq. (2.12).
2. As we mentioned in Section 2.2.1, you don't *have* to use the above rules of probability to calculate things. You can often instead just count up the various possibilities and solve the problem from scratch. Although, in many cases you're basically doing the same thing with either method, as we saw in the above examples with the die and the cards.
3. As with Eqs. (2.2) and (2.3), the rule in Eq. (2.10) for the "exclusive" case is a special case of the rule in Eq. (2.12) for the "non exclusive" case. This is true because if A and B are exclusive, then $P(A \text{ and } B) = 0$ (by definition). And Eq. (2.12) reduces to Eq. (2.10) when $P(A \text{ and } B) = 0$. Likewise, Fig. 2.4 is a special case of Fig. 2.5 in the case where the regions have zero overlap. There was therefore technically no need to introduce Eq. (2.10) first. We could have started with Eq. (2.12), which covers all possible scenarios, and then showed that it reduces to Eq. (2.10) when the events are exclusive. But as we did in Section 2.2.1, it's usually better to start with a special case and then work up to the more general case.
4. Two events are either independent or dependent, and they are also either exclusive or non exclusive. There are therefore $2 \cdot 2 = 4$ combinations of these characteristics. Let's see which combinations are possible. (You'll need to read this remark very slowly if you want to keep everything straight.)

   - Exclusive and Independent: This combination isn't possible. If two events are independent, then their probabilities are independent of each other, which means that there is a nonzero probability (namely the product of the individual probabilities) that both events happens. They therefore cannot be exclusive.

     Said in another way, if two events A and B are exclusive, then the probability of B given A is zero. But if they are also independent, then the probability of B is independent

of what happens with A. So the probability of B must be zero, period. But such a B should hardly be called an event, because it never happens.

- EXCLUSIVE AND DEPENDENT: This is possible. An example is the events $A = \{$rolling a 2 on one die$\}$ and $B = \{$rolling a 5 on the *same* die$\}$. Another example consists of A as one event and $B = \{$not $A\}$ as the other. In both of these cases the events can't happen at the same time, so they are exclusive. And furthermore the occurrence of one event affects the probability of the other occurring (in that the probability $P(B|A)$ takes the extreme value of zero, due to the exclusive nature of the events), so the events are therefore quite dependent (in a negative sort of way). In short, *all exclusive events are necessarily also dependent.*
- NON EXCLUSIVE AND INDEPENDENT: This is possible. An example is the events $A = \{$rolling a 2 on one die$\}$ and $B = \{$rolling a 5 on *another* die$\}$. Another example is the events $A = \{$getting a Heads on a coin flip$\}$ and $B = \{$getting a Heads on another coin flip$\}$. In both of these cases the events are clearly independent, because they involve different dice or coins. And the events *can* happen at the same time (a fact which is guaranteed by their independence, as mentioned in the "Exclusive and Independent" case above), so they are non exclusive. In short, *all independent events are necessarily also non exclusive.*
- NON EXCLUSIVE AND DEPENDENT: This is possible. An example is having a box with two red balls and two blue balls, with the events being $A = \{$picking a red ball$\}$ and $B = \{$then picking a blue ball *without replacement* after the previous pick$\}$. Another example is picking one card from a deck, with the events being $A = \{$the card is red$\}$ and $B = \{$the card is a heart$\}$. In both of these cases the events are dependent, since the occurrence of A affects the probability of B (in the second case, $P(B|A)$ takes on the extreme value of 1). And the events can happen at the same time, so they are non exclusive.

To sum up, we see that all exclusive events must be dependent, but non exclusive events can be either independent or dependent. Similarly, all independent events must be non exclusive, but dependent events can be either exclusive or non exclusive. These facts are summarized in Fig. 2.7, which indicates which combinations are possible.

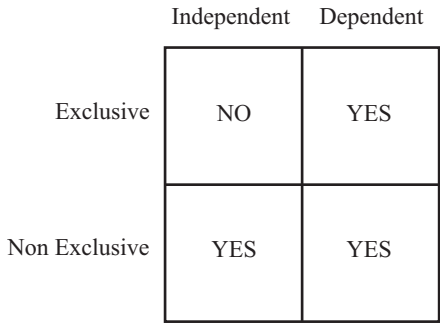


Figure 2.7

This remark was given for curiosity's sake only, in case you were wondering how the dependent/independent characteristic relates to the exclusive/non-exclusive characteristic. There is no need to memorize the results of this remark. Instead, you should think about each situation individually and determine its characteristics from scratch. ♣

2.3 Examples

Let's now do some examples. Introductory probability problems generally fall into a few main categories, so we've broken them up into the various subsections below. There is no

better way to learn how to do probability problems (or any kind of problem, for that matter) than to just sit down and do lots of them, so we've included a bunch!

If the statement of a given problem lists out the specific probabilities of the possible outcomes, then the rules in Section 2.2 are often called for. However, in most problems you encounter, you'll be calculating the probabilities from scratch (by counting things), and so the rules in Section 2.2 generally don't come into play. You simply have to do lots of counting. This will become clear in the examples below. In all of these examples, be sure to try the problem for a few minutes on your own before looking at the solution.

In virtually all of these examples, we will be dealing with situations in which the various possible outcomes are all equally likely. For example, we'll deal with tossing coins, picking cards, forming committees, forming permutations, etc. We will therefore be making copious use of Eq. (2.1),

$$p = \frac{\text{number of desired events}}{\text{total number of possible events}} \qquad (\textit{for equally likely events}) \tag{2.15}$$

We won't, however, specifically state each time that the different outcomes are all equally likely. Just remember that they are, and that this fact is necessary for Eq. (2.1) to be valid.

Before getting into all the examples, let's start off with a problem-solving strategy that comes in very handy in certain situations.

2.3.1 The art of "not"

There are many situations in which the easiest way to calculate the probability of a given event A is not to calculate it directly, but rather to calculate the probability of "not A" and then subtract the result from 1, because we know from Eq. (2.11) that $P(A) = 1 - P(\text{not } A)$.

The most common situation of this type involves a question along the lines of, "What is the probability of obtaining at least one of such-and-such?" The "at least" part makes things difficult, because it could mean one, or two, or three, etc. It will be at best rather messy, and at worst completely intractable, to calculate all of the individual probabilities and then add them up to obtain the answer. The "at least one" question is a far different one from the "exactly one" question.

The key point that simplifies matters is that the only way to *not* get at least one of something is to get exactly zero of it. This means that we can simply calculate the probability of getting zero, and then subtract the result from 1. We therefore need to calculate only *one* probability, instead of a potentially large number of probabilities.

Example (At least one 6): Three dice are rolled. What is the probability of getting at least one 6?

SOLUTION: We'll find the probability of getting zero 6's and then subtract the result from 1. In order to get zero 6's, we must get something other than a 6 on the first die (which happens with 5/6 probability), and likewise also on the second die (5/6 probability again), and likewise also on the third die (5/6 probability again). These are independent events, so the probability of getting zero 6's equals $(5/6)^3 = 125/216$. The probability of getting at least one 6 is therefore $1 - (5/6)^3 = 91/216$.

If you want to solve this problem the hard way, you can add up the probabilities of getting exactly one, two, or three 6's. This is the task of Problem 4.

REMARK: Beware of the following incorrect reasoning for this problem: There is a 1/6 chance of getting a 6 on each of the three rolls. The total probability of getting at least one 6 is therefore $3 \cdot (1/6) = 1/2$. This is incorrect because what we're trying to find is the probability of "a 6 on the

first roll" or "a 6 on the second roll" or "a 6 on the third roll." (This combination of or's is equivalent to there being at least one 6. Remember that when we write "or," we mean the "inclusive or.") But from Eq. (2.10) (or its extension to three events) it is legal to add up the individual probabilities *only if the events are exclusive.* These three events are clearly not exclusive, because it is possible to get, say, a 6 on both the first roll *and* the second roll. We have therefore double counted many of the outcomes, and this is why the incorrect answer of 1/2 is larger than the correct answer of 91/216. If you want to solve the problem in yet another way, you can use the result of Problem 2 to keep track of all the double (and triple) counting. This is the task of Problem 5.

Another way of seeing why the "$3 \cdot (1/6) = 1/2$" reasoning can't be correct is that it would imply that if we had, say, 12 dice, then the probability of getting at least one 6 would be $12 \cdot (1/6) = 2$. But probabilities larger than 1 are nonsensical. ♣

2.3.2 Picking seats

Situations often come up where we need to assign various things to various spots. We'll generally talk about assigning people to seats. There are two common ways to solve problems of this sort: (1) You can count up the number of desired outcomes along with the total number of outcomes, and then take their ratio via Eq. (2.1), or (2) you can imagine assigning the seats one at a time, finding the probability of success at each stage, and using the rules from Section 2.2, or their extensions to more than two events. It's personal preference which method you like to use. But it never hurts to solve a problem both ways, of course, because then you can double check your answer.

Example 1 (Middle in the middle): Three chairs are arranged in a line, and three people randomly take seats. What is the probability that the person with the middle height ends up in the middle position?

First solution: Let the people be labeled from tallest to shortest as 1, 2, and 3. Then the $3! = 6$ possible orderings (all equally likely) that they can take are:

$$123 \qquad 132 \qquad 213 \qquad 231 \qquad 312 \qquad 321 \tag{2.16}$$

We see that two of these (123 and 321) have the middle height in the middle seat. So the probability is $2/6 = 1/3$.

Second solution: Imagine assigning the people randomly to the seats, and let's assign the person with the middle height first (which we are free to do). There is a 1/3 chance that this person ends up in the middle position (or any other position, for that matter). So 1/3 is the desired answer. Nothing fancy going on here.

Third solution: If you insist on assigning the tallest person first, then there is a 1/3 chance that he ends up in the middle seat, in which case there is zero chance that the middle-height person ends up there. And there is a 2/3 chance that the tallest person *doesn't* end up in the middle seat, in which case there is a 1/2 chance that the middle-height person ends up there (because there are two seats remaining, and one yields success). So the total probability that the middle-height person ends up in the middle seat is

$$\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}\,. \tag{2.17}$$

Example 2 (Order of height in a line): Five chairs are arranged in a line, and five people randomly take seats. What is the probability that they end up in order of decreasing height, from left to right?

First solution: There are $5! = 120$ possible arrangements the five people can take in the seats. But there is only one arrangement where they end up in order of decreasing height. So the probability is $1/120$.

Second solution: If we randomly assign the tallest person to a seat, there is a $1/5$ chance that she ends up in the leftmost seat. Assuming that she ends up there, there is a $1/4$ chance that the second tallest person ends up in the second leftmost seat (since there are only four seats left). Likewise, the chances that the other people end up where we want them are $1/3$, then $1/2$, and then $1/1$. (If the first four people end up in the desired seats, the shortest person is guaranteed to end up in the rightmost seat.) So the probability is $1/5 \cdot 1/4 \cdot 1/3 \cdot 1/2 \cdot 1/1 = 1/120$.

The product of these five probabilities comes from the extension of Eq. (2.3) to five events (see Problem 2), which takes the form,

$$\begin{aligned} P(A \text{ and } B \text{ and } C \text{ and } D \text{ and } E) &= P(A) \cdot P(B|A) \cdot P(C|A \text{ and } B) \\ &\quad \cdot P(D|A \text{ and } B \text{ and } C) \\ &\quad \cdot P(E|A \text{ and } B \text{ and } C \text{ and } D). \end{aligned} \tag{2.18}$$

We will use similar extensions repeatedly in the examples below.

Example 3 (Order of height in a circle): Five chairs are arranged in a circle, and five people randomly take seats. What is the probability that they end up in order of decreasing height, going clockwise? (The decreasing sequence of people can start anywhere in the circle. That is, it doesn't matter which seat has the, say, tallest person.)

First solution: As in the previous example, there are $5! = 120$ possible arrangements the five people can take in the seats. But now there are *five* arrangements where they end up in order of decreasing height. This is true because the tallest person can take five possible seats, and once her seat is picked, the positions of the other people are are uniquely determined if they are to end up in order of decreasing height. The probability is therefore $5/120 = 1/24$.

Second solution: If we randomly assign the tallest person to a seat, it doesn't matter where she ends up, because all five seats in the circle are equivalent. But given that she ends up in a certain seat, the second tallest person needs to end up in the seat next to her in the clockwise direction. This happens with probability $1/4$. Likewise, the third tallest person has a $1/3$ chance of ending up in the next seat in the clockwise direction (given that the first two people ended up in the proper order). And then $1/2$ for the fourth tallest person, and $1/1$ for the shortest person. The probability is therefore $1/4 \cdot 1/3 \cdot 1/2 \cdot 1/1 = 1/24$.

If you want, you can preface this product with a "5/5" for the tallest person, because there are 5 possible spots she can take (this is the denominator), and also 5 "successful" spots she can take, because it doesn't matter where she ends up (this is the numerator).

Example 4 (Three girls and three boys): Six chairs are arranged in a line, and three girls and three boys randomly pick seats. What is the probability that the three girls end up in the three leftmost seats?

First solution: There are $3! = 6$ different ways that the three girls can be arranged in the three leftmost seats, and $3! = 6$ different ways that the three boys can be arranged in the other three (the rightmost) seats. So the total number of "successful" arrangements is

$3! \cdot 3! = 36$. Since total number of possible arrangements is $6! = 720$, the desired probability is $3!3!/6! = 36/720 = 1/20$.

SECOND SOLUTION: Let's assume that the girls pick their seats one at a time. The first girl has a $3/6$ chance of picking one of the three leftmost seats. Then, given that she is successful, the second girl has a $2/5$ chance of success, because only two of the remaining five seats are among the left three. And finally, given that she too is successful, the third girl has a $1/4$ chance of success, because only one of the remaining four seats is among the left three. The desired probability is therefore $3/6 \cdot 2/5 \cdot 1/4 = 1/20$.

THIRD SOLUTION: The $3!3!/6!$ result in the first solution looks suspiciously like the binomial coefficient $\binom{6}{3} = 6!/3!3!$, so it suggests that there is another way to solve this problem. And indeed, imagine randomly picking three of the six seats for the girls. There are $\binom{6}{3}$ ways to do this. Only one of these is the successful result of the three leftmost seats, so the desired probability is $1/\binom{6}{3} = 6!/3!3! = 1/20$.

2.3.3 Socks in a drawer

Picking colored socks from a drawer is a classic probabilistic setup. If you want to deal with such setups by counting things, then subgroups and binomial coefficients will come into play. If, however, you want to deal with them by picking the socks in succession, then you'll end up multiplying various probabilities and using the rules from Section 2.2.

Example 1 (Two blue and two red): A drawer contains two blue socks and two red socks. If you randomly pick two socks, what is the probability that you get a matching pair?

FIRST SOLUTION: There are $\binom{4}{2} = 6$ possible pairs you can pick. Of these, two are matching pairs. So the probability is $2/6 = 1/3$. If you want to list out all the pairs, they are (with 1 and 2 being the blue socks, and 3 and 4 being the red socks):

$$\mathbf{1,2} \qquad 1,3 \qquad 1,4 \qquad 2,3 \qquad 2,4 \qquad \mathbf{3,4} \tag{2.19}$$

The pairs in bold are the matching ones.

SECOND SOLUTION: After you pick the first sock, there is one sock of that color (whatever it may be) left in the drawer and two of the other color. So of the three socks left, one gives you a matching pair, and two don't. So the desired probability is $1/3$.

Example 2 (Four blue and two red): A drawer contains four blue socks and two red socks (Fig. 2.8). If you randomly pick two socks, what is the probability that you get a matching pair?

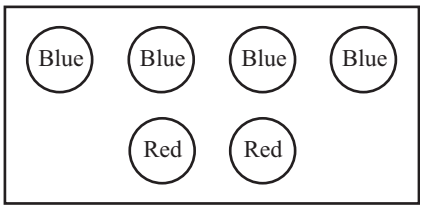


Figure 2.8

FIRST SOLUTION: There are $\binom{6}{2} = 15$ possible pairs you can pick. Of these, there are $\binom{4}{2} = 6$ blue pairs and $\binom{2}{2} = 1$ red pair. So the desired probability is

$$\frac{\binom{4}{2} + \binom{2}{2}}{\binom{6}{2}} = \frac{7}{15}\,. \tag{2.20}$$

SECOND SOLUTION: There is a 4/6 chance that the first sock you pick is blue. If this happens, there is a 3/5 chance that the second sock you pick is also blue (because there are three blue and two red socks left in the drawer). On the other hand, there is a 2/6 chance that the first sock you pick is red. If this happens, there is a 1/5 chance that the second sock you pick is also red (because there are one red and four blue socks left in the drawer). The probability that the socks match is therefore ("$B1$" stands for a blue sock on the first pick, etc.):

$$\begin{aligned} P(B1 \text{ and } B2) + P(R1 \text{ and } R2) &= P(B1)\cdot P(B2|B1) + P(R1)\cdot P(R2|R1) \\ &= \frac{4}{6}\cdot\frac{3}{5} + \frac{2}{6}\cdot\frac{1}{5} \\ &= \frac{14}{30} = \frac{7}{15}\,. \end{aligned} \tag{2.21}$$

2.3.4 Coins and dice

There's never any shortage of probability examples with flipping coins and rolling dice.

Example 1 (One of each number): Six dice are rolled. What is the probability of getting exactly one of each of the numbers 1 through 6?

FIRST SOLUTION: The total number of possible outcomes for what all six dice show is 6^6, because there are six possibilities for each of the dice. How many desired outcomes are there that have each number appearing once? Well, this is simply the question of how many permutations there are of six numbers, because we need all six numbers to appear, but it doesn't matter in what order. And we know from Section 1.2 that the answer to this question is 6!. So the probability is

$$\frac{6!}{6^6} = \frac{5}{324} \approx 1.5\%. \tag{2.22}$$

SECOND SOLUTION: Let's imagine rolling six dice in succession, with the goal of having each number appear once. On the first roll, we simply get what we get, and there's no way to fail. So the probability of success on the first roll is 1. However, on the second roll, we don't want to get a repeat of the number that appeared on the first roll (whatever that number may be). Since there are five "good" options left, the probability of success on the second roll is 5/6. On the third roll, we don't want to get a repeat of either of the numbers that appeared on the first and second rolls, so the probability of success on the third roll (given success on the first two rolls) is 4/6. Likewise, the fourth roll has a 3/6 chance of success, the fifth has 2/6, and the sixth has 1/6. The probability of complete success all the way through is therefore

$$1\cdot\frac{5}{6}\cdot\frac{4}{6}\cdot\frac{3}{6}\cdot\frac{2}{6}\cdot\frac{1}{6} = \frac{5}{324}\,, \tag{2.23}$$

in agreement with the first solution. Note that if we write the "1" here as 6/6, then this expression becomes $6!/6^6$, which makes it clear why it agrees with the first solution.

Example 2 (Three pairs): Six dice are rolled. What is the probability of getting three pairs (that is, three different numbers that each appear twice)?

SOLUTION: We'll count the total number of ways to get three pairs, and then we'll divide by the total number of possible rolls for all six dice together, which is 6^6.

There are two steps in the counting. First, how many different ways can we pick the three different numbers that show up? We need to pick three numbers from six numbers, so the number of ways is $\binom{6}{3} = 20$.

Second, given the three numbers that show up, how many different ways can we get two of each of them? Let's says the numbers are 1, 2, and 3. We can imagine plopping two of each of these numbers down on six blank spots on a piece of paper (which represent the six dice). There are $\binom{6}{2} = 15$ ways to pick the two spots where the 1's go. And then there are $\binom{4}{2} = 6$ ways to pick where the two 2's go in the four remaining spots. And then finally there is $\binom{2}{2} = 1$ way to pick where the two 3's go in the two spots remaining spots.

The total number of ways to get three pairs is therefore $\binom{6}{3}\cdot\binom{6}{2}\cdot\binom{4}{2}\cdot\binom{2}{2}$, and so the probability of getting three pairs is

$$p = \frac{\binom{6}{3}\cdot\binom{6}{2}\cdot\binom{4}{2}\cdot\binom{2}{2}}{6^6} = \frac{20\cdot 15\cdot 6\cdot 1}{6^6} = \frac{25}{648} \approx 3.9\%. \tag{2.24}$$

If you try to solve this problem in a manner analogous to the second solution in the previous example (that is, by multiplying probabilities for the successive rolls), things get very messy because there are many different scenarios that lead to three pairs.

Example 3 (Five coin flips): A coin is flipped five times. Calculate the probabilities of getting all the various possible numbers of Heads (0 through 5).

SOLUTION: We'll count the number of ways to get the different numbers of Heads, and then we'll divide by the total number of possible outcomes for the five rolls, which is 2^5.

There is only $\binom{5}{0} = 1$ way to get zero Heads, namely TTTTT. There are $\binom{5}{1} = 5$ ways to get one Heads (such as HTTTT), because there are $\binom{5}{1}$ ways to choose the one coin that shows Heads. There are $\binom{5}{2} = 10$ ways to get two Heads, because there are $\binom{5}{2}$ ways to choose the two coins that show Heads. And so on. Therefore, the various probabilities are (with the subscript denoting the number of Heads):

$$\begin{aligned} & P_0 = \frac{\binom{5}{0}}{2^5}\,, \quad P_1 = \frac{\binom{5}{1}}{2^5}\,, \quad P_2 = \frac{\binom{5}{2}}{2^5}\,, \quad P_3 = \frac{\binom{5}{3}}{2^5}\,, \quad P_4 = \frac{\binom{5}{4}}{2^5}\,, \quad P_5 = \frac{\binom{5}{5}}{2^5} \\ \implies & P_0 = \frac{1}{32}\,, \quad P_1 = \frac{5}{32}\,, \quad P_2 = \frac{10}{32}\,, \quad P_3 = \frac{10}{32}\,, \quad P_4 = \frac{5}{32}\,, \quad P_5 = \frac{1}{32}\,. \end{aligned} \tag{2.25}$$

The sum of all these probabilities equals 1, and this is true for any number of flips. The physical reason is that the number of Heads must be *something*, so the sum of all the probabilities must be 1. The mathematical reason is that the sum of the binomial coefficients (the numerators in the preceding fractions) equals 2^N (the common denominator). See Section 1.5.3 for the explanation of this.

2.3.5 Cards

We already did a lot of card counting in Chapter 1 (in particular in Problem 1), and some of those results will be applicable here. There is effectively an endless number of probability questions we can ask about cards.

Example 1 (Royal flush from seven cards): A few variations of poker involve being dealt seven cards (in one way or another) and forming the best 5-card hand that can be made from the seven cards. What is the probability of being able to form a Royal flush in this setup? (A Royal flush consists of 10,J,Q,K,A, all from the same suit.)

SOLUTION: The total number of possible 7-card hands is $\binom{52}{7} = 133,784,560$. The number of 7-card hands that contain a Royal flush is $4 \cdot \binom{47}{2} = 4,324$, because there are four ways to choose the five Royal flush cards (the four suits), and then $\binom{47}{2}$ ways to choose the other two cards in the hand from the remaining $52 - 5 = 47$ cards in the deck. The probability is therefore

$$\frac{4 \cdot \binom{47}{2}}{\binom{52}{7}} = \frac{4,324}{133,784,560} \approx 0.0032\%. \tag{2.26}$$

This is a bit larger than the result for 5-card hands. In that case, only four of the $\binom{52}{5} = 2,598,960$ hands are Royal flushes, so the probability is $4/2,598,960 \approx 0.00015\%$, which is about 20 times smaller.

Example 2 (Suit full house): In a 5-card poker hand, what is the probability of getting a "full house" of suits, that is, three cards of one suit and two of another? (This isn't an actual poker hand worth anything, but that won't stop us from calculating the probability!) How does your answer compare with the probability of getting an actual full house, that is, three cards of one value and two of another? Feel free to use the result from part (a) of Problem 1 in Chapter 1.

SOLUTION: There are four ways to choose the suit that appears three times, and $\binom{13}{3} = 286$ ways to choose the specific three cards from the 13 of this suit. And then there are three ways to choose the suit that appears twice from the remaining three suits, and $\binom{13}{2} = 78$ ways to choose the specific two cards from the 13 of this suit. The total number of suit-full-house hands is therefore $4 \cdot \binom{13}{3} \cdot 3 \cdot \binom{13}{2} = 267,696$. Since there are a total of $\binom{52}{5}$ possible hands, the desired probability is

$$\frac{4 \cdot \binom{13}{3} \cdot 3 \cdot \binom{13}{2}}{\binom{52}{5}} = \frac{267,696}{2,598,960} \approx 10.3\%. \tag{2.27}$$

From part (a) of Problem 1 in Chapter 1, the total number of actual full-house hands is 3,744, which yields a probability of $3,744/2,598,960 \approx 0.14\%$. It is therefore *much* more likely (by a factor of about 70) to get a full house of suits than an actual full house of values. This makes intuitive sense; there are more values than suits (13 compared with 4), so it is harder to have all five cards consist of just two values than just two suits.

Example 3 (Only two suits): In a 5-card poker hand, what is the probability of having all the cards be members of at most two suits (a single suit is allowed)? The suit full house in the previous example is a special case of "at most two suits." *Hint:* This is a little tricky, at least if you solve it a certain way; be careful about double counting some of the hands!

FIRST SOLUTION: There are $\binom{4}{2} = 6$ ways to pick the two suits that appear. For a given choice of two suits, there are $\binom{26}{5}$ ways to pick the five cards from the $2 \cdot 13 = 26$ cards in these two suits. It therefore seems that there should be $\binom{4}{2} \cdot \binom{26}{5} = 394,680$ different hands that consist of cards from at most two suits.
However, this isn't correct, because we double (or actually triple) counted the hands that involve only one suit (the flushes). For example, if all five cards are hearts, then we counted

such a hand in the heart/diamond set of $\binom{26}{5}$ hands, and also in the heart/spade set, and also in the heart/club set. We counted it three times when we should have counted it only once. Since there are $\binom{13}{5}$ hands that are heart flushes, we have included an extra $2 \cdot \binom{13}{5}$ hands, so we need to subtract these from our total. Likewise for the diamond, spade, and club flushes. The total number of hands that involve at most two suits is therefore $\binom{4}{2}\binom{26}{5} - 8 \cdot \binom{13}{5} = 394,680 - 10,296 = 384,384$. The desired probability is then

$$\frac{\binom{4}{2}\binom{26}{5} - 8 \cdot \binom{13}{5}}{\binom{52}{5}} = \frac{384,384}{2,598,960} \approx 14.8\%. \tag{2.28}$$

This is larger than the result in Eq. (2.27), as it should be, because suit full houses are a subset of the hands that involve at most two suits.

Second Solution: There are three general ways we can have at most two suits: (1) All five cards can be of the same suit (a flush), (2) Four cards can be of one suit, and one of another, or (3) Three cards can be of one suit, and two of another (this is the suit full house from the previous example). We will denote these types by $(5,0)$, $(4,1)$, and $(3,2)$, respectively.

There are $4 \cdot \binom{13}{5} = 5,148$ hands of the $(5,0)$ type (see part (c) of Problem 1 in Chapter 1). And from the previous example, there are $4 \cdot \binom{13}{3} \cdot 3 \cdot \binom{13}{2} = 267,696$ hands of the $(3,2)$ type. So we need to figure out only the number of hands of the $(4,1)$ type. From exactly the same kind of reasoning as in the previous example, this number is $4 \cdot \binom{13}{4} \cdot 3 \cdot \binom{13}{1} = 111,540$. Adding up these three results gives the total number of "at most two suits" hands as

$$\begin{aligned} 4 \cdot \binom{13}{5} + 4 \cdot \binom{13}{4} \cdot 3 \cdot \binom{13}{1} + 4 \cdot \binom{13}{3} \cdot 3 \cdot \binom{13}{2} &= 5,148 + 111,540 + 267,696 \\ &= 384,384, \end{aligned} \tag{2.29}$$

in agreement with the first solution.[6] The hands of the $(3,2)$ type account for about 2/3 of this total, consistent with the fact that the 10.3% result in Eq. (2.27) is about 2/3 of the 14.8% result in Eq. (2.28).

2.4 Two classic examples

No book on probability would be complete without a discussion of the "Birthday Problem" and the "Game Shown Problem." Both of these problems have answers that may seem counterintuitive at first, but they eventually make sense if you think about them long enough!

2.4.1 The Birthday Problem

Let's look at the Birthday Problem first. Aside from being a very interesting problem, its unexpected result allows you to take advantage of unsuspecting people and win money on bets at parties (as long as they're big enough parties, as we'll see). The problem is the following.

Problem: How many people need to be in a room in order for the probability to be greater than 1/2 that at least two of them have the same birthday? By "same birthday", we mean the same day of the year; the year may differ. Ignore leap years.

[6] The repetition of the "384" here is due in part to the factors of 13 and 11 in all of the terms on the lefthand side. These numbers are factors of 1001.

Solution: If there was ever a problem that called for the strategy in the "The art of 'not'" section above, this is it. There are many different ways for there to be at least one common birthday (one pair, two pairs, one triple, etc.), and it is essentially impossible to calculate each of these probabilities individually and add them up. It is *much* easier (and even with the italics, this is still a vast understatement) to calculate the probability that there *isn't* a common birthday, and then subtract this from 1 to obtain the probability that there *is* a common birthday.

The calculation of the probability that there *isn't* a common birthday proceeds as follows. Let's say there are n people in the room. We can imagine taking them one at a time and randomly plopping their names down on a calendar, with the present goal being that there are no common birthdays. The first name can go anywhere. But when we plop down the second name, there are only 364 "good" days left, because we don't want it to coincide with the first name. Then when we plop down the third name, there are only 363 "good" days left (assuming that the first two people don't have the same birthday), because we don't want it to coincide with either of the first two. Similarly, when we plop down the fourth name, there are only 362 "good" days left (given that the first three people don't have a common birthday), because we don't want it to coincide with any of the first three. And so on.

So if there are n people in the room, the probability that there *isn't* a common birthday (hence the superscript "not") among any of the people is

$$P_n^{\text{not}} = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdot \frac{361}{365} \cdot \ldots \cdot \frac{365-(n-1)}{365}. \tag{2.30}$$

If you want, you can write the "1" here as 365/365 to make things look nicer. Note that the last term involves $(n-1)$ and not n, because $(n-1)$ is the number of names that have already been plopped down. As a double check that this $(n-1)$ is correct, you can simply check that it works for a small number like $n = 2$ or $n = 3$. (You should always perform a simple check like this whenever you write down *any* formula!)

We now just have to multiply the above product out to the point where the product is less than 1/2 (so that the probability that there *is* a common birthday is greater than 1/2). This is tedious, but not horribly painful. We find that $P_{22}^{\text{not}} = 0.524$, and $P_{23}^{\text{not}} = 0.493$. If P_n^{yes} is the probability that there *is* a common birthday among n people, then $P_n^{\text{yes}} = 1 - P_n^{\text{not}}$, and so $P_{22}^{\text{yes}} = 0.476$ and $P_{23}^{\text{yes}} = 0.507$. Since our original goal was to have $P_n^{\text{yes}} > 1/2$ (or equivalently $P_n^{\text{not}} < 1/2$), we see that there must be at least 23 people in a room in order for the odds to favor at least two of them having the same birthday. And then the probability in that case is 50.7%.

The task of Problem 7 is to calculate the probability that among 23 people, *exactly* two of them have a common birthday. That is, there aren't two different pairs with common birthdays, or a triple with the same birthday, etc.

REMARK: This answer of $n = 23$ is much smaller than most people expect. So as we mentioned above, it provides a nice betting opportunity. For $n = 30$, the odds of a common birthday increase to 70.6%, and most people still find it hard to believe that among 30 people there are probably two who have the same birthday. The table below lists various values of n and the probabilities, $P_n^{\text{yes}} = 1 - P_n^{\text{not}}$, that at least two people have a common birthday.

n	10	20	23	30	50	60	70	100
P_n^{yes}	11.7%	41.1%	50.7%	70.6%	97.0%	99.4%	99.92%	99.9994%

Even for $n = 50$, most people would be happy to bet, at even odds, that no two people have the same birthday. But you'll win the bet 97% of the time.

One reason why many people can't believe the $n = 23$ result is that they're asking themselves a different question, namely, "How many people need to be present for there to be a 1/2 chance that someone else has *my* birthday?" The answer to this question is indeed much larger than 23. The probability that no one out of n people has a birthday on a given day is simply $(364/365)^n$, since each person has a 364/365 chance of not having that birthday. For $n = 252$, this is just over 1/2. And for $n = 253$, it is just under 1/2. Therefore, you need to come across 253 other people in order for there to be a greater than 1/2 chance that at least one of them has *your* birthday. ♣

2.4.2 The Game Show Problem

We'll now discuss the Game Show Problem. This problem not only has a variety of common incorrect solutions, it also has a long history of people arguing vehemently in favor of these incorrect solutions.

Problem: A game show host offers you the choice of three doors. Behind one of these doors is the grand prize, and behind the other two are goats. The host announces that after you select a door (without opening it), he will open one of the other two doors and purposefully reveal a goat (he knows what's behind each of the doors). You select a door. The host then opens one of the other doors and reveals the promised goat. He then offers you the chance to switch your choice to the remaining door. Should you switch or not? Or does it not matter?

Solution: We'll present three solutions, one right and two wrong. You should decide which one you think is correct before reading beyond the third solution. Cover up the page with a piece of paper so you don't inadvertently see which one is correct.

1. Reasoning 1: Once the host reveals a goat, the prize must be behind one of the two remaining doors. Since the prize was randomly placed to begin with, there must be equal chances of the prize being behind each of the doors. The probabilities are therefore both 1/2, so it doesn't matter if you switch.

   If you want, you can imagine a friend entering the room *after* the host opens the door, but he is still aware of the whole procedure of the host announcing that he will open a door to reveal a goat. This person sees two identical unopened doors and a goat, so there must be a 1/2 chance that the prize is behind each unopened door. The probabilities that you and your friend measure can't be any different, so you also say that each unopened door has a 1/2 chance of containing the prize. So it doesn't matter if you switch.

2. Reasoning 2: There is initially a 1/3 chance the prize is behind any of the three doors. So if you don't switch, your probability of winning equals 1/3. No actions taken by the host can change the fact that if you play a large number, N, of these games, then (roughly) $N/3$ of them will have the prize behind the door you initially pick.

   Likewise, if you switch to another door, there is a 1/3 chance that the prize is behind that door. (There is obviously a goat behind at least one of the other two doors, so the fact that the host reveals a goat doesn't tell you anything new.) Therefore, since the probability is 1/3 whether or not you switch, it doesn't matter if you switch.

3. Reasoning 3: As in the first paragraph in Reasoning 2, if you don't switch, your probability of winning equals 1/3.

   However, if you switch, your probability of winning turns out to be greater than 1/3. It increases to 2/3. This can be seen as follows. Without loss of generality, assume that

you pick the first door. (You can repeat the following reasoning again for the other doors if you wish. It gives the same result.) There are three equally likely possibilities for what is behind the three doors: PGG, GPG, and GGP, where P denotes prize, and G denotes goat. If you don't switch, then in only the first of these three cases do you win, so your odds of winning are 1/3 (this is consistent with the first paragraph of Reasoning 2). If you do switch, then in the first case you lose, but in the other two you win (because the door not opened by the host has the prize). Therefore, your odds of winning are 2/3. So you do in fact want to switch.

Which of these three reasonings is correct? Don't read any further until you've firmly decided which one you think is right.

The third reasoning is correct. The error in the first reasoning is the statement, "there must be equal chances of the prize being behind each of the doors." This is simply not true. The act of revealing a goat breaks the symmetry between the two remaining doors, as explained in the third reasoning. The fact that there are two possibilities doesn't mean that their probabilities have to be equal, of course!

The error in the supporting reasoning with your friend (who enters the room after the host opens the door) is the following. While it *is* true that the probabilities are both 1/2 for your friend, they aren't 1/2 for *you*. The statement, "the probabilities can't be any different for you," is false. You have information that your friend doesn't have, namely, you know which of the two unopened doors is the one you initially picked, and which is the door that the host chose to leave unopened. (And as seen in the third solution, this information yields probabilities of 1/3 and 2/3.) Your friend doesn't have this critical information. Both doors look the same to him. Probabilities can certainly be different for different people. If I flip a coin and peek and see a Heads, but I don't show you, then the probability of a Heads is 1/2 for you, but 1 for me.

The error in the second reasoning is that the act of revealing a goat *does* give you new information, as we just noted. This information tells you that the prize isn't behind that door, and it also distinguishes between the two remaining unopened doors (one is the door you initially picked, and one is among the two that you didn't pick). And as seen in the third solution, this information has the effect of increasing the probability that the goat is behind the other door. Note that another reason why this solution can't be correct is that the probabilities don't add up to 1.

To sum up, it should be no surprise that the probabilities are different for the switching and non-switching strategies *after* the host opens a door (the odds are obviously the same, equal to 1/3, whether or not a switch is made *before* the host opens a door), because the host gave you some of the information he had about the locations of things.

Remarks:

1. If you still doubt the validity of Reasoning 3, imagine a situation with 1000 doors containing one prize and 999 goats. After you pick a door, the host opens 998 other doors to reveal 998 goats (and he said beforehand that he was going to do this). In this setup, if you don't switch, your chances of winning are 1/1000. But if you do switch, your chances of winning are 999/1000, which can be seen by listing out (or imagining listing out) the 1000 cases, as we did with the three PGG, GPG, and GGP cases in Reasoning 3 above. It is clear that the switch should be made, because the *only* case where you lose after you switch is the case where you had initially picked the prize, and this happens only 1/1000 of the time.

   In short, a huge amount of information is gained by the revealing of 998 goats. There's a 999/1000 chance that the prize is somewhere behind the other 999 doors, and the host is kindly giving you the information of exactly which one it is.

2. The clause in the statement of the problem, "The host announces that after you select a door (without opening it), he will open one of the other two doors and purposefully reveal a goat," is crucial. If it is omitted, and it is simply stated that, "The host then opens one of the other doors and reveals a goat," then it is impossible to state a preferred strategy. If the host doesn't announce his actions beforehand, then for all you know, he *always* reveals a goat (in which case you should switch, as we saw above). Or he *randomly* opens a door and just happened to pick a goat (in which case it doesn't matter whether you switch, as you can show in Problem 8). Or he opens a door and reveals a goat if and only if your initial door has the prize (in which case you definitely should not switch). Or he could have one procedure on Tuesdays and another on Fridays, each of which depends on the color of socks he's wearing. And so on and so forth.

3. As mentioned above, this problem is infamous for the intense arguments it lends itself to. There's nothing bad about getting the wrong answer, nor is there anything bad about not believing the correct answer for a while. But concerning arguments that drag on and on, I think it should be illegal to argue about this problem for more than 15 minutes, because at that point everyone should simply stop and *play the game*. You can play a number of times with the switching strategy, and then a number of times with the non-switching strategy. Three coins with a dot on the bottom of one of them are all you need.[7] Not only will the actual game yield the correct answer (if you play enough times so that things average out), but the patterns that form when playing will undoubtedly convince you of the correct reasoning (or reinforce it, if you're already comfortable with it). Arguing endlessly about an experiment, when you can actually *do* the experiment, is as silly as arguing endlessly about what's behind a door, when you can simply open the door.

4. For completeness, there is one subtlety we should mention here. In Reasoning 2 above, we stated, "No actions taken by the host can change the fact that if you play a large number, N, of these games, then (roughly) $N/3$ of them will have the prize behind the door you initially pick." This part of the reasoning was correct; it was the "switching" part of Reasoning 2 that was incorrect. After doing Problem 8 (where the host randomly opens a door), you might disagree with this statement, because it will turn out in that problem that the actions taken by the host *do* affect this $N/3$ result. However, the above statement is still correct for "*these* games" (the ones governed by the original statement of this problem). See the third remark in the solution to Problem 8 for further discussion. ♣

2.5 Expectation value

The *expectation value* (or *expected value*) for a process is the expected average of a large number of trials of the process. So in some sense, the expectation value is simply a fancy name for the average. However, these two terms have different usages: The word "average" is generally associated with trials that have already taken place, whereas "expectation value" refers to the average that you would expect to obtain in trials yet to be carried out.

As an example, consider the roll of a die. Since the numbers 1 through 6 are all equally probable, the expectation value is just their average, which is 3.5. Of course, if you roll one die, there is no chance that you will actually obtain a 3.5, because you can roll only the integers 1 through 6. But this is irrelevant as far as the expectation value goes, because we're concerned only with the expected *average* value of a large number of trials. An expectation value of 3.5 is simply a way of saying that if you roll a die 1000 times and add up all the results, you should get a total of about 3500. Again, it's extremely unlikely (but not impossible in this case) that you will get a total of *exactly* 3500, but this doesn't matter for

[7] You actually don't need three objects (it's hard to find three identical coins anyway). The "host" can simply roll a die, without showing the "contestant" the result. A 1 or 2 can mean that the prize is placed behind the first door, a 3 or 4 the second, and a 5 or 6 the third. The game then basically involves calling out door numbers.

the expectation value.[8]

In order for an expectation value to exist, we need each possible outcome to be associated with a number, because we need to be able to take the average of the outcomes, or actually the *weighted average*; see Eq. (2.33) below. If there were no actual numbers around, it would be impossible to form an average. For example, let's say we draw a card from a deck, and let's assume that we're concerned only with its suit. Then it makes no sense to talk about the expected value of the suit, because it makes no sense to take an average of a heart, diamond, spade, and club. If, however, we assign the "suit values" of 1 through 4, respectively, to these suits, then it does make sense to talk about the expected value of the "suit value," and it happens to be 2.5 (the average of 1 though 4).

The above example with the rolled die consisted of six equally likely outcomes, so we found the expectation value by simply taking the average of the six outcomes. But what if the outcomes have different probabilities? For example, what if we have three balls in a box with two of them labeled with a "1" and one labeled with a "4"? If we pick a ball, what is the expectation value of the resulting number? To answer this, imagine performing a large number of trials of the process. Let's be general and denote this large number by N. Since the probability of picking a 1 is 2/3, we expect about $(2/3) \cdot N$ of the numbers to be a 1. Likewise, about $(1/3) \cdot N$ of the numbers should be a 4. The total sum of all the numbers should therefore be about $(2/3)N \cdot 1 + (1/3)N \cdot 4$. To obtain the expected average, we just need to divide this result by N, which gives

$$\text{expectation value} = \frac{(2/3)N \cdot 1 + (1/3)N \cdot 4}{N} = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 4 = 2. \tag{2.31}$$

Note that the N's canceled out, so the result is independent of N. This is how it should be, because the expected average value shouldn't depend on how many trials you do.

In general, if the probabilities are p_1 and p_2 instead of 2/3 and 1/3, and if the outcomes are R_1 and R_2 instead of 1 and 4, you can carry through the exact same reasoning as above to show that the expectation value is

$$\begin{aligned} \text{expectation value} &= \frac{(p_1 N) \cdot R_1 + (p_2 N) \cdot R_2}{N} \\ &= p_1 \cdot R_1 + p_2 \cdot R_2. \end{aligned} \tag{2.32}$$

What if we have more than two possible outcomes? The same reasoning works again, but now with more terms in the sum. You can quickly verify (by imagining a large number of trials, N) that if the probabilities are $p_1, p_2, \ldots, p_n$, and if the outcomes are $R_1, R_2, \ldots, R_n$, then the expectation value is

$$\boxed{\text{expectation value} = p_1 \cdot R_1 + p_2 \cdot R_2 + \cdots + p_n \cdot R_n} \tag{2.33}$$

This is the so-called *weighted average* of the outcomes, where each outcome is weighted (that is, multiplied) by its probability. This weighting has the effect of making outcomes with larger probabilities contribute more to the expectation value. This makes sense, because these outcomes occur more often, so they should influence the average more than outcomes that occur less often.

[8]The colloquial use of the word "expected" can cause some confusion, because you might think that the expected value is the value that is most likely to occur. This is *not* the case. If we have a process with four equally likely outcomes, 1,2,2,7, then even though 2 is the most likely value, the "expected value" is the average of the numbers, which is 3 (which happens to never occur).

Example 1 (Expected number of Heads): If you flip a coin four times, what is the expected value of the number of Heads you get?

SOLUTION: Without doing any work, we know that the expected number of Heads is 2, because on average half the coins will be Heads and half will be Tails.

But let's solve this again by using Eq. (2.33). By looking at the 16 equally likely outcomes in Table 1.8 in Section 1.4, the probabilities of getting 0, 1, 2, 3, or 4 Heads are, respectively, 1/16, 4/16, 6/16, 4/16, and 1/16. So Eq. (2.33) gives the expectation value of the number of Heads as

$$\frac{1}{16}\cdot 0 + \frac{4}{16}\cdot 1 + \frac{6}{16}\cdot 2 + \frac{4}{16}\cdot 3 + \frac{1}{16}\cdot 4 = \frac{32}{16} = 2. \tag{2.34}$$

Example 2 (Flip until Tails): If you flip a coin until you get a Tails, what is the expected total number of coins you flip?

SOLUTION: There is a 1/2 chance that you immediately get a Tails, in which case you flip only one coin. There is a $1/2 \cdot 1/2 = 1/4$ chance that you get a Heads then a Tails, in which case you flip two coins. There is a $1/2 \cdot 1/2 \cdot 1/2 = 1/8$ chance that you get a Heads, then another Heads, then a Tails, in which case you flip three coins. And so on. So the expectation value of the total number of coins is

$$\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{32}\cdot 5 + \cdots \tag{2.35}$$

This sum technically has an infinite number of terms, although they eventually become negligibly small. The sum is a little tricky to calculate (see Problem 9 if you're interested in the trick, but it's not important). However, if you use a calculator to add up the first dozen or so terms, it becomes clear that the sum approaches 2. You are encouraged to verify this experimentally by doing 50 (or so) trials.

The expectation value plays an important role in betting and decision making, because it is the amount on money you should be willing to pay up front in order to have a "fair game." By this we mean the following. Consider a game in which you can win various amounts of money based on the various possible outcomes. For example, let's say that you roll a die and that your winnings equal the resulting number (in dollars). How much money should you be willing to pay to play this game? Also, how much money should the "house" (the people running the game) be willing to charge you for the opportunity to play the game? You certainly shouldn't pay, say, \$6 each time you play it, because at best you will break even, and most of the time you will lose money. On average, you will win the average of the numbers 1 through 6, which is \$3.50, so this is the most that you should be willing to pay for each trial of the game. If you pay more than this, then you will lose money on average. Conversely, the "house" should charge you *at least* \$3.50 to play the game each time, because otherwise it will lose money on average.

Putting these two results together, we see that \$3.50 is the amount the game should cost *if the goal is to have a fair game*, that is, a game where neither side wins any money on average. Of course, in games run by casinos and such, things are arranged so that you pay *more* than the expectation value. So on average the house wins, which is consistent with the fact that casinos stay in business.

Note the italics in the previous paragraph. These are important, because when real-life considerations are taken into account, there might very well be goals that supersede the goal of having a fair game. The above discussion should therefore *not* be taken to imply

that you should *always* play a game if the fee is smaller than the expectation value, or that you should *never* play a game if the fee is larger than the expectation value. It depends on the circumstances. See Problem 10 for a discussion of this.

2.6 Stirling's formula

Stirling's formula gives an approximation to $N!$ which is valid for large N, in the sense that the larger N is, the better the approximation is. It is given by:

$$\boxed{N! \approx N^N e^{-N}\sqrt{2\pi N}} \qquad (Stirling's\ formula) \tag{2.36}$$

Here e is the "natural logarithm," equal to $e \approx 2.71828$ (see Appendix B for a discussion of e). There are various proofs of Stirling's formula, but they generally involve calculus, so we'll just accept the formula here. The formula does indeed give an accurate approximation to $N!$ (an extremely accurate one, if N is large), as you can see from Table 2.3, where $S(N)$ stands for the $N^N e^{-N}\sqrt{2\pi N}$ Stirling approximation. The approximation gets better as N gets larger. But even if N is just 10, the approximation is still off by only about 0.8%. And although there's never much need to use the formula for small numbers like 1 or 5, it works surprisingly well in those cases too.

N	$N!$	$S(N)$	$S(N)/N!$
1	1	0.922	0.922
5	120	118.0	0.983
10	$3.629 \cdot 10^6$	$3.599 \cdot 10^6$	0.992
100	$9.3326 \cdot 10^{157}$	$9.3249 \cdot 10^{157}$	0.9992
1000	$4.02387 \cdot 10^{2567}$	$4.02354 \cdot 10^{2567}$	0.99992

Table 2.3

You will note that for the powers of 10, the ratios of $S(N)$ to $N!$ all take the same form, namely decimals with an increasing number of 9's and then a 2 (it's actually not a 2, because we rounded off, but it's the same rounding off for all the numbers). This isn't a coincidence. It follows from a more accurate version of Stirling's formula, but we won't get into that here.

Stirling's formula will be critical in Chapter 3 when we talk about distributions. But for now, its usefulness arises in situations involving the binomial coefficients of large numbers, because these binomial coefficients in turn involve the factorials of large numbers. There are main two benefits to using Stirling's formula, both of which are illustrated in the example below:

- Depending on the type of calculator you have, you might get an error message when you plug in the factorial of a number that's too big. Stirling's formula lets you avoid this problem if you first simplify the expression that results from Stirling's formula (using the letter N to stand for the specific number you're dealing with), and *then* plug the simplified result into your calculator.

- If you use Stirling's formula (in terms of the letter N) and arrive at a simplified answer in terms of N (we'll call this a *symbolic* answer since it's written in terms of the symbol N instead of specific numbers), you can then plug in your specific value of N. Or any other value, for that matter. The benefit of having a symbolic answer in terms of N is that you don't have to solve the problem from scratch every time you're concerned

with a new value of N. You simply just plug the new value of N into your symbolic answer.

These two benefits are made clear in the following example.

Example (50 out of 100): A coin is flipped 100 times. Calculate the probabilities of getting *exactly* 50 Heads.

SOLUTION: There is a total of 2^{100} possible outcomes (all equally likely), and $\binom{100}{50}$ of these have exactly 50 Heads. So the probability of obtaining exactly 50 Heads is

$$P_{50} = \frac{\binom{100}{50}}{2^{100}} = \frac{100!}{50!\,50!} \cdot \frac{1}{2^{100}}\,. \tag{2.37}$$

Now, although this is the correct answer, your calculator might not be able to handle the large factorials. But even if it can, let's use Stirling's formula in order to produce a symbolic answer. To this end, we'll replace the number 50 with the letter N (and thus 100 with $2N$). In terms of N, we can write down the probability of getting exactly N Heads in $2N$ flips, and then we can use Stirling's formula (applied to both N and $2N$) to simplify the result. The first steps of this simplification will actually go in the wrong direction and turn things into a big mess, so you need to have faith that it will work out! We obtain:

$$\begin{aligned} P_N = \frac{\binom{2N}{N}}{2^{2N}} = \frac{2N!}{N!\,N!} \cdot \frac{1}{2^{2N}} &\approx \frac{(2N)^{2N} e^{-2N} \sqrt{2\pi(2N)}}{(N^N e^{-N} \sqrt{2\pi N}\,)^2} \cdot \frac{1}{2^{2N}} \\ &= \frac{2^{2N} N^{2N} e^{-2N} \cdot 2\sqrt{\pi N}}{N^{2N} e^{-2N} \cdot 2\pi N} \cdot \frac{1}{2^{2N}} \\ &= \frac{1}{\sqrt{\pi N}}\,. \end{aligned} \tag{2.38}$$

A simple answer indeed! And the "π" is a nice touch, too. In our specific case with $N = 50$, we have

$$P_{50} \approx \frac{1}{\sqrt{\pi \cdot 50}} \approx 0.7979 \approx 8\%. \tag{2.39}$$

This is small, but not negligible. If instead we have $N = 500$, we obtain $P_{500} \approx 2.5\%$. As noted above, we can just plug in whatever number we want, and not redo the whole calculation!

Eq. (2.38) is an extremely clean result, much simpler than the expression in Eq. (2.37), and *much* simpler than the expressions in the steps leading up to it. True, it's only an approximate expression, but it turns out that the exact result in Eq. (2.37) equals 0.07959. So for $N = 50$, the ratio of the approximate result to the exact result is about 1.0025. In other words, the approximation is off by only 0.25%. Plenty good for me. Even for a small number like $N = 5$, the error is only 2.5%.

When you derive symbolic approximations like Eq. (2.38), you gain something and you lose something. You lose some truth, of course, because your answer technically isn't correct anymore (although invariably its accuracy is more than sufficient). But you gain a great deal of information about how the answer depends on your input number, N. And along the same lines, you gain some aesthetics. Basically, the resulting symbolic answer is invariably nice and concise, so it allows you to easily see how the answer depends on N. For example, in the present problem, the expression in Eq. (2.38) is proportional to $1/\sqrt{N}$. This means that if we increase N by a factor of, say, 100, then P_N decreases by a factor of $\sqrt{100} = 10$. So without doing any work, we can quickly use the $P_{50} \approx 8\%$ result to deduce that $P_{5000} \approx 0.8\%$. In short, there is *far* more information contained in the symbolic result in Eq. (2.38) than in the numerical 8% result obtained directly from Eq. (2.37).

2.7 Summary

In this chapter we learned various things about probability. In particular, we learned:

1. The probability of an event is defined to be the fraction of the time the event occurs in a very large number of identical trials. In many situations the various events are all equally likely, in which case the probability of a certain class of events occurring is

$$p = \frac{\text{number of desired events}}{\text{total number of possible events}} \qquad (\textit{for equally likely events}) \tag{2.40}$$

2. The various "and" and "or" rules of probability are:

   - For any two (possibly dependent) events, we have

$$P(A \text{ and } B) = P(A) \cdot P(B|A). \tag{2.41}$$

   - In the special case of independent events, we have $P(B|A) = P(B)$, so Eq. (2.41) reduces to

$$P(A \text{ and } B) = P(A) \cdot P(B). \tag{2.42}$$

   - For any two (possibly non-exclusive) events, we have

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \tag{2.43}$$

   - In the special case of exclusive events, we have $P(A \text{ and } B) = 0$, so Eq. (2.43) reduces to

$$P(A \text{ or } B) = P(A) + P(B). \tag{2.44}$$

3. Two common ways to calculate probabilities are: (1) Count up the events and use Eq. (2.40), and (2) Imagine things happening in succession (for example, picking seats or rolling dice), and then multiply the relevant probabilities. The results of some problems, in particular the Birthday Problem and the Game Show Problem, might seem surprising at first, but you can avoid confusion by methodically using one (or both) of these strategies.

4. The *expectation value* for a process is the expected average of many trials of the process. It is given by

$$\text{expectation value} = p_1 \cdot R_1 + p_2 \cdot R_2 + \cdots + p_n \cdot R_n, \tag{2.45}$$

   where the R's are the possible outcomes and the p's are the associated probabilities.

5. Stirling's formula gives an approximation to $N!$. It is given by

$$N! \approx N^N e^{-N} \sqrt{2\pi N} \qquad (Stirling's\ formula) \tag{2.46}$$

   This is very helpful in simplifying binomial coefficients. We will be using it a great deal in future chapters.

2.8 Problems

1. **Odds** $*$

   If an event happens with probability p, the *odds* in favor of the event happening are defined to be "p to $(1-p)$." (And similarly, the odds *against* the event happening are defined to be "$(1-p)$ to p.") In other words, the odds are simply the ratio of the probabilities of the event happening (the p) to not happening (the $1-p$). It is customary to write "$p : (1-p)$" as shorthand for "p to $(1-p)$." [9] In practice, these two probabilities are usually multiplied through by the smallest number that turns them into integers. For example, odds of $1/3 : 2/3$ are generally written as $1 : 2$. Find the odds of the following events:

   (a) Getting a Heads on a coin toss.

   (b) Rolling a 5 on a die.

   (c) Rolling a multiple of 2 or 3 on a die.

   (d) Randomly picking a day of the week with more than six letters.

2. **Rules for three events** $**$

   (a) Consider three events, A, B, and C. If they are all independent of each other, show that

   $$P(A \text{ and } B \text{ and } C) = P(A) \cdot P(B) \cdot P(C). \tag{2.47}$$

   (b) If they are (possibly) dependent, show that

   $$P(A \text{ and } B \text{ and } C) = P(A) \cdot P(B|A) \cdot P(C|A \text{ and } B). \tag{2.48}$$

   (c) If they are all mutually exclusive, show that

   $$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C). \tag{2.49}$$

   (d) If they are (possibly) non exclusive, show that

   $$\begin{aligned} P(A \text{ or } B \text{ or } C) \quad = \quad & P(A) + P(B) + P(C) \\ & -P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) \\ & +P(A \text{ and } B \text{ and } C). \end{aligned} \tag{2.50}$$

3. **"Or" rule for four events** $***$

   Parts (a), (b), and (c) of the previous problem quickly generalize to more than three events, but part (d) is tricker. Derive the "or" rule for four (possibly) non exclusive events. That is, derive the rule analogous to Eq. (2.50). You should do the previous problem before this one.

4. **At least one 6** $**$

   Three dice are rolled. What is the probability of getting at least one 6? We solved this in the example in Section 2.3.1, but your task here is to solve it the long way, by adding up the probabilities of getting exactly one, two, or three 6's.

[9] The odds are sometimes also written as the ratio $p/(1-p)$, but this fraction can look like a probability and therefore cause confusion, so we'll avoid this notation.

5. **At least one 6, by the rules** $**$

   Three dice are rolled. What is the probability of getting at least one 6? We solved this in the example in Section 2.3.1, but your task here is to solve it by using Eq. (2.50) from Problem 2, with each of the three letters in that formula standing for a 6 on each of the three dice.

6. **Proofreading** $**$

   Two people each proofread a book. One finds 100 errors, and the other finds 60 errors. 20 of these are common to both people. Assuming that all errors are equally likely to be found (which is undoubtedly not true in practice), roughly how many errors does the book have? *Hint:* Draw a picture analogous to Fig. 2.1, and then find the probabilities of each person finding a given mistake.

7. **Exactly one pair** $**$

   If there are 23 people in a room, what is the probability that *exactly* two of them have a common birthday? The point here is that we don't want two different pairs with common birthdays, or a triple with a common birthday, etc.

8. **A random game show host** $**$

   Consider the following variation of the Game Show Problem we discussed in Section 2.4.2. A game show host offers you the choice of three doors. Behind one of these doors is the grand prize, and behind the other two are goats. The host announces that after you select a door (without opening it), he will *randomly* open one of the other two doors. You select a door. The host then randomly opens one of the other doors, and the result happens to be a goat. He then offers you the chance to switch your choice to the remaining door. Should you switch or not? Or does it not matter?

9. **Flip until Tails** $*$

   In Example 2 in Section 2.5, we found that if you flip a coin until you get a Tails, the expectation value of the total number of coins is

   $$\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{32}\cdot 5 + \cdots . \tag{2.51}$$

   We claimed that this sum equals 2. Show this by writing the sum as a geometric series starting with 1/2, plus another geometric series starting with 1/4, and so on. You can use the fact that the sum of a geometric series with first term a and ratio r is $a/(1-r)$.

10. **Playing "unfair" games** $*$

    (a) Assume that later on in life, things work out so that you have more than enough money in your retirement savings to take care of your needs and beyond, and that you truly don't have need for more money. Someone offers you the chance to play a one-time game where you have a 3/4 chance of doubling your money, and a 1/4 chance of losing it all. If you initially have N dollars, what is the expectation value of your resulting amount of money if you play the game? Would you want to play it?

    (b) Assume that you are stranded somewhere, and that you have only \$10 for a \$20 bus ticket. Someone offers you the chance to play a one-time game where you have a 1/4 chance of doubling your money, and a 3/4 chance of losing it all. What is the expectation value of your resulting amount of money if you play the game? Would you want to play it?

Many more problems will be added...

2.9 Solutions

1. **Odds**

   (a) The probabilities of getting a Heads and not getting a Heads are both $1/2$, so the desired odds are $1/2 : 1/2$, or equivalently $1 : 1$. These are know as "even odds."

   (b) The probabilities of getting a 5 and not getting a 5 are $1/6$ and $5/6$, respectively, so the desired odds are $1/6 : 5/6$, or equivalently $1 : 5$.

   (c) There are four desired outcomes (2,3,4,6), so the "for" and "against" probabilities are $4/6$ and $2/6$, respectively. The desired odds are therefore $4/6 : 2/6$, or equivalently $2 : 1$.

   (d) Tuesday, Wednesday, Thursday, and Saturday have more than six letters, so the "for" and "against" probabilities are $4/7$ and $3/7$, respectively. The desired odds are therefore $4/7 : 3/7$, or equivalently $4 : 3$.

   Note that to convert from odds to probability, the odds of $a : b$ are equivalent to a probability of $a/(a + b)$.

2. **Rules for three events**

   (a) The same type of reasoning that we used in Section 2.2.1 holds again here. If we have a large number of events, then A occurs in a fraction $P(A)$ of them. And then B occurs in a fraction $P(B)$ *of these* (because the events are independent, so the occurrence of A doesn't affect the probability of B). In other words, both A and B occur in a fraction $P(A) \cdot P(B)$ of the total. And then C occurs in a fraction $P(C)$ *of these* (because C is independent of A and B). In other words, all three of A, B, and C occur in a fraction $P(A) \cdot P(B) \cdot P(C)$ of the total. So the desired probability is $P(A) \cdot P(B) \cdot P(C)$. If you want to visualize this geometrically, you'll need to use a cube instead of the square in Fig. 2.1.

   This reasoning can easily be extended to an arbitrary number of independent events. The probability of all the events occurring is simply the product of all the individual probabilities.

   (b) The reasoning in part (a) works again, with only slight modifications. If we have a large number of events, then A occurs in a fraction $P(A)$ of them. And then B occurs in a fraction $P(B|A)$ *of these*, by definition. In other words, both A and B occur in a fraction $P(A) \cdot P(B|A)$ of the total. And then C occurs in a fraction $P(C|A \text{ and } B)$ *of these*, by definition. In other words, all three of A, B, and C occur in a fraction $P(A) \cdot P(B|A) \cdot P(C|A \text{ and } B)$ of the total. So the desired probability is $P(A) \cdot P(B|A) \cdot P(C|A \text{ and } B)$

   Again, this reasoning can easily be extended to an arbitrary number of independent events. For four events, we simply need to tack on a $P(D|A \text{ and } B \text{ and } C)$ factor, and so on.

   (c) Since the events are all mutually exclusive, we don't have to worry about any double counting. The total number of events in which A or B or C happens is simply the sum of the number of events where A happens, plus the number where B happens, plus the number where C happens. The same statement must be true if we substitute the word "fraction" for "number," because the fractions are related to the numbers by division by the total number of possible events. And since the fractions are simply the probabilities, we end up with the desired result, $P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$. If there are more events, we simply get more terms in the sum.

   (d) We can think of the probabilities in terms of areas, as we did in Section 2.2.2. The generic situation for three events is shown in Fig. 2.9 (we've chosen the three circles to be the same size for simplicity, but this of course doesn't have to be the case). The

various overlaps are shown, with the juxtaposition of two letters standing for their intersection. So "AB" means "A and B."[10]

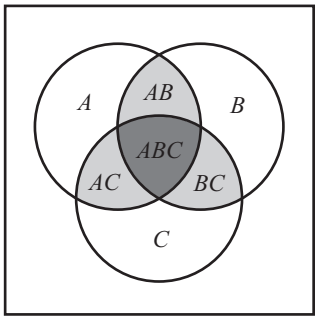


Figure 2.9

Our goal is to determine the total area contained in the three circles. We can add up the areas of the A, B, and C circles, but then we need to subtract off the areas we double counted. These areas are the pairwise overlaps of the circles, that is, AB, AC, and BC (remember that each of these regions includes the dark ABC region in the middle). At this point, we've correctly counted all of the white and light gray regions exactly once. But what about the ABC region in the middle? We counted it three times in the A, B, and C regions, but then we subtracted it off three times in the AB, AC, BC regions. So at the moment we actually haven't counted it at all. So we need to add it on once, and then every part of the union of the circles will be counted exactly once. The total area is therefore $A + B + C - AB - AC - BC + ABC$, where we're using the regions' labels to stand for their areas. Translating this from a statement about areas to a statement about probabilities yields the desired result,

$$\begin{aligned} P(A \text{ or } B \text{ or } C) &= P(A) + P(B) + P(C) \\ &\quad -P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) \\ &\quad +P(A \text{ and } B \text{ and } C). \end{aligned} \tag{2.52}$$

3. **"Or" rule for four events**

   As in the previous problem, we'll discuss things in terms of areas. If we add up the areas of the four regions, A, B, C, and D, we have double counted the pairwise overlaps, so we need to subtract these off. There are six of these regions: AB, AC, AD, BC, BD, and CD. But then what about the triple overlaps, such as ABC? Well, we counted ABC three times in A, B, and C, but then we subtracted it off three times in the regions AB, AC, and BC. So at this point we haven't counted is at all. We therefore need to add it on once (this is the same reasoning as in the previous problem). Likewise for ABD, ACD, and BCD. Finally what about the quadruple overlap region, $ABCD$? We counted this four times in the "single" regions (like A), and then we subtracted it off six times in the "double" regions (like AB), and then we added it on four times in the "triple" regions (like ABC). So at the moment we've counted it $4 - 6 + 4 = 2$ times. We only want to count it one time, so we therefore need to subtract it off once. The total area is therefore

   $$A+B+C+D-AB-AC-AD-BC-BD-CD+ABC+ABD+ACD+BCD-ABCD. \tag{2.53}$$

   Writing this in terms of probabilities gives the result:

   $$P(A \text{ or } B \text{ or } C \text{ or } D) = P(A) + P(B) + P(C) + P(D)$$

[10] The labels might appear to suggest otherwise, but remember that "A" includes the whole circle, and not just the white part. Similarly, "AB" includes the dark "ABC" region too, and not just the lighter region where the "AB" label is.

$$\begin{aligned}
&-P(A \text{ and } B) - P(A \text{ and } C) - P(A \text{ and } D) \\
&\quad -P(B \text{ and } C) - P(B \text{ and } D) - P(C \text{ and } D) \\
&+P(A \text{ and } B \text{ and } C) + P(A \text{ and } B \text{ and } D) \\
&\quad +P(A \text{ and } C \text{ and } D) + P(B \text{ and } C \text{ and } D) \\
&-P(A \text{ and } B \text{ and } C \text{ and } D). \qquad (2.54)
\end{aligned}$$

REMARK: You might think that it's a bit of a coincidence that at every stage, we either overcounted or undercounted each region *once*. Equivalently, the coefficient of every term in Eqs. (2.53) and (2.54) is a $+1$ or a -1. And the same thing is true in the case of three events in Eq. (2.50). And it is trivially true in the case of one or two events. Is it also true for larger numbers of events? Indeed it is, and the binomial expansion is the key to understanding why.

We won't go through the full reasoning, but if you want to think about it, the main points to realize are: First, the numbers 4, 6, and 4 in the above counting are actually the binomial coefficients $\binom{4}{1}$, $\binom{4}{2}$, $\binom{4}{3}$. This makes sense because, for example, the number of regions of double overlap (like AB) is simply the number of ways to pick two letters from four letters, which is $\binom{4}{2}$. Second, the "alternating sum" $\binom{4}{1} - \binom{4}{2} + \binom{4}{3}$ equals 2 (which means that we overcounted the ABCD region by one time) because this is what you get when you expand the right side of $0 = (1-1)^4$ with the binomial expansion (this is a nice little trick). And third, you can show how this generalizes to larger numbers of events, N. For even N, the "alternating sum" of the binomial coefficients is 2, as we just saw for $N = 4$. But for odd numbers, the $(1-1)^N$ expansion yields an alternating sum of zero, which means we undercount by one time. For example, $\binom{5}{1} - \binom{5}{2} + \binom{5}{3} - \binom{5}{4} = 0$. Food for thought if you want to think about it more. ♣

4. **At least one 6**

The probability of getting exactly one 6 equals $\binom{3}{1} \cdot (1/6)(5/6)^2$, because there are $\binom{3}{1} = 3$ ways to pick which die is the 6. And then given this choice, there is a $1/6$ chance that the die is in fact a 6, and a $(5/6)^2$ chance that both of the other dice are not 6's.

The probability of getting exactly two 6's equals $\binom{3}{2} \cdot (1/6)^2(5/6)$, because there are $\binom{3}{2} = 3$ ways to pick which two dice are the 6's. And then given this choice, there is a $(1/6)^2$ chance that they are in fact both 6's, and a $(5/6)$ chance that the other die isn't a 6.

The probability of getting exactly three 6's equals $\binom{3}{3} \cdot (1/6)^3$, because there is just $\binom{3}{3} = 1$ way for all three dice to be 6's. And then there is a $(1/6)^3$ chance that they are in fact all 6's.

The total probability of getting at least one six is therefore

$$\binom{3}{1} \cdot \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2 + \binom{3}{2} \cdot \left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) + \binom{3}{3} \cdot \left(\frac{1}{6}\right)^3 = \frac{75}{216} + \frac{15}{216} + \frac{1}{216} = \frac{91}{216}, \qquad (2.55)$$

in agreement with the result in Section 2.3.1.

REMARK: If we add this result to the probability of getting zero 6's, which is $(5/6)^3$, then the sum must be 1, because this encompasses every possible outcome. This fact was what we used to solve the problem the quick way in Section 2.3.1, after all. But let's pretend we don't know that the sum must be 1, and let's verify it explicitly. This will give us an excuse to show how the binomial expansion comes into play. If we write $(5/6)^3$ suggestively as $\binom{3}{0} \cdot (5/6)^3$, then our goal is to show that

$$\tbinom{3}{0} \cdot \left(\tfrac{5}{6}\right)^3 + \left(\tbinom{3}{1} \cdot \left(\tfrac{1}{6}\right)\left(\tfrac{5}{6}\right)^2 + \tbinom{3}{2} \cdot \left(\tfrac{1}{6}\right)^2\left(\tfrac{5}{6}\right) + \tbinom{3}{3} \cdot \left(\tfrac{1}{6}\right)^3 \right) = 1. \qquad (2.56)$$

And this is indeed true, because the lefthand side is simply the binomial-expanded version of $(5/6 + 1/6)^3 = 1$. This makes it clear why the sum of the probabilities of the various outcomes will still add up to 1, even if we have, say, an 8-sided die (again, forgetting that we know intuitively that the sum must be 1). The only difference is that we now have the expression $(7/8 + 1/8)^3 = 1$, which is still true. And any other exponent (that is, any other number of rolls) will also yield a sum of 1 (as we know it must). ♣

5. **At least one 6, by the rules**

We'll copy Eq. (2.50) here:

$$\begin{aligned} P(A \text{ or } B \text{ or } C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) \\ &\quad + P(A \text{ and } B \text{ and } C). \end{aligned} \tag{2.57}$$

The lefthand side here is the probability of getting at least one 6 (remember that the "or" is the inclusive "or"). So our task is to evaluate the righthand side.

The probability of getting a 6 on any given die (without caring what happens with the other two dice) is 1/6, so $P(A) = P(B) = P(C) = 1/6$. The probability of getting 6's on two given dice (without caring what happens with the third die) is $(1/6)^2$, so $P(A \text{ and } B) = P(A \text{ and } C) = P(B \text{ and } C) = 1/36$. The probability of getting 6's on all three dice is $(1/6)^3$, so $P(A \text{ and } B \text{ and } C) = 1/216$. Eq. (2.50) therefore gives the probability of getting at least one 6 as

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \frac{1}{36} - \frac{1}{36} - \frac{1}{36} + \frac{1}{216} = \frac{3}{6} - \frac{3}{36} + \frac{1}{216} = \frac{108 - 18 + 1}{216} = \frac{91}{216}, \tag{2.58}$$

in agreement with the result in Section 2.3.1.

6. **Proofreading**

The breakdown of the errors is shown in Fig. 2.10. If A and B represent the two people, then 20 errors are found by both A and B, 80 are found by A but not B, and 40 are found by B but not A.

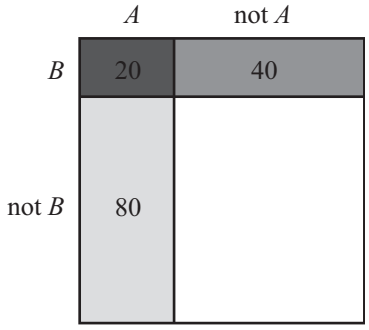


Figure 2.10

If we consider just the 100 errors that A found, we see that B found 20 of them, which is a fraction 1/5. We are assuming that all the errors are equally likely to be found, so if B finds 1/5 of the errors in a given subset (in particular, the ones found by A), then he must find 1/5 of the errors in *any* subset, on average. So 1/5 is the probability that B finds an error. Therefore, since we know that B found a total of 60 errors, the total number N must be given by $60/N = 1/5 \implies N = 300$. (The unshaded region in Fig. 2.10 therefore represents $300 - 80 - 20 - 40 = 160$ errors. This is the number that both people missed.)

We could also have done things the other way around: If we consider just the 60 errors that B found, we see that A found 20 of them, which is a fraction 1/3. By the same reasoning as above, this 1/3 is the probability that A finds an error. And then since we know that A found a total of 100 errors, the total number N must be given by $100/N = 1/3 \implies N = 300$, as above.

A quicker method (although in the end it's the same as the above methods) is the following. Let the area of the white region in Fig. 2.10 be x. Then if we look at how the areas of the two vertical rectangles are divided by the horizontal line, we see that the ratio of x to 40 must equal the ratio of 80 to 20. So $x = 160$, as we found above. Alternatively, if we look

at how the areas of the two horizontal rectangles are divided by the vertical line, we see that the ratio of x to 80 must equal the ratio of 40 to 20. So again, $x = 160$.

It's quite fascinating that you can get a sense of the total number of errors just by comparing the results of two readers' proofreadings. There is no need to actually find all the errors and count them up.

7. **Exactly one pair**

There are $\binom{23}{2}$ possible pairs that can have the common birthday. Let's look at a particular pair and calculate the probability that they have a common birthday *with no one else having a common birthday*. We'll then multiply this result by $\binom{23}{2}$ to account for all the possible pairs.

The probability that a given pair has a common birthday is 1/365, because the first person's birthday can be picked to be any day, and then the second person has a 1/365 chance of matching that day. We then need the other 21 people to have 21 different birthdays, none of which is the same as the pair's birthday. The first of these people can go in any of the remaining 364 days; this happens with probability 364/365. The second of these people can go in any of the remaining 363 days; this happens with probability 363/365. And so on, until the 21st of these people can go in any of the remaining 344 days; this happens with probability 344/365.

The total probability that exactly one pair has a common birthday is therefore

$$\binom{23}{2} \cdot \frac{1}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdot \ldots \cdot \frac{344}{365}. \tag{2.59}$$

Multiplying this out gives 0.363 = 36.3%. This is smaller than the "at least one common birthday" result of 50.7%, as it must be. The remaining 50.7% − 36.3% = 14.4% of the probability corresponds to occurrences of two different pairs with common birthdays, or a triple with a common birthday, etc.

8. **A random game show host**

We'll solve this problem by listing out the various possibilities. Without loss of generality, assume that you pick the first door. (You can repeat the following reasoning again for the other doors if you wish. It gives the same result.) There are three equally likely possibilities for what is behind the three doors: PGG, GPG, and GGP (where P denotes prize, and G denotes goat). For each of these cases, the host opens either the second or the third door (with equal probabilities), so there are six equally likely results of his actions (the bold letters signify the item he revealed):

P**G**G	$\boxed{\text{G}\mathbf{P}\text{G}}$	G**G**P
PG**G**	GP**G**	$\boxed{\text{GG}\mathbf{P}}$

We now note the critical fact that the two results where the prize is revealed (the boxed G**P**G and GG**P** results) are not relevant to this problem, because we are told in the statement of the problem that the host happens to reveal a goat. Therefore, only the other four results are possible:

P**G**G PG**G** GP**G** G**G**P

They are all still equally likely, so their probabilities must each be 1/4.[11] We therefore see that if you don't switch from the first door, you win on the first two of these results and lose on the second two. And if you do switch, you lose on the first two and win on the second two. So either way, your probability of winning is 1/2. So it doesn't matter if you switch.

[11] There is nothing wrong with these four probabilities jumping from 1/6 to 1/4 (and the other two probabilities falling from 1/6 to zero) due to the host's actions. He gave you information by picking a goat, so it's no surprise that the various probabilities change. An extreme example of having probabilities change due to new information is a setup where you look at the result of a coin toss and observe a Heads. This causes the probability of Heads to jump from 1/2 to 1, and the probability of Tails to fall from 1/2 to zero.

REMARKS:

1. In the original version of the problem in Section 2.4.2, the probability of winning was 2/3 if you switched. How can it possibly decrease to 1/2 in the present random version, when in both versions the exact same thing happened, namely the host revealed a goat? The difference is due to the two cases where the host revealed the prize in the random version (the G**P**G and GG**P** cases). You don't benefit from these cases in the random version, because we are told in the statement of the problem that they don't exist. But in the original version, they represent guaranteed success if you switch, because the host is forced to open the other door which is a goat.

   But still you may say, "If there are two identical setups, and if I pick, say, the first door in each, and if the host reveals a goat in each (by prediction in one case, and by random pick in the other), then the *exact same thing happens in both setups.* How can the resulting probabilities be any different?" The answer is that although these two setups happen to be identical, probabilities have nothing to do with *two* setups. Probabilities are defined only for a *large number* of setups. The point is that if you play a large number of these pairs of games, then in 1/3 of them the host will reveal different things (a goat in the original version and the prize in the random version). These cases yield success in the original version, but they don't even get mentioned in the random version. They are worthless there.

2. As with the original version of the problem, if you find yourself arguing about the result for more than 15 minutes, then just *play the game* a bunch of times (at least a few dozen, to get good enough statistics). The randomness can be determined by a coin toss.

3. We will now address the issue we mentioned in the fourth remark in Section 2.4.2. We correctly stated in Section 2.4.2 that "No actions taken by the host can change the fact that if you play a large number, N, of these games, then (roughly) $N/3$ of them will have the prize behind the door you initially pick." However, in the present random version of the problem, the actions of the host *do* affect the probability that the prize is behind the door you initially pick. It is now 1/2 instead of 1/3. So can the host affect this probability or not?

   Well, yes and no. If *all* of the N games are considered (as in the original version), then $N/3$ of them have the prize behind the initial door, and that's that. However, the random version of the problem involves throwing out 1/3 of the games (the ones where the host reveals the prize), because it is assumed in the statement of the problem that the host happens to reveal a goat. So of the *remaining games* (which are 2/3 of the initial total, so $2N/3$), 1/2 of them have the prize behind your initial door.

   If you play a large number, N, of games of each of these versions, the actual *number* of games that have the prize behind your initial door pick is the same. It's just that in the original version this number can be thought of as $N/3$, whereas in the random version it can be thought of as 1/2 of $2N/3$, which is still $N/3$. So in the end, the action of the random host that influences the probability and changes it from 1/3 to 1/2 isn't the opening of a door, but rather the throwing out of 1/3 of the games. Since no games are thrown out in the original version, the above statement in quotes is correct (with the key phrase being "*these* games"). ♣

9. **Flip until Tails**

   The given sum equals

$$\begin{aligned}
&\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\cdots \\
&\quad +\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\cdots \\
&\qquad +\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\cdots \\
&\qquad\quad +\frac{1}{16}+\frac{1}{32}+\cdots \\
&\qquad\qquad \vdots
\end{aligned} \tag{2.60}$$

   This has the correct number of each type of term. For example, the "1/16" appears four times. The first line here is a geometric series that sums to $a/(1-r) = (1/2)/(1-1/2) = 1$. The second line is also a geometric series, and it sums to $(1/4)/(1-1/2) = 1/2$. Likewise the third line sums to $(1/8)/(1-1/2) = 1/4$. And so on. The sum of the infinite number of lines in the above equation therefore equals

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\cdots. \tag{2.61}$$

But this itself is a geometric series, and its sum is $a/(1-r) = 1/(1-1/2) = 2$, as desired.

10. **Playing "unfair" games**

(a) The expectation value of your money after you play the game is $(3/4) \cdot 2N + (1/4) \cdot 0 = 3N/2$. So on average, you will gain $N/2$ dollars. It therefore seems like it would be a good idea to play the game. However, further thought shows that it would actually be a very bad idea. There is basically no upside, because you already have plenty of money; twice the money won't do you much good. But there is a huge downside; you might lose all your money.

The point here is that the important issue is your happiness, not the exact amount of money you have. And on the happiness scale, you stand to gain basically nothing (or perhaps a tiny bit). But you stand to lose a huge amount (not to say you can't be happy without money, but if you lose your entire savings, there's no doubt that it would put a damper on things). The expectation value of the level of your happiness (let's arbitrarily say it starts at 1) is essentially $(3/4) \cdot 1 + (1/4) \cdot 0 = 3/4$. This is less than the initial value of 1, so it suggests that you shouldn't play the game. (However, there is still another thing to consider; see the remark below.)

(b) The expectation value of your money after you play the game is $(3/4) \cdot 0 + (1/4) \cdot 20 = 5$. So on average, you stand to lose \$5. It therefore seems like it would be a bad idea to play the game. However, \$10 in your pocket is just as useless as \$0, because either way, you're guaranteed to be stuck at the bus station. You therefore *should* play the game. That way, at least there's a 1/4 chance that you'll make it home. (We'll assume that the overall money you have back home washes out any effect of gaining or losing \$10, in the long run.) The same argument we used above with the happiness level holds here. \$0 and \$10 yield the same level of happiness (or perhaps we should say misery), so there is basically no downside. But there is definitely an upside with the \$20, because you can then buy a ticket. The expectation value of the level of your happiness (let's arbitrarily say it starts at zero) is essentially $(3/4) \cdot 0 + (1/4) \cdot 1 = 1/4$. This is greater than the initial value of zero, so it suggests that you should play the game. (But see the following remark.)

REMARK: There is another consideration with these sorts of situations, in that they are *one-time events.* Even if we rig things so that the expectation value of your *happiness* level increases (or whatever measure you deem to be the important one), it's still not obvious whether or not you should play the game. Just as with any other probabilistic quantity, the expectation value has meaning only in the context of a *large number of identical trials.* You could imagine a situation where a group of many people play the game and the average happiness level increases. But *you* are only *one* person, and the increase in the overall happiness level of the group is of little comfort to you if you lose your shirt. Since you play the game only once, the expectation value is basically irrelevant to you. The decision mainly comes down to an assessment of the risk. Different people's reactions to risk are different, and you could imagine someone being very risk-averse and not playing any game with a significant downside, no matter what the upside is. ♣

Chapter 3

Distributions

Consider a variable that can take on certain values with certain probabilities. Such a variable is appropriately called a *random variable*. For example, the number of Heads that can arise in two coin tosses is a random variable, and it can take on the values of 0, 1, or 2. The probabilities for each of these possibilities are 1/4, 1/2, and 1/4, respectively, as you can quickly show. The collection of these probabilities is called the *probability distribution* for this particular process. A probability distribution is simply the collective information about how the total probability (which is always 1) is distributed among the various possible outcomes.

The outline of this chapter is as follows. In Section 3.1 we warm up with some examples of discrete distributions, and then in Section 3.2 we discuss continuous distributions. These involve the *probability density*, which is the main new concept in this chapter. It takes some getting used to, but we'll have plenty of practice with it. In Section 3.3 we derive a number of the more common and important distributions. We'll concentrate on the derivations of the distributions here, and for the most part we'll postpone the discussion of their various properties until the following chapter, when we start talking about actual statistics. Finally, in Section 3.4 we discuss the "law of large numbers" and why nearly every distribution you'll ever deal with reduces to the so-call Gaussian (or "normal") distribution when the number of trials becomes large.

Parts of this chapter are a bit mathematical, but there's no way around this if we want to do things properly. However, we've relegated some of the more technical issues to Appendices B and C. If you want to skip those and just accept the results that we derive there, that's fine. But you are strongly encouraged to at least take a look at Appendix B, where we derive many properties of the number e, which is the most important number in mathematics (and especially in probability and statistics).

3.1 Discrete distributions

In this section we'll give a few simple examples of discrete distributions. To start off, consider the results from Example 3 in Section 2.3.4, where we calculated the probabilities of obtaining the various possible numbers of Heads in five coin flips. We found:

$$P_0 = \frac{1}{32}\,, \quad P_1 = \frac{5}{32}\,, \quad P_2 = \frac{10}{32}\,, \quad P_3 = \frac{10}{32}\,, \quad P_4 = \frac{5}{32}\,, \quad P_5 = \frac{1}{32}\,. \tag{3.1}$$

These probabilities add up to 1, as they should. Fig. 3.1 shows a plot P_n versus n. The variable n here (the number of Heads) is the random variable, and it can take on the values

of 0 through 5 with the above probabilities.

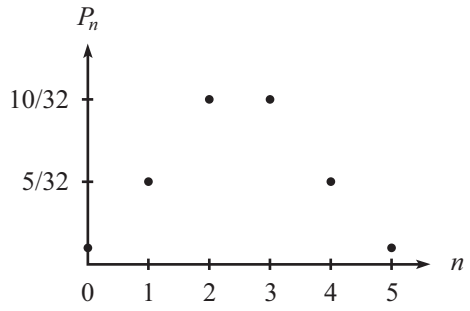


Figure 3.1

As we've done in Fig. 3.1, the convention is to plot the random variable on the horizontal axis and the probability on the vertical axis. The collective information given either visually in Fig. 3.1, or explicitly in the above list of probabilities, is called the *probability distribution*. A probability distribution simply tells you what all the probabilities are for the values that the random variable can take. Note that P_n here has meaning only if n takes on one of the *discrete* values, 0, 1, 2, 3, 4, or 5. It's a useless question to ask for the probability of getting 3.27 Heads, because n must of course be an integer, so the probability is trivially zero. Hence the word "discrete" in the title of this section.

A very simple example of a probability distribution is the one for the six possible outcomes of the roll of one die. The random variable in this setup is the number that faces up. If the die is fair, then all six numbers have equal probabilities, so the probability for each is 1/6, as shown in Fig. 3.2.

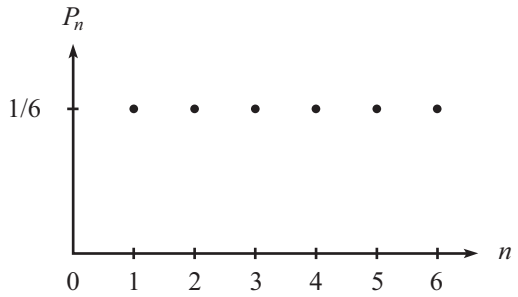


Figure 3.2

What if the die isn't fair? For example, what if we make the "1" side a little heavier than the others by embedding a small piece of lead in the center of the side, just below the surface? Then the die is more likely to land with the "1" side facing down. The "6" side is opposite to the 1, so the die is more likely to land with the 6 facing up. We therefore end up with a probability distribution looking something like the one in Fig. 3.3. The sum of the probabilities must still be 1, so P_6 lies the same distance above the 1/6 level as P_1 lies below.

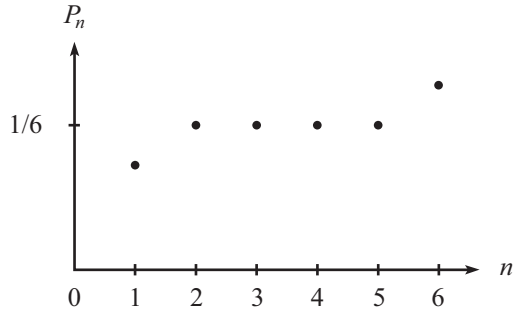


Figure 3.3

A note on terminology: A random variable is assumed to take on *numerical* values, by definition. So the outcomes of Heads and Tails for a coin flip technically aren't random variables. But it still makes sense to plot the probabilities as shown in Fig. 3.4, even though the outcomes on the horizontal axis aren't random variables. Of course, if we define a random variable to be the number of Heads, then the "Heads" in the figure turns into a 1, and the "Tails" turns into a 0. In most situations, however, the outcomes take on numerical values right from the start, so we can officially label them as random variables. We will therefore generally refer to the thing being plotted on the horizontal axis of a probability distribution as a random variable.

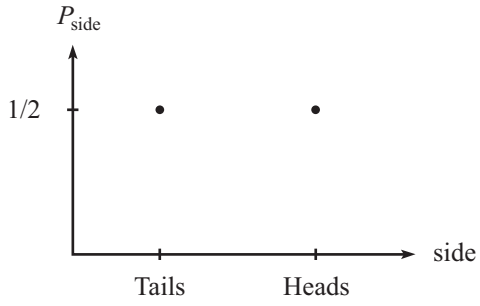


Figure 3.4

3.2 Continuous distributions

3.2.1 Motivation

Probability distributions are fairly straightforward when the random variable is discrete. You just list (or plot) the probabilities for each of the possible values of the random variable. And these probabilities will always add up to 1. However, not everything comes in discrete quantities. For example, the temperature outside your house takes on a continuous set of values, as does the amount of water in a glass (we'll ignore the atomic nature of matter!).

In finding the probability distribution for a continuous random variable, you might think that the procedure should be exactly the same as in the discrete case. That is, if our random variable is, say, the temperature at noon tomorrow, you might think that you simply have to find the answer to questions of the form: What is the probability that the temperature at noon tomorrow will be 70°?

But there is something wrong with this question, because it is too easy to answer. The answer is that the probability is *zero*, because there is simply no chance that the temperature

at a specified time will be *exactly* 70°. If it's 70.1°, that's not good enough. And neither is 70.01°, nor even 70.00000001°. Basically, since the temperature takes on a continuous set of values (and hence an infinite number of values), the probability of a specific value is (roughly speaking) $1/\infty$, which is zero.

However, the fact that this was a useless question to ask doesn't mean that we should throw in the towel and conclude that probability distributions don't exist for continuous random variables. They do in fact exist, because there *are* some useful questions we can ask. These useful questions take the general form of: What is the probability that the temperature at noon lies in the range of 69° to 71°? This question has a nontrivial answer, in the sense that it isn't automatically zero. And depending on what the forecast is for tomorrow, the answer might be, say, 20%.

We could also ask: What is the probability that the temperature at noon lies somewhere between 69.5° and 70.5°? The answer to this question is clearly smaller than the answer to the previous one, because it involves a range of only one degree instead of two degrees. If we assume that the chance of being somewhere in the range of 69° and 71° is roughly uniform (which is probably a reasonable approximation although undoubtedly not exactly correct), and if the previous answer was 20%, then the present answer is (roughly) 10%, simply because the range is half the size.

The point is that the smaller the range, the smaller the chance that the temperature lies in that range. Conversely, the bigger the range, the bigger the chance that the temperature lies in that range. Taken to an extreme, if we ask for the probability that the temperature at noon lies somewhere between $-100°$ and 200°, then the answer is exactly equal to 1, for all practical purposes.

In addition to depending on the size of the range, the probability also of course depends on where the range is located on the temperature scale. For example, the probability that the temperature at noon lies somewhere between 69° and 71° is undoubtedly different from the probability that it lies somewhere between 11° and 13°. Both ranges have a span of two degrees, but if the day happens to be in late summer, the temperature is much more likely to be around 70° than to be sub-freezing (let's assume we're in, say, Boston). To actually figure out the probabilities, many different pieces of data would have to be considered. In the present problem, the data would be of the meteorological type. But if we were interested in, say, the probability that a random person is between 69 and 71 inches tall, then we'd need to consider a whole different set of data.

The lesson to take away from all this is that if we're looking at a random variable that can take on a continuous set of values, then the probability that this random variable falls into a given range depends on two things. It depends on:

- the size of the range, and
- the specifics of the situation we're dealing with.

This second of these is what determines the so-called *probability density*, which we will now discuss.

3.2.2 Probability density

Consider the plot in Fig. 3.5, which gives a hypothetical probability distribution for the temperature example we've been discussing. This plot shows the probability distribution on the vertical axis, as a function of the temperature (the random variable) on the horizontal axis. We've arbitrarily chosen to measure the temperature in Fahrenheit. We're denoting the probability distribution by $\rho(T)$ to distinguish it from the type of probability distribution we've been talking about for discrete variables. The reason for this new notation is that $\rho(T)$ is a *probability density* and not an actual probability. We'll talk about this below.

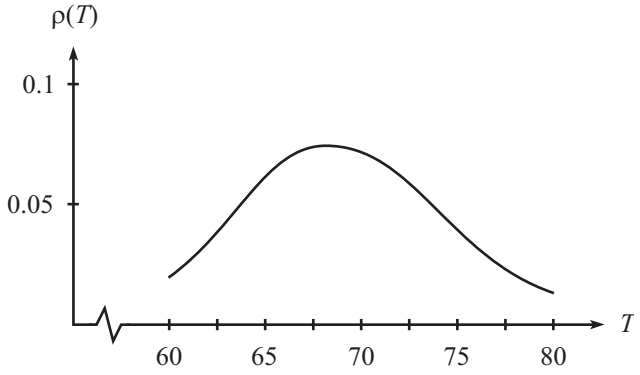


Figure 3.5

We haven't said yet exactly what we mean by $\rho(T)$. But in any case, it's clear from the plot that the temperature is more likely to be near 70° than near 60°. The following definition of $\rho(T)$ allows us to be precise about what we mean by this.

- Definition of the probability density, $\rho(T)$:

  $\rho(T)$ is the quantity that, when multiplied by a small interval ΔT, gives the probability that the temperature falls between T and $T + \Delta T$. That is,

$$P(\text{temp lies between } T \text{ and } T + \Delta T) = \rho(T) \cdot \Delta T. \tag{3.2}$$

Note that on the lefthand side we have an actual probability P, whereas on the righthand side we have the probability *density*, $\rho(T)$, which is something that needs to be multiplied by a range of T (or whatever quantity we're dealing with) in order to get an actual probability.

Eq. (3.2) might look a little scary, but a few examples should clear things up. From the figure, it looks like $\rho(70°)$ is about 0.07, so if we pick $\Delta T = 1°$, we find that the probability of the temperature falling between 70° and 71° is about $\rho(T) \cdot \Delta T = (0.07)(1) = 0.07 = 7\%$. If we instead pick a smaller ΔT, say 0.5°, we find that the probability of the temperature falling between 70° and 70.5° is about $(0.07)(0.5) = 3.5\%$. And if we pick an even smaller ΔT, say 0.1°, we find that the probability of the temperature falling between 70° and 70.1° is about $(0.07)(0.1) = 0.7\%$.

We can do the same thing with any other value of T. For example, it looks like $\rho(60°)$ is about 0.02, so if we pick $\Delta T = 1°$, we find that the probability of the temperature falling between 60° and 61° is about $(0.02)(1) = 2\%$. And as above, we can pick other values of ΔT too.

Remember that *two* quantities are necessary to find the probability that the temperature falls into a specified range. One is the size of the range, ΔT, and the other is the probability density, $\rho(T)$. These are the two quantities on the righthand side of Eq. (3.2). Knowing only one of these quantities isn't enough to give you a probability.

There is a very important difference between probability distributions for continuous random variables and those for discrete random variables. For continuous variables, the probability distribution consists of *probability densities*. But for discrete variables, it consists of *actual probabilities*. We plot densities for continuous distributions, because it wouldn't make sense to plot actual probabilities, since they're all zero. This is true because the probability of *exactly* obtaining a particular value is zero, since there is an infinite number of possible values. And conversely, we plot actual probabilities for discrete distributions, because it wouldn't make sense to plot densities, since they're all infinite. This is true because, for example, there is a 1/6 chance of rolling a die and obtaining a number between,

say, 4.9999999 and 5.0000001. The probability density, which from Eq. (3.2) equals the probability divided by the interval length, is then $(1/6)/(.0000002)$, which is huge. And this interval can be made arbitrarily small, which means that the density is arbitrarily large. To sum up, the term "probability distribution" applies to both continuous and discrete variables, but the term "probability density" applies to only continuous variables.

REMARKS:

1. The function $\rho(T)$ is a function of T, so it depends critically on what units we're measuring T in. We used Fahrenheit above, but what if we instead wanted to use Celsius? Problem 1 addresses this issue.

2. Note the inclusion of the word "small" in the above definition of the probability density. The reason for this word is that we want $\rho(T)$ to be (roughly) constant over the specified range. If ΔT is small enough, then this is approximately true. If $\rho(T)$ varied greatly over the range of ΔT, then it wouldn't be clear which value of $\rho(T)$ we should multiply by ΔT to obtain the probability. The point is that if ΔT is small enough, then all of the $\rho(T)$ values are roughly the same, so it doesn't matter which one we pick.

   An alternate definition of the density $\rho(T)$ is

   $$P\big(\text{temp lies between } T - (\Delta T)/2 \text{ and } T + (\Delta T)/2\big) = \rho(T) \cdot \Delta T. \tag{3.3}$$

   The only difference between this definition and the one in Eq. (3.2) is that we're now using the value of $\rho(T)$ at the midpoint of the temperature range, as opposed to the value at the left end we used in Eq. (3.2). Both definitions are equally valid, because they give essentially the same result for $\rho(T)$, provided that ΔT is small.

   How small do we need ΔT to be? The answer to this will be evident when we talk below about probability in terms of areas. In short, we need the change in $\rho(T)$ over the span of ΔT to be small compared with the values of $\rho(T)$ in that span.

3. Note that the probability density is a function only of the specifics of the situation at hand (meteorological data in the above temperature example, etc). The density is completely independent of the arbitrary value of ΔT that you choose. This is how things work with any kind of density. For example, consider the mass density of gold. This mass density is a property of the gold itself. More precisely, it is a function of each point in the gold. For pure gold, the density is constant throughout the volume, but we could imagine impurities which would make the mass density a varying function of position, just as the above probability density was a varying function of temperature. Let's call the mass density $\rho(V)$, where V signifies the (possible) dependence on where the given point is located in the volume. And let's call the small volume we're concerned with ΔV. Then the mass in the small volume ΔV is given by the product of the density and the volume, that is, $\rho(V) \cdot \Delta V$. This is directly analogous to the fact that the probability in the above temperature example is given by the product of the probability density and the temperature span, that is, $\rho(T) \cdot \Delta T$. The correspondence among the various quantities is

   $$\begin{aligned} \text{Mass in volume } \Delta V \text{ around location } V &\iff \text{Prob that temp lies in } \Delta T \text{ around } T \\ \rho(V) &\iff \rho(T) \\ \Delta V &\iff \Delta T. \quad \clubsuit \end{aligned} \tag{3.4}$$

3.2.3 Probability equals area

Graphically, the interpretation of the product $\rho(T) \cdot \Delta T$ in Eq. (3.2) is that it is the area of the rectangle shown in Fig. 3.6. This is true because ΔT is the base of the rectangle, and $\rho(T)$ is the height.

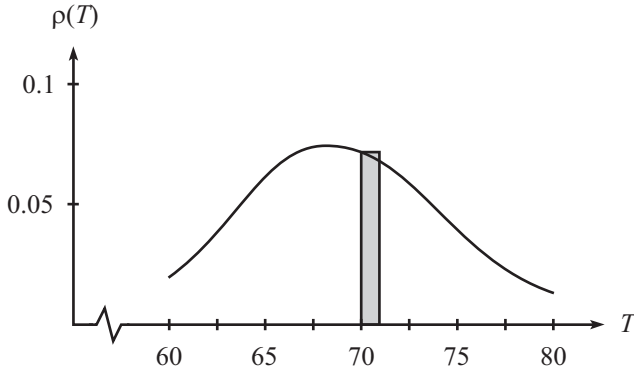


Figure 3.6

We have chosen ΔT to be 2°, so the area of the rectangle, which is $\rho(70^\circ) \cdot (2^\circ)$, gives the probability that the temperature falls between 70° and 72°. But not exactly, because $\rho(T)$ isn't constant over this 2° interval. A better approximation of the probability that the temperature falls between 70° and 72° is achieved by splitting the interval into two intervals of 1° each, and then adding up the probabilities of falling into these two intervals. These two probabilities are $\rho(70^\circ) \cdot (1^\circ)$ and $\rho(71^\circ) \cdot (1^\circ)$, and the two corresponding rectangles are shown in Fig. 3.7,

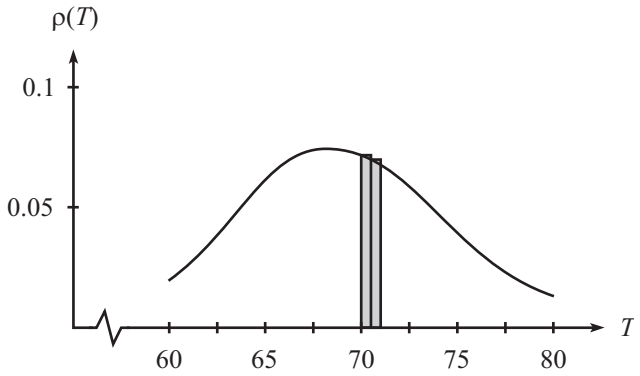


Figure 3.7

But again, the sum of the areas of these two rectangles is still only an approximate result for the probability that the temperature falls between 70° and 72°, because $\rho(T)$ isn't constant over the 1° intervals either. A better approximation is achieved by splitting the 1° intervals into smaller intervals, and then again into even smaller ones. And so on. When we get to the point of having 100 or 1000 extremely thin rectangles, the sum of their areas will essentially be the area shown in Fig. 3.8.

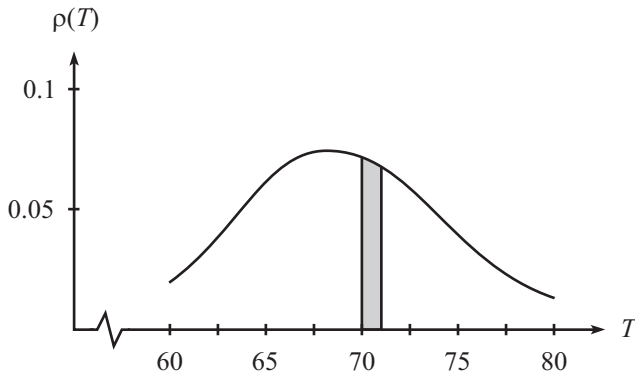


Figure 3.8

We therefore arrive at a more precise definition of the probability density, $\rho(T)$:

- Improved definition of the probability density, $\rho(T)$:

  $\rho(T)$ is the quantity for which the area under the $\rho(T)$ curve between T and $T + \Delta T$ gives the probability that the temperature (or whatever quantity we're dealing with) falls between T and $T + \Delta T$.

This is an exact definition, and there is no need for the word "small," as there was in the definition involving Eq. (3.2).

Note that the total area under any probability density curve must be 1, because this area represents the probability that the temperature takes on some value between $-\infty$ and $+\infty$. (Although, in any realistic case, the density is essentially zero outside a reasonably small region, so there is essentially no contribution to the area outside that small region.) Since the temperature must take on *some* value, the total probability (and hence area) must be 1. And indeed, the total area under the preceding curves (including the tails on either side, which we haven't bothered to draw) equals 1. Well, at least roughly; the curves were drawn by hand.

3.3 Specific types of distributions

We'll now spend a fair amount of time on some of the more common types of probability distributions. There is technically an infinite number of possible distributions, although only a hundred or so come up frequently enough to have names. And even many of these are rather obscure. A handful, however, come up again and again in a variety of settings, so we'll concentrate on these.[1] As we mentioned in the introduction to this chapter, we'll derive the distributions here, but we'll generally postpone the discussion of their various properties until the following chapter.

3.3.1 Bernoulli

The Bernoulli distribution is very simple. It deals with a process in which only two possible outcomes can occur, with probabilities p and $1 - p$ (they must add up to 1, of course). If the two outcomes of the random variable (whatever it may be) are generically labeled as

[1] If you randomly look through a large set of books and make a note of how many times you encounter the various types of distributions, you'll end up with a probability distribution of types of probability distributions, with your random variable being the type of probability distribution! Sorry if that confuses you, but I couldn't resist.

A and B, then the plot of the probability distribution is shown in Fig. 3.9. The Bernoulli distribution is a discrete one, of course.

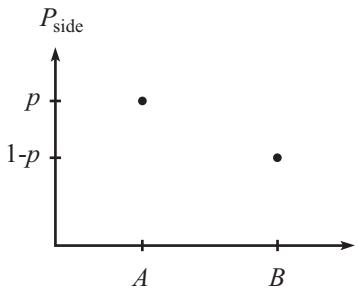


Figure 3.9

A special case of a Bernoulli distribution is the distribution for a coin toss, where the two probabilities for Heads and Tails are both equal to 1/2. The Bernoulli distribution is the simplest of all the distributions, with the exception of the trivial case where only one possible outcome can occur, which therefore has a probability of 1.

3.3.2 Binomial

An example of a binomial distribution is the probability distribution for the number of Heads in five coin tosses, which we discussed in Section 3.1. The technical definition of a binomial distribution is that it is the probability distribution for the total number of successes that arise from a collection of Bernoulli processes (that is, processes with only two possible outcomes). It is a discrete distribution.

In the case of the five coin tosses, each coin toss is a Bernoulli process, and when we put all five together and look at the total number of successes (which we'll define to be Heads), then we get a binomial distribution. In this specific example, there are $N = 5$ Bernoulli processes, with each having a $p = 1/2$ probability of success. The probability distribution is plotted above in Fig. 3.1. For the case of general N and p, the probability distribution can be found as follows. We'll change notation from P_n to $P(n)$. So our goal is to find the value of $P(n)$ for all the different possible values (from 0 to N) of the total number of Heads, n.

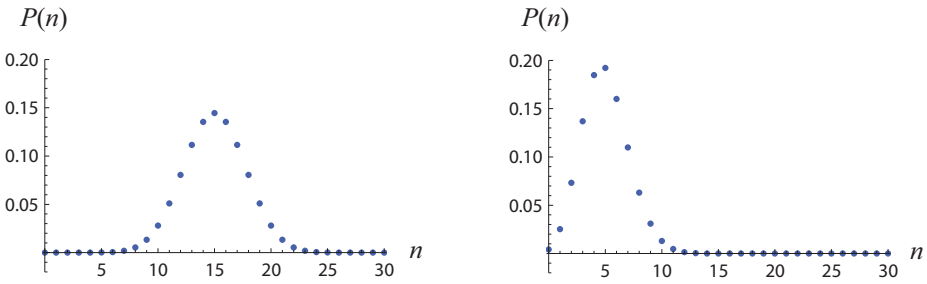
The probability that a *specific set* of n of the N Bernoulli processes all yield success is p^n, because each of the n processes has a p chance of yielding success. We then need the other $N - n$ processes to *not* yield success (because we want exactly n successes). This happens with probability $(1-p)^{N-n}$, because each of the $N-n$ processes has a $1-p$ chance of yielding a failure. So the probability that this specific set of n processes (and no others) all yield success is $p^n \cdot (1-p)^{N-n}$. Finally, since there are $\binom{N}{n}$ ways to pick this specific set of n processes, we see that the probability that exactly n of the N processes yield success is

$$\boxed{P(n) = \binom{N}{n} p^n (1-p)^{N-n}} \tag{3.5}$$

This is the desired binomial distribution. Coin tosses yield a special case of this. If we define Heads to be success, then $p = 1/2$, and Eq. (3.5) reduces to

$$P(n) = \frac{1}{2^N}\binom{N}{n}. \tag{3.6}$$

To recap: In Eq. (3.5), N is the total number of Bernoulli processes, p is the probability of success in each Bernoulli process, and n is the random variable representing the total number of successes in the N processes (so n can be anything from 0 to N). Fig. 3.10 shows the binomial distribution for the case where $N = 30$ and $p = 1/2$ (which arises from 30 coin flips), and also where $N = 30$ and $p = 1/6$ (which arises from 30 die rolls, with one of the six numbers representing success).



$p = 1/2,\ N = 30$ $p = 1/6,\ N = 30$

Figure 3.10

Having derived the binomial distribution in Eq. (3.5), there is a simple double check we can perform on the result. Since the number of successes can be any number from 0 to N, the sum of the $P(n)$ probabilities from $n = 0$ to $n = N$ must equal 1. The $P(n)$ expression in Eq. (3.5) does indeed satisfy this requirement, due the binomial expansion. The binomial formula gives

$$\Big(p + (1-p)\Big)^N = \sum_{n=0}^{N} \binom{N}{n} p^n (1-p)^{N-n}. \tag{3.7}$$

The lefthand side is simply $1^N = 1$. And each term in the sum on the righthand side is a $P(n)$ term from Eq. (3.5). So Eq. (3.7) becomes

$$1 = \sum_{n=0}^{N} P(n), \tag{3.8}$$

as we wanted to show. You are encouraged to verify this result for the probabilities in, say, the left plot in Fig. 3.10. Feel free to make rough estimates of the probabilities when reading them off from the plot. You will find that the sum is indeed 1, up to the rough estimates you make.

The task of Problem 2 is to use Eq. (2.33) to explicitly demonstrate that the expectation value of the binomial distribution in Eq. (3.5) equals pN (which must be true, of course, because on average a fraction p of the N trials yields success, by the definition of p).

REMARK: We should emphasize what is meant by a probability distribution. Let's say that you want to experimentally verify that the left plot in Fig. 3.10 is the correct distribution for the total number of Heads that show up in 30 coin flips. You of course can't do this by flipping a coin just once. And you can't even do it by flipping a coin 30 times, because all you'll get from that is just one number for the total number of Heads, for example, 17. In order to experimentally verify the distribution, you need to perform *a large number of sets of 30 coin flips,* and you need to record the total number of Heads you get in each 30-flip set. The result will be a long string of numbers such as 13,16,15,16,18,14,12,15,.... If you then calculate the fractions of the time that each number

appears, these fractions should (roughly) agree with the probabilities given in the plot. And the longer the string of numbers, the better the agreement, in general. The main point here is that the distribution does't say much about *one particular* set of 30 flips. Rather, is says what the expected distribution of outcomes is for a *large number* of sets of 30 flips. ♣

3.3.3 Uniform

The above Bernoulli and binomial distributions are discrete distributions. Let's now look at a very simple continuous probability distribution, namely one that is uniform over a given interval, and zero otherwise. Such a distribution might look like the one shown in Fig. 3.11.

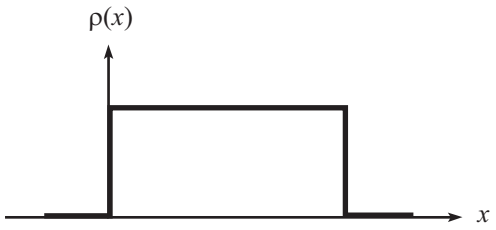


Figure 3.11

This distribution could arise, for example, from a setup where a rubber ball bounces around in an empty rectangular room. When it finally comes to rest, we measure its distance, x, from a particular one of the walls. If you initially throw the ball hard enough, then it's a pretty good approximation to say that x is equally likely to take on any value between 0 and L, where L is the length of the room in the relevant direction.

The random variable here is x, so we plot that on the horizontal axis. On the vertical axis we plot the probability *density* (not the probability!).[2] If L equals 10 feet, then outside the region $0 < x < 10$, the probability density, $\rho(x)$, equals zero. Inside this region, the density equals the total probability divided by the total interval, which gives 1 per 10 feet, or equivalently 1/10 per foot. If we want to find the actual probability that the ball ends up between, say, $x = 6$ and $x = 8$, then we just multiply $\rho(x)$ by the interval length, which is 2 feet. The result is (1/10 per foot)(2 feet), which equals $2/10 = 1/5$. This makes complete sense, of course, because the interval is 1/5 of the total distance.

The reason why this is so simple is because the density is uniform, which means that the area under a given part of the curve (which equals the desired probability, as explained in Section 3.2.3) is simply a rectangle. And the area of a rectangle is just the base times the height, which here is the interval length times the density, which is exactly the product we formed above. When the density isn't unform, it can be very difficult sometimes to find the area under a given part of the curve.

Note that the larger the region of nonzero $\rho(x)$ in a uniform distribution, the smaller the value of $\rho(x)$. This follows from the fact that the total area under the density "curve" must equal 1. So if the base gets longer, the height must get shorter.

3.3.4 Exponential

Let's now look at some probability distributions that are a little more complicated than the three above. We'll start with the exponential distribution, which takes the general form,

$$\boxed{\rho(t) = Ae^{-bt}} \tag{3.9}$$

[2] See the discussion in Section 3.2.2.

where A and b are quantities that depend on the specific situation at hand (we will find that they must be related in a certain way if the total probability is to be 1), and t stands for whatever the relevant random variable is. This is a continuous distribution, so $\rho(t)$ is a probability density. The most common type of situation where this distribution arises is the following.

Consider a repeating event that happens completely randomly in time. By "completely randomly" we mean that there is a uniform probability that the event happens at any given instant (or more precisely, in any small time interval), independent of what has already happened. That is, the process has no "memory." Time here is a continuous quantity, and it will require some formalism to analyze this situation. So before tackling this, let's consider the slightly easier case where time is assumed to be discrete. The main result we'll eventually arrive at below is the expression for the probability distribution (for the continuous-time case) of the waiting time until the next event occurs. We will find that it takes the form of an exponential distribution; see Eq. (3.21).

Discrete case

Consider a process where we roll a hypothetical 10-sided die once every second. So time is discretized into 1-second intervals.[3] If the die shows a "1," we consider that a success. The other nine numbers represent failure. There are two reasonable questions we can ask: What is the average waiting time between successes? And what is the probability distribution of the waiting times between successes?

Average waiting time

It's fairly easy to determine the average waiting time. There are 10 possible numbers, so on average we can expect 1/10 of them to be 1's. For example, if we run the process for an hour, which consists of 3600 seconds, then we can expect to get about 360 1's. So the average waiting time is (3600 seconds)/360 = 10 seconds.

More generally, if the probability of success for each trial is p, then the waiting time is $1/p$ (assuming that the trials happen at 1-second intervals). This can be seen by the same reasoning as above. If we perform N trials of the process, then on average pN of them will yield success. The average waiting time between these successes is then $N/(pN) = 1/p$, as desired.

Distribution of waiting times

Determining the probability distribution of the waiting times is more difficult. For the 10-sided die example, the question we're trying to answer is: What is the probability that if we consider two successive 1's, the time between them will be 6 seconds? Or 30 seconds? Or 1 second? And so on. Although the average waiting time is 10 seconds, this certainly doesn't mean that it will always be 10 seconds. In fact, we will find below that the probability that the waiting time is exactly 10 seconds in quite small.

Let's be general and say that the probability of success is p (so $p = 1/10$ here). Then the question is: What is the probability, P_n, that we will have to wait exactly n iterations (1 second here) to obtain the next success? Well, in order for the next success to happen on the nth iteration, there must be failure (which happens with probability $1 - p$) on the next

[3] It's actually not necessary to introduce time here at all. We could simply talk about the number of iterations of the process. But it's much easier to talk abut things like "waiting time" than "the number of iterations you need to wait for." So for convenience, we'll discuss things in the context of time.

$n-1$ iterations, and then success on the nth one. The probability of this happening is

$$\boxed{P_n = (1-p)^{n-1}p} \tag{3.10}$$

This is the desired (discrete) probability distribution for the waiting time. We see that the probability that the next success comes on the next iteration is p, the probability that it comes on the second iteration is $(1-p)p$, the probability that it comes on the third iteration is $(1-p)^2p$, and so on. A plot of this distribution for $p = 1/10$ is shown in Fig. 3.12. Note that it is maximum at $n = 1$ and falls off from that value. Even though $n = 10$ is the average waiting time, the probability of the waiting time being *exactly* $n = 10$ is only $P_{10} = (0.9)^9(0.1) \approx 0.04 = 4\%$.

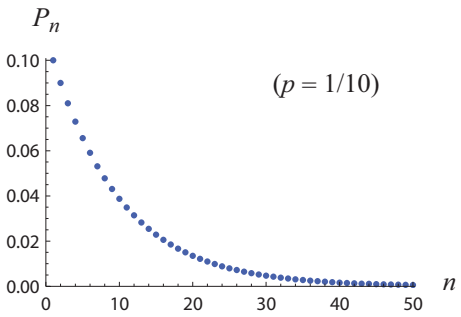


Figure 3.12

As a double check on result in Eq. (3.10), we know that the next success has to happen *sometime*, sooner or later, so the sum of all the P_n probabilities must be 1. These P_n probabilities form a geometric series whose first term is p and whose ratio is $1-p$. The general formula for the sum of a geometric series with first term a and ratio r is $a/(1-r)$, so we have

$$\begin{aligned} P_1 + P_2 + P_3 + P_4 + \cdots &= p + p(1-p) + p(1-p)^2 + p(1-p)^3 + \cdots \\ &= \frac{p}{1-(1-p)} \qquad (\textit{sum of a geometric series}) \\ &= 1, \end{aligned} \tag{3.11}$$

as we wanted to show. As another check, we can verify that the expectation value (the average) of the waiting times for the distribution given in Eq. (3.10) is $1/p$, as we already found above. The is the task of Problem 3.

You are encouraged to use a die to experimentally verify Eq. (3.10), or equivalently the plot analogous to Fig. 3.12, for the case of $p = 1/6$. Just roll the die as many times as you can in five minutes or so, and write down a "+" if you get a 1, and a "−" if you get anything else. Then make a long list of the waiting times between 1's. Then count up the number of one-roll waits, two-roll waits, and so on. Then divide by the total number of waits (not the total number of rolls!) to find the probability of each waiting length. The results should be (roughly) consistent with Eq. (3.10) for $p = 1/6$.

Rates, expectation values, and probabilities

We'll now consider the case where time is a continuous quantity. That is, we'll assume that we can have a "successful" event at *any* instant, not just on the evenly-spaced 1-second

marks as above. A continuous process that is uniform in time can be completely described by just *one* number: the average rate of success, which we'll call r. We generally won't bother writing the word "average," so we'll just call r the "rate." Before getting into the derivation of the continuous exponential distribution below, we'll need to talk a little about rates.

The rate r can be determined by counting the number of successful events that occur during a long time interval, and then dividing by this time. For example, if 300 events happen during 100 minutes, then the rate r is 3 events per minute.[4] You can also write this as 1 event per 20 seconds, or 1/20 of an event per second. There is an infinite number of ways to write r, and it's personal preference which one you pick. Just remember that you have to state the time interval you're using. If you just say that the average rate is 3, then that is a meaningless statement.

What is the expectation value of the number of events that happen during a time t? This average number equals the product rt, simply from the definition of r. If it were anything else, then if we divided it by t to get the rate, we wouldn't get r. So we have:

$$(\text{Expected number of events in time } t) = rt. \tag{3.12}$$

In the above setup where r equals 3 events per minute, the expected number of events that happen in, say, five minutes is

$$rt = (3 \text{ events per minute})(5 \text{ minutes}) = 15 \text{ events}. \tag{3.13}$$

Does this mean that we are guaranteed to get exactly 15 events during a particular 5-minute span? Absolutely not. We can theoretically get any number of events, although there is essentially zero chance that the number will differ significantly from 15.[5] But the *expectation value* is 15. That is, if we perform a huge number of 5-minute trials and then calculate the average number of events that occur in each trial, the result will be very close to 15.

A trickier question to ask is: What is the probability that *exactly* one event happens during a time t? Since r is the rate, you might think that you could just multiply r by t again to say that the probability is rt. But this certainly can't be right, because it would imply a probability of 15 for a 5-minute interval. This is nonsense, because probabilities can't be larger than 1. Even if we picked a time interval of 20 seconds (1/3 of a minute), we would obtain an rt value of 1. This doesn't have the fatal flaw of being larger than 1, but it has another issue, in that it says that exactly one event is *guaranteed* to happen during this 20-second interval. This can't be right either, because it's certainly possible for zero (or two or three, etc.) events to occur. We'll figure out the exact probability of these in Section 3.3.5.

The strategy of multiplying r by t to obtain a probability doesn't seem to work. However, there is one special case in which it *does* work. If the time interval is very small (let's call it ϵ, which is the standard letter to use for something that is very small), then it *is* true that the probability of exactly one event occurring during the ϵ time interval is essentially equal to $r\epsilon$. We're using the word "essentially" here because although this statement is technically not true, it becomes arbitrarily close to being true in the limit where ϵ approaches zero. In the present example with $r = 1/20$ events per second, the statement, "rt is the probability that exactly one event happens during a time t," is a lousy approximation if $t = 20$ seconds, a decent approximation if $t = 2$ seconds, and an excellent approximation if $t = 0.2$ seconds.

[4] Of course, if you happen to count the number of events in a different span of 100 minutes, you'll most likely get a slightly different number, perhaps 312 or 297. But in the limit of a very long time interval, you will find essentially the same rate, independent of which specific interval you use. This is a consequence of the results we'll derive in Section 3.4.

[5] The probability of obtaining the various numbers of events is governed by the Poisson distribution, which we'll discuss in Section 3.3.5.

And it only gets better as the time interval gets smaller. We'll explain why in the first remark below.

So if $P_\epsilon(1)$ stands for the probability that exactly one event happens during a small time interval ϵ, then we can say that

$$P_\epsilon(1) \approx r\epsilon \qquad \text{(if } \epsilon \text{ is very small)} \tag{3.14}$$

The smaller ϵ is, the better this approximation is. When we deal with continuous time below, we'll actually be taking the $\epsilon \to 0$ limit. In this mathematical limit, the "$\approx$" sign in Eq. (3.14) becomes an exact "=" sign.

To sum up, if t is very small, then rt is both the expected number of events that occur during the time t and also (essentially) the probability that exactly one event occurs during the time t. But if t isn't very small, then rt is only the expected number of events.

Remarks:

1. The reason why rt equals the probability of exactly one event occurring only if t is very small is because if t *isn't* small, then there is the possibility of *multiple* events occurring during the time t. We can be explicit about this as follows. Since we know from Eq. (3.12) that the expected number of events during any time t is rt, we can use the expression for the expectation value in Eq. (2.33) to write

   $$rt = P_0 \cdot 0 + P_1 \cdot 1 + P_2 \cdot 2 + P_3 \cdot 3 + \cdots, \tag{3.15}$$

   where P_n indicates the probability of obtaining exactly n events during the time t. Solving for P_1 gives

   $$P_1 = rt - P_2 \cdot 2 - P_3 \cdot 3 + \cdots. \tag{3.16}$$

   We see that P_1 is smaller than rt due to all the P_2 and P_3, etc. probabilities. So P_1 doesn't equal rt. However, if all of the probabilities of multiple events occurring (P_2, P_3, etc.) are very small, then P_1 is *essentially* equal to rt. And this is exactly what happens if the time interval is very small. For small times, there is hardly any chance of the event even occurring once. So it is even less likely that it will occur twice, and even less for three times, etc. Roughly speaking, if the probability that exactly one event occurs during a small time ϵ is $r\epsilon$, then the probability that exactly two events occur should be proportional to $(r\epsilon)^2$.[6] The important point here is that $(r\epsilon)^2$ is quadratic in ϵ, so if ϵ is sufficiently small, then $(r\epsilon)^2$ is negligible compared with $r\epsilon$. In other words, we can completely ignore the scenarios where multiple events occur. Eq. (3.16) then gives $P_1 \approx rt$, in agreement with Eq. (3.14), in slightly different notation.

2. The area under the r vs. t "curve" (which we're assuming is just a constant flat line) that corresponds to a time interval Δt is equal to $r\Delta t$ (since it's a rectangular region). So from Eq. (3.14), this area gives the probability that an event occurs during a time Δt, provided that Δt is very small. This might make you think that r can be interpreted as a probability distribution, because we found in Section 3.2.3 that the area under a distribution curve gives the probability. However, the r "curve" *cannot* be interpreted as a probability distribution, because this area-equals-probability result holds *only for very small* Δt. The area under a probability distribution curve has to give the probability for *any* interval on the horizontal axis. The r "curve" doesn't satisfy this property. Said in a different way, the total area under the r "curve" is infinite, whereas actual probability distributions must have a total area of 1.

3. Since only *one* quantity, r, is needed to describe everything about a random process that is uniform in time, any other quantity that we might want to determine must be able to be written in terms of r. This will become evident below. ♣

[6] As we'll see in Section 3.3.5, there's actually a factor of 1/2 involved here, but that is irrelevant for the present argument.

Continuous case

In the above case of discrete time, we asked two questions: What is the average waiting time between successes? And what is the probability distribution of the waiting times between successes? We'll now answer these two questions for the case where time is a continuous quantity.

Average waiting time

As in the discrete case, the first of these questions is fairly easy to answer. Let the average rate of success be r, and consider a large time t. We know from Eq. (3.12) that the average number of events that occur during the time t is rt. Let's label this number as n_t. The average waiting time (which we'll call T) is simply the total time divided by the number of occurrences. So we have

$$T = \frac{t}{n_t} = \frac{t}{rt} \qquad \Longrightarrow \qquad \boxed{T = \frac{1}{r}} \tag{3.17}$$

We see that the average waiting time is simply the reciprocal of the rate. It makes sense that r is in the denominator, because if r is small, the average waiting time is large. And if r is large, the average waiting time is small. And as promised in the third remark above, T depends on r.

Distribution of waiting times

Now let's answer the second (more difficult) question: what is the probability distribution of the waiting times between successes? Equivalently, what is the probability that the time from a given event to the next one is between t and $t + \Delta t$, where Δt is small? To answer this, we'll use the same general strategy that we used above in the discrete case, except that now the time intervals will be a very small time ϵ instead of 1 second. We will then take the limit $\epsilon \to 0$, which will make time be essentially continuous.

The division into time intervals is summarized in Fig. 3.13. From time zero (which is when we'll assume the first event happens) to time t, we'll break time up into a very large number of very small intervals of length ϵ (which means that there are t/ϵ of these intervals). And then the interval of Δt sits at the end. Both ϵ and Δt are assumed to be very small, but they need not have anything to do with each other. ϵ exists as a calculational tool only, and Δt is the arbitrarily-chosen small time interval that appears in Eq. (3.2).

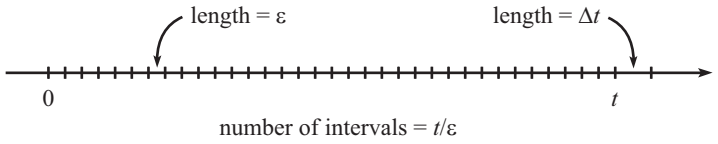


Figure 3.13

In order for the next success to happen between t and $t+\Delta t$, there must be failure during every one of the ϵ intervals shown in Fig. 3.13, and then there must be success between t and $t+\Delta t$. From Eq. (3.14), the latter happens with probability $r\,\Delta t$. Also, Eq. (3.14) says that the probability of success in a given one of the ϵ intervals is $r\epsilon$, so the probability of failure is $1 - r\epsilon$. And since there are t/ϵ of these intervals, the probability of failure in all of them is $(1 - r\epsilon)^{t/\epsilon}$. The probability that the next success happens between t and $t + \Delta t$ is therefore

$$P_{t,\Delta t} = \Big((1 - r\epsilon)^{t/\epsilon}\Big)\big(r\,\Delta t\big). \tag{3.18}$$

It's now time to use one of the results from Appendix B, namely the approximation given in Eq. (4.7), which says that for small a we can write[7]

$$(1+a)^n \approx e^{an}. \tag{3.19}$$

This works for negative a as well as positive a. Here e is the "natural logarithm" which has a value of $e \approx 2.71828$. (If you want to know more about e, there's plenty of information in Appendix B!) For the case at hand, a comparison of Eqs. (3.18) and (3.19) shows that we want to define $a \equiv -r\epsilon$ and $n \equiv t/\epsilon$, which yields $an = (-r\epsilon)(t/\epsilon) = -rt$. Eq. (3.19) then gives $(1-r\epsilon)^{t/\epsilon} \approx e^{-rt}$, and so Eq. (3.18) becomes

$$P_{t,\Delta t} = e^{-rt} r\,\Delta t. \tag{3.20}$$

The probability distribution (or density) is obtained by simply erasing the Δt, because Eq. (3.2) says that the density is obtained by dividing the probability by the interval length. So we see that the desired probability distribution of the waiting time between successes is $\rho_{\text{wait}}(t) = re^{-rt}$. Note that ρ_{wait} is a function of t, as expected. And as promised in the third remark on page 81, it depends on r. Although this is the answer, it's generally more natural to think in terms of the waiting time T than the rate r, so we'll write this result as (using $r = 1/T$ from Eq. (3.17))

$$\boxed{\rho_{\text{wait}}(t) = \frac{e^{-t/T}}{T}} \tag{3.21}$$

The "exponential" name of this distribution comes from the exponential function $e^{-t/T}$. In the notation of Eq. (3.9), both A and b are equal to $1/T$. So they are in fact related, as we noted right after Eq. (3.9).

Fig. 3.14 shows plots of $\rho_{\text{wait}}(t)$ for a few different values of the average waiting time, T. The two main properties of each of these curves are the starting value at $t = 0$, and the rate of decay as t increases. From Eq. (3.21), the starting value at $t = 0$ is $e^0/T = 1/T$. So the bigger T is, the smaller the starting value. This makes sense, because if the average waiting time T is large (equivalently, the rate r is small), then there is a small chance that the next event will happen right away.

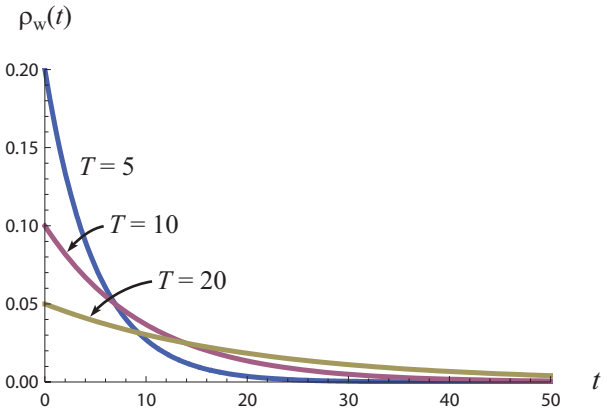


Figure 3.14

[7] You are strongly encouraged to read Appendix B at this point, if you haven't already. But if you want to take this equation on faith, that's fine too. But you should at least verify with a calculator that it works for, say, $a = 0.01$ and $n = 100$.

How fast does the curve decay? This is governed by the denominator of the exponent in Eq. (3.21). For every T units that t increases by, $\rho_{\text{wait}}(t)$ decreases by a factor of $1/e$. This can be seen by plugging a time of $t+T$ into Eq. (3.21), which gives

$$\rho_{\text{wait}}(t+T) = \frac{e^{-(t+T)/T}}{T} = \frac{(e^{-t/T} \cdot e^{-1})}{T} = \frac{1}{e} \cdot \frac{e^{-t/T}}{T} = \frac{1}{e}\,\rho_{\text{wait}}(t). \tag{3.22}$$

So $\rho_{\text{wait}}(t+T)$ is $1/e$ times as large as $\rho_{\text{wait}}(t)$, and this holds for any value of t. A few particular values of $\rho_{\text{wait}}(t)$ are

$$\begin{aligned} \rho_{\text{wait}}(0) &= \frac{1}{T}\,, \\ \rho_{\text{wait}}(T) &= \frac{1}{eT}\,, \\ \rho_{\text{wait}}(2T) &= \frac{1}{e^2T}\,, \\ \rho_{\text{wait}}(3T) &= \frac{1}{e^3T}\,, \end{aligned} \tag{3.23}$$

and so on. If T is large, the curve takes longer to decrease by a factor of $1/e$. This is consistent with Fig. 3.14, where the large-T curve falls off slowly, and the small-T curve falls off quickly.

To sum up, if T is large, then the $\rho_{\text{wait}}(t)$ curve starts off low and decays slowly. And if T is small, the curve starts off high and decays quickly.

Remarks:

1. In comparing Eq. (3.18) with Eq. (3.10), we see in retrospect that we could have obtained Eq. (3.18) by simply replacing the first p in Eq. (3.10) with $r\epsilon$ (because $r\epsilon$ is the probability of success at each intermediate step), the second p with $r\,\Delta t$ (this is the probability of success at the last step), and $n-1$ with t/ϵ (this is the number of steps). But you might find these replacements a bit mysterious without the benefit of the reasoning preceding Eq. (3.18).

2. The area under any of the curves in Fig. 3.14 must be 1, because the waiting time has to be *something*, so the the sum of all the probabilities must be 1. The proof of this fact is very quick, but it requires calculus, so we'll skip it here. (But we demonstrated it for the discrete case in Eq. (3.11).) Likewise, the expectation value of the waiting time must be T, because that's how T was defined. Again, the proof is quick but requires calculus. (The demonstration for the discrete case is the task of Problem 3.)

3. We've been referring to $\rho_{\text{wait}}(t)$ as the probability distribution of waiting times from one event to the next. However, it's actually the distribution of waiting times from *any point in time* to the occurrence of the next event. That is, you can start your stopwatch at any time, not just at the occurrence of an event. If you go back through the above discussion, you'll see that nowhere did we use the fact that an event actually occurred at $t = 0$.

   But beware of the following incorrect reasoning. Let's say that an event happens at $t = 0$, but that you don't start your stopwatch until, say, $t = 1$. The fact that the next event after $t = 1$ doesn't happen on average until $t = 1+T$ (from the previous paragraph) seems to imply that the average waiting time from $t = 0$ is $1 + T$. But it better not be, because we know from above that it's just T. The error here is that we forgot about the scenarios where the next event after $t = 0$ happens *between* $t = 0$ and $t = 1$. When these events are considered, the average waiting time ends up correctly being T. Basically, the waiting time from $t = 1$ is still T, but the next event (after the $t = 0$ one) might have already happened before $t = 1$.

4. In a sense, all the curves for different values of T in Fig. 3.14 are really the same curve. They're just stretched or squashed in the horizontal and vertical directions. The general form of the curve described by the expression in Eq. (3.21) is shown in Fig. 3.15.

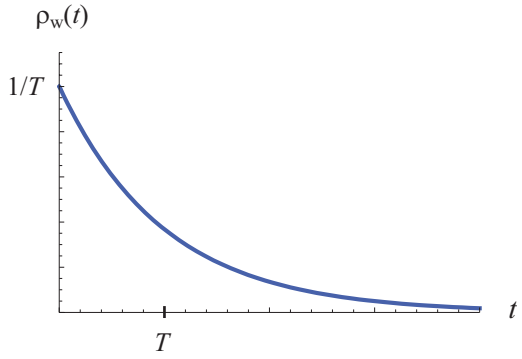


Figure 3.15

As long as we change the scales on the axes so that T and $1/T$ are always located at the same positions, then the curves will look the same for any T. For example, as we saw in Eq. (3.23), no matter what the value of T is, the value of the curve at $t = T$ is always $1/e$ times the value at $t = 0$. Of course, when we plot things, we generally keep the scales fixed, in which case the T and $1/T$ positions move along the axes, as shown in Fig. 3.16 (these are the same curves as in Fig. 3.14).

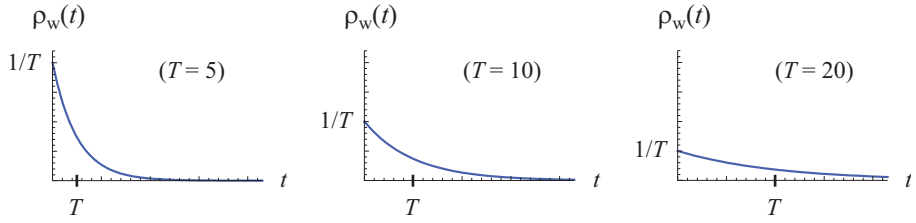


Figure 3.16

5. The fact that any of the curves in Fig. 3.16 can be obtained from any of the other curves by stretching and squashing the two directions by inverse factors implies that every curve has the same area. (This is consistent with the fact that all the areas must be 1.) To see how these inverse factors work together to keep the area constant, imagine the area being broken up into a large number of thin vertical rectangles, stacked side by side under the curve. The stretching and squashing of the curve does the same thing to each rectangle. All the widths get stretched by a factor of f, and all the heights get squashed by the same factor of f (or $1/f$, depending on your terminology). So the area of every rectangle remains the same. The same thing must then be true for the area under the whole curve.
6. Note that the distribution for the waiting time is a discrete distribution in the case of discrete time, and a continuous distribution for continuous time. This might sound like a tautological statement, but it actually isn't, as we'll see in the Poisson case below. ♣

3.3.5 Poisson

The Poisson probability distribution takes the general form,

$$\boxed{P_{\text{Poisson}}(n) = \frac{a^n e^{-a}}{n!}} \tag{3.24}$$

where a is a quantity that depends on the specific situation at hand, and n is the random variable, which is the number of events that occur in a certain region of time (or space, or

whatever), as we'll discuss below. The most common type of situation where this distribution arises is the following.

As with the exponential distribution in the previous section, consider a repeating event that happens completely randomly in time. The main result of this section is that the probability distribution of the *number of events that happen during a given time interval* takes the form of a Poisson distribution (see Eq. (3.33) below). As in the case of the exponential distribution, our strategy for deriving this will be to first consider the case of discrete time, and then the case of continuous time.

Discrete case

Consider a process that is repeated each second (so time is discretized into 1-second intervals), and let the probability of success for each iteration be p (the same for all iterations). For example, as in the previous section, we can roll a hypothetical 10-sided die once every second, and if the die shows a "1," then we consider that a success. The other nine numbers represent failure.[8]

The question we will answer here is: What is the probability distribution of the number of successes that happen in a time interval of length N seconds? In other words, what is the probability, $P(n)$, that exactly n events happen during a time span of N seconds? It turns out that this is *exactly* the same question we answered in Section 3.3.2 when we derived the binomial distribution in Eq. (3.5). So we can basically just copy over the reasoning here. We'll formulate things in the language of rolls of a die, but the setup could be anything with a probability p of success.

The probability that a *specific set* of n of the N rolls all yield a 1 is p^n, because each of the n rolls has a p chance of yielding a 1. We then need the other $N-n$ rolls to *not* yield a 1 (because we want exactly n successes). This happens with probability $(1-p)^{N-n}$, because each of the $N-n$ rolls has a $1-p$ chance of being something other than a 1. So the probability that this specific set of n rolls (and no others) all yield success is $p^n \cdot (1-p)^{N-n}$. Finally, since there are $\binom{N}{n}$ ways to pick this specific set of n rolls, we see that the probability that exactly n of the N rolls yield a 1 is

$$\boxed{P(n) = \binom{N}{n} p^n (1-p)^{N-n}} \tag{3.25}$$

This is exactly the same as the binomial distribution in Eq. (3.5), so there's really nothing new here. But there will be when we discuss the continuous case below.

Example (Balls in boxes): Let N balls be thrown at random into B boxes. What is the probability, $P(n)$, that a given box has exactly n balls in it?

Solution: This is a restatement of the problem we just solved. Imagine throwing one ball each second into the boxes,[9] and consider one particular box. If a ball ends up in that box, then we'll label that as success. This happens with probability $1/B$, because there are B boxes. So the p above is $1/B$. Since we're throwing N balls into the boxes, we're simply performing N iterations of a process that has a probability $p = 1/B$ of success. Eq. (3.25) is therefore applicable, and it gives the probability of obtaining exactly n successes (that is,

[8] As in the case of the exponential distribution, it isn't necessary to introduce time here at all. We could simply talk about the number of iterations of the process, as we do in the balls-in-boxes example below.

[9] But as mentioned in the previous footnote, the time interval of one second is irrelevant. All that matters is that we perform (sooner or later) N iterations of the process.

exactly n balls in one particular box) as

$$P(n) = \binom{N}{n} \left(\frac{1}{B}\right)^n \left(1 - \frac{1}{B}\right)^{N-n}. \tag{3.26}$$

That's the answer to the problem, but let's see if it makes sense. As a concrete example, consider the case where we have $N = 1000$ balls and $B = 100$ boxes. On average, we expect to have $N/B = 10$ balls in each box. But many (in fact most) of the boxes will have other numbers of balls. Intuitively, we expect most of the boxes to to have *roughly* 10 balls (say, between 5 and 15 balls). We certainly don't expect many boxes to have, say, 2 or 50 balls.

Fig. 3.17 shows a plot of $P(n)$ for the case where $N = 1000$ and $B = 100$. As expected, it is peaked near the average value, $N/B = 10$, and it becomes negligible far enough away from $n = 10$. There is essentially no chance of having fewer than 3 or more than 20 balls in a given box (the probability of having 21 is less than 0.1%). We've arbitrarily chopped the plot off at $n = 30$ because the probabilities between $n = 30$ (or even earlier) and $n = 1000$ are indistinguishable from zero. (But technically all of these probabilities are nonzero. For example $P(1000) = (1/100)^{1000}$, because all of the 1000 balls need to end up in the given box, and each one ends up there with probability 1/100. But this number is completely negligible.)

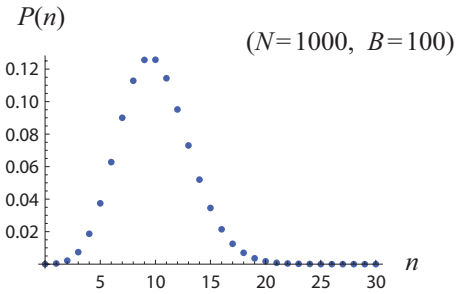


Figure 3.17

Continuous case

As with the exponential distribution in Section 3.3.4, we'll now consider the case where time is continuous. That is, we'll assume that we can have a "successful" event at *any* instant, not just on the evenly-spaced 1-second marks as above.

As in Section 3.3.4, such a process can be completely described by just *one* number: the rate of success, which we'll call r. Eq. (3.14) tells us that the actual probability that an event occurs in a very small time interval, ϵ, equals $r\epsilon$. The smaller ϵ, the smaller the probability that the event occurs. We're assuming that r is constant in time, that is, the event is just as likely to occur at one time as any other.

Our goal here is to answer the question: What is the probability, $P(n)$, that exactly n events happen during a given time span of τ? (We're using τ here instead of the N above, because we're saving N for another purpose.) To answer this, we'll use the same general strategy that we used above in the discrete case, except that now the time interval will be a very small time ϵ instead of 1 second. We will then take the $\epsilon \to 0$ limit, which will make time be essentially continuous. The division into time intervals is summarized in Fig. 3.18. We're dividing the time interval τ into a very large number of very small intervals of length ϵ. This means that there are τ/ϵ of these intervals, which we'll label as N.

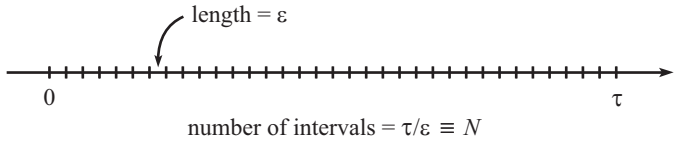


Figure 3.18

The strategy here will be the same as in the discrete case, except that the probability of success in each interval is now $r\epsilon$ instead of p. So we can basically just repeat the above derivation (which itself was a repetition of the derivation in Section 3.3.2; you're probably getting tired of it by now!):

The probability that a *specific set* of n of the N intervals all yield success is $(r\epsilon)^n$, because each of the n intervals has a $r\epsilon$ chance of yielding success. We then need the other $N - n$ intervals to *not* yield success (because we want exactly n successes). This happens with probability $(1 - r\epsilon)^{N-n}$, because each of the $N - n$ intervals has a $1 - r\epsilon$ chance of yielding failure. So the probability that this specific set of n intervals (and no others) all yield success is $(r\epsilon)^n \cdot (1 - r\epsilon)^{N-n}$. Finally, since there are $\binom{N}{n}$ ways to pick this specific set of n intervals, we see that the probability that exactly n of the N yield success is

$$P(n) = \binom{N}{n}(r\epsilon)^n(1 - r\epsilon)^{N-n}. \tag{3.27}$$

This is simply Eq. (3.25) with $r\epsilon$ in place of p.

Now the fun begins (well, assuming you like math). Let's see what this expression reduces to in the $\epsilon \to 0$ limit, which will give us the desired continuous nature of time. Note that $\epsilon \to 0$ implies that $N \equiv \tau/\epsilon \to \infty$. If we write out the binomial coefficient and expand things a little, Eq. (3.27) becomes

$$P(n) = \frac{N!}{(N - n)!\, n!}(r\epsilon)^n(1 - r\epsilon)^N(1 - r\epsilon)^{-n}. \tag{3.28}$$

Of the various letters in this equation, N is huge, ϵ is tiny, and r and n are just "normal" numbers, not assumed to be huge or tiny. r is determined by the setup, and n is the number of successes we're concerned with. (We'll see below that the relevant n's are roughly of the same size as the product $r\tau$.) In the limit $\epsilon \to 0$ (and hence $N \to \infty$) we can make three approximations to Eq. (3.28):

- First, in the $N \to \infty$ limit, we can say that

$$\frac{N!}{(N - n)!} \approx N^n, \tag{3.29}$$

  at least in a multiplicative sense (we don't care about an additive sense). This follows from the fact that $N!/(N - n)!$ is the product of the n numbers from N down to $N - n + 1$. And if N is huge compared with n, then all of these n numbers are essentially equal to N (multiplicatively). Therefore, since there are n of them, we simply get N^n. You can verify this for, say, the case of $N = 1,000,000$ and $n = 10$. The product of the numbers from $1,000,000$ down to $999,991$ equals $1,000,000^{10}$ to within an error of .005%

- Second, we can apply the $(1 + a)^n \approx e^{an}$ approximation from Eq. (4.7) in Appendix B, which we already used once in the derivation of the exponential distribution; see Eq. (3.19). We can use this approximation to simplify the $(1 - r\epsilon)^N$ term. We obtain

$$(1 - r\epsilon)^N \approx e^{-r\epsilon N}. \tag{3.30}$$

- Third, in the $\epsilon \to 0$ limit, we can use the $(1+a)^n \approx e^{an}$ approximation again to simplify the $(1-r\epsilon)^{-n}$ term. The result is

$$(1-r\epsilon)^{-n} \approx e^{-r\epsilon n} \approx e^{-0} = 1, \tag{3.31}$$

  because for any given values of r and n, the exponent here becomes infinitesimally small as $\epsilon \to 0$. Basically, we're forming a finite power of a number that's essentially equal to 1. For any given value of n, if you make ϵ smaller and smaller, $(1-r\epsilon)^{-n}$ will get closer and closer to 1. Note that this reasoning doesn't apply to the $(1-r\epsilon)^N$ term in Eq. (3.30) because N isn't a given number. It changes with ϵ, in that it becomes large as ϵ becomes small.

In the $\epsilon \to 0$ and $N \to \infty$ limits, the "$\approx$" signs in all of these approximations turn into exact "=" signs. So applying these three approximations to Eq. (3.28) gives

$$\begin{aligned} P(n) &= \frac{N!}{(N-n)!\,n!}(r\epsilon)^n(1-r\epsilon)^N(1-r\epsilon)^{-n} \\ &= \frac{N^n}{n!}(r\epsilon)^n e^{-r\epsilon N}\cdot 1 \qquad (\textit{using the three approximations}) \\ &= \frac{1}{n!}(r\epsilon N)^n e^{-r\epsilon N} \\ &\equiv \frac{1}{n!}(r\tau)^n e^{-r\tau} \qquad (\textit{using } N \equiv \tau/\epsilon) \end{aligned} \tag{3.32}$$

But from Eq. (3.12), $r\tau$ is simply the average number of events that you expect to happen in the time τ. Let's label this average number of events as $r\tau = a$. We can then write Eq. (3.32) as

$$\boxed{P_{\text{Poisson}}(n) = \frac{a^n e^{-a}}{n!}} \tag{3.33}$$

where a is the average number of successes in the time interval under consideration. This is the desired Poisson distribution. It gives the probability of obtaining exactly n successes during a period of time for which the expected number is a. The random variable here is n.

This result technically holds only in the limit of a continuous process, but it also provides a very good approximation to discrete processes, as long as the numbers involved are fairly large. Take the above balls-in-box setup, for example. Since $N = 1000$ and $B = 100$, the average number of balls in a box is $a = N/B = 10$. Since N and B are fairly large,[10] we expect that the Poisson distribution in Eq. (3.33) with $a = 10$ should provide a fairly good approximation to the exact binomial distribution in Eq. (3.25) with $N = 1000$ and $p = 1/100$ (there are $B = 100$ boxes, so $p = 1/100$). The right plot in Fig. 3.19 shows superimposed plots of these two distributions.[11]

[10] There is a certain condition that N and B need to satisfy in order for the Poisson result in Eq. (3.33) to be a good approximation to the exact discrete result in Eq. (3.25) or Eq. (3.26). Roughly speaking, the condition is $B^2 \gg N$. But we won't worry about this, because we'll generally be using the Poisson distribution in the continuum case where it applies exactly.

[11] We've drawn the Poisson distribution as a continuous curve, because it would be difficult to tell what's going on in the figure if we plotted two sets of points nearly on top of each other. (The $n!$ in Eq. (3.33) can be extrapolated for non-integer values of n.) But you should remember that we're really only concerned with integer values of n, since the n in Eq. (3.33) is the number of times something occurs. We've plotted the whole curve for visual convenience only.

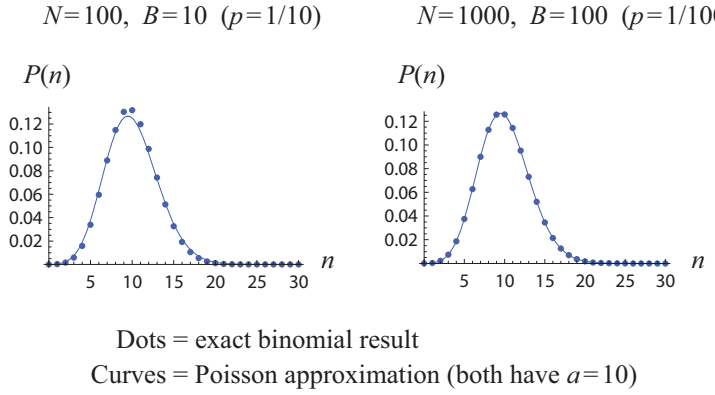


Figure 3.19

The points pretty much lie on the curve, so the approximate Poisson probabilities in Eq. (3.33) (the curve) are essentially the same as the exact binomial probabilities in Eq. (3.25) (the dots). In other words, the approximation is a very good one. However, the left plot in Fig. 3.19 shows the Poisson and binomial probabilities for the smaller pair of numbers, $N = 100$ and $B = 10$. The average $a = N/B$ still equals 10, so the Poisson curve is the same. But the exact binomial probabilities in Eq. (3.25) are changed from the $N = 1000$ and $B = 100$ case, because N is now 100, and p is now $1/10$. The Poisson approximation doesn't work as well here, although it's still reasonably good. For a given value of $a = N/B$, the larger N and B are, the better the approximation.

REMARKS:

1. The $P_{\text{Poisson}}(n)$ result in Eq. (3.33) depends on a only (along with n, of course). In the context of the balls-in-box example, this implies, as we just noted, that the $(N, B) = (1000, 100)$ combination yields the same $P_{\text{Poisson}}(n)$ distribution as, say, the $(N, B) = (100, 10)$ combination, because they both have $a = 10$.

   In the context of a continuous-time process, this a-only dependence implies the following. Let's say we have two different processes with different rates r. And let's say we consider a time interval for one process where the expected number of events is a, and also another time interval for the other process where the expected number of events is also a (so this time interval will be longer if the rate is lower, and shorter if the rate is higher). Then the probability distributions for the number of events, n, that happen in the two intervals are exactly the same. They both have the same value of $P(5)$, or $P(16)$, or anything else. The two processes have the same value of a, and that's all that matters in finding the Poisson $P(n)$.

2. For what n is $P_{\text{Poisson}}(n)$ maximum? A convenient way to find this particular n is to set $P_{\text{Poisson}}(n) = P_{\text{Poisson}}(n+1)$. This will tell us where the maximum is, because this relation can be true only if n and $n+1$ are on either side of the maximum.[12] So we have

$$\begin{aligned} P_{\text{Poisson}}(n) = P_{\text{Poisson}}(n+1) &\implies \frac{a^n e^{-a}}{n!} = \frac{a^{n+1} e^{-a}}{(n+1)!} \\ &\implies \frac{1}{1} = \frac{a}{n+1} \qquad (\textit{canceling common factors}) \\ &\implies n+1 = a. \end{aligned} \tag{3.34}$$

[12] The reason for this statement is the following. The relation $P_{\text{Poisson}}(n) = P_{\text{Poisson}}(n+1)$ can't be true on the right side of the curve, because the curve is decreasing there, so all those points have $P_{\text{Poisson}}(n) > P_{\text{Poisson}}(n+1)$. Similarly, all the points on the left side of the curve have $P_{\text{Poisson}}(n) < P_{\text{Poisson}}(n+1)$. The only remaining possibility is that n is on the left side and $n+1$ is on the right side. That is, they're on either side of the maximum.

So $n = a - 1$. The two relevant points on either side of the maximum, namely n and $n + 1$, are therefore $a - 1$ and a. So the maximum of the $P_{\text{Poisson}}(n)$ plot falls between $a - 1$ and a. Since we're concerned only with integer values of n, the maximum is located at the integer that lies between $a-1$ and a (or at both of these values if a is an integer). In situations where a is large (which is often the case), the distinction between $a - 1$ and a isn't too important, so we generally say that the maximum of the probability distribution occurs at a.

3. Eq. (3.33) works perfectly well for small a, even $a < 1$. It's just that in such a case, the plot of $P_{\text{Poisson}}(n)$ isn't a bump as in Fig. 3.19. Instead, it starts high and falls off as n increases. Fig. 3.20 shows the plot of $P_{\text{Poisson}}(n)$ for various values of a.[13] As a increases, the bump (once it actually becomes a bump) shifts to the right (because it is centered around a), decreases in height (due to the following remark), and becomes wider (due to the result in Section 3.4.2 below). The last two of these properties are consistent with each other, in view of the fact that the sum of all the probabilities must equal 1, for any value of a.

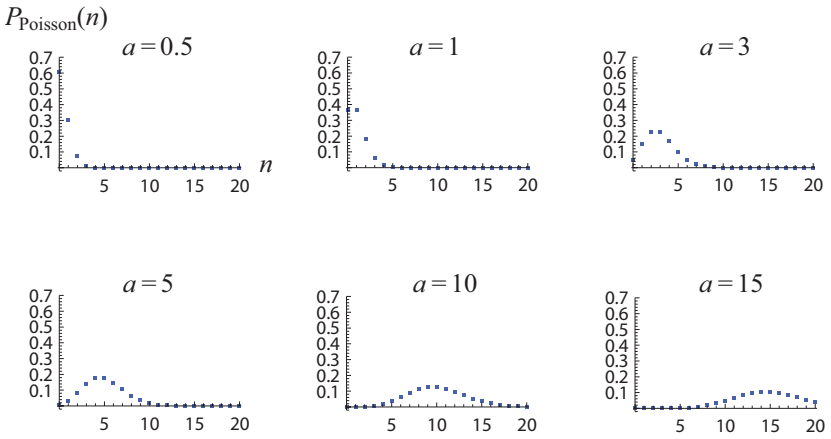


Figure 3.20

4. If a is fairly large (say, larger than 10), what is the height of the bump in the $P_{\text{Poisson}}(n)$ graph? Since we know that the peak occurs essentially at $n = a$ (from the second remark above), this is the same question as: What is the value of $P_{\text{Poisson}}(a)$? It turns out that Stirling's formula allows us to make a quick approximation to this value. Stirling's formula is given in Eq. (2.36) in Section 2.6 as[14]

$$N! \approx N^N e^{-N} \sqrt{2\pi N}. \tag{3.35}$$

Plugging $n = a$ into Eq. (3.33) gives

$$\begin{aligned} P_{\text{Poisson}}(a) &= \frac{a^a e^{-a}}{a!} \qquad (\textit{letting } n = a) \\ &\approx \frac{a^a e^{-a}}{a^a e^{-a} \sqrt{2\pi a}} \qquad (\textit{using Stirling's formula for } a!) \\ &= \frac{1}{\sqrt{2\pi a}}\,. \end{aligned} \tag{3.36}$$

We see that the height is proportional to $1/\sqrt{a}$. So if a goes up by a factor of, say, 4, then the height goes down by a factor of 2. Values of n that are close to a all have roughly the same value of $P_{\text{Poisson}}(n)$, and this (almost) common value is $1/\sqrt{2\pi a}$. It's easy to make

[13] We've arbitrarily decided to cut off the plots at $n = 20$, even though they technically go on forever. But the probabilities are pretty much zero by that point anyway, except in the $a = 15$ case.

[14] You might want to review that section now if you're not comfortable with Stirling's formula. We'll be using this formula a great deal in the remainder of this chapter.

quick estimates using this result. If $a = 15$, then $1/\sqrt{2\pi(15)}$ is about $1/\sqrt{100}$ (very roughly), which is $1/10$. So the $P_{\text{Poisson}}(n)$ values for $n = 14, 15, 16$ and also maybe 13 and 17, should be relatively close to $1/10$. From the last plot in Fig. 3.20, we see that this is indeed the case.

5. The sum of the discrete-time probabilities in Eq. (3.25), for n from 0 to N, must equal 1. Since this distribution is simply the binomial distribution, we already verified this sum-equals-1 property in Section 3.3.2; see Eqs. (3.7) and (3.8). Let's now verify this property for the continuous-time Poisson probabilities given in Eq. (3.33), for n from 0 to ∞. (With continuous time, we can technically have an arbitrarily large number of events occur during the time τ, although if n is much larger than a, then $P_{\text{Poisson}}(n)$ is negligibly small.) We have

$$\begin{aligned} \sum_{n=0}^{\infty} P_{\text{Poisson}}(n) &= \sum_{n=0}^{\infty} \frac{a^n e^{-a}}{n!} \\ &= e^{-a} \sum_{n=0}^{\infty} \frac{a^n}{n!} \\ &= e^{-a} e^{a} \qquad (\textit{using Eq. (4.11)}) \\ &= 1, \end{aligned} \tag{3.37}$$

as desired. We invoked Eq. (4.11) from Appendix B. (You are encouraged to read the derivation of that equation, but it isn't critical.)

6. In the last remark in Section 3.3.4, we noted that the exponential distribution for the waiting time, t, is a discrete distribution in the case of discrete time, and a continuous distribution for continuous time. This seems reasonable. But in the Poisson case, the distribution for the number of events, n, is a discrete distribution in the case of discrete time, and *also a discrete distribution for continuous time.* It is simply always a discrete distribution, because the random variable is the number of events, n, which is discrete. The fact that time might be continuous is irrelevant as far as the discreteness of n goes. The point is that in the case of the exponential distribution, *time itself* is the random variable (because we're considering waiting times), so if we make time continuous, then by definition we're also making the random variable continuous, which means that we have a continuous distribution. ♣

3.3.6 Gaussian

The Gaussian probability distribution (also known as the "normal distribution" or the "bell curve") takes the general form,

$$\boxed{P_{\text{Gaussian}}(x) = Ae^{-b(x-x_0)^2}} \tag{3.38}$$

where A, b, and x_0 are quantities that depend on the specific situation at hand (we will find that A and b must be related in a certain way if the total probability is to be 1), and x stands for whatever the random variable is. The Gaussian is the most important of all the probability distributions. The reason, as we'll see below in Section 3.4, is that in the limit of large numbers (we'll say what we mean by this below), many other distributions reduce to the Gaussian. But for now, we'll just examine the mathematical properties of the curve. We'll discuss how it relates to other distributions in Section 3.4.

Fig. 3.21 shows a plot of the Gaussian distribution for the arbitrary choice of $A = 5$, $b = 10$, and $x_0 = 6$. So we're plotting the function, $P_{\text{Gaussian}}(x) = 5e^{-10(x-6)^2}$. Of the three parameters in Eq. (3.38), x_0 and A are easy to interpret, but b takes a little more thought. Let's look at each of these in turn.

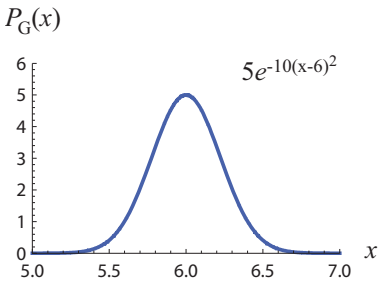


Figure 3.21

x_0 is the location of the maximum of the curve. This is true because the factor of $e^{-b(x-x_0)^2}$ has an exponent that is either zero of negative (because a square is always zero or positive). So this factor is always less than or equal to 1. Its maximum value occurs when the exponent is zero, that is, when $x = x_0$. So the peak is located at $x = x_0$. If we increase x_0, the whole curve just shifts to the right, keeping the same shape.

A is the maximum value of the curve, because x equals x_0 at the peak, which means that the $e^{-b(x-x_0)^2}$ factor equals 1. So the value of the curve at the maximum point is A. If we increase A, the whole curve just expands upward, remaining centered around the same point, x_0.

It turns out that b is related to the width of the curve. Fig. 3.22 shows a plot of a similar Gaussian, with the only change being that b is now 40 instead of 10. So we're plotting the function, $P_{\text{Gaussian}}(x) = 5e^{-40(x-6)^2}$.

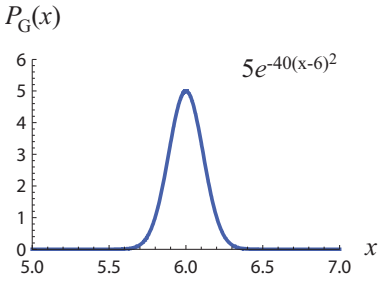


Figure 3.22

We see that the bump is narrower than in Fig. 3.21, but that it has the same height and midpoint, since we haven't changed A or x_0. The factor by which it has shrunk in the horizontal direction appears to be about $1/2$. And in fact, it is *exactly* $1/2$. It turns out that the width of a Gaussian curve is proportional to $1/\sqrt{b}$. This means that since we increased b by a factor of 4, we decreased the width by a factor of $1/\sqrt{4} = 1/2$. Let's now show that the width is in fact proportional to $1/\sqrt{b}$.

First of all, what do we mean by "width"? A vertical rectangle has a definite width, but a Gaussian curve doesn't, because the "sides" are tilted. We could arbitrarily define the width to be how wide the curve is at a height equal to half the maximum height. Or instead of half, we could say a third. Or a tenth. We can define it however we want, but the nice thing is that however we arbitrarily define it, the above "proportional to $1/\sqrt{b}$" result will hold, as long as we pick one definition and stick with it for whatever curves we're looking at.

The definition we'll choose is: The width of a curve is the width at a height equal to $1/e$ (which happens to be about 0.37) times the maximum height. This $1/e$ choice is a

natural one, because the x values that generate this height are easy to find. They are simply $x_0 \pm 1/\sqrt{b}$, because

$$\begin{aligned} P_{\text{Gaussian}}(x_0 \pm 1/\sqrt{b}) &= Ae^{-b[(x_0 \pm 1/\sqrt{b}) - x_0]^2} \\ &= Ae^{-b(\pm 1/\sqrt{b})^2} \\ &= Ae^{-b/b} \\ &= A/e, \end{aligned} \tag{3.39}$$

as desired. Since the difference between the points $x_0 + 1/\sqrt{b}$ and $x_0 - 1/\sqrt{b}$ is $2/\sqrt{b}$, this means that the width of the curve (by our arbitrary definition) is $2/\sqrt{b}$. But again, any other definition would also yield the $\sqrt{b}$ in the denominator. That's the important part. The 2 in the numerator here doesn't have much significance.

The fact that the width is proportional to $1/\sqrt{b}$ and not, say, $1/b$ has *huge* consequences in the study of statistics. Suffice it to say, if the width were proportional to $1/b$, then the world around us wouldn't look anything like what it does. The reasons for this will become clear in the "Law of large numbers" part of Section 3.4.1 below.

1. When we get into statistics in later chapters, we'll change the notation to a more conventional one and write b in terms of the so-called "standard deviation." But this isn't important for the present purposes.
2. The Gaussian distribution can be discrete or continuous. We'll find in Section 3.4 that the Gaussian is a good approximation to the binomial and Poisson distributions if the numbers involved are large. In these cases it is discrete. (You can still draw the continuous curve described by Eq. (3.38), but it's relevant only for certain discrete values of x.) However, the Gaussian distribution also applies (at least approximately) to a nearly endless list of processes with continuous random variables such as length, time, light intensity, affinity for butternut squash, etc. We'll discuss many examples in future chapters.
3. We mentioned above that A and b must be related, due to the fact that the total probability must be 1. We'll see in Section 3.4.1 what this relation is, but for now we can just note that since the width is proportional to $1/\sqrt{b}$, the height must be proportional to $\sqrt{b}$. This is true because if you increase b by a factor of, say, 100 and thereby squash the curve by a factor of $\sqrt{100} = 10$ in the horizontal direction, then you also have to *stretch* the curve by a factor of 10 in vertical direction, if you want to keep the area the same. (See the fifth remark on page 85.) But note that this reasoning tells us only that A is proportional to $\sqrt{b}$, and not what the constant of proportionality is. We'll determine what this constant is in Section 3.4.1.
4. Note that two parameters are needed to describe the Gaussian distribution, namely x_0 and either A or b (because one is determined by the other, as we noted in the preceding remark). This should be contrasted with the Poisson distribution, where only one parameter, a, is needed. In the Poisson case, not only does, say, the width determine the height, but it also determines the location of the bump. We'll be quantitative about this in Section 3.4.2. ♣

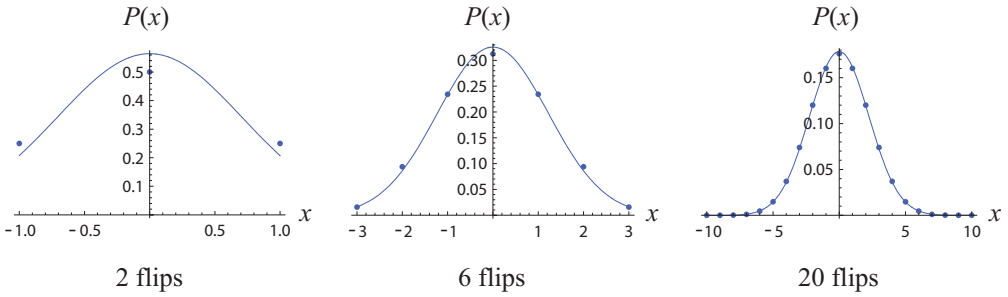
3.4 Gaussians for large numbers

3.4.1 Binomial and Gaussian

In Section 3.3.2 we discussed the binomial distribution that arises from a series of coin flips. The probability distribution for the total number of Heads in, say, 30 flips takes the form of the left plot in Fig. 3.10. The shape of this plot looks suspiciously similar to the shape of the Gaussian plot in Fig. 3.21, so you might wonder if the binomial distribution is actually a Gaussian distribution. It turns out that for small numbers of coin flips, this isn't quite

true. But for large numbers of flips, a binomial distribution essentially takes the form of a Gaussian distribution. The larger the number of flips, the closer it comes to a Gaussian.

Fig. 3.23 shows some comparisons, for a few different numbers of coin flips, between the binomial distribution (the dots) and the Gaussian distribution that we'll derive below in Eq. (3.52) (the curves). The coordinate on the x axis is the number of Heads relative to the average (which is half the number of flips). The Gaussian approximation is clearly very good for 20 flips, and it only gets better as the number of flips increases.



2 flips 6 flips 20 flips

Dots = actual Binomial distribution
Curves = Gaussian approximation
x = number of Heads relative to average

Figure 3.23

We will now demonstrate why the binomial distribution essentially takes the form of a Gaussian distribution when the number of flips is large. For convenience, we'll let the number of flips be $2N$, just to keep some factors of $1/2$ from cluttering things up. N is assumed to be large.

We'll need two bits of mathematical machinery for this derivation. The first is Stirling's formula, which we introduced in Section 2.6. It says that if N is large, then $N!$ is approximately given by

$$N! \approx N^N e^{-N} \sqrt{2\pi N} \qquad (Stirling's\ formula) \tag{3.40}$$

It's a good idea at this point to go back and review Section 2.6. The proof of Stirling's formula involves calculus, so we'll just have to accept it here. The second thing we'll need is the approximation,

$$(1+a)^n \approx e^{an} e^{-na^2/2}. \tag{3.41}$$

This is derived in Appendix C. You are encouraged to read that Appendix (after reading Appendix B) to see where this approximation comes from. However, feel free to just accept it if you want. But in that case you should at least verify with a calculator that it works fairly well for, say, $n = 30$ and $a = 1/10$. And larger numbers work even better in general.

The following derivation will be a bit mathematical, but the result will be well worth it. We'll start with the expression in Eq. (3.6), which gives the probability of obtaining n Heads in N coin flips. However, since we're letting the number of coin flips be $2N$ here, the N in Eq. (3.6) gets replaced by $2N$. Also, let's replace n by $N + x$, which just means that we're defining x to be the number of Heads relative to the average (which is N). Writing the number of heads as $N + x$ will make our calculations *much* simpler than if we had stuck with n. With these adjustments, Eq. (3.6) becomes

$$P_{\text{Binomial}}(x) = \frac{1}{2^{2N}} \binom{2N}{N+x}. \tag{3.42}$$

We will now show that if N is large, $P_{\text{Binomial}}(x)$ takes the approximate form,

$$P_{\text{Binomial}}(x) \approx \frac{e^{-x^2/N}}{\sqrt{\pi N}}\,, \tag{3.43}$$

which is the desired Gaussian.

The first step is to use the Stirling approximation to rewrite each of the three factorials in the binomial coefficient. This gives

$$\begin{aligned}\binom{2N}{N+x} &= \frac{(2N)!}{(N+x)!(N-x)!} \qquad (3.44)\\ &\approx \frac{(2N)^{2N}e^{-2N}\sqrt{2\pi(2N)}}{\big[(N+x)^{N+x}e^{-(N+x)}\sqrt{2\pi(N+x)}\,\big]\cdot\big[(N-x)^{N-x}e^{-(N-x)}\sqrt{2\pi(N-x)}\,\big]}\,.\end{aligned}$$

Yes, this is a big mess, but don't worry, it will simplify! Canceling all the e's and rewriting things a little gives

$$\binom{2N}{N+x} \approx \frac{(2N)^{2N}\sqrt{4\pi N}}{(N+x)^{N+x}(N-x)^{N-x}(2\pi)\sqrt{N^2-x^2}}\,. \tag{3.45}$$

Let's now divide both the numerator and denominator by N^{2N}. In the denominator, we'll do this by dividing the first and second factors by N^{N+x} and N^{N-x}, respectively. We can also cancel a factor of $2\sqrt{\pi}$. The result is

$$\binom{2N}{N+x} \approx \frac{2^{2N}\sqrt{N}}{\big(1+\frac{x}{N}\big)^{N+x}\big(1-\frac{x}{N}\big)^{N-x}\sqrt{\pi}\sqrt{N^2-x^2}}\,. \tag{3.46}$$

It's now time to apply the approximation in Eq. (3.41). With the a and n in that formula defined to be $a \equiv x/N$ and $n \equiv N+x$, we have (using the notation $\exp(y)$ to mean e^y, just to avoid writing lengthy exponents)

$$\Big(1+\frac{x}{N}\Big)^{N+x} \approx \exp\left((N+x)\Big(\frac{x}{N}\Big) - \frac{1}{2}(N+x)\Big(\frac{x}{N}\Big)^2\right). \tag{3.47}$$

When we multiply things out here, we find that there is a $-x^3/2N^2$ term. However, we'll see below that the x's we'll be dealing with are much smaller than N, so this term is much smaller than the others, so we'll ignore it. We are then left with

$$\Big(1+\frac{x}{N}\Big)^{N+x} \approx \exp\Big(x+\frac{x^2}{2N}\Big). \tag{3.48}$$

Although the $x^2/2N$ term here is much smaller than the x term (assuming $x \ll N$), we will in fact need to keep it, because the x term will cancel in Eq. (3.51) below. (The $-x^3/(2N^2)$ term would actually cancel for the same reason, too.) In a similar manner, we obtain

$$\Big(1-\frac{x}{N}\Big)^{N-x} \approx \exp\Big(-x+\frac{x^2}{2N}\Big). \tag{3.49}$$

Using these results in Eq. (3.46), we find

$$\binom{2N}{N+x} \approx \frac{2^{2N}\sqrt{N}}{\exp\Big(x+\frac{x^2}{2N}\Big)\exp\Big(-x+\frac{x^2}{2N}\Big)\sqrt{\pi}\sqrt{N^2-x^2}}\,. \tag{3.50}$$

When combining (adding) the exponents, the x and $-x$ cancel. Also, under the assumption that $x \ll N$, we can say that $\sqrt{N^2 - x^2} \approx \sqrt{N^2 - 0} = N$.[15] The previous equation then becomes

$$\binom{2N}{N+x} \approx \frac{2^{2N}\sqrt{N}}{e^{x^2/N}\sqrt{\pi}\,N}\,. \tag{3.51}$$

Finally, if we substitute Eq. (3.51) into Eq. (3.42), the 2^{2N} factors cancel, and we're left with the desired result,

$$\boxed{P_{\text{Binomial}}(x) \approx \frac{e^{-x^2/N}}{\sqrt{\pi N}} \equiv P_{\text{Gaussian}}(x)} \qquad \text{(for } 2N \text{ coin flips)} \tag{3.52}$$

This is the probability of obtaining $N + x$ Heads in $2N$ coin flips. The most important part of this result is the N in the denominator of the exponent, because this determines the width of the distribution. We'll talk about this below, but first some remarks.

Remarks:

1. We claimed at various points in the above derivation that the values of x that we're concerned with are much less than N, and this allowed us to simplify some expressions by ignoring terms. This claim is valid because of the exponential factor, $e^{-x^2/N}$, in $P_{\text{Gaussian}}(x)$. If x is much larger than $\sqrt{N}$, then this factor is essentially zero. So only x values up to order $\sqrt{N}$ yield non-negligible probabilities. And since we're assuming that N is large, we have $\sqrt{N} \ll N$. Putting these two results together gives $x \sim \sqrt{N} \ll N$. So we conclude that $x \ll N$ for any x's that yield non-negligible values of $P_{\text{Gaussian}}(x)$.
2. In our terminology where the number of coin flips is $2N$, the plots in Fig. 3.23 correspond to N values of 1, 3, and 10. So in the third graph, for example, the continuous curve is a plot of $P_{\text{Gaussian}}(x) = e^{-x^2/10}/\sqrt{\pi(10)}$.
3. $P_{\text{Gaussian}}(x)$ is an "even" function of x. That is, it is symmetric around $x = 0$. Or in mathematical terms, x and $-x$ yield the same value of the function. This is true because x appears only through its square. This evenness makes intuitive sense, because we're just as likely to get, say, 4 Heads above the average as 4 Heads below the average.
4. We saw in Eq. (2.38) in Section 2.6 that the probability that exactly half (that is, N) of the $2N$ coin flips come up Heads is $1/\sqrt{\pi N}$. This result is a special case of the $P_{\text{Gaussian}}(x)$ result in Eq. (3.52) because if we plug $x = 0$ (which corresponds to $n = N$ Heads) into Eq. (3.52), we obtain $P_{\text{Gaussian}}(x) = e^{-0}/\sqrt{\pi N} = 1/\sqrt{\pi N}$.
5. Note that we really did need the $e^{-na^2/2}$ factor in the approximation in Eq. (3.41). If we had used the less accurate version, $(1 + a)^n \approx e^{an}$, from Eq. (4.7) in Appendix B, we wouldn't have had the $x^2/2N$ terms in Eqs. (3.48) and (3.49). And since the $\pm x$ terms in these equations canceled, the exponent in the final result in Eq. (3.52) would simply have been "0." So there wouldn't have been any x dependence at all. The probability would have just been given by $1/\sqrt{\pi N}$. This is indeed an approximation, but it's a rather poor one. It makes the assumption that all the values are the same as the value at the peak.[16] This is a fine approximation near the peak, but lousy elsewhere. By using the improved formula with the $e^{-na^2/2}$ factor, we obtained an approximation that works far away from the peak too.
6. The sum of the $P_{\text{Gaussian}}(x)$ probabilities in Eq. (3.52), for x from $-N$ to N (that is, for the number of Heads from 0 to $2N$), must equal 1. Or at least approximately, given that Eq. (3.52) is an approximation. Assuming that we didn't make any mistakes in the above derivation, this has to be true, of course, because $P_{\text{Gaussian}}(x)$ is essentially equal to $P_{\text{Binomial}}(x)$, and we showed in Section 3.3.2 that the sum of the $P_{\text{Binomial}}(x)$ probabilities equals 1.

[15] As with any approximation claims, if you don't trust this, you should try it with some numbers, say $N = 10,000$ and $x = 100$, which satisfy the $x \ll N$ relation.

[16] This is the so-called "zeroth order approximation," because the highest power of x that appears in the formula is x^0, which is a fancy way of saying that there are no x's.

However, it would be nice to verify independently, without any reference to the binomial distribution, that the sum of the $P_{\text{Gaussian}}(x)$ probabilities equals 1. If N is large (which we are assuming), the plot of the $P_{\text{Gaussian}}(x)$ points is essentially a continuous curve. Therefore, showing that the sum of the probabilities equals 1 is equivalent to showing that the area under the $P_{\text{Gaussian}}(x)$ curve is 1. This can in fact be demonstrated, but unfortunately the calculation involves calculus, so we'll just have to accept it here.

But on the bright side, it should be noted that we would need to demonstrate this for only one such curve, because the argument in Remark 3 in Section 3.3.6 explains why the area is the same for all curves of the form given in Eq. (3.52) (the b in that remark is $1/N$ here).

7. If we compare Eqs. (3.52) and (3.38), we see that $b = 1/N$ and $A = 1/\sqrt{\pi N}$. The first of these gives $N = 1/b$, and then plugging this into the second yields $A = \sqrt{b/\pi}$. Introducing an x_0 term doesn't affect this relation, so the general form of a Gaussian distribution is $\sqrt{b/\pi}\, e^{-b(x-x_0)^2}$. This has the correct relation between A and b to make the area under the curve equal to 1 (although calculus is needed to demonstrate this).

8. If the two probabilities involved in the binomial distribution are p and $1-p$ instead of the two 1/2's in the case of coin tosses, then the probability of n successes in N trials is given in Eq. (3.5) as $P(n) = \binom{N}{n} p^n (1-p)^{N-n}$. (Note that we've gone back to using N here to represent the total number of trials instead of the $2N$ we used in Eq. (3.42).) For example, if we're concerned with the number of 5's we roll in N rolls of a die, then $p = 1/6$.

   It turns out that for large N, the binomial $P(n)$ distribution is essentially a Gaussian distribution for *any* value of p, not just the $p = 1/2$ value we discussed above. And the Gaussian is centered around the average value of n (namely pN), as you would expect. The derivation of this Gaussian form follows exactly the same steps as above. But it gets a bit messy, so we'll just state the result: For large N, the probability of obtaining $pN + x$ successes in N trials is approximately equal to

$$P(x) \approx \frac{e^{-x^2/[2Np(1-p)]}}{\sqrt{2\pi Np(1-p)}}. \tag{3.53}$$

   If you replace N with $2N$ (because we defined $2N$ to be the total number of trials in the coin case above) and if you let $p = 1/2$, then this reduces to the result in Eq. (3.52), as it should. Eq. (3.53) implies that the bump is symmetric around $x = 0$ (or equivalently, around $n = pN$), even for $p \neq 1/2$, which isn't so obvious. (Well, the tail extends farther to one side, but $P(x)$ is essentially zero in the tails.) For $p \neq 1/2$, the bump isn't centered around $N/2$, so you might think that the shape of the bump should be lopsided too. But it isn't. Fig. 3.24 shows a plot of Eq. (3.53) for $p = 1/6$ and $N = 60$ (which corresponds to rolling a die 60 times and seeing how many, say, 5's you get). The $x = 0$ point corresponds to having $pN = (1/6)(60) = 10$ rolls of a 5. ♣

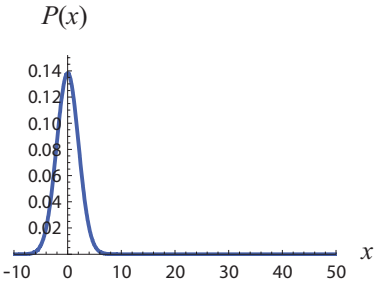


Figure 3.24

The law of large numbers

Let's now take a look at how the N in the denominator of the exponent in $P_{\text{Gaussian}}(x)$ in Eq. (3.52) influences the width of the curve. We found in Section 3.3.6 that the width of

a Gaussian curve of the form Ae^{-bx^2} is $1/\sqrt{b}$. (We're arbitrarily defining the "width" of a curve to be the width where the height is $1/e$ times the maximum height.) The $P_{\text{Gaussian}}(x)$ distribution in Eq. (3.52) has $b = 1/N$, so this means that the width is $\sqrt{N}$.

Fig. 3.25 shows plots of $P_{\text{Gaussian}}(x)$ for $N = 10$, 100, and 1000. As N gets larger, the curve's height shrinks (because Eq. (3.52) says that it is proportional to $1/\sqrt{N}$), and its width expands (because it equals $\sqrt{N}$). Note that because these two factors are reciprocals of each other, this combination of shrinking and expanding doesn't change the area under the curve (see the fifth remark on page 85). This is consistent with the fact that the area is always equal to 1.

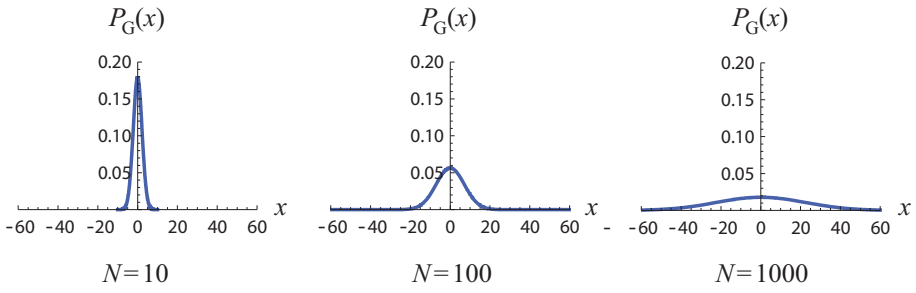
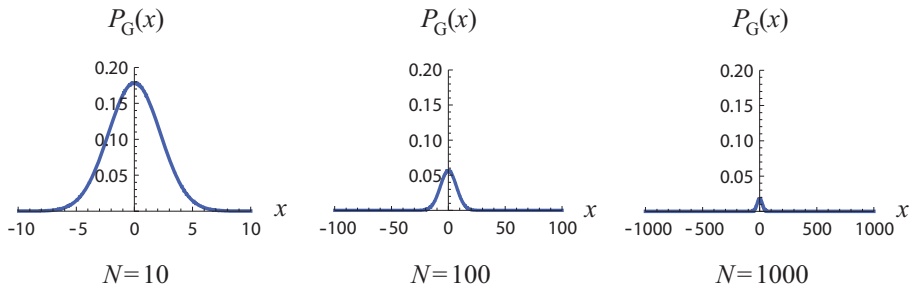


Figure 3.25

The critical fact about the $\sqrt{N}$ expansion factor is that although it increases as N increases, *it doesn't increase as fast as N does.* In fact, compared with N, it actually *decreases* by a factor of $1/\sqrt{N}$. This means that if we plot $P_{\text{Gaussian}}(x)$ with the horizontal axis running from $-N$ to N (instead of it being fixed as in Fig. 3.25), the width of the curve actually *shrinks* by a factor of $1/\sqrt{N}$ (relative to N). Fig. 3.26 shows this effect. In this figure, both the width (relative to N) and the height of the curves shrink by a factor of $1/\sqrt{N}$ (the height behaves the same as in Fig. 3.25), so all the curves have the same shape. They just have different sizes (they differ successively by a factor of $1/\sqrt{10}$). The area under each curve is technically still equal to 1, though, because of the different scales on the x axis.

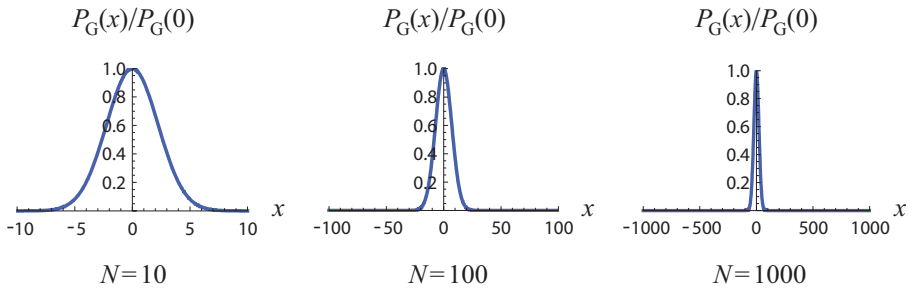


(Note differerent scales on x axis)

Figure 3.26

A slightly more informative curve to plot is the ratio of $P_{\text{Gaussian}}(x)$ to its maximum height at $x = 0$. This modified plot makes it easier to see what's happening with the width. Since the maximum height is $1/\sqrt{\pi N}$, we're now just plotting $e^{-x^2/N}$. So all the curves have the same value of 1 at $x = 0$. If we let the horizontal axis run from $-N$ to N as in

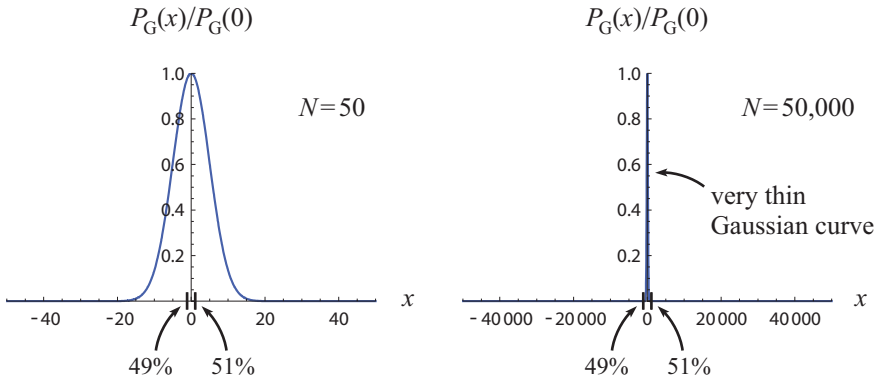
Fig. 3.26, we obtain the plots in Fig. 3.27. These are simply the plots Fig. 3.26, except that they're stretched in the vertical direction so that they all have the same height. We see that the bump gets thinner and thinner (on the scale of N) as N increases. This implies that the *percentage* deviation from the average of N Heads gets smaller and smaller as N increases.



(Note differerent scales on x axis)

Figure 3.27

We can now understand the reason behind the claim we made at the end of Section 2.1, when we said that the observed fraction of Heads gets closer and closer to the actual probability of 1/2, as the number of trials gets larger and larger. We stated that if you flip a coin 100 times (which corresponds to $N = 50$), the probability of obtaining 49, 50, or 51 Heads is only about 24%. This is consistent with the first plot in Fig. 3.28, where we've indicated the 49% and 51% marks on the x axis (which correspond to $x = \pm 1$). A fairly small portion of the area under the curve lies between these marks.[17] We also stated that if you flip a coin 100,000 times (which corresponds to $N = 50,000$), the probability of obtaining Heads between 49% and 51% of the time is 99.99999997%. This is consistent with the second plot in Fig. 3.28, because essentially all of the area under the curve lies between the 49% and 51% marks (which correspond to $x = \pm 1000$).



(Note differerent scales on x axis)

Figure 3.28

[17] If we made a histogram of the probabilities, the natural thing would be to have the "bin" for 49 go from 48.5 to 49.5, etc. So we're actually concerned with (approximately) the area between 48.5 and 51.5 if we're looking at 49, 50, or 51 Heads. This distinction becomes inconsequential when N becomes large, because the distribution becomes effectively continuous.

This effect is known as the "law of large numbers." The law can be stated in various ways, but we'll go with:

- The law of large numbers:

  If you repeat a random process a large number of times, and if you write down the observed faction of times that a certain outcome happens, then this observed fraction will be very close to the theoretically predicted probability.

More precisely, consider the probability, P, that the observed probability differs from the theoretical probability by more than a particular specified number, say $\delta = 0.01$ or 0.001. Then P goes to zero as the number of trials becomes large. This is just what we saw in Fig. 3.28, with $\delta = 0.01$. Said in a more down to earth manner, if you perform enough trials, the observed probability will be pretty much what it "should" be.

This is an extremely important result, and it all comes down to the fact that although the width (which is $\sqrt{N}$) grows with N, it *shrinks* in comparison to the full spread of outcomes (which is $2N$). Said a different way, the width grows in an *additive* sense (this is sometimes called an "absolute" sense), but decreases in a *multiplicative* sense (compared with N). It is the latter of these effects that is relevant when calculating percentages.

The law of large numbers is what makes polls more accurate if more people are interviewed, and why casinos always come out ahead. It is what makes it prohibitively unlikely for all the air molecules in a room to end up in one half of it, and why a piece of paper on your desk doesn't spontaneously combust. The list of applications is essentially endless, and it would be an understatement to say that the world would be a very different place without the law of large numbers.

3.4.2 Poisson and Gaussian

We showed in Section 3.4.1 that the binomial distribution becomes a Gaussian distribution in the limit where the number of trials is large. We will now show that the Poisson distribution in Eq. (3.33) also approaches a Gaussian distribution. The limit that will produce this result is the limit of large a, where a is the expected number of successes.[18]

As in the binomial case, we will need to use the two approximations in Eqs. (3.40) and (3.41). Applying Stirling's formula to the $n!$ in Eq. (3.33) gives

$$\begin{aligned} P_{\text{Poisson}}(n) &= \frac{a^n e^{-a}}{n!} \\ &\approx \frac{a^n e^{-a}}{n^n e^{-n}\sqrt{2\pi n}}\,. \end{aligned} \tag{3.54}$$

Now, we saw in the second remark following Eq. (3.33) that the maximum of $P_{\text{Poisson}}(n)$ occurs at a (or technically between $a-1$ and a, but for large a this distinction is inconsequential). So let's see how $P_{\text{Poisson}}(n)$ behaves near $n = a$. To this end, let's define x by $n \equiv a + x$. So x is the number of successes relative to the average. This is analogous to the $n \equiv N + x$ definition that we used in Section 3.4.1. As it did there, the use of x instead of n will make our calculations much simpler. In terms of x, Eq. (3.54) becomes

$$P_{\text{Poisson}}(x) \approx \frac{a^{a+x} e^{-a}}{(a+x)^{a+x} e^{-a-x}\sqrt{2\pi(a+x)}}\,. \tag{3.55}$$

[18] It wouldn't make sense to take the limit of a large number of trials here, as we did in the binomial case, because the number of trials isn't specified in the Poisson case. The only parameter that appears is a. But a large number of trials in the binomial case implies a large expected number of successes, so the large-a limit in the Poisson case is analogous to the large-trial-number limit in the binomial case.

We can cancel a factor of e^{-a}. And we can also divide both the numerator and denominator by a^{a+x}. And we can ignore the x in the square root, because we'll find below that the x's we're concerned with are small compared with a. The result is

$$P_{\text{Poisson}}(x) \approx \frac{1}{\big(1+(x/a)\big)^{a+x} e^{-x}\sqrt{2\pi a}}. \tag{3.56}$$

It's now time to use the approximation in Eq. (3.41). (The procedure from here on will be very similar to the binomial case.) With the "a" in Eq. (3.41) defined to be x/a (a means two completely different things here), and the "n" defined to be $a+x$, Eq. (3.41) gives

$$\Big(1+\frac{x}{a}\Big)^{a+x} \approx \exp\left((a+x)\Big(\frac{x}{a}\Big) - \frac{1}{2}(a+x)\Big(\frac{x}{a}\Big)^2\right). \tag{3.57}$$

Multiplying this out and ignoring the small $-x^3/2a^2$ term (because we'll find below that $x \ll a$), we obtain

$$\Big(1+\frac{x}{a}\Big)^{a+x} \approx \exp\left(x + \frac{x^2}{2a}\right). \tag{3.58}$$

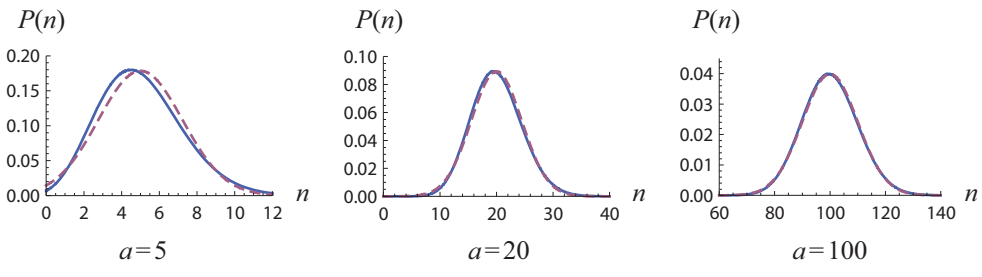
Substituting this into Eq. (3.56) gives

$$\begin{aligned} P_{\text{Poisson}}(x) &\approx \frac{1}{e^x e^{x^2/2a} e^{-x}\sqrt{2\pi a}} \\ &= \frac{e^{-x^2/2a}}{\sqrt{2\pi a}} \equiv P_{\text{Gaussian}}(x), \end{aligned} \tag{3.59}$$

which is the desired Gaussian. If you want to write this result in terms of n instead of x, then the definition $n \equiv a+x$ gives $x = n-a$, so we have

$$\boxed{P_{\text{Poisson}}(n) \approx \frac{e^{-(n-a)^2/2a}}{\sqrt{2\pi a}} \equiv P_{\text{Gaussian}}(n)} \tag{3.60}$$

Fig. 3.29 shows a comparison between the exact $P_{\text{Poisson}}(n)$ function in the first line of Eq. (3.54), and the approximate $P_{\text{Gaussian}}(n)$ function in Eq. (3.60). The approximation works quite well for $a = 20$, and by the time $a = 100$, it works so well that it's hard to tell that there are actually two different curves.



(Note differerent scales on axes)

Solid curve = exact Poisson

Dashed curve = approximate Gaussian

Figure 3.29

REMARKS:

1. As with the Gaussian approximation to the binomial distribution, the Gaussian approximation to the Poisson distribution is likewise symmetric around $x = 0$ (equivalently, $n = a$).
2. With the definition of the width we gave in Section 3.3.6, the width of the Gaussian curve in Eq. (3.59) is $\sqrt{2a}$.
3. As we noted in the last remark in Section 3.3.6, the Poisson distribution (and hence the Gaussian approximation to it) depends on only one parameter, a. Comparing Eq. (3.60) with Eq. (3.38), we see that $A = 1/\sqrt{2\pi a}$, $b = 1/2a$, and $x_0 = a$. So all three quantities can be written in terms of one parameter.
4. The $P_{\text{Gaussian}}(x)$ expression in Eq. (3.59) equals the expression in Eq. (3.53) in the limit of small p, with $a = pN$. Basically, in Eq. (3.53) just replace pN with a, and $(1 - p)$ with 1, and the result is Eq. (3.59). As an exercise, you can think about why this is true. In short, the result in Eq. (3.59) was obtained by taking the small-p limit (to derive the Poisson distribution in the first place, which is what we started with in Eq. (3.54)) and then the large-a (which implies large-N) limit. The route to Eq. (3.59) via Eq. (3.53) simply takes these limits in the reverse order, first large-N and then small p. ♣

3.4.3 Binomial, Poisson, and Gaussian

We've seen in various places in this chapter (Sections 3.3.5, 3.4.1, and 3.4.2) how the binomial, Poisson, and Gaussian distributions are related to each other in various limits. The summary of these relations is shown in Fig. 3.30.

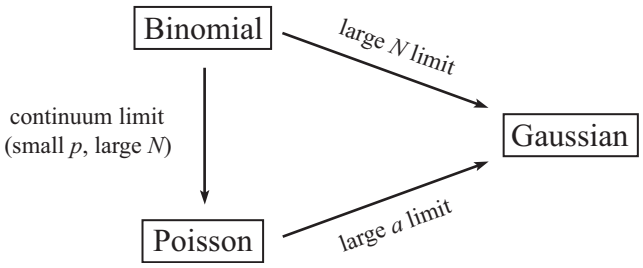


Figure 3.30

The detailed descriptions of the three relations are the following.

- The arrow on the left side indicates that the Poisson distribution is obtained from the binomial distribution by taking the continuum limit. By this we mean that instead of having, say, N trials of a process happening each second with probability p of success, we have $10 \cdot N$ trials happening each 1/10 of a second with probability $p/10$ of success. Or $100 \cdot N$ trials happening each 1/100 of a second with probability $p/100$ of success. And so on. All of these scenarios have the same average of $a = pN$ events occurring. And all of them are governed by the binomial expansion. But the more that time is subdivided, the closer the probability distribution (for the number of observed events) comes to the Poisson distribution given in Eq. (3.33), with $a = pN$.

- The upper right arrow indicates that the Gaussian distribution is obtained from the binomial distribution by taking the large-N limit, where N is the number of trials performed. We derived this result in Section 3.4.1 for $p = 1/2$, and then we stated the general result for any p in Eq. (3.53).

- The lower right arrow indicates that the Gaussian distribution is obtained from the Poisson distribution by taking the large-a limit, where a is the expected number of events that happen during the particular number of trials you are performing. We derived this result in Section 3.4.2, and the result was Eq. (3.59). The large-a limit in the Poisson-to-Gaussian case is consistent with the large-N limit in the binomial-to-Gaussian case, because the only way to have a large expected number of events, a, is to perform a large number of trials.

Note that there are two paths in Fig. 3.30 that go from the binomial distribution to the Gaussian distribution. One goes directly by taking the large-N limit. The other goes via the Poisson distribution by first taking the continuum limit, and then taking the large-a limit (which implies large N). The reason why these two routes end up at the same place is because when taking the large-N limit, you are essentially taking the continuum limit. To see why this is true, imagine that you have a process that happens once every second, and that you run these processes for a year. On the time scale of a year, the processes essentially take place continuously, because a second is so small compared with a year.

The fact that all the arrows in Fig. 3.30 eventually end up at the Gaussian (equivalently, that no arrows point away from the Gaussian) is effectively the statement of the *central limit theorem*, which says:

- Central limit theorem:

  If you perform a large number of trials of a random process, then the probability distribution for the sum of the random-variable values will be approximately a Gaussian (or a "normal") distribution. And the greater the number of trials, the better the Gaussian approximation.

In the examples above where we were counting successful events, we effectively assigned a value of 1 to a success, and a value of 0 to a failure. So there were two possible values for the random variable, and the sum of the random variables equaled the number of successes. We found that the probability distribution for the number of successes approached a Gaussian when the number of trials was large, so this is consistent with the central limit theorem.

Another example consists of rolling a large number of dice and looking at the probability distribution for their sum. If you roll 100 dice, then the distribution will be (essentially) a Gaussian centered around 350, since the average for each roll is 3.5. (If you wanted to verify this, you would have to perform a large number of processes, each consisting of 100 rolls.) Note that (as stated in the theorem) we need the number of rolls to be large. If you roll just one die, then the probability distribution for the sum is simply a flat line, which is certainly *not* a Gaussian. The "sum" is simply the reading on the one die, so it has equal 1/6 chances of taking on the values of 1 through 6. (And again, you could verify this by performing a large number of processes, each consisting of one die roll.) If you instead roll 2 dice, then as an exercise you can show that Table 1.7 implies that the distribution for the sum takes the shape of a triangle that is peaked around 7. This isn't a Gaussian either. A Gaussian arises only if the number of trials (rolls here) within the process is large.

REMARK: Not to belabor the point, but we should emphasize the distinction between the two large numbers mentioned in the previous paragraph. The first is the number of trials, N_{t}, within the process (for example, the 100 rolls we considered). The sum of the random variables for these N_{t} trials takes the form of an (approximate) Gaussian distribution only if N_{t} is large. (The nature of the setup and the desired accuracy in the approximation determine what exactly is meant by "large.") The second number is the number of processes, N_{p}, that you need to perform (each of which consists of N_{t} trials) in order to verify the distribution for the sum. This number N_{p} must be large, period. (And again, various things determine what is meant by "large.") Otherwise you

won't get good statistics. As an extreme example, if $N_{\text{p}} = 1$ then you'll just have one data point for the sum, and one data point of course doesn't look anything like a whole distribution curve. ♣

3.5 Summary

In this chapter we learned about probability distributions. In particular, we learned:

1. A *probability distribution* is the collective information about how the total probability (which is always 1) is distributed among the various possible outcomes for the random variable.

2. Probability distributions for continuous random variables are given in terms of a *probability density*. To obtain an actual probability, this density must be multiplied by an interval of the random variable. More generally, the probability equals the area under the density curve.

3. We discussed six specific probability distributions:

   - *Bernoulli:* (Discrete) The random variable can take on only two values, with probabilities p and $1 - p$. An example is a coin toss.
   - *Binomial:* (Discrete) The random variable is the number of "successes" for a collection of Bernoulli processes. An example is the total number of Heads in a given number of coin tosses. The general form is

     $$P_{\text{Binomial}}(n) = \binom{N}{n} p^n (1-p)^{N-n}, \tag{3.61}$$

     where N is the total number of Bernoulli processes (such as coin tosses).
   - *Uniform:* (Continuous) The probability density is uniform over a given span of random-variable values, and zero otherwise. An example is the location of an object that is constrained to be in a given region, with equal likelihood of being anywhere in the region.
   - *Exponential:* (Continuous) This is the probability distribution for the waiting time until the next successful event, for a completely random process. We derived this by taking the continuum limit of the analogous discrete result. The general form is

     $$\rho_{\text{Exponential}}(t) = \frac{e^{-t/T}}{T}\,, \tag{3.62}$$

     where T is the average waiting time.
   - *Poisson:* (Discrete) This is the probability distribution for the number of events that happen in a given region (of time, space, etc) for a completely random process. We derived this by taking the continuum limit of the analogous discrete result, which was simply the binomial distribution. The general form is

     $$P_{\text{Poisson}}(n) = \frac{a^n e^{-a}}{n!}\,, \tag{3.63}$$

     where a is the expected number of events in the given region.
   - *Gaussian:* (Discrete or Continuous) This distribution takes the form,

     $$P_{\text{Gaussian}} = A e^{-b(x-x_0)^2}\,. \tag{3.64}$$

     We must have $A = \sqrt{b/\pi}$ if the total probability is to be 1.

4. The Gaussian distribution has the property that the larger the number of trials, the thinner the distribution's bump, relative to the whole span of possible outcomes. This is essentially the statement of the *law of large numbers*, which says that the measured probability over many trials will be essentially equal to the theoretical probability.

5. For a large number of trials, the binomial and Poisson distributions reduce to the Gaussian distribution. This is consistent with the *central limit theorem*, which says that if many trials are performed, the sum of the values of the random variables has a Gaussian distribution.

3.6 Problems

1. **Fahrenheit and Celsius** $*$

   Fig. 3.5 shows the probability density with the temperature measured in Fahrenheit. Draw a reasonably accurate plot of the same probability density, but with the temperature measured in Celsius. (The conversion formula from Fahrenheit to Celsius is $C = (5/9)(F - 32)$. So it takes a ΔF of $9/5 = 1.8$ to create a ΔC of 1.)

2. **Expectation for binomial** $**$

   Use Eq. (2.33) to explicitly demonstrate that the expectation value of the binomial distribution in Eq. (3.5) equals pN. This must be true, of course, because on average a fraction p of the N trials yields success, by the definition of p. *Hint:* The goal is to produce a result of pN, so try to factor a pN out of the sum. You'll eventually need to use an expression analogous to Eq. (3.7).

3. **Expectation for discrete exponential** $*$

   Verify that the expectation value of the discrete probability distribution given in Eq. (3.10) equals $1/p$. (This is the waiting time we found by an easier method in Section 3.3.4 prior to Eq. (3.10).) This involves a math trick, so you should do Problem 9 in Chapter 2 before solving this one.

4. **Expectation for Poisson** $*$

   Verify that the expectation value for the Poisson distribution in Eq. (3.33) is a. This must be the case, of course, because a is defined to be the average number of successful events.

Many more problems will be added...

3.7 Solutions

1. **Fahrenheit and Celsius**

   Density is always given in terms of "something per something else." In the temperature example in Section 3.2, the "units" of probability density were probability per degree Fahrenheit. These units are equivalent to saying that we need to multiply the density by a certain number of degrees Fahrenheit (the ΔT) to obtain a probability. Analogously, we need to multiply a mass density (mass per volume) by a volume to obtain a mass.

   If we instead want to write the probability density in terms of probability per degree *Celsius*, we can't simply use the same function $\rho(T)$ that appears in Fig. 3.5. Since there are 1.8 degrees Fahrenheit in each degree Celsius, the correct plot of $\rho(T)$ is shown in either of the graphs in Fig. 3.31. Since the peak of the curve in Fig. 3.5 was at about 68 degrees Fahrenheit, it is now at about $(5/9)(68 - 32) = 20$ degrees Celsius in Fig. 3.31 (both graphs have this property).

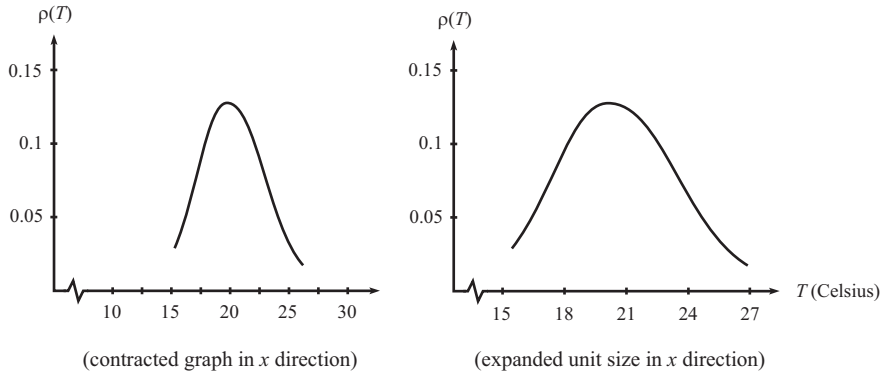


Figure 3.31

   But the more important change is that compared with Fig. 3.5, the curve in Fig. 3.31 is *contracted* by a factor of 1.8 in the horizontal direction due to the conversion from Fahrenheit to Celsius. The span is only about 11 degrees Celsius, compared with a span of about 20 degrees Fahrenheit in Fig. 3.5. This contraction can be signified in two ways: In the first graph in Fig. 3.31, we kept the x-axis unit size the same and squeezed the plot, and in the second graph we expanded the x-axis unit size but didn't squeeze the plot. Both of these procedures yield the correct span of about 11 degrees Celsius.

   However, since the the area under the entire curve in Fig. 3.31 *must still be 1*, the curve must also be *expanded* by a factor of 1.8 in the vertical direction. The maximum value is about 0.13, compared with the maximum value of about 0.07 in Fig. 3.5.

   These counter effects allow everything to be consistent. For example, we found in Section 3.2 that that probability of the temperature falling between 70° F and 71° F is about 7%. Now, 70° F and 71° F correspond to 21.11° C and 21.67° C, as you can show using $C = (5/9)(F - 32)$. So the probability of the temperature falling between 21.11° C and 21.67° C had better also be 7%. It's the same temperature interval, we're just describing it in different ways. And indeed, from the Celsius plot, the value of the density near 21° is about 0.12. So the probability of being between 21.11° C and 21.67° C, which equals the density times the interval, is $(0.12)(21.67 - 21.11) = 0.067 \approx 7\%$, in agreement with the Fahrenheit calculation (up to the rough readings we made from the plots). If we forgot to expand the vertical axis by a factor of 1.8, we would have obtained only about half of this probability, and therefore a different answer to exactly the same question (asked in a different language). That wouldn't be good.

2. **Expectation for binomial**

The $n = 0$ term doesn't contribute anything to the sum in Eq. (2.33), so we can start the sum with the $n = 1$ term (and it goes up to $n = N$). Plugging the probabilities from Eq. (3.5) into Eq. (2.33) gives an expectation value of

$$\sum_{n=1}^{N} n \cdot P(n) = \sum_{n=1}^{N} n \cdot \binom{N}{n} p^n (1-p)^{N-n}. \tag{3.65}$$

If the factor of n weren't on the righthand side, we would know how to do this sum; see Eq. (3.7). So let's somehow try to get rid of the n and create a sum that looks like Eq. (3.7). The steps are the following.

$$\begin{aligned}
\sum_{n=1}^{N} n \cdot P(n) &= \sum_{n=1}^{N} n \cdot \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \qquad (\textit{expanding the binomial coeff.}) \\
&= pN \sum_{n=1}^{N} n \cdot \frac{(N-1)!}{n!(N-n)!} p^{n-1} (1-p)^{N-n} \qquad (\textit{factoring out } pN) \\
&= pN \sum_{n=1}^{N} \frac{(N-1)!}{(n-1)!(N-n)!} p^{n-1} (1-p)^{N-n} \qquad (\textit{canceling the } n) \\
&= pN \sum_{n=1}^{N} \binom{N-1}{n-1} p^{n-1} (1-p)^{(N-1)-(n-1)} \qquad (\textit{rewriting}) \\
&= pN \sum_{m=0}^{N-1} \binom{N-1}{m} p^{m} (1-p)^{(N-1)-m} \qquad (\textit{defining } m \equiv n-1) \\
&= pN\big(p + (1-p)\big)^{N-1} \qquad (\textit{using the binomial expansion}) \\
&= pN \cdot 1,
\end{aligned} \tag{3.66}$$

as desired. Note that in the fifth line, the sum over m goes from 0 to $N-1$, because the sum over n went from 1 to N. Even though we know that the expectation value has to be pN, it's nice to see that the math does in fact work out.

3. **Expectation for discrete exponential**

From Eq. (3.10), the probability that we wait just one iteration for the next success is p, for two iterations it is $(1-p)p$, for three iterations it is $(1-p)^2 p$, and so on. So the expectation value of the number of iterations (that is, the waiting time) is

$$1 \cdot p + 2 \cdot (1-p)p + 3 \cdot (1-p)^2 p + 4 \cdot (1-p)^3 p + \cdots. \tag{3.67}$$

To calculate this sum, we'll use the trick introduced in Problem 9 in Chapter 2 and write the sum as a geometric series starting with p, plus another geometric series starting with $(1-p)p$, and so on. And we'll use the fact that the sum of a geometric series with first term a and ratio r is $a/(1-r)$. So the expectation value in Eq. (3.67) becomes

$$\begin{aligned}
&p + (1-p)p + (1-p)^2 p + (1-p)^3 p + \cdots \\
&\qquad (1-p)p + (1-p)^2 p + (1-p)^3 p + \cdots \\
&\qquad\qquad (1-p)^2 p + (1-p)^3 p + \cdots \\
&\qquad\qquad\qquad (1-p)^3 p + \cdots \\
&\qquad\qquad\qquad\qquad \vdots
\end{aligned} \tag{3.68}$$

This has the correct number of each type of term. For example, the $(1-p)^2 p$ term appears three times. The first line here is a geometric series that sums to $a/(1-r) = p/(1-(1-p)) = 1$. The second line is also a geometric series, and it sums to $(1-p)p/(1-(1-p)) = 1-p$.

Likewise the third line sums to $(1-p)^2 p/(1-(1-p)) = (1-p)^2$. And so on. The sum of the infinite number of lines in the above equation therefore equals

$$1 + (1-p) + (1-p)^2 + (1-p)^3 + \cdots. \tag{3.69}$$

But this itself is a geometric series, and its sum is $a/(1-r) = 1/(1-(1-p)) = 1/p$, as desired.

4. **Expectation for Poisson**

The expectation value is the sum of $n \cdot P_{\text{Poisson}}(n)$, from $n = 0$ to $n = \infty$. However, the $n = 0$ term contributes nothing, so we can start the sum with the $n = 1$ term. Using Eq. (3.33), the expectation value is therefore

$$\begin{aligned}
\sum_{n=1}^{\infty} n \cdot P_{\text{Poisson}}(n) &= \sum_{n=1}^{\infty} n \cdot \frac{a^n e^{-a}}{n!} \\
&= \sum_{n=1}^{\infty} \frac{a^n e^{-a}}{(n-1)!} \qquad (\textit{canceling the } n) \\
&= a \cdot \sum_{n=1}^{\infty} \frac{a^{n-1} e^{-a}}{(n-1)!} \qquad (\textit{factoring out an } a) \\
&= a \cdot \sum_{m=0}^{\infty} \frac{a^m e^{-a}}{m!} \qquad (\textit{defining } m \equiv n-1) \\
&= a \cdot \sum_{m=0}^{\infty} P_{\text{Poisson}}(m) \\
&= a \cdot 1, \qquad (\textit{using Eq. (3.37)})
\end{aligned} \tag{3.70}$$

as desired. In the fourth line, we used the fact that since $m \equiv n-1$, the sum over m starts with the $m = 0$ term (because the sum over n started with the $n = 1$ term).

Chapter 4

Appendices

4.1 Appendix A: Subtleties about probability

There are a number of subtle issues with probability, so let's list them out here. This appendix isn't necessary for the material in this book, so it can be omitted on a first reading.

Determining probabilities

How do you determine the probability that a given event occurs? There are two ways: You can calculate it theoretically, or you can estimate it experimentally by performing a large number of trials of the process.

We can use a theoretical argument to determine, for example, the probability of getting Heads on a coin toss. There is no need to actually *do* a coin toss, because it suffices to just think about it and note that the two possibilities of Heads and Tails are equally likely (assuming a fair coin), so each one must occur half of the time. So the probability is 1/2, and that's that. Similarly for the probabilities of 1/6 for each of the six possible rolls of a die (assuming a fair die).

However, there are certainly many situations where we don't have enough information to calculate the probability by theoretical means. In these cases we have no choice but to simply perform a large number of trials and then assume that the true probability is roughly equal to the fraction of events that occurred. For example, let's say that you take a bus to work or school, and that sometimes the bus is early and sometimes it's late. What is the probability that it's early? There are countless things that influence the bus's timing: traffic (which itself depends on countless things), weather, engine issues, delays caused by other passengers, slow service at a restaurant the night before which caused the driver to see a later movie than planned which caused him to go to bed later than usual and hence get up later than usual which caused him to start the route two minutes late, and so on and so forth. It is clearly hopeless to try to incorporate all these effects into some sort of theoretical reasoning that produces a result that can be trusted. The only option then, is to observe what happens during a reasonably large number of days, and to assume that the fraction of early arrivals that you observe is roughly the probability. If the bus is early for 20 out of 50 days, then we can say that the probability of being early is probably about 40%.

Of course, having established this result of 40%, it just might happen that a construction project starts the next day a few blocks up the route, which makes the bus late every day

for the next two months. So probabilities based on observation should be taken with a grain of salt!

A similar situation arises with, say, basketball free-throw percentages. There is absolutely no hope of theoretically calculating the probability of a certain player hitting a free throw, because it would require knowing everything that's going on from the thoughts in her head to the muscles in her fingers to the air currents en route to the basket. All we can say is that the player has hit a certain fraction of the free throws she's already attempted, and that's our best guess for the probability of hitting free throws in the future.

True randomness

We stated above that the probability of a coin toss resulting in Heads is 1/2. The reasoning was that Heads and Tails should have equal probabilities if everything is random, so they each must be 1/2. *But is the toss truly random?* What if we know the exact torque and force that you apply to the coin? We can then know exactly how fast it spins and how long it stays in the air (let's assume we let it fall to the ground). And if we know the makeup of the ground, we can determine exactly how it bounces, and therefore we can predict which side will land facing up. And even if we *don't* know all these things, they all have definite values, independent of our knowledge of them. So once the coin leaves our hand, it is completely determined which side will land up. The "random" nature of the toss is therefore nothing more than a result of our ignorance of the initial properties of the coin.

The question then arises: *How do we create a process that is truly random?* It's a good bet that if you try to create a random process, you'll discover that it actually isn't random. Instead, it just appears to be random due to your lack of knowledge of various inputs at the start of the process. You might try to make a coin toss random by having a machine flip the coin, where the force and torque that it applies to the coin take on random values. But how do we make *these* things random? We've done nothing but shift the burden of proof back a step, so we haven't really accomplished anything.

This state of affairs is particularly prevalent when computers are used to generate random numbers. By various processes, computers can produce numbers that seem to be fairly random. However, there is actually no way that they can be truly random, because the output is completely determined by the input. And if the input isn't random (which we're assuming is the case, because otherwise we wouldn't need a random number generator!), then the output isn't either.

In all of the above scenarios, the issue at hand is that our definition of probability in Section 2.1 involved the phrase, "a very large number of *identical* trials." In none of the above scenarios are the trials identical. They all have (slightly) different inputs. So it's no surprise that things aren't truly random.

This then brings up the question: If we have *truly* identical processes, then shouldn't they give exactly identical results? If we flip a coin in exactly the same way each time, then we should get exactly the same outcome each time. So our definition of probability seems to preclude true randomness! This makes us wonder if there are actually *any* processes that can be truly identical and at the same time yield different results.

Indeed there are. It turns out that in quantum mechanics, this is exactly what happens. It is possible to have two exactly identical process that yield different results. Things are truly random; you can't trace the different outcomes to different inputs. A great deal of effort has gone into investigating this randomness, and unless our view of the universe of severely off-kilter, there are processes in quantum mechanics that involve bona fide randomness. If you think about this hard enough, it should make your head hurt. Our experiences in everyday life tell us that things happen *because* other things happened. But not so in quantum mechanics. There is no causal structure in certain settings. Some things just

happen. Period.

But even without quantum mechanics, there are plenty of other physical processes in the world that are essentially random, for all practical purposes. The ingredient that makes these processes essentially random is generally either (1) the sheer largeness of the numbers (of, for example, molecules) involved, or (2) the phenomenon of "chaos," which turns small uncertainties into huge ones. Using these effects, it is possible to create methods for generating nearly random things. For example, the noise in the radio frequency range in the atmosphere generates randomness due to the absurdly large number of input bits of data. And the pingpong balls bouncing around in a box used for picking lottery numbers generate randomness due to the chaotic nature of the ball collisions.

Different information

Let's say that I flip a coin and then look at the result and see a Heads, but I don't show you. Then for you, the probability of the coin being Heads is 1/2. But for me, the probability is 1. So if someone asks for the probability of the coin showing Heads, which number is it, 1/2 or 1? Well, there isn't a unique answer to this question, because the question is an incomplete one. The correct question to ask is, "What is the probability of the coin showing Heads, as measured by such-and-such a person?" You have to state who is calculating the probability, because different people have different information, and this affects the probability.

But you might argue that it's the same process, so it should have a uniquely-defined probability, independent of who is measuring it. But it actually *isn't* the same process for the two of us. The process for me involves looking at the coin, while the process for you doesn't. Said in another way, our definition of probability involved the phrase, "a very large number of *identical* trials." As far as you're concerned, if we do 1000 trials of this process, they're all identical to you. But they certainly aren't identical to me, because for some of them I observe Heads, and for some I observe Tails. This is about as nonidentical as they can be.

4.2 Appendix B: The natural logarithm, e

Consider the expression,

$$\left(1+\frac{1}{n}\right)^n. \tag{4.1}$$

Admittedly, this comes a bit out of the blue, but let's not worry about the motivation for now. After we derive a number of cool results below, you'll see why we chose to consider this particular expression. Table 4.1 gives the value of $(1+1/n)^n$ for various integer values of n. (Non-integer values are fine to consider, too.)

n	1	2	5	10	10^2	10^3	10^4	10^5	10^6
$(1+1/n)^n$	2	2.25	2.49	2.59	2.705	2.717	2.71815	2.71827	2.7182804

Table 4.1

Apparently, the values converge to a number somewhere around 2.71828. This can also be seen from Fig. 4.1, which shows a plot of $(1+1/n)^n$ vs. $\log(n)$. The $\log(n)$ here simply means that the "0" on the x-axis corresponds $n = 10^0 = 1$, the "1" corresponds $n = 10^1 = 10$, the "2" corresponds $n = 10^2 = 100$, and so on.

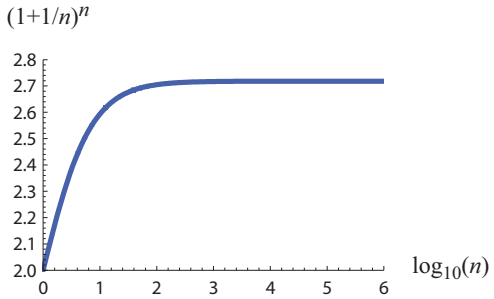


Figure 4.1

It's clear that by the time we reach the "6" (that is, $n = 10^6 = 1,000,000$), the curve has essentially leveled off to a constant value. This value happens to be $2.7182818285\ldots$. It turns out that the digits go on forever, with no overall pattern. However, the fortuitous double appearance of the "1828" makes it fairly easy to remember to 10 digits (even though you'll rarely ever need more accuracy than, say, 2.718). This number is known as the *natural logarithm*, and it is denoted by the letter e. The precise definition of e in terms of the expression in Eq. (4.1) is

$$\boxed{e \equiv \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n \approx 2.71828} \tag{4.2}$$

The "lim" notation simply means that we're taking the limit of this expression as n approaches infinity. If you don't like dealing with limits or infinity, just set n equal to a really large number like 10^{10}, and then you've pretty much got the value of e.

Remember that Eq. (4.2) is a *definition*. There's no actual content in it. All we did was take the quantity $(1 + 1/n)^n$ and look at what value it approaches as n became very large, and then we decided to call the result "e." We will, however, derive some actual results below, which aren't just definitions.

REMARK: If we didn't use a log plot in Fig. 4.1 and instead just plotted $(1 + 1/n)$ vs. n, the graph would stretch far out to the right if we wanted to go up to a large number like $n = 10^6$. Of course, we could shrink the graph in the horizontal direction, but then the region at small values of n would be squeezed down to essentially nothing. For example, the region up to $n = 100$ would take up only 0.01% of the graph. We would therefore be left with basically just a horizontal line. Even if we go up to only $n = 10^4$, we end up with the essentially straight line shown in Fig. 4.2.

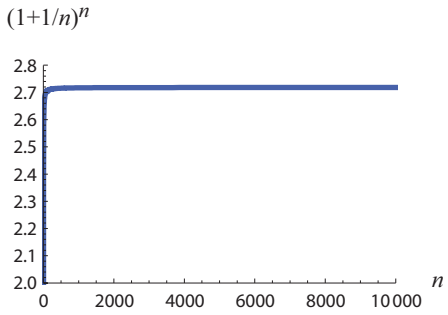


Figure 4.2

The features in the left part of the plot in Fig. 4.1 aren't too visible here. You can barely see the bend in the curve. And the rise up from 2 is basically a vertical line. In short, log plots are used to prevent the larger numbers from dominating the graph. This issue isn't so critical here, since we're actually only concerned with what $(1 + 1/n)^n$ looks like for large n, but nevertheless it's generally more informative to use a log plot in certain settings. ♣

It is quite interesting that $(1 + 1/n)^n$ approaches a definite value as n gets larger and larger. On one hand, you might think that because the $1/n$ term gets smaller and smaller (which means that $(1 + 1/n)$ gets closer and closer to 1), the whole expression should get closer and closer to 1. On the other hand, you might think that because the exponent n gets larger and larger, the whole expression should get larger and larger and approach infinity, because we're raising something to an ever-increasing power. It turns out that it does neither of these things. Instead, these two effects cancel each other, and the result ends up somewhere between 1 and ∞ at the particular value of about 2.71828.

As mentioned above, we introduced $(1 + 1/n)^n$ a bit out of the blue. But we've already found one interesting feature of it, namely that it approaches a definite number (which we labeled as "e") as $n \to \infty$. There are many other features, too. So many, in fact, that e ends up being arguably the most important number in mathematics, with the possible exception of π (but my vote is for e). From the nearly endless list of interesting facts about e, here are four:

1. Raising e to a power

What do we get when we raise e to a power? That is, what is the value of e^x? There are (at least) two ways to answer this. The simple way is to just use your calculator to raise $e = 2.71828$ to the power x. A number will pop out, and that's that.

However, there is another way which turns out to be immensely useful in the study of probability and statistics. If we relabel n in Eq. (4.2) as m (strictly for convenience), and if we then define $n \equiv mx$ in the fourth line below, we obtain

$$\begin{aligned} e^x &= \lim_{m\to\infty} \left(\left(1 + \frac{1}{m} \right)^m \right)^x \qquad (\textit{using } m \textit{ instead of } n \textit{ in Eq. (4.2)}) \\ &= \lim_{m\to\infty} \left(1 + \frac{1}{m} \right)^{mx} \qquad (\textit{multiplying exponents}) \\ &= \lim_{m\to\infty} \left(1 + \frac{x}{mx} \right)^{mx} \qquad (\textit{multiplying by 1 in the form of } x/x) \\ &= \lim_{n\to\infty} \left(1 + \frac{x}{n} \right)^{n} \qquad (\textit{using } n \equiv mx) \end{aligned} \tag{4.3}$$

If n is large but not infinite, then we can simply replace the "=" sign with a "≈" sign:

$$\boxed{e^x \approx \left(1 + \frac{x}{n} \right)^n} \qquad (\textit{for large } n) \tag{4.4}$$

The bigger n, the better the approximation. The condition under which this is a good approximation is

$$x \ll \sqrt{n}. \tag{4.5}$$

This will invariably hold for the situations we'll be dealing with. We'll just accept this condition here, but see the last paragraph in Appendix C if you want to know where it comes from.

Eq. (4.4) is a rather nice result. The x that appears in the numerator of the fraction is simply the exponent of e. It almost seems like too simple of a generalization of Eq. (4.2)

to be true. (Eq. (4.2) is a special case of Eq. (4.4) with $x = 1$.) Let's check that Eq. (4.4) does indeed hold for, say, $x = 2$. If we pick $n = 10^6$ (which certainly satisfies the $x \ll \sqrt{n}$ condition), we obtain $(1 + x/n)^n = (1 + 2/10^6)^{10^6} = 7.389041$. This is very close to the true value of e^2, which is about 7.389056. Larger values of n would make it even closer.

Example 1 (Compound interest): Assume that you have a bank account for which the interest rate is 5% per year. If this 5% is simply applied as a one-time addition at the end of the year, then after one year you will have 1.05 times the amount of money you started with. However, another way for the interest to be applied is for it to be compounded daily, with (5%)/365 being the daily rate (which happens to be about 0.014%). That is, your money at the end of each day equals $1 + (.05)/365$ times what you had at the beginning of the day. In this scenario, by what factor does your money increase after one year?

SOLUTION: Your money gets multiplied by a factor of $1 + (.05)/365$ each day, so at the end of one year (365 days), it increases by a factor of

$$\left(1 + \frac{.05}{365}\right)^{365}. \tag{4.6}$$

But this has exactly the same form as the expression in Eq. (4.4), with $x = .05$ and $n = 365$ (which certainly satisfies the $x \ll \sqrt{n}$ condition). So Eq. (4.4) tells us that after one year your money increases by a factor $e^{.05}$, which happens to be about 1.051. The effective interest rate is therefore 5.1%. That is, someone who had a 5.1% interest rate that was applied as a one-time addition at the end of the year would end up with the same amount of money as you. This effective interest rate of 5.1% is called the *yield*. So an annual rate of 5% has a yield of 5.1%. The reason why the yield ends up being larger than 5% is because the interest rate each day is being applied not only to your initial amount, but also to all the interest you've received in the preceding days. In short, you're earning interest on your interest.

The increase by .1% isn't so much. But if the annual interest rate is instead 10%, and if it is compounded daily, then you can show that you will end up with a yearly factor of $e^{.10} = 1.105$, which means that the yield is 10.5%. And an annual rate of 20% produces a yearly factor of $e^{.20} = 1.22$, which means that the yield is 22%.

Example 2 (Doubling your money): The extra compound-interest benefit of .1% (for the rate of 5%) we found in the previous example is quite small, so you can't go too wrong if you just ignore it. However, the effect of compound interest *cannot* be ignored in the following question: If the annual interest rate is 5%, and if it is compounded daily, how many years will it take to double your money?

SOLUTION: First, note the incorrect line of reasoning: If you start with N dollars, then doubling your money means that you eventually need to increase it by another N dollars. Since it increases by $(.05)N$ each year, you need 20 of these increases (because $20(.05) = 1$) to obtain the desired increase of N. So it takes 20 years. But this is incorrect, because it ignores the fact that you have more money each year and are therefore earning interest on a larger and larger amount of money. The correct line of reasoning is the following.

We saw in the previous example that at the end of each year, your money increases by a factor of $e^{.05}$ compared with what it was at the beginning of the year. So after n years it increases by n of these factors, that is, by $(e^{.05})^n$ which equals $e^{(.05)n}$. Now, we want to find the value of n for which this overall factor equals 2. A little trial and error in your calculator shows that $e^{.7} \approx 2$. (In the language of logs, this is the statement that $\log_e 2 \approx 0.7$. But this terminology isn't important here.) So we need the $(.05)n$ exponent to equal .7, which in turn implies that $n = (.7)/(.05) = 14$. So it takes 14 years to double your money.

You can think of this result for n as 70 divided by 5. For a general interest rate of r%, the exact same reasoning shows that the number of years required to double your money is 70

divided by r. So in remembering this rule, you simply need to remember one number: 70. Equivalently, the time it takes to double your money is 70% of the naive answer that ignores the effects of compound interest.

Unlike the previous example where the interest earned was small, the interest earned in this example is large (it equals N dollars by the end), so the effects of earning interest on your interest (that is, the effects of compound interest) cannot be ignored. ♣

2. A handy formula

Expressions of the form $(1+a)^n$ come up often in mathematics, especially in probability and statistics. It turns out that if a is small enough or if n is large enough (which is invariably true for the situations we'll be dealing with), then the following very nice approximate formula holds:

$$\boxed{(1+a)^n \approx e^{an}} \tag{4.7}$$

This formula was critical in our discussion of the exponential and Poisson distributions in Section 3.4. The condition under which this approximation is a good one is

$$na^2 \ll 1, \qquad \text{or equivalently} \quad a \ll 1/\sqrt{n}. \tag{4.8}$$

Feel free to just accept this, but the explanation is given in the last paragraph in Appendix C if you're interested.

Although it looks different, Eq. (4.7) says essentially the same thing as Eq. (4.4), in that the derivation of Eq. (4.7) from Eq. (4.4) is hardly a derivation at all. It takes only two lines:

$$\begin{aligned} (1+a)^n &= \left(1+\frac{an}{n}\right)^n \qquad (\textit{multiplying by 1 in the form of } n/n) \\ &\approx e^{an} \qquad\quad (\textit{using Eq. (4.4) with } x \equiv an) \end{aligned} \tag{4.9}$$

And that's all there is to it. A special case of Eq. (4.7) is $a = 1/n$ (and so $an = 1$), in which case Eq. (4.7) gives $(1 + 1/n)^n \approx e^1$, which is equivalent to Eq. (4.2) if n is large. Another special case is $n = 1$ with a being very small. Eq. (4.7) then gives

$$\boxed{1 + a \approx e^a} \qquad (\textit{if } a \textit{ is very small}) \tag{4.10}$$

As a more numerical example, if $a = 0.001$ and $n = 10,000$, we have $an = 10$, and so Eq. (4.7) gives $(1.001)^{10,000} \approx e^{10}$. This is indeed roughly true. The lefthand side equals $21,917$ and the righthand side equals $22,026$, so the error is only about 0.5%.[1] Note that all of these examples and special cases do indeed satisfy the $a \ll 1/\sqrt{n}$ condition stated in Eq. (4.8).

Although Eq. (4.7) is very handy in many situations, it turns out that for some purposes (as we saw in Section 3.4) it isn't quite a good enough approximation. So we'll present a more accurate version of Eq. (4.7) in Appendix C.

[1] Whenever we use a "$\approx$" sign, we use it in a *multiplicative* (equivalently, a ratio) sense, and not an *additive* sense. $21,917$ and $22,026$ differ by 109, which you might consider to be a large number, but that is irrelevant. The ratio of the numbers is essentially equal to 1, so they are "approximately equal" in that sense.

3. The infinite series for e^x

There is a very cool alternative expression for e^x that we can derive. This expression is[2]

$$\boxed{e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots} \tag{4.11}$$

The first two terms can technically be written as $x^0/0!$ and $x^1/1!$, respectively, so all the terms take the form of $x^n/n!$, where n runs from zero to infinity.

A special case of Eq. (4.11) occurs when $x = 1$, in which case we have

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots. \tag{4.12}$$

These terms get very small very quickly, so you don't need to include many of them to get a good approximation to e. Even just going out to the 10! term gives $e \approx 2.71828180$, which is accurate to the seventh digit beyond the decimal point. Eq. (4.12) provides a much less labor-intensive approximation to e than the original $(1 + 1/n)^n$ approximation. Even with $n = 10^6$, Table 4.1 shows that the latter is accurate only to the fifth digit.

We can derive Eq. (4.11) by using Eq. (4.4) along with our good old friend, the binomial expansion. Expanding Eq. (4.4) via the binomial expansion gives

$$\begin{aligned} e^x &= \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n \\ &= \lim_{n\to\infty} \left[1 + \binom{n}{1}\left(\frac{x}{n}\right) + \binom{n}{2}\left(\frac{x}{n}\right)^2 + \binom{n}{3}\left(\frac{x}{n}\right)^3 + \cdots\right] \\ &= \lim_{n\to\infty} \left[1 + x\left(\frac{n}{n}\right) + \frac{x^2}{2!}\left(\frac{n(n-1)}{n^2}\right) + \frac{x^3}{3!}\left(\frac{n(n-1)(n-2)}{n^3}\right) + \cdots\right]. \end{aligned} \tag{4.13}$$

This looks roughly like what we're trying to show in Eq. (4.11), if only we could make the terms in parentheses go away. And indeed we can, because in the limit $n \to \infty$, all these terms equal 1. This is true because if $n \to \infty$, then both $n - 1$ and $n - 2$ are essentially equal to n (multiplicatively). More precisely, the ratios $(n - 1)/n$ and $(n - 2)/n$ are both equal to 1 if $n = \infty$. So we have

$$\lim_{n\to\infty} \left(\frac{n(n-1)}{n^2}\right) = 1, \qquad \text{and} \qquad \lim_{n\to\infty} \left(\frac{n(n-1)(n-2)}{n^3}\right) = 1, \tag{4.14}$$

and likewise for the terms involving higher-order powers of x. Eq. (4.13) therefore reduces to Eq. (4.11).[3] If you have any doubts about Eq. (4.11), you should verify with a calculator that it holds for, say, $x = 2$. Going out to the 10! term should convince you that it works.

REMARK: Another way to convince yourself that Eq. (4.11) is correct is the following. Consider what e^x looks like if x is a very small number, say, $x = 0.0001$. We have

$$e^{0.0001} = 1.0001000050001666\ldots \tag{4.15}$$

[2] For those of you who know calculus, this expression is known as the *Taylor series* for e^x. But that's just a name, so ignore it if you've never heard of it.

[3] For any large but finite n, the terms in parentheses far out in the series in Eq. (4.13) will eventually differ from 1, but by that point the factorials in the denominators will make the terms negligible, so we can ignore them. Even if x is large, so that the powers of x in the numerators become large, the factorials in the denominators will dominate after a certain point in the series, making the terms negligible. But we're assuming $n \to \infty$ anyway, so any of these issues relating to finite n are irrelevant.

This can be written more informatively as

$$\begin{aligned} e^{0.0001} &= 1.0 \\ &\quad + 0.0001 \\ &\quad + 0.000000005 \\ &\quad + 0.0000000000001666\ldots \\ &= 1 + (0.0001) + \frac{(0.0001)^2}{2!} + \frac{(0.0001)^3}{3!} + \cdots, \end{aligned} \tag{4.16}$$

in agreement with Eq. (4.11). If you made x even smaller (say, 0.000001), then the same pattern would form, but just with more zeros between the numbers than in Eq. (4.15).

Eq. (4.16) shows that if e^x can be expressed as a sum of powers of x (that is, in the form of $a + bx + cx^2 + dx^3 + \cdots$), then a and b must equal 1, c must equal 1/2, and d must equal 1/6. If you kept more digits in Eq. (4.15), you could verify the $x^4/4!$ and $x^5/5!$, etc., terms in Eq. (4.11) too. But things aren't quite as obvious at this point, because we don't have all the nice zero's as we do in the first 12 digits of Eq. (4.15). ♣

4. The slope of e^x

Perhaps the most interesting and important property of e (although for the specific purposes of this book, the second property above is the most important one) is that if we plot the function $f(x) = e^x$, the slope of the curve at any point equals the value of the function at that point, namely e^x. For example, in Fig. 4.3 the slope at $x = 0$ is $e^0 = 1$, and the slope at $x = 2$ is $e^2 \approx 7.39$.[4] (Note the different scales on the x and y axes, which makes these slopes appear different on the page.) The number e is the one special number for which this is true. That is, the same thing is *not* true for, say, 2^x or 10^x. The derivation of this property is by no means necessary for an understanding of the material in this book, but we'll present it in Appendix D, just for the fun of it.

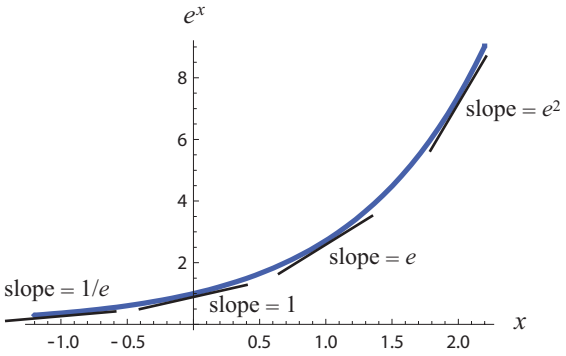


Figure 4.3

Actually, any function of the form Ae^x (where A is some constant) has the property that the slope at any point equals the value of the function at that point. This is true because both the value and the slope differ by the same factor of A from the corresponding quantities in the e^x case. So if the property holds for e^x (which it does), then it also holds for Ae^x.

[4] By "slope" we mean the slope of the line that is tangent to the curve at the given point. You can imagine the curve being made out of an actual piece of wire, and if you press a straight stick up against it, the stick will form the tangent to the curve at the point of contact.

4.3 Appendix C: More accuracy for $(1+a)^n$

In Appendix B, we derived the "handy formula" in Eq. (4.7),

$$(1+a)^n \approx e^{an}. \tag{4.17}$$

This formula was critical in the derivations of the Exponential and Poisson distributions in Sections 3.3.4 and 3.3.5. However, when we derived the Gaussian approximations in Section 3.4, we saw that a more accurate approximation was needed, namely

$$\boxed{(1+a)^n \approx e^{an}e^{-na^2/2}} \tag{4.18}$$

If a is sufficiently small, this extra factor of $e^{-na^2/2}$ is irrelevant, because it is essentially equal to $e^{-0} = 1$. So Eq. (4.18) reduces to Eq. (4.17). But if a isn't so small, this extra factor is necessary if we want to have a good approximation.[5] For example, let's say we have $n = 100$ and $a = 1/10$. Then

$$(1+a)^n = (1+1/10)^{100} \approx 13,781, \qquad \text{and} \qquad e^{an} = e^{10} \approx 22,026. \tag{4.19}$$

So $(1+a)^n \approx e^{an}$ is not a good approximation at all. However, the $e^{-na^2/2}$ factor in this case equals $e^{-1/2} \approx 0.6065$, which gives

$$e^{an}e^{-na^2/2} \approx (22,026)(0.6065) \approx 13,359. \tag{4.20}$$

So $(1+a)^n \approx e^{an}e^{-na^2/2}$ is a rather good approximation, with an error of only about 3%. As as exercise, you can show that if we had picked larger numbers, say, $n = 10,000$ and $a = 1/100$, then Eq. (4.17) would be a similarly poor approximation, but Eq. (4.18) would be an excellent one, off by only 0.3%.

Let's now derive Eq. (4.18). We'll start by considering the $e^a \approx 1 + a$ approximation in Eq. (4.10). This is a decent approximation if a is small, but we know for a fact that it can't be exactly correct, because Eq. (4.11) tells us what the value of e^a actually is, namely

$$e^a = 1 + a + \frac{a^2}{2!} + \cdots, \tag{4.21}$$

where the dots indicate higher powers of a which are small compared with the a and $a^2/2$ terms if a is small. So we see that the error in the $e^a \approx 1 + a$ approximation is mainly due to the $a^2/2$ term in Eq. (4.21). We would therefore like to get rid of this term.

So the question is: what power should we raise e to, in order to get rid of the $a^2/2$ term in Eq. (4.21)? It's possible to answer this question with a little trial and error, but let's be systematic about it. Since we're looking to get rid of the $a^2/2$ term, it's reasonable to tack on a term involving a^2 to the exponent in Eq. (4.21). So let's try an exponent of the form $a + ka^2$, where k is a number yet to be determined. Plugging $a + ka^2$ in for the x in Eq. (4.11) gives the following result (the dots here indicate terms that are of order at least a^3, and thus negligible if a is small):

$$\begin{aligned} e^{a+ka^2} &= 1 + (a+ka^2) + \frac{1}{2!}(a+ka^2)^2 + \cdots \\ &= 1 + (a+ka^2) + \frac{1}{2}(a^2 + \cdots) + \cdots \\ &= 1 + a + \left(k + \frac{1}{2}\right)a^2 + \cdots. \end{aligned} \tag{4.22}$$

[5] Of course, if a is too big, then even the inclusion of this extra factor won't be enough to yield a good approximation. Another factor needs to be tacked on. We won't worry about that here, but see Problem 1 if you're interested.

We want the coefficient of a^2 to be zero, so we need k to equal $-1/2$. With $k = -1/2$, Eq. (4.22) then takes the form of $e^{a-a^2/2} \approx 1 + a$. But $e^{a-a^2/2} = e^a e^{-a^2/2}$, so we obtain $e^a e^{-a^2/2} \approx 1 + a$. Raising both side of this equation to the nth power then yields Eq. (4.18), as desired.

We can now see why, as we claimed in Eq. (4.8), that the condition for Eq. (4.17) to be a good approximation is $na^2 \ll 1$, or equivalently $a \ll 1/\sqrt{n}$. This is the condition that makes the extra factor of $e^{-na^2/2}$ in Eq. (4.18) be essentially equal to $e^{-0} = 1$, thereby leaving us with only the e^{an} term that appears in Eq. (4.17).

4.4 Appendix D: The slope of e^x

(Note: This Appendix is for your entertainment only. We won't be using any of these results in this book. But the derivation of the slope of the function e^x gives us an excuse to play around with some of the properties of e, and also to present some of the foundational concepts of calculus.)

First derivation

In the fourth property of e in Appendix B, we stated that the slope of the $f(x) = e^x$ curve at any point equals the value of the function at that point, namely e^x. We'll now show why this is true. (In the language of calculus, this is the statement that the *derivative* (the slope) of e^x equals itself, e^x.)

There are two main ingredients in the derivation. The first is Eq. (4.10) from the Appendix B. To remind ourselves that the a in that equation is assumed to be very small, let's relabel it as ϵ, which is the customary letter that mathematicians use for a very small quantity. We then have, switching the sides of the equation,

$$e^\epsilon \approx 1 + \epsilon \qquad (\textit{if } \epsilon \textit{ is very small}) \tag{4.23}$$

You should verify this with a calculator, letting ϵ be 0.1 or 0.01, etc. The number e is the one special number for which this is true. It is *not* the case that $2^\epsilon \approx 1 + \epsilon$ or $10^\epsilon \approx 1 + \epsilon$ (which you should also verify with a calculator).

The second main ingredient is the strategy of finding the slope of the function $f(x) = e^x$ (or any function, for that matter) at a given point, by first finding an *approximate* slope, and by then making the approximation better and better. This proceeds as follows.

An easy approximate way to determine the slope at a value of x, say $x = 2$, is to find the *average* slope between $x = 2$ and a nearby point, say $x = 2.1$. This average slope is

$$\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{e^{2.1} - e^2}{0.1} \approx 7.77. \tag{4.24}$$

From Fig. 4.4, however, we see that this approximate slope is larger than the true slope.[6] To get a better approximation, we can use a closer point, say $x = 2.01$. And then an even better approximation can be achieved with $x = 2.001$. These two particular values of x yields slopes of

$$\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{e^{2.01} - e^2}{0.01} \approx 7.43, \qquad \text{and} \qquad \frac{e^{2.001} - e^2}{0.001} \approx 7.392. \tag{4.25}$$

[6]The curve is this figure is just an arbitrary curve and not the specific e^x one, but the general features are the same. All that really matters is that the curve is concave upward like the e^x curve. The reason we're not using the actual e^x curve here is that $x = 2.1$ is so close to $x = 2$ that we wouldn't be able to see the important features.

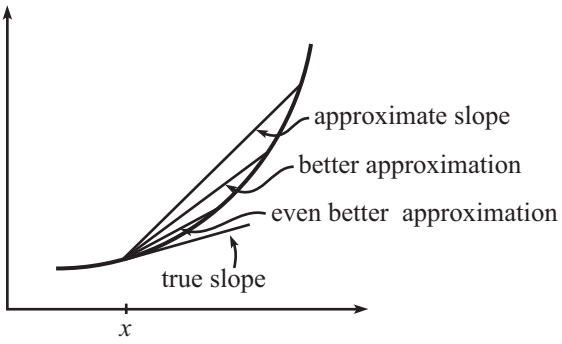


Figure 4.4

If we kept going with smaller and smaller differences from 2, we would find that the slopes converge to a certain value, which happens to be about 7.389. It is clear from Fig. 4.4 (which, again, is just a picture of a generic-looking curve) that the approximate slopes swing down and get closer and closer to the actual tangent-line slope. This number of 7.389 must therefore be the slope of the e^x curve at $x = 2$.

Now, our goal in all of this is to show that the slope of e^x equals e^x. We just found that the slope at $x = 2$ equals 7.389, so it had better be true that e^2 also equals 7.389. And indeed it does, as you can verify. So at least in the case of $x = 2$, we have demonstrated that the slope of e^x equals e^x.

Having learned how to calculate the slope at the specific value of $x = 2$, we can now address the case of general x. To find the slope, we can imagine taking a very small number ϵ and finding the average slope between x and $x + \epsilon$ (as we did with, say, 2 and 2.01), and then letting ϵ become smaller and smaller. Written out explicitly, the formal definition of the slope of a general function $f(x)$ at the value x is

$$\boxed{\text{slope} = \frac{\text{rise}}{\text{run}} = \lim_{\epsilon \to 0} \left(\frac{f(x+\epsilon) - f(x)}{\epsilon} \right)} \tag{4.26}$$

This might look a little scary, but it's simply saying with an equation what Fig. 4.4 says with a picture: you can get a better and better approximation to the slope by looking at the average slope between two points and having the points get closer and closer together.

For the case at hand where our function $f(x)$ is e^x, we have (with the understanding that we're concerned with the $\epsilon \to 0$ limit in all of these steps)

$$\begin{aligned} \text{slope} = \frac{\text{rise}}{\text{run}} &= \frac{e^{x+\epsilon} - e^x}{\epsilon} \\ &= e^x \left(\frac{e^\epsilon - 1}{\epsilon} \right) \qquad (\textit{factoring out } e^x) \\ &\approx e^x \left(\frac{(1+\epsilon) - 1}{\epsilon} \right) \qquad (\textit{Using Eq. (4.23)}) \\ &= e^x \left(\frac{\epsilon}{\epsilon} \right) \\ &= e^x, \end{aligned} \tag{4.27}$$

as we wanted to show. Since we're concerned with the $\epsilon \to 0$ limit (that's how the true slope is obtained), the "$\approx$" sign in the third line actually becomes an exact "$=$" sign. So we are correct to say that the slope of the e^x curve is *exactly* equal to e^x.

Note that Eq. (4.23) was critical in this derivation. Eq. (4.27) holds only for the special number e, because the $e^\epsilon \approx 1+\epsilon$ result from Eq. (4.23) that we used in the third line of Eq. (4.27) holds only for e. The slope of, say, 2^x is *not* equal to 2^x, because Eq. (4.23) doesn't hold if e is replaced by 2 (or any other number).

Given that we're concerned with the $\epsilon \to 0$ limit, you might be worried about having an ϵ in the denominator in Eq. (4.27), since division by zero isn't allowed. But there is also an ϵ in the numerator, so you can cancel them first, and *then* take the limit $\epsilon \to 0$.

Second derivation

Having introduced the strategy of finding slopes by finding approximate slopes via a small number ϵ, let's take advantage of this strategy and also find the slope of a general power-law function, $f(x) = x^n$, where n is a nonnegative integer. We'll then use this result to give an alternate derivation of the fact that the slope of e^x equals itself, e^x.

We claim that the slope of the function x^n takes the form,

$$\boxed{\text{slope of } x^n \text{ equals } nx^{n-1}} \tag{4.28}$$

which you can quickly verify for the cases of $n = 0$ and $n = 1$. To demonstrate this for a general integer n, we can (as we did in the first derivation above) imagine taking a very small number ϵ and finding the average slope between x and $x+\epsilon$, and then letting ϵ become smaller and smaller (that is, taking the $\epsilon \to 0$ limit); see Eq. (4.26). To get a feel for what's going on, let's start with a specific value of n, say, $n = 2$. In the same manner as above, we have (using Eq. (4.26) along with our trusty friend, the binomial expansion)

$$\begin{aligned} \text{slope} = \frac{\text{rise}}{\text{run}} &= \frac{(x+\epsilon)^2 - x^2}{\epsilon} \\ &= \frac{(x^2 + 2x\epsilon + \epsilon^2) - x^2}{\epsilon} \\ &= \frac{2x\epsilon + \epsilon^2}{\epsilon} \\ &= 2x + \epsilon. \end{aligned} \tag{4.29}$$

If we now take the $\epsilon \to 0$ limit, the ϵ term goes away, leaving us with only the $2x$ term. So we've shown that the slope of the x^2 curve equals $2x$, which is consistent with the nx^{n-1} expression in Eq. (4.28).

Let's try the same thing with $n = 3$. Again using the binomial expansion, we have

$$\begin{aligned} \text{slope} = \frac{\text{rise}}{\text{run}} &= \frac{(x+\epsilon)^3 - x^3}{\epsilon} \\ &= \frac{(x^3 + 3x^2\epsilon + 3x\epsilon^2 + \epsilon^3) - x^3}{\epsilon} \\ &= \frac{3x^2\epsilon + 3x\epsilon^2 + \epsilon^3}{\epsilon} \\ &= 3x^2 + 3x\epsilon + \epsilon^2. \end{aligned} \tag{4.30}$$

When we take the $\epsilon \to 0$ limit, *both* the $3x\epsilon$ and ϵ^2 terms go away, leaving us with only the $3x^2$ term. Basically, anything with an ϵ in it goes away when we take the $\epsilon \to 0$ limit. So we've shown that the slope of the x^3 curve equals $3x^2$, which is again consistent with the nx^{n-1} expression in Eq. (4.28).

You can see how this works for the case of general n. The goal is to calculate

$$\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{(x+\epsilon)^n - x^n}{\epsilon}\,. \tag{4.31}$$

Using the binomial expansion, the first few values of $(x+\epsilon)^n$ are (you'll see below why we've added the parentheses in the second terms on the righthand side):

$$\begin{aligned}
(x+\epsilon)^1 &= x + (1)\epsilon, \\
(x+\epsilon)^2 &= x^2 + (2x)\epsilon + \epsilon^2, \\
(x+\epsilon)^3 &= x^3 + (3x^2)\epsilon + 3x\epsilon^2 + \epsilon^3, \\
(x+\epsilon)^4 &= x^4 + (4x^3)\epsilon + 6x^2\epsilon^2 + 4x\epsilon^3 + \epsilon^4, \\
(x+\epsilon)^5 &= x^5 + (5x^4)\epsilon + 10x^3\epsilon^2 + 10x^2\epsilon^3 + 5x\epsilon^4 + \epsilon^5.
\end{aligned} \tag{4.32}$$

When we substitute these into Eq. (4.31), the first term disappears when we subtract off the x^n. Then when we perform the division by ϵ as Eq. (4.31) indicates, we simply reduce the power of ϵ by one in every term. So at this stage, for each of the expansions in Eq. (4.32), the first term has disappeared, the second term involves no ϵ's, and the third and higher terms involve at least one power of ϵ. Therefore, when we take the $\epsilon \to 0$ limit, the third and higher terms all go to zero, so we're left with only the second term (without the ϵ). In other words, in each line of Eq. (4.32) we're left with just the term in the parentheses. And this term has the form of nx^{n-1}, as desired. We have therefore proved Eq. (4.28). The multiplicative factor of n here is simply the $\binom{n}{1}$ binomial coefficient.

We can now provide a second derivation of the fact that the slope of e^x equals itself, e^x. This derivation involves writing e^x in the form given in Eq. (4.11), which we'll copy here,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots. \tag{4.33}$$

and then finding the slope by applying Eq. (4.28) to each term.

REMARK: We'll need to use the fact that the slope of the sum of two functions equals the sum of the slopes of the two functions. And also that the "two" here can be replaced by any number. This might seem perfectly believable and not necessary to prove, but let's prove it anyway. We're setting this proof off in a Remark, just in case you want to ignore it.

Consider a function $F(x)$ that can be written as the sum of two other functions, $F(x) = f_1(x) + f_2(x)$. We claim that the slope of $F(x)$ at a particular value of x is simply the sum of the slopes of $f_1(x)$ and $f_2(x)$ at that value of x. This follows from the expression for the slope in Eq. (4.26). We have

$$\begin{aligned}
\text{slope of } F(x) = \frac{\text{rise}}{\text{run}} &= \lim_{\epsilon\to 0}\left(\frac{F(x+\epsilon) - F(x)}{\epsilon}\right) \\
&= \lim_{\epsilon\to 0}\left(\frac{\big(f_1(x+\epsilon) + f_2(x+\epsilon)\big) - \big(f_1(x) + f_2(x)\big)}{\epsilon}\right) \\
&= \lim_{\epsilon\to 0}\left(\frac{f_1(x+\epsilon) - f_1(x)}{\epsilon}\right) + \lim_{\epsilon\to 0}\left(\frac{f_2(x+\epsilon) - f_2(x)}{\epsilon}\right) \\
&= \big(\text{slope of } f_1(x)\big) + \big(\text{slope of } f_2(x)\big).
\end{aligned} \tag{4.34}$$

The main point here is that in the third line we grouped the f_1 terms together, and likewise for the f_2 terms. We can do this with any number of functions, of course, so that's why the above "two" can be replaced with any number. We can even have an infinite number of terms, as in the case in Eq. (4.33). ♣

So we now know that the slope of e^x is the sum of the slopes of all the terms in Eq. (4.33), of which there are an infinite number. And Eq. (4.28) tells us how to find the slope of each term. Let's look at the first few.

The slope of the first term in Eq. (4.33) (the 1) is zero. The slope of the second term (the x) is 1. The slope of the third term (the $x^2/2!$) is $(2x)/2! = x$. The slope of the fourth term

(the $x^3/3!$) is $(3x^2)/3! = x^2/2!$. It appears that when finding the slope, each term turns into the one preceding it in the series (a fact that is due to the factorials in the denominators). So the infinite series that arises after finding the slope is the same as the original series. In other words, the derivative of e^x equals itself, e^x. Written out explicitly, we have

$$\begin{aligned} \text{Slope of } e^x &= \text{Slope of } \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots\right) \\ &= 0 + 1 + \frac{2x}{2!} + \frac{3x^2}{3!} + \frac{4x^3}{4!} + \frac{5x^4}{5!} + \cdots \\ &= 0 + 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \\ &= e^x, \end{aligned} \tag{4.35}$$

as we wanted to show.

4.5 Problems

1. **Third term in $(1+a)^n$** $**$

   Following the strategy used in Eq. (4.22), find the values of k_2 and k_3 that make the a^2 and a^3 terms vanish in the expansion of $e^{a+k_2a^2+k_3a^3}$. That is, find the values of k_2 and k_3 that make $e^{a+k_2a^2+k_3a^3} = 1+a$, up to corrections of order a^4. You can then raise both sides of this equation to the nth power to obtain $(1+a)^n \approx e^{an}e^{k_2a^2}e^{k_2a^3}$, which is the improved version of Eq. (4.18).

Many more problems will be added...

4.6 Solutions

1. **Third term in $(1+a)^n$**
   We have

$$\begin{aligned} e^{a+k_1a^2+k_2a^3} &= 1 + (a + k_1a^2 + k_2a^3) + \frac{1}{2!}(a + k_1a^2 + k_2a^3)^2 + \frac{1}{3!}(a + k_1a^2 + k_2a^3)^3 + \cdots \\ &= 1 + (a + k_1a^2 + k_2a^3) + \frac{1}{2}(a^2 + 2k_1a^3 + \cdots) + \frac{1}{6}(a^3 + \cdots) + \cdots \\ &= 1 + a + \left(k_1 + \frac{1}{2}\right)a^2 + \left(k_2 + k_1 + \frac{1}{6}\right)a^3 + \cdots, \end{aligned} \tag{4.36}$$

   where the dots stand for terms that are of order at least a^4. We want the coefficients of a^2 and a^3 to be zero, so we need $k_1 + 1/2 = 0$ and $k_2 + k_1 + 1/6 = 0$. The first of these equations gives $k_1 = -1/2$, and then the second gives $k_2 = 1/3$. As mentioned in the statement of the problem, we then have the following improvement to the approximation in Eq. (4.18):

$$(1+a)^n \approx e^{an}e^{-na^2/2}e^{na^3/3}. \tag{4.37}$$

   Using this method, you can find the next-order correction, and then the next, and so on. But it gets to be a calculational pain. The next factor happens to be $e^{-na^4/4}$, so the pattern is fairly clear: the factors are the reciprocals of the integers, with alternating signs.[7]

[7] It turns out that a method from calculus (using a certain Taylor series) enables this full result to be derived in about two lines. So if you wanted to explicitly calculate all the factors out to the one involving, say, na^6, it would probably be quicker to just go and learn calculus than to fight your way through a long and tedious calculation.