Just Enough Probability*

Alan R. Rogers†

January 3, 2023

Chapter Contents

Preface

This pamphlet is for students who need enough probability to get through an undergraduate course on such subjects as population genetics or ecology. It assumes that readers are comfortable with algebra. There is a little calculus too, but it does not appear until section 3.2, at the very end.

†Dept. of Anthropology, 260 Central Campus Dr, University of Utah, Salt Lake City, UT 84112, USA. I thank Elizabeth Cashdan and Jon Seger for advice.

Chapter 1. Probability

Probability theory is about things that cannot be predicted with certainty. It contributes to science in two ways. First, there is uncertainty in the phenomena we study. Even when we know the genes of the parents, we cannot predict with certainty those of their offspring. Neither can we predict with certainty the path of a molecule through a gas, or how long anyone will live. All such phenomena need theories with probabilistic components. Probability also contributes to science in a second way. We often study populations of things that are too large to examine in their entirety. Instead we study incomplete samples, which reflect only imperfectly the properties of the larger whole. Thus, probability theory also underlies statistics, the science of sampling.

There are two versions of probability theory. The first studies *statistical probability*, the relative frequency with which an event occurs in the long run. The second studies *subjective (or Bayesian) probability*, which measures one's degree of belief. Bayesian probability has recently become important in statistics. The focus here is on statistical probability, which has long played a central role in many disciplines.

This primer is short, but you should not expect to read it quickly. Quantitative material takes time to digest. Read it with pencil in hand and do the exercises as you go. But do not spend too much time on any one exercise. If you get stuck, consult the answers in the back. Read these answers in any case, because they contain some of the material I am trying to teach.

1.1. Probability

All of us have an intuitive understanding of statements such as "the chances are three in five." This is exactly

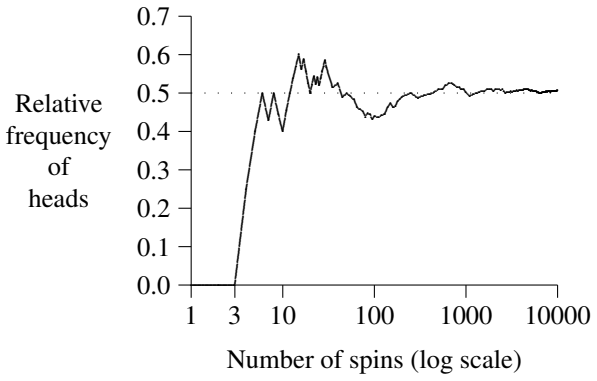


Figure 1: Fraction of heads in 10,000 spins of a coin [2]

what is meant by the statement "the (statistical) probability is $3/5$." In a large number of repeated trials we would expect the event in question to occur about $3/5$ of the time. In other words, its *relative frequency* should be about $3/5$. The relative frequency of an event is simply the number of trials on which it occurs expressed as a fraction of the total. It is a cumbersome term for a simple idea, and people often shorten it to "frequency." Unfortunately, *that* word is also used for the raw count of events. For example, if we spin a coin 100 times and observe 30 heads, then the relative frequency of heads is 0.3. This is also referred to simply as the frequency of heads. In other contexts, however, the "frequency of heads" might refer to 30, the number of heads. You have to catch meaning from context.

The larger the number of trials, the smaller will be the difference between the relative frequency and the probability. My favorite illustration of this idea comes from World War II. Just before the war, the mathematician John Kerrich [2] was visiting Denmark. Two days before his scheduled departure, the German army overran Denmark, and Kerrich was interned for the duration of the war. He must have had a lot of spare time, for he undertook several lengthy experiments on probability. In one of them, he spun a coin 10,000 times and kept a record of the outcomes.

The resulting data are shown in Figure 1. There, the vertical axis shows the relative frequency of heads. Its value bounced around at first but got closer and closer to $1/2$ as the number of spins increased. When we say that the probability of heads on a single spin is $1/2$, this is what we have in mind. There is nothing special about the number $1/2$. Had Kerrich's coin been bent, it might have tended to favor one side. Even so, in 10,000 tosses it would have converged toward some particular value. This value is called the *probability* of "heads."

A coin that is equally likely to fall on either side is said to be "fair," but nothing guarantees that coins must be fair. When we say that one is fair, we are stating a hypothesis, not a fact. One can estimate a probability, as Kerrich did by spinning his coin. But estimates always have error, and the exact probability is never known. We often make judgments about probabilities even without this sort of evidence. I have never spun any of the coins in my pocket, yet I am confident that all of them are very nearly fair. Why? I'm not sure. Perhaps because people have told me so, and perhaps because coins are nearly (although not quite) symmetrical. Neither of these prove that my coins are exactly fair. When we assume that probabilities have particular values, we are building a model. The fit of that model to the real world is always a matter to be tested against data.

If Kerrich were doing his experiment today, he might have automated the process. Most computer languages provide a way to generate "uniform random numbers," which we'll discuss more fully on p. 12. The gist is that these generators deliver numbers that are (in theory) equally likely to fall anywhere in the interval from 0 to 1. Hence, the value falls between 0 and $1/2$ with probability $1/2$. This makes it easy to simulate the spin of a fair coin. Simply get one number from the random number generator. Interpret it as heads if less than $1/2$ and as tails if greater.

⋆ EXERCISE **1–1** Describe a method for simulating the spin of an unfair coin, for which heads has probability 0.3.

⋆ EXERCISE **1–2** If you know how to write computer programs, write one that replicates Kerrich's experiment.

1.2. An urn experiment

Texts on probability theory frequently discuss experiments that involve drawing balls out of urns. During his captivity, Kerrich did his own version of an urn experiment. Instead of an urn, he used a box and four ping-pong balls—two red and two green. The experiment consisted of 5000 identical trials, each of which began with all four balls in the box. Kerrich's assistant then shook the box, looked away, and drew out first one ball and then (without replacing the first) another. They wrote down the colors of the two balls. On each trial, there were four possible outcomes. Kerrich counted the trials in which each occurred, with results as shown in Table 1.

1.3. A model

There is a clear pattern in these data: the second ball was usually red if the first was green and usually green if the

Table 1: The results of 5000 repetitions of Kerrich's [2] urn experiment

First ball	Second ball: Red	Second ball: Green	sum
Red	756	1689	2445
Green	1688	867	2555
sum	2444	2556	5000

first was red. What features of the experiment might account for this pattern?

To answer such a question, we need to build a model. This involves making assumptions that seem plausible and then using these to calculate the probabilities of the various events in the data. Finally, we compare the predicted values to those in the data.

I built my own model using the tree diagram in Fig. 2. Each trial begins at the root and progresses through the tree to one of the tips. For example, if two red balls are drawn, we take the upper path at each node. There are labels at the tips of the branches, which correspond to events: RR for red–red, RG for red–green, and so forth.

The numbers in the figure are all probabilities. To explain them, I begin at the root of the tree, which corresponds to the beginning of a trial. At that point, the box contains two red balls and two green ones. If each ball is equally likely to be chosen, then the first ball is equally likely to be red or green. Thus, the two paths from the root each have probability $1/2$. So far, the model is consistent with the data, since the first ball was red on $2445/5000$—very nearly $1/2$—of the trials.

Suppose now that the first ball was red and you are about to draw a second ball. Three balls are left: two green and one red. If these are still equally likely to be

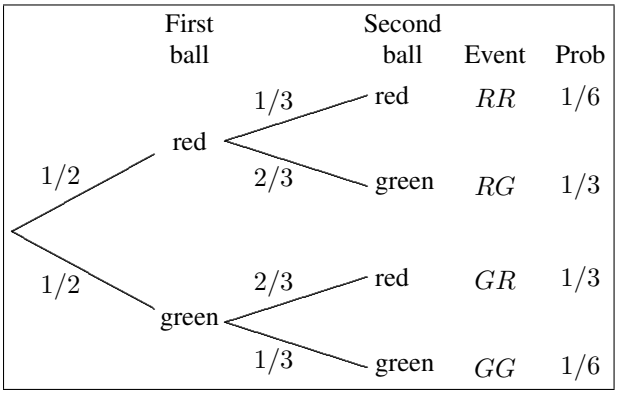


Figure 2: Tree showing the calculation of probabilities in urn model.

Table 2: Theoretical probabilities and observed relative frequencies of events in Kerrich's urn experiment.

Event	Prob.	Rel. freq.
RR	0.167	0.151
RG	0.333	0.338
GR	0.333	0.338
GG	0.167	0.173

chosen, then the probability of drawing a second red ball is $1/3$ and that of drawing a green ball is $2/3$. If the first ball was green then the same argument applies, except that now the red ball has probability $2/3$ and the green ball $1/3$. This accounts for the probabilities of the remaining branches.

We can now use the tree to calculate the probabilities of the events (RR, RG, GR, and GG) associated with the branch tips. The trick is to start at the root and trace a path through the tree, multiplying together the probabilities of the branches in the path. Take for example event RR. Its probability is $1/2$ (the probability that the first ball is red) times $1/3$ (probability that the second is red given that the first one was). The result, $1/6$, is shown in the right column of Fig. 2 along with the probabilities of the other three events.

These probabilities also appear in Table 2, where they are compared with the relative frequencies in Kerrich's data. I have re-expressed everything as a decimal fraction. For example, the theoretical probability of RR is $1/6 \approx 0.167$ and its observed relative frequency is $756/5000 \approx 0.151$. The relative frequencies do not quite equal the probabilities, but the differences are small. A careful statistical analysis would probably conclude that this model is consistent with the data. (We can't know without doing such an analysis, and that is beyond my scope here.)

1.4. Conditional and joint probability: the multiplication law

In the tree diagram, the probabilities involving the second ball deserve comment. They are called *conditional probabilities* because their values depend on (i.e. are conditioned by) the color of the first ball. The conditional probability of B given A is written $\Pr[B|A]$. In multiplying along paths within the tree, we were using something called the *law of multiplication of probabilities*:

$$\Pr[A \,\&\, B] = \Pr[A] \Pr[B|A] \tag{1}$$

We are interested in the probability of event "A & B." The relative frequency of this event is $n(A \ \& \ B)/N$, where $n(A \ \& \ B)$ is the number of trials on which this event occurred and N is the total number of trials. Multiply by $n(A)/n(A) = 1$ (which changes nothing) to obtain

$$\frac{n(A \ \& \ B)}{N} = \frac{n(A)}{N} \cdot \frac{n(A \ \& \ B)}{n(A)}$$

As the number of trials grows large, relative frequencies become closer and closer to the corresponding probabilities. Thus, the left side approaches $\Pr[A \ \& \ B]$, the "joint probability" of "A & B." Meanwhile, on the right $n(A)/N$ approaches $\Pr[A]$. But what are we to make of the remaining fraction, $n(A \ \& \ B)/n(A)$? It is the relative frequency of B among trials in which A occurred. As N grows large, this ratio converges to the *conditional probability of B given A*, or $\Pr(B|A)$. Thus, in the equation above all three ratios converge to probabilities as N grows large, and the equation itself converges to the multiplication law (Eqn. 1).

Box 1: Deriving the multiplication law

Here, "$\Pr[A \ \& \ B]$" is called the *joint probability* of events A and B; it is the probability that A and B both happened on the same trial. For example, suppose A is the event that the first ball is red and B the event that the second is green. Using the tree diagram, we calculated $\Pr[A \ \& \ B]$ by multiplying along the relevant path. In terms of the multiplication law, $\Pr[A] = 1/2$, $\Pr[B|A] = 2/3$, and $\Pr[A \ \& \ B] = 1/2 \times 2/3 = 1/3$.

The multiplication law is important because it explains *why* we multiply along paths within the tree. To me, this seems intuitive and obvious. But some things that are obvious are also wrong, so Box 1 explains why it works.

1.5. The addition law

We turn now to events of form "A or B" (or both). The probabilities of such events are especially easy when the two events are *mutually exclusive*, i.e. when they could not have happened on a single trial. Take for example the event that the second ball is green. This happens in either of two cases: RG and GG. The probability of this event is the probability of "RG or GG," which (according to the tree diagram) equals $1/3 + 1/6 = 1/2$. To get this answer, we simply summed the probabilities of RG and GG.

⋆ EXERCISE **1–3** In Kerrich's urn data, show that the relative frequency of "RG or GG" equals the sum of the frequencies of RG and GG.

The calculation is a little harder when the two events are *not* mutually exclusive. To see why, consider the event that either the first ball is red or the second is green (or both). This is also of form "A or B," but if we try summing $\Pr[$1st ball red$]$ and $\Pr[$2nd ball green$]$ we get $1/2 + 1/2 = 1$. This *can't* be right.

⋆ EXERCISE **1–4** In Kerrich's urn data, calculate the relative frequencies of events A, B, and "A or B." Show that the sum of the first two does not equal the third.

To see what went wrong, let us look under the hood. The first ball is red in either of two cases: RR and RG. Similarly, the second is green in cases RG and GG. Thus, our incorrect calculation can be expanded as:

$$\overbrace{\Pr[RR] + \Pr[RG]}^{\Pr[A]} + \overbrace{\Pr[RG] + \Pr[GG]}^{\Pr[B]}$$

We have (incorrectly) summed $\Pr[RG]$ twice. To fix this, we must subtract $\Pr[RG]$. This illustrates the *law of addition of probabilities*:

$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \ \& \ B] \qquad (2)$$

When the two events are mutually exclusive (as in the preceding example), the probability of "A & B" is zero, and there is no need to subtract it off.

⋆ EXERCISE **1–5** Use the addition law to calculate the probability that the first ball is red or the second green.

1.6. Statistical independence

If one event does not influence another, this fact should be reflected in their probabilities. Two events A and B are said to be *statistically independent* if $\Pr[B|A] = \Pr[B]$.

For an example in which this condition is not met, we return to Kerrich's urn experiment. The probabilities of red and green in the second ball depend on the color of the first ball. Thus, intuition suggests that the balls are not independent.

⋆ EXERCISE **1–6** Define A as the event that the first ball is is red and B as the event that the second is green. Use the tree diagram in Fig. 2 to calculate $\Pr[B]$ and $\Pr[B|A]$ and thus to decide whether the balls in the model are statistically independent.

⋆ EXERCISE **1–7** Use Kerrich's data to estimate the same probabilities.

	White						
Red	1	2	3	4	5	6	sum
1	547	587	500	462	621	690	3407
2	609	655	497	535	651	684	3631
3	514	540	468	438	587	629	3176
4	462	507	414	413	509	611	2916
5	551	562	499	506	658	672	3448
6	563	598	519	487	609	646	3422
sum:	3246	3449	2897	2841	3635	3932	20000

Table 3: The results of 20,000 throws with two dice (Wolf 1850, cited in [1])

Dice give us the other sort of example. In 1850, the astronomer Rudolf Wolf described the results of 20,000 throws of two dice, one red and one white. The results are shown in Table 3. We can use these data to ask whether the two dice were independent. Consider the event "red 4" of a 4 on the red die. The unconditional frequency of this event was $2916/20000 \approx 0.15$. Conditional on a 5 on the white die, the frequency of "red 4" was $509/3635 \approx 0.14$. These two numbers are nearly the same, as they ought to be if the red and white dice are independent.

⋆ EXERCISE **1–8** Use Wolf's data to estimate (a) the unconditional probability of "red 2," and (b) the conditional probability of "red 2" given "white 4." Use the results to comment on whether Wolf's dice were statistically independent.

If these dice were fair, each of the row and column sums in Table 3 should be close to $20000/6$, or 3333. Instead, both dice show an excess of 2s and 5s and a deficit of 3s and 4s. In addition, the white die shows an excess of 6s. Michael Bulmer [1] discusses several plausible causes: the dice may not be cubes, their corners may be rounded unevenly, or the process of cutting pips (dots) into their faces may have altered their centers of gravity. Whatever the explanation, these data show that the model of a "fair die" is only an approximation.

⋆ EXERCISE **1–9** Suppose that Kerrich had placed the first ball back into the box and then shaken it again before drawing the second. Draw a decision tree to represent this experiment and use it to calculate the probabilities of RR, RG, GR, and GG. Use the tree to show that the two balls are independent. (By the way, this new version of the experiment involves what is called *sampling with replacement*.)

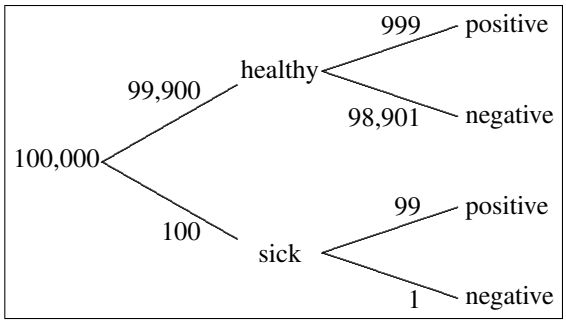


Figure 3: Bayes's rule example.

1.7. Bayes's rule

Suppose your doctor orders a test to see whether you have some hypothetical disease. Experiments have shown that the test was positive in 99% of patients who have the disease. But it was also positive in 1% of patients who *didn't*. These, of course, are the wrong numbers. You want the probability that *you have the disease* given that your test was positive. To figure this out, you or your doctor will need to use what is called Bayes's rule.

It is easier to understand the principles involved if we think in terms of counts rather than probabilities. Consider the data in Fig. 3. This tells us that in 100,000 people of your sex and age, 99,900 will be healthy and 100 will be sick, with some hypothetical disease. We get positive test results from nearly all (99/100, or 99%) of the sick individuals, but also from a small fraction (999/99,900 or 1%) of the healthy ones. Of those who test positive, the sick fraction is

$$\frac{99}{99+999} \approx 0.09$$

Fewer than 1/10 of those who test positive are really sick!

Let us now rephrase this result in terms of probabilities. According to the multiplication law (Eqn. 1)

$$\Pr[A\&B] = \Pr[B]\Pr[A|B] = \Pr[A]\Pr[B|A]$$

Divide through by $\Pr[B]$ to get

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]} \tag{3}$$

This is called *Bayes's Rule*. In the context of our example, B is the event that an individual got positive test, and A is the event that the individual is really sick.

To illustrate Bayes's Rule in action, let us revisit Fig. 3. We use the figure to calculate relative frequencies and then interpret these as probabilities. If A is the event that you are sick and B the event that your test was positive, then $\Pr[A] = 100/100,000 = 1/1000$, $\Pr[B] = (999 +$

$99)/100,000 = 1098/100,000$, and $\Pr[B|A] = 99/100$. Eqn. 3 gives

$$\begin{aligned}\Pr[A|B] &= \frac{1/1000 \times 99/100}{1098/100,000} \\ &= \frac{99}{1098} \approx 0.09\end{aligned}$$

This is the same answer we got above using counts.

Chapter 2. Random Variables and Expectations

It is time to introduce some vocabulary. For any experiment, the set of possible outcomes is called the *sample space*. For example, there were four possible outcomes (RR, RG, GR, and GG) in Kerrich's urn experiment. These constitute the sample space of that experiment. There is nothing new here except the term itself.

We will also need the idea of a *random variable*. A variable X is called a *random variable* if the values it takes are numbers and it takes each value with a specified probability. For example, we might define X as the result of one roll of a die. The sample space of X is $\{1, 2, 3, 4, 5, 6\}$. If the die is fair, it takes each of these values with probability $1/6$. Another random variable is $Y = X^2$. The sample space of Y is $\{1, 4, 9, 16, 25, 36\}$ and again it takes each value with probability $1/6$. X and Y are distinct random variables even though the underlying experiment—rolling a die—is the same in both cases.

⋆ EXERCISE **2–1** Consider the experiment of throwing two dice, one red and one white. What is the sample space? If the dice are fair, what are the associated probabilities? Do not enumerate the entire sample space; just describe it.

⋆ EXERCISE **2–2** What is the sample space of X, the number of heads in two spins of a fair coin? (Hint: the events HT and TH both yield the same number of heads.)

2.1. Averages and expectations

There are two ways to calculate an average, one using relative frequencies and the other using the method you learned in grade school. Take for example the following toy data set: $[0, 0, 0, 1, 1, 2, 2, 2]$. There are 8 numbers here, and their sum is also 8, so their average is 1. Let us now repeat this calculation using the relative frequencies of these data, which are shown in Table 4.

In "sigma notation," the average (or mean) is[1]

$$\text{mean} = \sum_x x f_x \tag{4}$$

[1] If you are unfamiliar with the "Σ" symbol, see appendix A.

Table 4: Frequency distribution of toy data. f_i is the relative frequency of value i.

Value	Relative frequency
0	$f_0 = 3/8$
1	$f_1 = 2/8$
2	$f_2 = 3/8$

where x is a sample value and f_x is the relative frequency of that value. Using the relative frequencies from Table 4, this gives $(0 \times 3/8) + (1 \times 2/8) + (2 \times 3/8) = 1$, just as we calculated using the grade-school method. If the data set is large, relative frequencies often make the problem easier.

⋆ EXERCISE **2–3** Calculate the mean of the numbers 1, 1, and 3 using both methods.

As the sample size grows large, the relative frequencies in Eqn. 4 get closer and closer to the corresponding probabilities. As this happens, the mean converges toward what is called the *expected value* of the corresponding random variable. The expected value of X is written "$E[X]$" and is calculated just as you calculate a mean:

$$E[X] = \sum_x x \Pr[X = x] \tag{5}$$

Note the similarity between Eqns. 4 and 5. Relative frequencies (f_x) have been replaced by probabilities $(\Pr[X = x])$; the two formulas are otherwise the same.

Example If you spin a coin twice, the number X of heads must equal either 0, 1, or 2. If the coin is fair, then these events have probabilities $1/4$, $1/2$, and $1/4$, respectively. The expectation of X is

$$E[X] = (0 \times 1/4) + (1 \times 1/2) + (2 \times 1/4) = 1$$

⋆ EXERCISE **2–4** What is the expected value of X^2?

⋆ EXERCISE **2–5** What is the expected value of $X + X^2$?

Expected values have several properties that make them easy to manipulate. If X and Y are random variables and a is a constant, then

$$E[a] = a \tag{6}$$
$$E[aX] = aE[X] \tag{7}$$
$$E[X + Y] = E[X] + E[Y] \tag{8}$$

If X and Y are statistically independent, it is also true that

$$E[XY] = E[X]E[Y] \tag{9}$$

These are really properties of averages. They apply to expectations because expectations are a kind of average. Rather than proving them, I will illustrate them using averages.

The average of a constant The average of 4, 4, and 4 is 4. This is why $E[a] = a$ when a is a constant.

The average of aX Start with the numbers 1, 3, and 5. The sum of these numbers is 9, and their average is 3. Now multiply each number by a constant a. The average of the resulting numbers is

$$(a + 3a + 5a)/3 = a \times (1 + 3 + 5)/3 = 3a$$

which is a times the original average. This illustrates that $E[aX] = aE[X]$.

The average of $X + Y$ Consider the following table

	X	Y	$X+Y$
	0	2	2
	2	3	5
	7	7	14
sum	9	12	21
average	3	4	7

The average of X is 3 and that of Y is 4. The sum of these is 7, which is also the average of $X + Y$. This illustrates that $E[X + Y] = E[X] + E[Y]$.

⋆ EXERCISE **2–6** What is $E[3]$?

⋆ EXERCISE **2–7** If $E[X] = 5$, then what is $E[2X]$?

⋆ EXERCISE **2–8** If $E[X] = 5$ and $E[Y] = 6$, then what is $E[2X + 3Y]$?

⋆ EXERCISE **2–9** What is $E[aX + bY^2]$, assuming that a and b are constant and the values of $E[X]$ and $E[Y^2]$ are unknown?

⋆ EXERCISE **2–10** Prove that $E[(X + Y)^2] = E[X^2] + 2E[X]E[Y] + E[Y^2]$ if X and Y are statistically independent. Hint: Begin by expanding $(X + Y)^2 = X^2 + 2XY + Y^2$. Then use Eqns. 8 and 9.

2.2. Variance

We are often interested in quantities that vary. There are several ways to measure variation, of which the most important is the *variance*. We can measure variance either in a data set or in a random variable. The procedures are similar so let's begin with data.

The variance is the *average squared difference from the mean*. Take for example the numbers 10, 12, 10, and 8. Their mean is 10, so their variance is $V = ((10 - 10)^2 + (12 - 10)^2 + (10 - 10)^2 + (8 - 10)^2)/4 = 2$. There are several ways to write the variance, including

$$\begin{aligned} V &= N^{-1} \sum_i (x_i - \bar{x})^2 && (10) \\ &= \sum_x (x - \bar{x})^2 f_x && (11) \\ &= \sum_x x^2 f_x - \bar{x}^2 && (12) \end{aligned}$$

Here, $\bar{x}$ is the average of the x_i and N^{-1} means $1/N$.

⋆ EXERCISE **2–11** Verify that the formulas Eqn. 10–12 are equivalent, using the numbers 10, 12, 10, and 8.

⋆ EXERCISE **2–12** What are the mean and variance of the numbers 3, 9, 15, and 8?

If X is a random variable (rather than data), its variance is

$$V[X] = E\left[(X - E[X])^2\right] \tag{13}$$

Note the similarity between this expression and Eqn. 11. The variance can also be written in either of the following ways:

$$\begin{aligned} V[X] &= E\big[X^2\big] - E[X]^2 && (14) \\ &= E\big[X(X - E[X])\big] && (15) \end{aligned}$$

⋆ EXERCISE **2–13** Suppose that the random variable X takes the values 0, 1, and 2 with probabilities 1/3, 1/2, and 1/6. What are the mean and variance of X?

⋆ EXERCISE **2–14** In a previous exercise, you verified that Eqns. 11–12 were equivalent, using as data the numbers 8, 10, 10, and 12. This illustrates that Eqns. 13 and 14 are equivalent, since they are the same formulas in a different notation. Now use the same method and data to verify the equivalence of Eqn. 15.

⋆ EXERCISE **2–15** Prove that if a is a constant and X a random variable, then $V[aX] = a^2V[X]$.

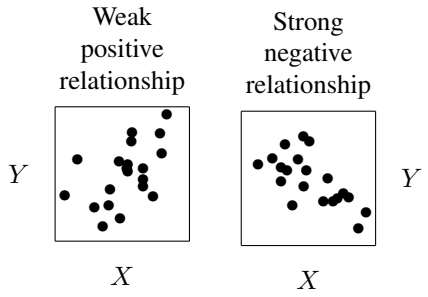


Figure 4: Examples of positive and negative relationship between variables

2.3. Covariance

In addition to variation, we are often interested in the *relationship* between variables. Fig. 4 illustrates this idea. The scatterplot on the left illustrates a positive relationship: one in which Y tends to increase when X increases. On the right we see the opposite case: Y decreases as X increases, so the relationship is negative. The two relationships differ not only in direction but also in strength. The one on the right is the stronger of the two.

These two ideas—strength and direction of relationships—come up all the time, and we need ways to measure them. Several statistics have been invented for this purpose, but most are based on the same idea: if the relationship is positive, the X and Y will often be on the same side of their respective means, so $(X - E[X])(Y - E[Y])$ is positive for most (X, Y) pairs. For a negative relationship, this product tends to be negative. We can measure a relationship by

$$C[X, Y] = E\big[(X - E[X])(Y - E[Y])\big] \qquad (16)$$

which is called the *covariance*. It is positive for positive relationships but negative for negative ones. It is far from zero for strong relationships but near zero for weak ones. Thus, it measures both the strength and direction of relationships.

The covariance, like the variance, can be written in several different ways:

$$\begin{aligned} C[X, Y] &= E[XY] - E[X]E[Y] \qquad (17) \\ &= E[X(Y - E[Y])] \qquad (18) \end{aligned}$$

In calculations, it is often most convenient to use (17).

To get familiar with covariances, consider the two random variables in Table 5. The probabilities imply that we should see lots of (X, Y) pairs like $(0, 0)$ or $(1, 1)$ but few like $(0, 1)$ or $(1, 0)$. For the most part then, X and Y will vary in the same direction, and their relationship should be positive.

Table 5: A bivariate probability distribution

X	Y	P_{XY}	$(X - E[X])(Y - E[Y])$
0	0	0.4	+0.25
0	1	0.1	–0.25
1	0	0.1	–0.25
1	1	0.4	+0.25

Here, P_{XY} is the probability of the pair (X, Y), and $E[X] = E[Y] = 0.5$.

The first step is to calculate $E[X]$ and $E[Y]$. I leave the details to you, but you should find that both equal 0.5. Next, calculate $(X - 0.5)(Y - 0.5)$ for each (X, Y) pair. These values appear in the right column of the table. Finally, we take the expectation by multiplying column 3 by column 4 and summing the results. (If this seems mysterious, consult Eqn. 5.) Here is the calculation in detail:

$$\begin{aligned} C[X, Y] &= \sum_{x,y} P_{xy}(x - 0.5)(y - 0.5) \\ &= (0.4 \times 0.25) - (0.1 \times 0.25) \\ &\quad - (0.1 \times 0.25) + (0.4 \times 0.25) \\ &= 0.15 \end{aligned}$$

The covariance is positive, just as expected.

⋆ EXERCISE **2–16** In Table 5, suppose that the P_{XY} values were 0.3, 0.2, 0.2, and 0.3. First use your intuition to figure out what this would do to the relationship between X and Y. Is it still positive? Does it become stronger or weaker? Then calculate $C[X, Y]$ to check your intuition.

⋆ EXERCISE **2–17** Construct a bivariate probability distribution in which the relationship between X and Y is negative, and use it to calculate $C[X, Y]$.

⋆ EXERCISE **2–18** So far we have been dealing with the covariance between random variables. To deal with data, we need formulas analogous to Eqns, 10–12. Write these formulas down and use them to estimate the covariance of X and Y in the following collection of (X, Y) values: (0,0), (0,0), (1,0), (0,1), (1,1), (1,1).

You may have been wondering what the magnitude of the covariance really means. If the covariance is 0.25, is the relationship a strong one or a weak one? It is impossible to say. The problem is that the magnitude of $C[X, Y]$ depends not only on the strength of the relationship but also on the units of measurement. For example, the value of $C[X, Y]$ would change if we decided to measure X and Y in millimeters rather than centimeters. To avoid such

effects, data analysts often normalize their covariances to obtain what is called a *correlation coefficient*. This topic is not covered here.

Chapter 3. Probability distributions

By now you are familiar with distributions of relative frequencies, and you know that as N grows large, relative frequencies converge to probabilities. It is thus easy to see that a frequency distribution will converge to a distribution of probabilities. Section 1.3 above described a model of Kerrich's urn experiment, which led to the probabilities in the right column of Fig. 2. These constitute a *probability distribution*.

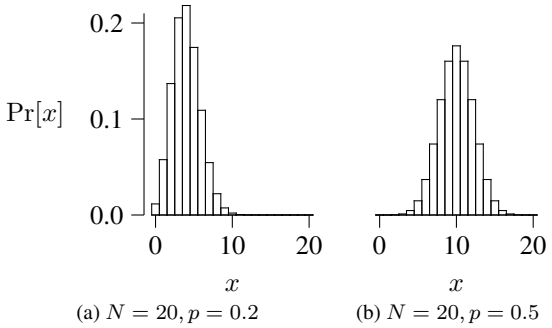
In that case, it was easy to list the events in the sample space and their probabilities. This is the simplest and most obvious way to describe a probability distribution. There is also another approach, which involves thinking of probability distributions as functions. A function, you may remember, is a translation rule. The function $f(x) = x^2$ for example would translate 2 into 4, or 3 into 9. Similarly, the probability distribution from the urn model in Fig. 2 translates the event RR into the probability $1/4$, and RG into $1/2$. It is thus a function too. Every probability distribution is a function that translates events into probabilities.

⋆ EXERCISE **3–1** What is the probability distribution of the number X of heads in two spins of a fair coin? (You wrote down the sample space in a previous exercise.)

In the case of random variables, we are translating numbers into numbers, and the function can often be expressed in mathematical form. This chapter will cover several probability functions that are widely used in science. These fall into two categories: *discrete* and *continuous*. The random variables that we have discussed thus far are discrete, but it will be easier to define continuous ones first.

A continuous random variable is one whose sample space is a continuum, such as space or time. The remarkable thing about a continuum is that between any two points there is an infinity of other points. In biology, continuous random variables are used to model such things as lifespan and body size. Section 3.2 will describe the methods to describe them.

If a sample space is not continuous, it is discrete. This does not make it finite. For example, there is an infinity of integers, but they are not continuous. There is no integer, for example, between 2 and 3. Discrete random variables are used to describe things that you can count: the number of heads in N spins of a coin, and so forth. The methods used to describe these random variables are described below in section 3.1.



(a) $N = 20, p = 0.2$ (b) $N = 20, p = 0.5$

Figure 5: Binomial distribution functions.

3.1. Discrete random variables

For discrete random variables, the *probability distribution function*, P_x, gives the probability associated with each point x in the sample space. In genetics and many other parts of biology, the discrete distributions used most often are the binomial and the Poisson.

3.1.1. The binomial distribution

Kerrich's coin experiment is a good example of a binomial experiment. There are N independent trials (in Kerrich's case, $N = 10,000$), and in each trial we observe some event with probability p. (In Kerrich's case, the event was "heads" and p was apparently close to $1/2$.) The number x of events is a binomial random variable. Its distribution function is

$$P_x = \binom{N}{x} p^x (1-p)^{N-x} \tag{19}$$

Here, $\binom{N}{x}$ is pronounced "N choose x" and represents the number of ways of obtaining x heads and $N - x$ tails. For example, there are are two ways (HT and TH) of obtaining one head in two spins, so $\binom{2}{1} = 2$. The expression $p^x(1-p)^{N-x}$ is the probability of obtaining any given sequence of x heads and $N - x$ tails. The binomial distribution is illustrated in Fig. 5 for $N = 20$ and two values of p. The mean and variance of the binomial distribution are $E[X] = Np$ and $V[X] = Np(1-p)$.

The form of the binomial distribution function is not hard to understand. To see why, consider the probability

of the following outcome:

$$\overbrace{\overbrace{H\ H\ \ldots H}^{x\text{ tosses}}\ \overbrace{T\ T\ \ldots T}^{N-x\text{ tosses}}}^{N\text{ tosses}}$$

Here, the coin has been tossed N times, producing heads on the first x tosses and tails on the remaining $N - x$. Since each heads is an event of probability p and each tails is an event of probability $1 - p$, the probability of the outcome observed on this sequence of tosses is $p^x(1 - p)^{N-x}$. But this is not the only outcome that would yield x heads in N tosses. No matter what order the heads and tails appear in, if there are x heads in N tosses we have observed an event of probability $p^x(1 - p)^{N-x}$. If we don't know the order in which the heads and tails appear, we have to sum across all the ways in which x heads and $N - x$ tails can be re-ordered. This sum accounts for the term $\binom{N}{x}$ in equation 19.

Example Population geneticists use the binomial distribution to model the random component of evolutionary change—genetic drift. Suppose that in the parental generation there are N diploid individuals. At each diploid locus, the population contains $2N$ genes, of which a fraction p are copies of allele A_1 and $1 - p$ are copies of A_2. The model assumes that each of the $2N$ genes in the offspring generation is (in effect) drawn at random with replacement from an urn with $2Np$ copies of allele A_1 and $2N(1 - p)$ copies of A_2. The number of copies of A_1 among the offspring is binomial with mean $2Np$ and variance $2Np(1 - p)$.

⋆ EXERCISE **3–2** Suppose that a population contains 1000 diploid individuals and that the relative frequency of A_1 is $1/1000$. If this population produces 1000 offspring, what is the probability (under the binomial model) that allele A_1 will not be represented among the offspring?

3.1.2. The Bernoulli distribution

The binomial distribution has an important special case: that in which we observe just one trial. For example, suppose we toss a coin a single time, and let $x = 1$ if the result is "heads" or 0 if the result is "tails." The distribution function has just two values:

$$P_x = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \tag{20}$$

This is just a special case of Eqn. 19.

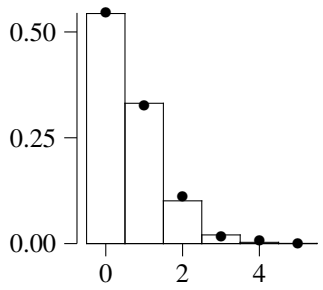


Figure 6: Frequency of deaths caused by mule kicks in the Prussian army. Bullets (•) show data, and bars show the Poisson distribution function.

⋆ EXERCISE **3–3** Prove Eqn. 20 by substituting $N = 1$ into Eqn. 19.

It is also easy to derive the mean and variance of the Bernoulli distribution, simply by setting $N = 1$ in the corresponding formulas for the Binomial: $E[X] = p$, and $V[X] = p(1 - p)$.

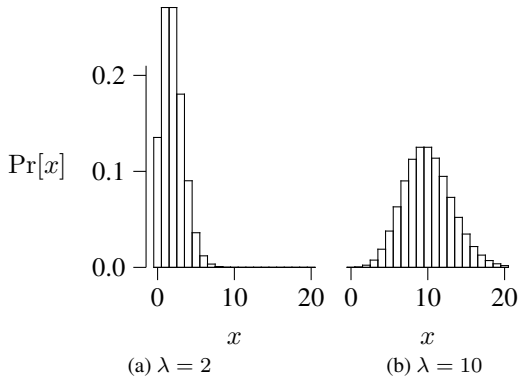
⋆ EXERCISE **3–4** Consider the experiment of tossing a fair coin a single time, and recording $X = 1$ if the result is "heads," or $X = 0$ if "tails." What are the mean and variance of this random variable?

⋆ EXERCISE **3–5** Consider the experiment of drawing a copy of a single gene at random from some population, and scoring $X = 1$ if the result is allele A_1, or $X = 0$ otherwise. If allele A_1 has frequency p, then what are the mean and variance of random variable X?

3.1.3. The Poisson distribution

This distribution comes up a lot when we are interested in counts. How many prey items will a forager encounter during one hour? How many gamma rays will strike the tube of a Geiger counter in one minute? How many rain drops will strike your roof during one second, in the middle of a storm? How many mutations occurred along the lineage that connects you to your mother's mother's mother's mother's... mother, who lived 10,000 years ago? In each case, if the events in question are independent and occur at a constant rate, then the random variable is Poisson.

At the end of the 19th century, Ladislaus von Bortkiewicz fit the Poisson to some peculiar data involving deaths in the Prussian Army. In those days, the army's supply train involved wagons pulled by mules. Mules



(a) $\lambda = 2$ (b) $\lambda = 10$

Figure 7: Poisson distribution functions.

have tempers and are seldom eager to pull wagons. Every now and then, a soldier was killed by the kick of a mule. Military records list the number of soldiers killed this way in each year within each army corps. These data, as tabulated by von Bortkiewicz, are shown as bullets (•) in Fig. 6. The histogram shows the corresponding Poisson distribution function.

The distribution has one parameter, the mean (λ). The probability that x events occur is

$$P_x = \frac{\lambda^x e^{-\lambda}}{x!} \tag{21}$$

Here, $e \approx 2.718$ is the base of natural logarithms; $x!$ is pronounced "x factorial" and represents the number of ways of rearranging x items. (See appendix B for details.) The variance of the Poisson is the same as the mean: $E[X] = V[X] = \lambda$.

The shape of the Poisson varies in response to the parameter λ, as shown in Fig. 7. When λ is small (as shown in the left-hand graph), the distribution function is asymmetric, with a high left shoulder. When λ is large (as shown in the right-hand graph), the function becomes symmetrical. For large λ, the Poisson is very nearly identical to the normal distribution (discussed below).

Because P_x depends in such a simple fashion on the mean, λ, the Poisson is among the easiest distributions to fit to data. For example, in von Bortkiewicz's data there are on average 0.61 mule-kick deaths per corps per year. To fit these data to the Poisson, we simply set $\lambda = 0.61$. That is all there is to it. With λ known we can calculate numerical probabilities. For example, the probability of a single death is $P_1 = \lambda e^{-\lambda} = 0.331$. In other words, we expect a death in any given corp 1 year in 3. There were 200 corps-years in von Bortkiewicz's data, so the number of these with a single mule-kick death should have been $0.331 \times 200 = 66.2$. There were 65 in the real data.

⋆ EXERCISE **3–6** Use the same procedure to calculate the expected number of corps-years with 2 mule-kick deaths. Compare this expected value to the real value, 22.

Example Consider the lineage that connects me to an ancestor who lived t generations ago. The expected number of mutations along that lineage is $\lambda = ut$, where u is the mutation rate. The probability that x mutations occurred is given by the Poisson distribution function. If $u = 10^{-3}$ and $t = 2000$ generations, then $\lambda = 2$. The probability that 1 mutation occurred is $\lambda e^{-\lambda} = 0.271$.

⋆ EXERCISE **3–7** What is the probability that no mutation occurred?

⋆ EXERCISE **3–8** What is the probability that at least one mutation occurred?

3.2. Continuous random variables

Board games often come with a device for generating random numbers. One type consists of a flat piece of cardboard to which a needle is attached. You spin the needle, and it ends up pointing in a random direction. In the real world, these devices probably have irregularities that make the needle more likely to land in some positions than others. But let's ignore that. In our hypothetical world, the needle is equally likely to point in any direction. What is the probability that it stops exactly 87.729543328 degrees clockwise of where it started?

This is a trick question. The problem is that there is a continuum of possible outcomes, all equally probable. The probability of any particular outcome, such as 87.729543328, is zero. Why? With an infinity of equally probable outcomes, the probability of each must be something like $1/\infty$.

It makes more sense to talk about the probability that the random variable will lie within some range of values. Let us define a function $f(x)$ such that[2]

$$\int_a^b f(x)dx$$

is the probability that the random variable lies between a and b. Here, $f(x)$ is called the *probability density function (pdf)*. Loosely speaking, $f(x)dx$ is the probability that the random variable lies within the small range from x to $x + dx$.

In biology, the most widely used continuous distributions are the uniform, the exponential, and the normal.

[2]The symbol "$\int$" is from calculus, which you will need from here on.

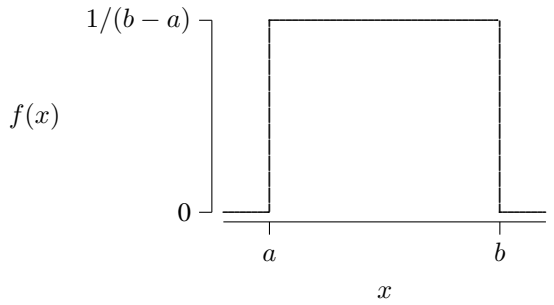


Figure 8: Uniform density function

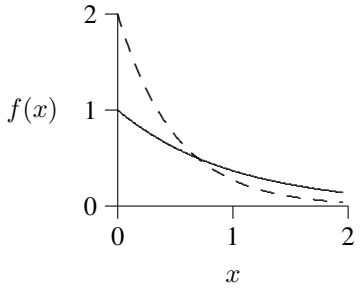


Figure 9: Exponential density functions with $h = 1$ (solid line) and $h = 2$ (dashed).

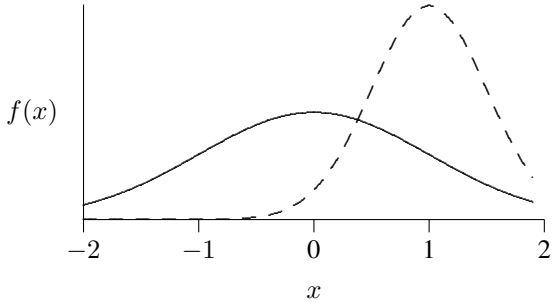


Figure 10: Normal density functions with $\mu = 0$, $\sigma = 1$ (solid line) and $\mu = 1$, $\sigma = 1/2$ (dashed).

3.2.1. The uniform distribution

A *uniform distribution* (which we encountered above on p. 2) is equally likely to take any value between two constants a and b but never takes values outside this range. Thus, its density function is very simple: $f(x) = 1/(b-a)$, as shown in Fig. 8. The mean and variance are $E[X] = (a+b)/2$ and $V[X] = (b-a)^2/12$. An important special case is the *standard uniform distribution*, for which $a = 0$, $b = 1$, and $f(x) = 1$.

⋆ EXERCISE **3–9** Make a graph of the standard uniform distribution function, and shade the area that corresponds to the range from 0.2 to 0.3. What is the area of this shaded region? What is the probability that a standard uniform r.v. will lie between these values?

⋆ EXERCISE **3–10** Solve the same problem using calculus.

3.2.2. The exponential distribution

We are often interested in the waiting time until some event. If these events happen at a constant rate (or hazard) h, then the waiting time has an exponential distribution. The density function is

$$f(x) = he^{-hx} \tag{22}$$

for values of x between 0 and ∞. As shown in Fig. 9, the density is greatest at $x = 0$ and declines smoothly with increasing x. The rate of decline increases with h. The mean of the exponential is $E[X] = 1/h$ and its variance is $V[x] = 1/h^2$. For this distribution, the standard deviation (the square root of the variance) is equal to the mean.

⋆ EXERCISE **3–11** What is the probability that $X < 1$?

⋆ EXERCISE **3–12** In Europe, the crude death rate (including individuals of all ages) is close to 0.01 deaths per individual per year. If this rate were constant throughout life, what would be the average lifespan?

⋆ EXERCISE **3–13** The mutation rate in autosomal DNA is roughly 10^{-9} per nucleotide site per year. If we follow a single copy of a single nucleotide forward across the generations, how long must we wait on average until it mutates? What is the variance of this number?

3.2.3. The normal distribution

The density function of the normal distribution is the familiar bell-shaped curve, two examples of which are shown in Fig. 10. The normal distribution has two parameters, the mean μ and the variance σ^2. As Fig. 10 illustrates, μ controls the location of the center (peak) of the distribution and σ controls its width. The density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{23}$$

The normal distribution is widely used in statistics. There are several reasons, but chief among them is this: any variable that is the sum of many other random variables tends to look normal. The larger the number of variables, the more normal their sum will look. We have already passed over two special cases of this. A binomial r.v. is a sum of N smaller r.v.s, one for each toss of the coin. If you look closely at the binomial distributions in

Fig. 5, you will see that they closely resemble the normal. The Poisson is also approximately normal if λ is large, as you can see from panel b of Fig. 7. To understand why, recall that the Poisson describes the number of events that occur in a fixed interval of time. But we can think of this as the sum of the numbers of events that occur in a series of sub-intervals. Thus, the Poisson is also a sum.

In addition to these technical concerns, there is a very practical reason for wide interest in the normal distribution: many of the variables studied in biology, agriculture, and medicine seem to be approximately normal. Why should the same pattern appear so often and in so many different contexts? The answer to this question is very likely the same business about sums that we discussed just above. Many of the variables we study are affected by a multitude of causes. Many genetic loci, for example, contribute to variation in human stature. To the extent that these loci act additively, stature is a sum. Stature, of course, is affected by environmental causes as well as genetic ones. To the extent that these environmental factors act additively, they contribute to the sum. In this sense, many of the variables we study are sums of a sort, and it makes sense that their distributions should look normal.

Appendix A. Sums and sigma notation

You learned in grade school to calculate sums such as $10+12+10+8=40$. To generalize this calculation, suppose we have 4 arbitrary numbers, x_1, x_2, x_3, and x_4. Their sum is

$$x_1 + x_2 + x_3 + x_4$$

This sum can also be written using "sigma notation" as

$$\sum_{i=1}^{4} x_i$$

The "Σ" symbol is a Greek sigma and indicates summation. The subscript "$i = 1$" indicates that the sum begins with x_1, and the superscript "4" indicates that the sum ends with x_4.

More generally, if the number of numbers is an unknown value, N, then their sum is

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots + x_N$$

Sometimes sums are written without limits, as in

$$\sum_{i} x_i.$$

This means the sum over all terms, however many there may be. When sums are written within the text of a paragraph, the limits look like subscripts and superscripts, as in $\sum_{i=1}^{N} x_i$.

Appendix B. Factorials and binomial coefficients

The *factorial* of x is written $x!$ and equals

$$x! = x \cdot (x-1) \cdot (x-1) \cdots 1$$

It is pronounced "x factorial." For example, $3! = 3 \cdot 2 \cdot 1 = 6$. As a special case, $0!$ is defined to equal 1. Factorials arise in problems that involve rearrangements of the items in a list. For example, the letters "ABC" can be arranged in six different orders: (1) ABC, (2) ACB, (3) BAC, (4) BCA, (5) CAB, and (6) CBA. These rearrangements are called *permutations*. More generally, suppose we have a string of x letters. How many permutations does it have? There are x ways to choose the first letter. Having chosen the first, there are then $x-1$ ways to choose the second, $x-2$ ways to choose the third, and so on. There is only one way to choose the last, for by then all the other letters have been chosen. Thus, the number of permutations of x items is $x!$.

A *binomial coefficient* is written $\binom{N}{x}$ and pronounced "N choose x." It equals

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}$$

and can be interpreted as the number of ways of choosing x items out of a list of N. For example, consider the number of pairs of letters in the string ABC. According to the formula, there should be $\binom{3}{2} = 3!/(2! \cdot 1!) = 3$ pairs. We get the same answer by listing the pairs: AB, AC, and BC.

Appendix C. Answers to Exercises

⋆ EXERCISE **1–1** There are many correct answers. Here are two: (1) Interpret numbers as heads if less than 0.3 but as tails if greater. (2) Interpret numbers as heads if between 0.2 and 0.5 but as tails otherwise.

⋆ EXERCISE **1–2**

```
# Python program that simulates 10000
# spins of a fair coin

from random import random
```

```
for i in xrange(10000):
    u = random()
    if u < 0.5:
        print 'heads'
    else:
        print 'tails'
```

⋆EXERCISE **1–3** The relative frequency of "RG or GG" is $2556/5000$, that of RG is $1689/5000$, and that of GG is $867/5000$. The sum of the last two is $1689/5000 + 867/5000 = 2556/5000$.

⋆EXERCISE **1–4** The relative frequency of the event (A) that the first ball is red equals $2445/5000$; that of the event (B) that the second is green equals $2556/5000$; that of event "A or B" is $(756 + 1689 + 867)/5000 = 3312/5000$. The sum of the first two relative frequencies is $2445/5000 + 2556/5000 = 5001/5000$, which is much larger than the relative frequency of "A or B."

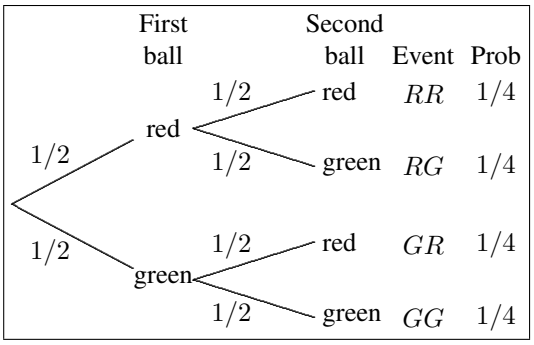
⋆EXERCISE **1–5** The first ball is red with probability $\Pr[A] = 1/2$, and the second is green with probability $\Pr[B] = 1/2$. The probability that both events happened is $\Pr[A \ \& \ B] = 1/3$. All this is from Fig. 2. Using these values, the addition law gives $\Pr[A \text{ or } B] = 1/2 + 1/2 - 1/3 = 2/3$.

⋆EXERCISE **1–6** As shown on p. 4, the tree diagram implies that the unconditional probability of B is $\Pr[B] = 1/2$. On the other hand, the conditional probability is $\Pr[B|A] = 2/3$. These probabilities are not equal, so A and B are not independent.

⋆EXERCISE **1–7** According to Table 1, $\Pr[B]$ is estimated by $2556/5000 \approx 0.51$, and $\Pr[B|A]$ by $1689/2445 \approx 0.69$. The two numbers are only estimates, so we cannot conclude that the probabilities differ merely because the estimates do. However, the difference between these estimates is large, and so is the sample. Even without a careful statistical analysis, these results suggest that the two balls were not statistically independent.

⋆EXERCISE **1–8** The relative frequency of "red 2" is $3631/20,000 \approx 0.18$. This estimates the unconditional probability of rolling "2" with the red. If we restrict attention to those trials on which the white die rolled "4," the relative frequency of "red 2" is $535/2841 \approx 0.19$. This estimates the conditional probability of "red 2" given "white 4." The numbers are pretty nearly equal, as they should be if the dice are statistically independent.

⋆EXERCISE **1–9** Under sampling with replacement, the decision tree is as shown below.



Using this decision tree, we can test for statistical independence as follows: the second ball is green with unconditional probability $\Pr[RG \text{ or } GG] = 1/2$. If the first ball is red, the second is green with conditional probability $1/2$. These numbers are the same, so the two balls are statistically independent.

⋆EXERCISE **2–1** The sample space consists of all possible (X, Y) pairs, where X is the number on the red die and Y the number on the white. X and Y are both integers between 1 and 6, so there are 36 possible outcomes. If the dice are fair, then each outcome has probability $1/36$.

⋆EXERCISE **2–2** The sample space is $\{0, 1, 2\}$.

⋆EXERCISE **2–3** The mean is $5/3$, since there are 3 numbers that sum to 5. Using relative frequencies, the problem becomes $1 \times (2/3) + 3 \times (1/3) = 2/3 + 3/3 = 5/3$.

⋆EXERCISE **2–4**

$$E[X^2] = (0 \times 1/4) + (1 \times 1/2) + (4 \times 1/4) = 1.5$$

⋆EXERCISE **2–5**

$$\begin{aligned} E[X + X^2] &= ((0+0) \times 1/4) + ((1+1) \times 1/2) \\ &\quad + ((2+4) \times 1/4) \\ &= 2.5 \end{aligned}$$

Compare this answer to that of the preceding exercise, and you will see that $E[X + X^2] = E[X] + E[X^2]$.

⋆EXERCISE **2–6** $E[3] = 3$

⋆EXERCISE **2–7** $E[2X] = 10$

⋆EXERCISE **2–8** $E[2X + 3Y] = 10 + 18 = 28$

⋆EXERCISE **2–9** $E[aX + bY^2] = aE[X] + bE[Y^2]$

⋆EXERCISE **2–10** First expand the squared term:

$$E[(X + Y)^2] = E[X^2 + 2XY + Y^2]$$

Next, use Eqn. 8 to turn the expectation of a sum into a sum of expectations:

$$E[X^2 + 2XY + Y^2] = E[X^2] + E[2XY] + E[Y^2]$$

Finally, re-express the middle term using Eqns. 7 and 9:

$$E[X^2] + E[2XY] + E[Y^2] = E[X^2] + 2E[X]E[Y] + E[Y^2]$$

⋆EXERCISE **2–11** The text used Eqn 10 to calculate that $m = 10$ and $V = 2$. For the other versions, we need relative frequencies: $f_8 = 1/4$, $f_{10} = 1/2$, and $f_{12} = 1/4$. Eqn. 11 gives

$$\begin{aligned} V &= (1/4)(8-10)^2 + (1/2)(10-10)^2 \\ &\quad + (1/4)(12-10)^2 \\ &= 1 + 0 + 1 = 2 \end{aligned}$$

For the Eqn. 12, we need

$$\begin{aligned} \sum x^2 f_x &= 8^2/4 + 10^2/2 + 12^2/4 \\ &= 16 + 50 + 36 = 102 \end{aligned}$$

We also need $m^2 = 10^2 = 100$. Subtracting gives $V = 102 - 100 = 2$.

⋆EXERCISE **2–12** The mean and variance are 8.75 and 18.1875.

⋆EXERCISE **2–13** The mean is

$$E[X] = 0 \times 1/3 + 1 \times 1/2 + 2 \times 1/6 = 0.833.$$

The variance is

$$\begin{aligned} V[X] &= (0-0.833)^2 \times 1/3 \\ &\quad + (1-0.833)^2 \times 1/2 \\ &\quad + (2-0.833)^2 \times 1/6 \\ &= 0.472. \end{aligned}$$

⋆EXERCISE **2–14** In the data, the relative frequencies are $f_8 = 1/4$, $f_{10} = 2/4$, and $f_{12} = 1/4$. The mean is 10. According to Eqn. 15, the variance is

$$\begin{aligned} V &= (1/4) \times 8 \times (8-10) \\ &\quad + (1/2) \times 10 \times (10-10) \\ &\quad + (1/4) \times 12 \times (12-10) \\ &= (1/4) \times (-16) \\ &\quad + (1/2) \times 0 \\ &\quad + (1/4) \times 24 \\ &= -4 + 6 = 2 \end{aligned}$$

⋆EXERCISE **2–15** First, $E[aX] = aE[X]$ by equation 7. Next,

$$\begin{aligned} V[aX] &= E[(aX - aE[x])^2] \\ &= E[a^2(X - E[x])^2] \\ &= a^2 E[(X - E[x])^2] \qquad \text{using (7) again} \\ &= a^2 V[X] \end{aligned}$$

⋆EXERCISE **2–16** With the new probabilities, we expect fewer $(0,0)$ and $(1,1)$ pairs but more $(0,1)$ and $(1,0)$. There is still a tendency for X and Y to vary in the same direction, but that tendency is weaker. In other words, the relationship is weaker but still positive. The covariance turns out to be $C[X,Y] = 0.05$.

⋆EXERCISE **2–17** One way is to replace the P_{XY} values in Table 5 with 0.1, 0.4, 0.4, and 0.1. This yields $C[X,Y] = -0.15$.

⋆EXERCISE **2–18** The corresponding formulas are

$$\begin{aligned} C[X,Y] &= N^{-1} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \qquad (24) \\ &= \sum_{x,y} (x - \bar{x})(y - \bar{y}) f_{xy} \qquad (25) \\ &= \sum_{x,y} xy f_{xy} - \bar{x}\bar{y} \qquad (26) \end{aligned}$$

where f_{xy} is the frequency of (X,Y) pairs in the data for which $X = x$ and $Y = y$. Using any of these formulas, the given data imply that $C[X,Y] = 0.0833$.

⋆EXERCISE **3–1** The sample space is $\{0,1,2\}$; the corresponding probabilities are $1/4$, $1/2$, and $1/4$.

⋆EXERCISE **3–2** The allele disappears with probability

$$P_0 = \binom{2000}{0} p^0 (1-p)^{2000} \approx 0.135$$

In this calculation, there is no need to calculate $\binom{2000}{0}$; it must equal 1 because there is only 1 way to choose 0 of something. In addition, $p^0 = 1$ because *anything* raised to the zeroth power equals 1. The only part that needs calculating is $(1-p)^{2000}$.

⋆EXERCISE **3–3** For the Bernoulli distribution, there is only one event. Consequently, $N = 1$ and Eqn. 19 becomes

$$P_x = \frac{1!}{1! \times 0!} p^x (1-p)^{1-x}$$

Recall that $0! = 1! = 1$. (See appendix B for details.) Consequently, $1!/(1! \times 0!) = 1$ and drops out of the equation. If $x = 0$, then $p^x = p^0 = 1$, and this term drops out. Eqn. 19 becomes $P_0 = 1 - p$. On the other hand, if $x = 1$, then $(1-p)^{1-x} = (1-p)^0 = 1$ and drops out. We are then left with $P_1 = p$.

⋆EXERCISE **3–4** $E[X] = 1/2$ and $V[X] = 1/4$.

⋆EXERCISE **3–5** Since A_1 has frequency p, that is also the probability that we have drawn a copy of this allele. It follows that $E[X] = p$ and $V[X] = p(1-p)$.

⋆EXERCISE **3–6** For $x = 2$ and $\lambda = 0.61$, the Poisson formula gives $P_2 = 0.101$. We therefore expect to see about $0.101 \times 200 = 20.2$ corps-years with 2 mule-kick deaths. This is close to the real number of 22.

⋆ EXERCISE **3–7** The probability of no mutation is $P_0 = e^{-2} \approx 0.135$. Note that this answer is identical to that of the preceding exercise. This illustrates the fact that, when N is large and p is small, the Poisson is a good approximation to the binomial.

⋆ EXERCISE **3–8** The probability of "at least one mutation" is $\Pr[X > 0]$. There are two ways to think about this problem. The hard way sums across all non-zero entries of the Poisson distribution:

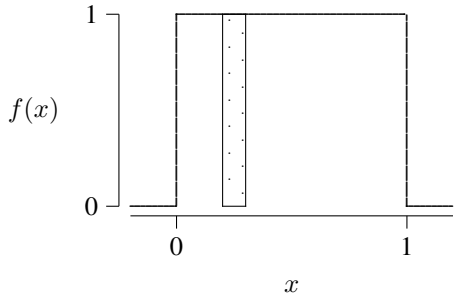
$$\Pr[X > 0] = \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!}$$

As I said, that is the hard way and is not recommended. The easy solution proceeds from the observation that all probability distributions sum to 1. For a non-negative random variable such as the Poisson, this implies that $\Pr[X = 0] + \Pr[X > 0] = 1$. Thus,

$$\Pr[X > 0] = 1 - \Pr[X = 0] = 1 - e^{-\lambda}$$

In the current question, $\lambda = 2$, so $\Pr[X > 0] \approx 0.865$.

⋆ EXERCISE **3–9** The graph is:



The area of the shaded rectangle is $(0.3 - 0.2) \times 1 = 0.1$. This is also the probability that $0.2 < X < 0.3$.

⋆ EXERCISE **3–10** For any continuous r.v., the probability that X lies between two values a and b is $\int_a^b f(x)dx$. In this problem, $a = 0.2$, $b = 0.3$, and $f(x) = 1$. The integral is thus equal to 0.1.

⋆ EXERCISE **3–11** An exponential variable is < 1 with probability $\int_0^1 he^{-hx}dx = 1 - e^{-h}$.

⋆ EXERCISE **3–12** 100 years, because the mean is $1/h$, and the problem says that $h = 0.01$.

⋆ EXERCISE **3–13** The mean is 10^9 years; the variance is 10^{18}.

References

[1] Michael G. Bulmer. *Principles of Statistics*. Dover, New York, 1967.

[2] John E. Kerrich. *An Experimental Introduction to the Theory of Probability*. Munksgaard, Copenhagen, 1946.