Contents

CHAPTER 1

Notation and conventions

A dataset as a collection of d-tuples (a d-tuple is an ordered list of d elements). Tuples differ from vectors, because we can always add and subtract vectors, but we cannot necessarily add or subtract tuples. There are always N items in any dataset. There are always d elements in each tuple in a dataset. The number of elements will be the same for every tuple in any given tuple. Sometimes we may not know the value of some elements in some tuples.

We use the same notation for a tuple and for a vector. Most of our data will be vectors. We write a vector in bold, so $\mathbf{x}$ could represent a vector or a tuple (the context will make it obvious which is intended).

The entire data set is $\{\mathbf{x}\}$. When we need to refer to the i'th data item, we write $\mathbf{x}_i$. Assume we have N data items, and we wish to make a new dataset out of them; we write the dataset made out of these items as $\{\mathbf{x}_i\}$ (the i is to suggest you are taking a set of items and making a dataset out of them). If we need to refer to the j'th component of a vector $\mathbf{x}_i$, we will write $x_i^{(j)}$ (notice this isn't in bold, because it is a component not a vector, and the j is in parentheses because it isn't a power). Vectors are always column vectors.

Terms:

- mean $(\{x\})$ is the mean of the dataset $\{x\}$ (definition **??**, page **??**).
- std (x) is the standard deviation of the dataset $\{x\}$ (definition **??**, page **??**).
- var $(\{x\})$ is the standard deviation of the dataset $\{x\}$ (definition **??**, page **??**).
- median $(\{x\})$ is the standard deviation of the dataset $\{x\}$ (definition **??**, page **??**).
- percentile$(\{x\}, k)$ is the $k\%$ percentile of the dataset $\{x\}$ (definition **??**, page **??**).
- iqr$\{x\}$ is the interquartile range of the dataset $\{x\}$ (definition **??**, page **??**).
- $\{\hat{x}\}$ is the dataset $\{x\}$, transformed to standard coordinates (definition **??**, page **??**).
- Standard normal data is defined in definition **??**, page **??**).
- Normal data is defined in definition **??**, page **??**).
- corr $(\{(x, y)\})$ is the correlation between two components x and y of a dataset (definition **??**, page **??**).
- $\emptyset$ is the empty set.
- Ω is the set of all possible outcomes of an experiment.

- Sets are written as $\mathcal{A}$.
- $\mathcal{A}^c$ is the complement of the set $\mathcal{A}$ (i.e. $\Omega - \mathcal{A}$).
- $\mathcal{E}$ is an event (page 121).
- $P(\{\mathcal{E}\})$ is the probability of event $\mathcal{E}$ (page 121).
- $P(\{\mathcal{E}\}|\{\mathcal{F}\})$ is the probability of event $\mathcal{E}$, conditioned on event $\mathcal{F}$ (page 121).
- $p(x)$ is the probability that random variable X will take the value x; also written $P(\{X = x\})$ (page 121).
- $p(x, y)$ is the probability that random variable X will take the value x and random variable Y will take the value y; also written $P(\{X = x\} \cap \{Y = y\})$ (page 121).

Background information:

- *Cards:* A standard deck of playing cards contains 52 cards. These cards are divided into four suits. The suits are: spades and clubs (which are black); and hearts and diamonds (which are red). Each suit contains 13 cards: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack (sometimes called Knave), Queen and King. It is common to call Jack, Queen and King *court cards*.
- *Dice:* If you look hard enough, you can obtain dice with many different numbers of sides (though I've never seen a three sided die). We adopt the convention that the sides of an N sided die are labeled with the numbers $1 \ldots N$, and that no number is used twice. Most dice are like this.
- *Fairness:* Each face of a fair coin or die has the same probability of landing upmost in a flip or roll.

1.1 ACKNOWLEDGEMENTS

Typos spotted by: Han Chen,

CHAPTER 2

Basic ideas in probability

We will perform experiments — which could be pretty much anything, from flipping a coin, to eating too much saturated fat, to smoking, to crossing the road without looking — and reason about the outcomes (mostly bad, for the examples I gave). But these outcomes are uncertain, and we need to weigh those uncertainties against one another. If I flip a coin, I could get heads or tails, and there's no reason to expect to see one more often than the other. If I eat too much saturated fat or smoke, I will very likely have problems, though I might not. If I cross the road without looking, I may be squashed by a truck or I may not. Our methods need also to account for information. If I look before I cross the road, I am much less likely to be squashed.

Probability is the machinery we use to predict what will happen in an experiment. Probability measures the tendency of events to occur frequently or seldom when we repeat experiments. Building this machinery involves a formal model of potential outcomes and sets of outcomes. Once we have this, a set of quite simple axioms allows us, at least in principle, to compute probabilities in most situations.

2.1 EXPERIMENTS, OUTCOMES, EVENTS, AND PROBABILITY

If we flip a fair coin many times, we expect it to come up heads about as often as it comes up tails. If we toss a fair die many times, we expect each number to come up about the same number of times. We are performing an experiment each time we flip the coin, and each time we toss the die. We can formalize this experiment by describing the set of **outcomes** that we expect from the experiment. Every run of the experiment produces one of the set of possible outcomes. There are never two or more outcomes, and there is never no outcome.

In the case of the coin, the set of outcomes is:

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

In the case of the die, the set of outcomes is:

$$\{1, 2, 3, 4, 5, 6\}.$$

We are making a modelling choice by specifying the outcomes of the experiment. For example, we are assuming that the coin can only come up heads or tails (but doesn't stand on its edge; or fall between the floorboards; or land behind the bookcase; or whatever). We write the set of all outcomes Ω; this is usually known as the **sample space**.

Worked example 2.1 *Find the lady*

We have three playing cards. One is a queen; one is a king, and one is a knave. All are shown face down, and one is chosen at random and turned up. What is the set of outcomes?

Solution: Write Q for queen, K for king, N for knave; the outcomes are $\{Q, K, N\}$

Worked example 2.2 *Find the lady, twice*

We play Find the Lady twice, replacing the card we have chosen. What is the sample space?

Solution: We now have $\{QQ, QK, QN, KQ, KK, KN, NQ, NK, NN\}$

Worked example 2.3 *Children*

A couple decides to have children until either (a) they have both a boy and a girl or (b) they have three children. What is the set of outcomes?

Solution: Write B for boy, G for girl, and write them in birth order; we have $\{BG, GB, BBG, BBB, GGB, GGG\}$.

Worked example 2.4 *Monty Hall (sigh!)*

There are three boxes. There is a goat, a second goat, and a car. These are placed into the boxes at random. The goats are indistinguishable for our purposes; equivalently, we do not care about the difference between goats. What is the sample space?

Solution: Write G for goat, C for car. Then we have $\{CGG, GCG, GGC\}$.

Worked example 2.5 *Monty Hall, different goats (sigh!)*

There are three boxes. There is a goat, a second goat, and a car. These are placed into the boxes at random. One goat is male, the other female, and the distinction is important. What is the sample space?

Solution: Write M for male goat, F for female goat, C for car. Then we have $\{CFM, CMF, FCM, MCF, FMC, MFC\}$. Notice how the number of outcomes has increased, because we now care about the distinction between goats.

Worked example 2.6 *A poor choice of strategy for planning a family*

A couple decides to have children. As they know no mathematics, they decide to have children until a girl then a boy are born. What is the sample space? Does this strategy bound the number of children they could be planning to have?

Solution: Write B for boy, G for girl. The sample space looks like any string of B's and G's that (a) ends in GB and (b) does not contain any other GB. In regular expression notation, you can write such strings as B^*G^+B. There is a lower bound (two), but no upper bound. As a family planning strategy, this is unrealistic, but it serves to illustrate the point that sample spaces don't have to be finite to be tractable.

2.1.1 The Probability of an Outcome

We represent our model of how often a particular outcome will occur in a repeated experiment with a **probability**, a non-negative number. This number gives the relative frequency of the outcome of interest, when an experiment is repeated a very large number of times.

Assume an outcome A has probability P. Assume we repeat the experiment a very large number of times N, and assume that the coins, dice, whatever don't communicate with one another from experiment to experiment (or, equivalently, that experiments don't "know" about one another). Then, for about $N \times P$ of those experiments the outcome will occur. Furthermore, as N gets larger, the fraction where the outcome occurs will get closer to P. We write $\#(A)$ for the number of times outcome A occurs. We interpret P as

$$\lim_{N\to\infty} \frac{\#(A)}{N}.$$

We can draw two important conclusions immediately.

- For any outcome A, $0 \leq P(A) \leq 1$.
- $\sum_{A_i \in \Omega} P(A_i) = 1$.

Remember that every run of the experiment produces exactly one outcome. The probabilities add up to one because each experiment must have one of the outcomes in the sample space.

Worked example 2.7 *A biased coin*

Assume we have a coin where the probability of getting heads is $P(H) = \frac{1}{3}$, and so the probability of getting tails is $P(T) = \frac{2}{3}$. We flip this coin three million times. How many times do we see heads?

Solution: $P(H) = \frac{1}{3}$, so we expect this coin will come up heads in $\frac{1}{3}$ of experiments. This means that we will very likely see very close to a million heads.

Some problems can be handled by building a set of outcomes and reasoning about the probability of each outcome. This is particularly useful when the outcomes *must* have the same probability, which happens rather a lot.

Worked example 2.8 *Find the Lady*

Assume that the card that is chosen is chosen fairly — that is, each card is chosen with the same probability. What is the probability of turning up a Queen?

Solution: There are three outcomes, and each is chosen with the same probability, so the probability is 1/3.

Worked example 2.9 *Monty Hall, indistinguishable goats, again*

Each outcome has the same probability. We choose to open the first box. With what probability will we find a goat (any goat)?

Solution: There are three outcomes, each has the same probability, and two give a goat, so 2/3

Worked example 2.10 *Monty Hall, yet again*

Each outcome has the same probability. We choose to open the first box. With what probability will we find the car?

Solution: There are three places the car could be, each has the same probability, so 1/3

Worked example 2.11 *Monty Hall, with distinct goats, again*

Each outcome has the same probability. We choose to open the first box. With what probability will we find a female goat?

Solution: Using the reasoning of the previous example, but substituting "female goat" for "car", 1/3. The point of this example is that the sample space matters. If you care about the gender of the goat, then it's important to keep track of it; if you don't, it's a good idea to omit it from the sample space.

2.1.2 Events

Assume we run an experiment and get an outcome. We know what the outcome is (that's the whole point of a sample space). This means we can tell whether the outcome we get belongs to some particular known *set* of outcomes. We just look in the set and see if our outcome is there. This means that we should be able to predict the probability of a *set* of outcomes from any reasonable model of an experiment. For example, we might roll a die and ask what the probability of getting an even

number is. As another example, we might flip a coin ten times, and ask what the probability of getting three heads is.

An **event** is a set of outcomes. The set of all outcomes, which we wrote Ω, must be an event. It is not a particularly interesting event, because we must have $P(\Omega) = 1$ (because we said that every run of an experiment produces one outcome, and that outcome must be in Ω). In principle, there could be no outcome, although this never happens. This means that the empty set, which we write $\emptyset$, is an event, and we have $P(\emptyset) = 0$. The space of events has a rich structure, which makes it possible to compute probabilities for other, more interesting, events.

Notation: Generally, we write sets like $\mathcal{A}$; in principle, you could confuse this notation with the matrix notation, but it's clear from context which is meant. We write $\mathcal{A} \cup \mathcal{B}$ for the union of two sets, $\mathcal{A} \cap \mathcal{B}$ for the intersection of two sets, and $\mathcal{A} - \mathcal{B}$ for the set theoretic difference (i.e. $\mathcal{A} - \mathcal{B} = \{x \in \mathcal{A} | x \notin \mathcal{B}\}$). We will write $\Omega - \mathcal{U}$ as $\mathcal{U}^c$; read "the complement of $\mathcal{U}$".

Events have three important properties that follow from their nature as sets of outcomes:

- If $\mathcal{U}$ and $\mathcal{V}$ are events — sets of outcomes — then so is $\mathcal{U} \cap \mathcal{V}$. You should interpret this as the event that we have an outcome that is in $\mathcal{U}$ and also in $\mathcal{V}$.
- If $\mathcal{U}$ and $\mathcal{V}$ are events, then $\mathcal{U} \cup \mathcal{V}$ is also an event. You should interpret this as the event that we have an outcome that is either in $\mathcal{U}$ or in $\mathcal{V}$ (or in both).
- If $\mathcal{U}$ is an event, then $\mathcal{U}^c = \Omega - \mathcal{U}$ is also an event. You should think of this as the event we get an outcome that is not in $\mathcal{U}$.

This means that the set of all possible events Σ has a very important structure.

- $\emptyset$ is in Σ.
- Ω is in Σ.
- If $\mathcal{U} \in \Sigma$ and $\mathcal{V} \in \Sigma$ then $\mathcal{U} \cup \mathcal{V} \in \Sigma$.
- If $\mathcal{U} \in \Sigma$ and $\mathcal{V} \in \Sigma$ then $\mathcal{U} \cap \mathcal{V} \in \Sigma$.
- If $\mathcal{U} \in \Sigma$ then $\mathcal{U}^c \in \Sigma$.

This means that the space of events can be quite big. For a single flip of a coin, the only possible space of events looks like:

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

Many experiments admit more than one space of events. For example, if we flip two coins, one natural event space is

$$\left\{ \begin{array}{l} \emptyset, \Omega, \\ \{HH\}, \{HT\}, \{TH\}, \{TT\}, \\ \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TT\}, \{HT, TH\}, \{TT, HH\}, \\ \{HT, TH, TT\}, \{HH, TH, TT\}, \{HH, HT, TT\}, \{HH, HT, TH\} \end{array} \right\}$$

which can represent any possible event that can be built out of two coin flips. But this is not the only event space possible with these outcomes.

Worked example 2.12 *The structure of event spaces*

I flip two coins. Is the following collection of sets an event space?

$$\Sigma = \{\emptyset, \Omega, \{HH\}, \{TT, HT, TH\}\}$$

Solution: Yes, because: $\emptyset \in \Sigma$; $\Omega \in \Sigma$; if $\mathcal{A} \in \Sigma$, $\mathcal{A}^c \in \Sigma$; if $\mathcal{A} \in \Sigma$ and $\mathcal{B} \in \Sigma$, $\mathcal{A} \cup \mathcal{B} \in \Sigma$; and if $\mathcal{A} \in \Sigma$ and $\mathcal{B} \in \Sigma$, $\mathcal{A} \cap \mathcal{B} \in \Sigma$.

So, from example 12, we can have different consistent collections of events built on top of the same set of outcomes. This makes sense, because it allows us to reason about different kinds of result obtained with the same experimental equipment. You can interpret the event space in example 12 as encoding the events "two heads" and "anything other than two heads".

For a single throw of the die, the set of every possible event is

$$\left\{\begin{array}{llllll}
\emptyset, & \{1,2,3,4,5,6\}, & & & & \\
\{1\}, & \{2\}, & \{3\}, & \{4\}, & \{5\}, & \{6\}, \\
 & \{1,2\}, & \{1,3\}, & \{1,4\}, & \{1,5\}, & \{1,6\}, \\
 & & \{2,3\}, & \{2,4\}, & \{2,5\}, & \{2,6\}, \\
 & & & \{3,4\}, & \{3,5\}, & \{3,6\}, \\
 & & & & \{4,5\}, & \{4,6\}, \\
 & & & & & \{5,6\}, \\
 & \{1,2,3\}, & \{1,2,4\}, & \{1,2,5\}, & \{1,2,6\}, & \\
 & & \{1,3,4\}, & \{1,3,5\}, & \{1,3,6\}, & \\
 & & & \{1,4,5\}, & \{1,4,6\}, & \\
 & & & & \{1,5,6\}, & \\
 & & \{2,3,4\}, & \{2,3,5\}, & \{2,3,6\}, & \\
 & & \{2,4,5\}, & \{2,4,6\}, & \{2,5,6\}, & \\
 & & & \{3,4,5\}, & \{3,4,6\}, & \\
 & & & & \{3,5,6\}, & \\
 & & & & \{4,5,6\}, & \\
 & & \{1,2,3,4\}, & \{1,2,3,5\}, & \{1,2,3,6\}, & \\
 & & & \{1,3,4,5\}, & \{1,3,4,6\}, & \\
 & & & \{2,3,4,5\}, & \{2,3,4,6\}, & \\
 & & & & \{3,4,5,6\}, & \\
\{2,3,4,5,6\}, & \{1,3,4,5,6\}, & \{1,2,4,5,6\}, & \{1,2,3,5,6\}, & \{1,2,3,4,6\}, & \{1,2,3,4,5\}
\end{array}\right\}$$

(which gives some explanation as to why we don't usually write out the whole space of events). In fact, it is seldom necessary to explain which event space one is working with. We usually assume that the event space consists of all events that can be obtained with the outcomes (as in the event space shown for a die).

2.1.3 The Probability of Events

So far, we have described the probability of each outcome with a non-negative number. This number represents the relative frequency of the outcome. Because we can tell when an event has occurred, we can compute the relative frequency of events, too. Because it is a relative frequency, the probability of an event is a

non-negative number, and is no greater than one. But the probability of events must be consistent with the probability of outcomes. This implies a set of quite straightforward properties:

- **The probability of every event is non-negative**, which we write $P(\mathcal{A}) \geq 0$ for all $\mathcal{A}$ in the collection of events.
- **Every experiment has an outcome**, which we write $P(\Omega) = 1$.
- **The probability of disjoint events is additive**, which requires more notation. Assume that we have a collection of events $\mathcal{A}_i$, indexed by i. We require that these have the property $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ when $i \neq j$. This means that there is no outcome that appears in more than one $\mathcal{A}_i$. In turn, if we interpret probability as relative frequency, we must have that $P(\cup_i \mathcal{A}_i) = \sum_i P(\mathcal{A}_i)$.

Any function P taking events to numbers that has these properties is a probability. These very simple properties imply a series of other very important properties.

Useful Facts 2.1 *The probability of events*

- $P(\mathcal{A}^c) = 1 - P(\mathcal{A})$
- $P(\emptyset) = 0$
- $P(\mathcal{A} - \mathcal{B}) = P(\mathcal{A}) - P(\mathcal{A} \cap \mathcal{B})$
- $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B})$
- $P(\cup_1^n \mathcal{A}_i) = \sum_i P(\mathcal{A}_i) - \sum_{i<j} P(\mathcal{A}_i \cap \mathcal{A}_j) + \sum_{i<j<k} P(\mathcal{A}_i \cap \mathcal{A}_j \cap \mathcal{A}_k) + \ldots (-1)^{(n+1)} P(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \ldots \cap \mathcal{A}_n)$

I prove each of these below. Looking at the useful facts should suggest a helpful analogy. Think about the probability of an event as the "size" of that event. This "size" is relative to Ω, which has "size" 1. I find this a good way to remember equations.

For example, $P(\mathcal{A}) + P(\mathcal{A}^c) = 1$ has to be true, by this analogy, because $\mathcal{A}$ and $\mathcal{A}^c$ don't overlap, and together make up all of Ω. Similarly, $P(\mathcal{A} - \mathcal{B}) = P(\mathcal{A}) - P(\mathcal{A} \cap \mathcal{B})$ is easily captured — the "size" of the part of $\mathcal{A}$ that isn't $\mathcal{B}$ is obtained by taking the "size" of $\mathcal{A}$ and subtracting the "size" of the part that is also in $\mathcal{B}$. Similarly, $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B})$ says — you can get the "size" of $\mathcal{A} \cup \mathcal{B}$ by adding the two "sizes", then subtracting the "size" of the intersection because otherwise you would count these terms twice. Some people find Venn diagrams a useful way to keep track of this argument, and Figure 2.1 is for them.

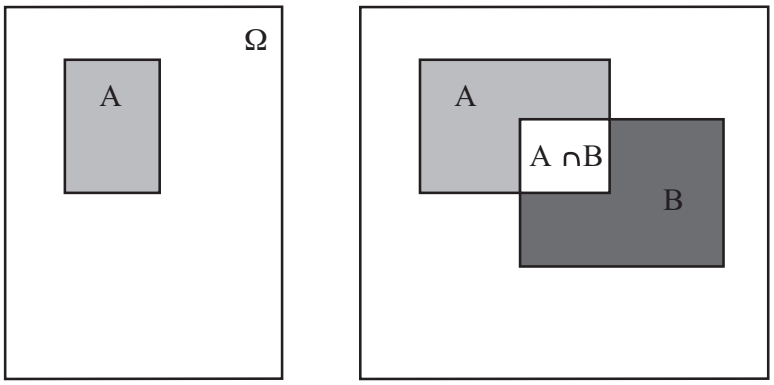

FIGURE 2.1: *If you think of the probability of an event as measuring its "size", many of the rules are quite straightforward to remember. Venn diagrams can sometimes help. On the* **left**, *a Venn diagram to help remember that* $P(\mathcal{A}) + P(\mathcal{A}^c) = 1$. *The "size" of* Ω *is 1, outcomes lie either in* $\mathcal{A}$ *or* $\mathcal{A}^c$, *and the two don't intersect. On the* **right**, *you can see that* $P(\mathcal{A} - \mathcal{B}) = P(\mathcal{A}) - P(\mathcal{A} \cap \mathcal{B})$ *by noticing that* $P(\mathcal{A} - \mathcal{B})$ *is the "size" of the part of* $\mathcal{A}$ *that isn't* $\mathcal{B}$. *This is obtained by taking the "size" of* $\mathcal{A}$ *and subtracting the "size" of the part that is also in* $\mathcal{B}$, *i.e. the "size" of* $\mathcal{A} \cap \mathcal{B}$. *Similarly, you can see that* $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B})$ *by noticing that you can get the "size" of* $\mathcal{A} \cup \mathcal{B}$ *by adding the "sizes" of* $\mathcal{A}$ *and* $\mathcal{B}$, *then subtracting the "size" of the intersection to avoid double counting.*

Proposition: $P(\mathcal{A}^c) = 1 - P(\mathcal{A})$

Proof: $\mathcal{A}^c$ and $\mathcal{A}$ are disjoint, so that $P(\mathcal{A}^c \cup \mathcal{A}) = P(\mathcal{A}^c) + P(\mathcal{A}) = P(\Omega) = 1$.

Proposition: $P(\emptyset) = 0$

Proof: $P(\emptyset) = P(\Omega^c) = P(\Omega - \Omega) = 1 - P(\Omega) = 1 - 1 = 0$.

Proposition: $P(\mathcal{A} - \mathcal{B}) = P(\mathcal{A}) - P(\mathcal{A} \cap \mathcal{B})$

Proof: $\mathcal{A} - \mathcal{B}$ is disjoint from $P(\mathcal{A} \cap \mathcal{B})$, and $(\mathcal{A} - \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{B}) = \mathcal{A}$. This means that $P(\mathcal{A} - \mathcal{B}) + P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A})$.

Proposition: $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B})$

Proof: $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A} \cup (\mathcal{B} \cap \mathcal{A}^c)) = P(\mathcal{A}) + P((\mathcal{B} \cap \mathcal{A}^c))$. Now $\mathcal{B} = (\mathcal{B} \cap \mathcal{A}) \cup (\mathcal{B} \cap \mathcal{A}^c)$. Furthermore, $(\mathcal{B} \cap \mathcal{A})$ is disjoint from $(\mathcal{B} \cap \mathcal{A}^c)$, so we have $P(\mathcal{B}) = P((\mathcal{B} \cap \mathcal{A})) + P((\mathcal{B} \cap \mathcal{A}^c))$. This means that $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P((\mathcal{B} \cap \mathcal{A}^c)) = P(\mathcal{A}) + P(\mathcal{B}) - P((\mathcal{B} \cap \mathcal{A}))$.

Proposition: $P(\cup_1^n \mathcal{A}_i) = \sum_i P(\mathcal{A}_i) - \sum_{i<j} P(\mathcal{A}_i \cap \mathcal{A}_j) + \sum_{i<j<k} P(\mathcal{A}_i \cap \mathcal{A}_j \cap \mathcal{A}_k) + \ldots (-1)^{(n+1)} P(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \ldots \cap \mathcal{A}_n)$

Proof: This can be proven by repeated application of the previous result. As an example, we show how to work the case where there are three sets (you can get the rest by induction).

$$\begin{aligned}
P(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3) &= P(\mathcal{A}_1 \cup (\mathcal{A}_2 \cup \mathcal{A}_3)) \\
&= P(\mathcal{A}_1) + P(\mathcal{A}_2 \cup \mathcal{A}_3) - P(\mathcal{A}_1 \cap (\mathcal{A}_2 \cup \mathcal{A}_3)) \\
&= P(\mathcal{A}_1) + (P(\mathcal{A}_2) + P(\mathcal{A}_3) - P(\mathcal{A}_2 \cap \mathcal{A}_3)) - \\
&\quad P((\mathcal{A}_1 \cap \mathcal{A}_2) \cup (\mathcal{A}_1 \cap \mathcal{A}_3)) \\
&= P(\mathcal{A}_1) + (P(\mathcal{A}_2) + P(\mathcal{A}_3) - P(\mathcal{A}_2 \cap \mathcal{A}_3)) - \\
&\quad P(\mathcal{A}_1 \cap \mathcal{A}_2) - P(\mathcal{A}_1 \cap \mathcal{A}_3) \\
&\quad -(-P((\mathcal{A}_1 \cap \mathcal{A}_2) \cap (\mathcal{A}_1 \cap \mathcal{A}_3))) \\
&= P(\mathcal{A}_1) + P(\mathcal{A}_2) + P(\mathcal{A}_3) - \\
&\quad P(\mathcal{A}_2 \cap \mathcal{A}_3) - P(\mathcal{A}_1 \cap \mathcal{A}_2) - P(\mathcal{A}_1 \cap \mathcal{A}_3) + \\
&\quad P(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3)
\end{aligned}$$

2.1.4 Computing Probabilities by Counting Outcomes

The rule $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B})$ yields a very useful procedure for computing the probability of some events. Imagine $\mathcal{A}$ and $\mathcal{B}$ are disjoint (so $\mathcal{A} \cap \mathcal{B} = \emptyset$). Then $P(\mathcal{A} \cup \mathcal{B})$ is just $P(\mathcal{A}) + P(\mathcal{B})$. So imagine some event $\mathcal{E}$ that consists of a set of outcomes. Each singleton set — containing just one outcome — is disjoint from each other one. I can make the set $\mathcal{E}$ by: starting with an empty set; unioning this with one of the outcomes in $\mathcal{E}$; then repeatedly unioning the result a new outcome until I get $\mathcal{E}$. I am always unioning disjoint sets, so the probability of this event is the sum of probabilities of the outcomes. So we have

$$P(\mathcal{E}) = \sum_{O \in \mathcal{E}} P(O)$$

where O ranges over the outcomes in $\mathcal{E}$.

This is particularly useful when you know each outcome in Ω has the same probability. In this case, you can show

$$P(\mathcal{E}) = \frac{\text{Number of outcomes in } \mathcal{E}}{\text{Total number of outcomes in } \Omega}$$

(exercises). Such problems become, basically, advanced counting exercises.

Worked example 2.13 *Odd numbers with fair dice*

We throw a fair (each number has the same probability) die twice, then add the two numbers. What is the probability of getting an odd number?

Solution: There are 36 outcomes. Each has the same probability (1/36). 18 of them give an odd number, and the other 18 give an even number, so the probability is $18/36 = 1/2$

Worked example 2.14 *Drawing a red ten*

I shuffle a standard pack of cards, and draw one card. What is the probability that it is a red ten?

Solution: There are 52 cards, and each is an outcome. Two of these outcomes are red tens; so 1/26.

Worked example 2.15 *Numbers divisible by five with fair dice*

We throw a fair (each number has the same probability) die twice, then add the two numbers. What is the probability of getting a number divisible by five?

Solution: There are 36 outcomes. Each has the same probability (1/36). For this event, the spots must add to either 5 or to 10. There are 4 ways to get 5. There are 3 ways to get 10, so the probability is 7/36.

Worked example 2.16 *Children - 1*

This example is a version of of example 1.12, p44, in Stirzaker, "Elementary Probability". A couple decides to have children. They decide simply to have three children. Assume that each gender is equally likely at each birth. Let B_i be the event that there are i boys, and C be the event there are more girls than boys. Compute $P(B_1)$ and $P(C)$.

Solution: There are eight outcomes. Each has the same probability. Three of them have a single boy, so $P(B_1) = 3/8$. $P(C) = P(C^c)$ (because C^c is the event there are more boys than than girls, AND the number of children is odd), so that $P(C) = 1/2$; you can also get this by counting outcomes.

Sometimes a bit of fiddling with the space of outcomes makes it easy to compute what we want.

Worked example 2.17 *Children - 2*

This example is a version of of example 1.12, p44, in Stirzaker, "Elementary Probability". A couple decides to have children. They decide to have children until the first girl is born, or until there are three, and then stop. Assume that each gender is equally likely at each birth. Let B_i be the event that there are i boys, and C be the event there are more girls than boys. Compute $P(B_1)$ and $P(C)$.

Solution: In this case, we could write the outcomes as $\{G, BG, BBG\}$, but if we think about them like this, we have no simple way to compute their probability. Instead, we could use the sample space from the previous answer, but assume that some of the later births are fictitious. This gives us natural collection of events for which it is easy to compute probabilities. Having one girl corresponds to the event $\{Gbb, Gbg, Ggb, Ggg\}$, where I have used lowercase letters to write the fictitious later births; the probability is $1/2$. Having a boy then a girl corresponds to the event $\{BGb, BGg\}$ (and so has probability $1/4$). Having two boys then a girl corresponds to the event $\{BBG\}$ (and so has probability $1/8$). Finally, having three boys corresponds to the event $\{BBB\}$ (and so has probability $1/8$). This means that $P(B_1) = 1/4$ and $P(C) = 1/2$.

Worked example 2.18 *Children - 2*

This example is a version of of example 1.12, p44, in Stirzaker, "Elementary Probability". A couple decides to have children. They decide to have children until there is one of each gender, or until there are three, and then stop. Assume that each gender is equally likely at each birth. Let B_i be the event that there are i boys, and C be the event there are more girls than boys. Compute $P(B_1)$ and $P(C)$.

Solution: We could write the outcomes as $\{GB, BG, GGB, GGG, BBG, BBB\}$. Again, if we think about them like this, we have no simple way to compute their probability; so we use the sample space from the previous example with the device of the fictitious births again. The important events are $\{GBb, GBg\}$; $\{BGb, BGg\}$; $\{GGB\}$; $\{GGG\}$; $\{BBG\}$; and $\{BBB\}$. Like this, we get $P(B_1) = 5/8$ and $P(C) = 1/4$.

Worked example 2.19 *Birthdays in succession*

We stop three people at random, and ask the day of the week on which they are born. What is the probability that they are born on three days of the week in succession (for example, the first on Monday; the second on Tuesday; the third on Wednesday; or Saturday-Sunday-Monday; and so on).

Solution: We assume that births are equally common on each day of the week. The space of outcomes consists of triples of days, and each outcome has the same probability. The event is the set of triples of three days in succession (which has seven elements, one for each starting day). The space of outcomes has 7^3 elements in it, so the probability is

$$\frac{\text{Number of outcomes in the event}}{\text{Total number of outcomes}} = \frac{7}{7^3} = \frac{1}{49}.$$

Worked example 2.20 *Shared birthdays*

We stop two people at random. What is the probability that they were born on the same day of the week?

Solution: The day the first person was born doesn't matter; the probability the second person was born on that day is 1/7. Or you could count outcomes explicitly to get

$$\frac{\text{Number of outcomes in the event}}{\text{Total number of outcomes}} = \frac{7}{7 \times 7} = \frac{1}{7}.$$

An important feature of this class of problem is that your intuition can be quite misleading. This is because, although each outcome can have very small probability, the number of outcomes in an event can be big.

2.1.5 Computing Probabilities by Reasoning about Sets

The rule $P(\mathcal{A}^c) = 1 - P(\mathcal{A})$ is occasionally useful for computing probabilities on its own; more commonly, you need other reasoning as well.

Worked example 2.21 *Shared birthdays*

What is the probability that, in a room of 30 people, there is a pair of people who have the same birthday?

Solution: We simplify, and assume that each year has 365 days, and that none of them are special (i.e. each day has the same probability of being chosen as a birthday). This model isn't perfect (there tend to be slightly more births roughly 9 months after: the start of spring; blackouts; major disasters; and so on) but it's workable. The easy way to attack this question is to notice that our probability, $P(\{\text{shared birthday}\})$, is

$$1 - P(\{\text{all birthdays different}\}).$$

This second probability is rather easy to compute. Each outcome in the sample space is a list of 30 days (one birthday per person). Each outcome has the same probability. So

$$P(\{\text{all birthdays different}\}) = \frac{\text{Number of outcomes in the event}}{\text{Total number of outcomes}}.$$

The total number of outcomes is easily seen to be 365^{30}, which is the total number of possible lists of 30 days. The number of outcomes in the event is the number of lists of 30 days, all different. To count these, we notice that there are 365 choices for the first day; 364 for the second; and so on. So we have

$$P(\{\text{shared birthday}\}) = 1 - \frac{365 \times 364 \times \ldots 336}{365^{30}} = 1 - 0.2937 = 0.7063$$

which means there's really a pretty good chance that two people in a room of 30 share a birthday.

If we change the birthday example slightly, the problem changes drastically. If you stand up in a room of 30 people and bet that two people in the room have the same birthday, you have a probability of winning of about 0.71. If you bet that there is someone else in the room who has the same birthday that you do, your probability of winning is very different.

Worked example 2.22 *Shared birthdays*

You bet there is someone else in a room of 30 people who has the same birthday that you do. Assuming you know nothing about the other 29 people, what is the probability of winning?

Solution: The easy way to do this is

$$P(\{\text{winning}\}) = 1 - P(\{\text{losing}\}).$$

Now you will lose if everyone has a birthday different from you. You can think of the birthdays of the others in the room as a list of 29 days of the year. If your birthday is on the list, you win; if it's not, you lose. The number of losing lists is the number of lists of 29 days of the year such that your birthday is not in the list. This number is easy to get. We have 364 days of the year to choose from for each of 29 locations in the list. The total number of lists is the number of lists of 29 days of the year. Each list has the same probability. So

$$P(\{\text{losing}\}) = \frac{364^{29}}{365^{29}}$$

and

$$P(\{\text{winning}\}) \approx 0.0765.$$

There is a wide variety of problems like this; if you're so inclined, you can make a small but quite reliable profit off people's inability to estimate probabilities for this kind of problem correctly (examples 21 and 22 are reliably profitable; you could probably do quite well out of examples 19 and 20).

The rule $P(\mathcal{A}-\mathcal{B}) = P(\mathcal{A})-P(\mathcal{A}\cap\mathcal{B})$ is also occasionally useful for computing probabilities on its own; more commonly, you need other reasoning as well.

Worked example 2.23 *Dice*

You flip two fair six-sided dice, and add the number of spots. What is the probability of getting a number divisible by 2, but not by 5?

Solution: There is an interesting way to work the problem. Write $\mathcal{D}_n$ for the event the number is divisible by n. Now $P(\mathcal{D}_2) = 1/2$ (count the cases; or, more elegantly, notice that each die has the same number of odd and even faces, and work from there). Now $P(\mathcal{D}_2 - \mathcal{D}_5) = P(\mathcal{D}_2) - P(\mathcal{D}_2 \cap \mathcal{D}_5)$. But $\mathcal{D}_2 \cap \mathcal{D}_5$ contains only two outcomes ($6,4$ and $4,6$), so $P(\mathcal{D}_2 - \mathcal{D}_5) = 18/36 - 2/36 = 4/9$

Sometimes it is easier to reason about unions than to count outcomes directly.

Worked example 2.24 *Two fair dice*

I roll two fair dice. What is the probability that the result is divisible by either 2 or 5, or both?

Solution: Write $\mathcal{D}_n$ for the event the number is divisible by n. We want $P(\mathcal{D}_2 \cup \mathcal{D}_5) = P(\mathcal{D}_2) + P(\mathcal{D}_5) - P(\mathcal{D}_2 \cap \mathcal{D}_5)$. From example 23, we know $P(\mathcal{D}_2) = 1/2$ and $P(\mathcal{D}_2 \cap \mathcal{D}_5) = 2/36$. By counting outcomes, $P(\mathcal{D}_5) = 6/36$. So $P(\mathcal{D}_2 \cup \mathcal{D}_5) = (18 + 6 - 2)/36 = 22/36$.

2.1.6 Independence

Some experimental results do not affect others. For example, if I flip a coin twice, whether I get heads on the first flip has no effect on whether I get heads on the second flip. As another example, I flip a coin; the outcome does not affect whether I get hit on the head by a falling apple later in the day. We refer to events with this property as independent.

Useful Facts 2.2 *Independent events*

Two events $\mathcal{A}$ and $\mathcal{B}$ are **independent** if and only if

$$P(\mathcal{A} \cap \mathcal{B}) = P(A)P(B)$$

The "size" analogy helps motivate this expression. We think of $P(A)$ as the "size" of $\mathcal{A}$ relative to Ω, and so on. Now $P(\mathcal{A} \cap \mathcal{B})$ measures the "size" of $\mathcal{A} \cap \mathcal{B}$ — that is, the part of $\mathcal{A}$ that lies inside $\mathcal{B}$. But if $\mathcal{A}$ and $\mathcal{B}$ are independent, then the size of $\mathcal{A} \cap \mathcal{B}$ relative to $\mathcal{B}$ should be the same as the size of $\mathcal{A}$ relative to Ω (Figure 2.2). Otherwise, $\mathcal{B}$ affects $\mathcal{A}$, because $\mathcal{A}$ is more (or less) likely when $\mathcal{B}$ has occurred.

So for $\mathcal{A}$ and $\mathcal{B}$ to be independent, we must have

$$\text{"Size" of } \mathcal{A} = \frac{\text{"Size" of piece of } \mathcal{A} \text{ in } \mathcal{B}}{\text{"Size" of } \mathcal{B}},$$

or, equivalently,

$$P(\mathcal{A}) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{B})}$$

which yields our expression. Independence is important, because it is straightforward to compute probabilities for sequences of independent outcomes.

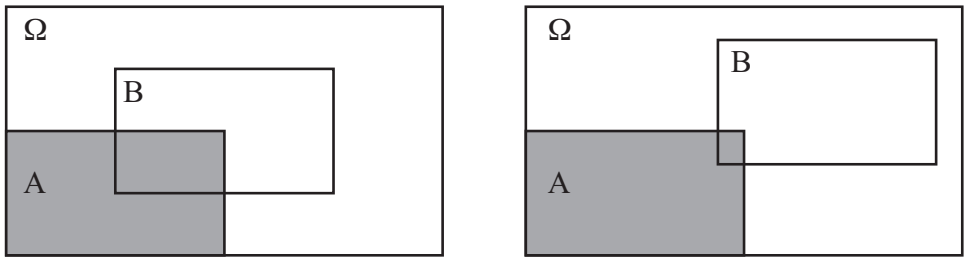

FIGURE 2.2: *On the* **left**, $\mathcal{A}$ *and* $\mathcal{B}$ *are independent.* $\mathcal{A}$ *spans* $1/4$ *of* Ω, *and* $\mathcal{A} \cap \mathcal{B}$ *spans* $1/4$ *of* $\mathcal{B}$. *This means that* $\mathcal{B}$ *can't affect* $\mathcal{A}$. $1/4$ *of the outcomes of* Ω *lie in* $\mathcal{A}$, *and* $1/4$ *of the outcomes in* $\mathcal{B}$ *lie in* $\mathcal{A} \cap \mathcal{B}$. *On the* **right**, *they are not. Very few of the outcomes in* $\mathcal{B}$ *lie in* $\mathcal{B} \cap \mathcal{A}$, *so that observing* $\mathcal{B}$ *means that* $\mathcal{A}$ *becomes less likely, because very few of the outcomes in* $\mathcal{B}$ *also lie in* $\mathcal{A} \cap \mathcal{B}$.

Worked example 2.25 *A fair coin*

A **fair** coin (or die, or whatever) is one where each outcome has the same probability. A fair coin has two sides, so the probability of each outcome must be $1/2$. We flip this coin twice - what is the probability we see two heads?

Solution: The flips are independent. Write $\mathcal{H}_1$ for the event that the first flip comes up heads. We have $\mathcal{H}_1 = \{HH, HT\}$ and $\mathcal{H}_2 = \{HH, HT\}$. We seek $P(\mathcal{H}_1 \cap \mathcal{H}_2) = P(\mathcal{H}_1)P(\mathcal{H}_2)$. The coin is fair, so $P(\mathcal{H}_1) = P(\mathcal{H}_2) = 1/2$. So the probability is $\frac{1}{4}$.

The reasoning of example 25 is moderately rigorous (if you want real rigor, you should specify the event space, etc.; I assumed that it would be obvious), but it's a bit clumsy for everyday use. The rule to remember is this: the probability that both events occur is $P(\mathcal{A} \cap \mathcal{B})$ and, if $\mathcal{A}$ and $\mathcal{B}$ are independent, then this is $P(\mathcal{A})P(\mathcal{B})$.

Worked example 2.26 *A fair die*

The space of outcomes for a fair six-sided die is

$$\{1, 2, 3, 4, 5, 6\}.$$

The die is fair means that each outcome has the same probability. Now we toss two fair dice — with what probability do we get two threes?

Solution:

$$\begin{aligned} P(\text{first toss yields 3} \cap \text{second toss yields 3}) &= P(\text{first toss yields 3}) \times \\ &\quad P(\text{second toss yields 3}) \\ &= (1/6)(1/6) \\ &= 1/36 \end{aligned}$$

Worked example 2.27 *Find the Lady, twice*

Assume that the card that is chosen is chosen fairly — that is, each card is chosen with the same probability. The game is played twice, and the cards are reshuffled between games. What is the probability of turning up a Queen and then a Queen again?

Solution: The events are independent, so 1/9.

Worked example 2.28 *Children*

A couple decides to have two children. Genders are assigned to children at random, fairly, independently and at birth (our models have to abstract a little!). What is the probability of having a boy and then a girl?

Solution:

$$\begin{aligned} P(\text{first is boy} \cap \text{second is girl}) &= P(\text{first is boy})P(\text{second is girl}) \\ &= (1/2)(1/2) \\ &= 1/4 \end{aligned}$$

You can use the expression to tell whether events are independent or not. Quite small changes to a problem affect whether events are independent. For example, simply removing a card from a deck can make some events dependent.

Worked example 2.29 *Independent cards*

We shuffle a standard deck of 52 cards and draw one card. The event $\mathcal{A}$ is "the card is a red suit" and the event $\mathcal{B}$ is "the card is a 10". Are they independent?

Solution: $P(\mathcal{A}) = 1/2$, $P(\mathcal{B}) = 1/13$ and in example 14 we determined $P(\mathcal{A} \cap \mathcal{B}) = 2/52$. But $2/52 = 1/26 = P(\mathcal{A})P(\mathcal{B})$, so they are independent.

Worked example 2.30 *Independent cards*

We take a standard deck of cards, and remove the ten of hearts. We now shuffle this deck, and draw one card. The event $\mathcal{A}$ is "the card is a red suit" and the event $\mathcal{B}$ is "the card is a 10". Are they independent?

Solution: These are not independent because $P(\mathcal{A}) = 25/51$, $P(\mathcal{B}) = 3/51$ and $P(\mathcal{A} \cap \mathcal{B}) = 1/51 \neq P(\mathcal{A})P(\mathcal{B}) = 75/(51^2)$

The probability of a sequence of independent events can become very small very quickly, and this often misleads people.

Worked example 2.31 *Accidental DNA Matches*

I search a DNA database with a sample. Each time I attempt to match this sample to an entry in the database, there is a probability of an accidental chance match of $1e-4$. Chance matches are independent. There are 20, 000 people in the database. What is the probability I get at least one match, purely by chance?

Solution: This is $1 - P(\text{no chance matches})$. But $P(\text{no chance matches})$ is much smaller than you think. We have

$$\begin{aligned} P(\text{no chance matches}) &= P\left(\begin{array}{c} \text{no chance match to record } 1 \cap \\ \text{no chance match to record } 2 \cap \\ \ldots \cap \\ \text{no chance match to record } 20,000 \end{array}\right) \\ &= P(\text{no chance match to a record})^{20,000} \\ &= (1 - 1e-4)^{20,000} \end{aligned}$$

so the probability is about 0.86 that you get at least one match by chance. Notice that if the database gets bigger, the probability grows; so at 40, 000 the probability of one match by chance is 0.98.

2.1.7 Permutations and Combinations

Counting outcomes in an event can require pretty elaborate combinatorial arguments. One form of argument that is particularly important is to reason about permutations. You should recall that the number of permutations of k items is $k!$.

Worked example 2.32 *Counting outcomes with permutations*

I flip a coin k times. How many of the possible outcomes have exactly r heads?

Solution: Here is one natural way to think about this problem. Each outcome we are interested in is a string of r H's and $k-r$ T's, and we need to count the number of such strings. We can do so with permutations. Write down any string of r H's and $k-r$ T's. Any other string of r H's and $k-r$ T's is a permutation of this one. However, many of these permutations simply swap one H with another, or one T with another. The total number of permutations of a string of k entries is $k!$. The number of permutations that swap H's with one another is $r!$ (because there are r H's), and the number of permutations that swap T's with one another is $(k-r)!$. The total number of strings must then be

$$\frac{k!}{r!(k-r)!} = \left(\begin{array}{c} k \\ r \end{array} \right)$$

There is another way to think about example 32, which is more natural if you've seen combinations before. You start with a string of k T's, then you choose r distinct elements to turn into H's. The number of choices of r distinct elements in k items is:

$$\frac{k!}{r!(k-r)!} = \left(\begin{array}{c} k \\ r \end{array} \right),$$

so there must be this number of strings. We can use this result to investigate our model of probability as frequency.

Worked example 2.33 *A fair coin, revisited*

We conduct N experiments, where each experiment is to flip a fair coin twice. In what fraction of these experiments do we see both sides of the coin?

Solution: The sample space is $\{HH, HT, TH, TT\}$. The coin is fair, so each outcome has the same probability. This means that $(1/4)N$ experiments produce HT; $(1/4)N$ produce TH; and so on. We see both sides of the coin in an experiment for about $(1/2)N$ experiments.

Example 33 might seem like a contradiction to you. I claimed that we could interpret P as the relative frequency of outcomes, and that my coin was fair; but, in only half of my experiments did I see both sides of the coin. In the other half, the coin behaved as if the probability of seeing one side is 1, and the probability of seeing the other side is 0. This occurs because the number of flips *in each experiment* is very small. If the number of flips is larger, you are much more likely to see about the right frequencies.

Worked example 2.34 *A fair coin, yet again*

We conduct N experiments, where each experiment is to flip a fair coin six times. In what fraction of these experiments do we get three heads and three tails?

Solution: There are $64 = 2^6$ outcomes in total for six coin flips. Each has the same probability. By the argument of example 32, there are

$$\frac{6!}{3!3!} = 20$$

outcomes with three heads. So the fraction is $20/64 = 1/3$.

At first glance, example 34 suggests that making the number of experiments get bigger doesn't help. But in the definition of probability, I said that there would be "about" $N \times P$ experiments with the outcome of interest.

Worked example 2.35 *A fair coin, yet again*

We conduct N experiments, where each experiment is to flip a fair coin 10 times. In what fraction of these experiments do we get between 4 and 6 heads?

Solution: There are $1024 = 2^{10}$ outcomes for an experiment. We are interested in the four head outcomes, the five head outcomes and the six head outcomes. Using the argument of example 34 gives

$$\begin{aligned} \text{total outcomes} &= \text{4H outcomes} + \text{5H outcomes} + \text{6H outcomes} \\ &= \frac{10!}{4!6!} + \frac{10!}{5!5!} + \frac{10!}{6!4!} \\ &= 210 + 252 + 210 \\ &= 692 \end{aligned}$$

so in $692/1024 \approx 0.68$ of the experiments, we will see between four and six heads

Worked example 2.36 *A fair coin, and a lot of flipping*

We conduct N experiments, where each experiment is to flip a fair coin 100 times. In what fraction of these experiments do we get between 45 and 65 heads?

Solution: There are 2^{100} outcomes for an experiment. Using the argument of example 35 gives

$$\text{total outcomes} = \sum_{i=45}^{65} \frac{100!}{i!(100-i)!}$$

which isn't particularly easy to evaluate.

As these examples suggest, if an experiment consists of flipping a large number of coins, then a high fraction of those experiments will show heads with a frequency very close to the right number. We will establish this later, with rather more powerful machinery.

Worked example 2.37 *Overbooking - 1*

An airline has a regular flight with six seats. It always sells seven tickets. Passengers turn up for the flight with probability p, and do so independent of other passengers. What is the probability that the flight is overbooked?

Solution: This is like a coin-flip problem; think of each passenger as a biased coin. With probability p, it comes up T (for *t*urn up) and with probability $(1-p)$, it turns up N (for *n*o-show). This coin is flipped seven times, and we are interested in the probability that there are seven T's. This is p^7, because the flips are independent.

Worked example 2.38 *Overbooking - 2*

An airline has a regular flight with six seats. It always sells eight tickets. Passengers turn up for the flight with probability p, and do so independent of other passengers. What is the probability that six passengers arrive? (i.e. the flight is not overbooked or underbooked).

Solution: Now we flip the coin eight times, and are interested in the probability of getting exactly six T's. Two flips must be N; there are eight choices for the first flip that is N, and seven choices for the second. So there are a total of $8 \times 7 = 56$ strings of 6 T's and 2 N's. The probability of any one string is $p^6(1-p)^2$; so the probability that six passengers arrive is

$$56p^6(1-p)^2$$

Worked example 2.39 *Overbooking - 3*

An airline has a regular flight with six seats. It always sells eight tickets. Passengers turn up for the flight with probability p, and do so independent of other passengers. What is the probability that the flight is overbooked?

Solution: Now we flip the coin eight times, and are interested in the probability of getting more than six T's. This is the union of two disjoint events (seven T's and eight T's). For the case of seven T's, one flip must be N; there are eight choices. For the case of eight T's, all eight flips must be T, and there is only one way to achieve this. So the probability the flight is overbooked is

$$\begin{aligned} P(\text{overbooked}) &= P(7\ T\text{'s} \cup 8\ T\text{'s}) \\ &= P(7\ T\text{'s}) + P(8\ T\text{'s}) \\ &= 7p^7(1-p) + p^8 \end{aligned}$$

Worked example 2.40 *Overbooking - 4*

An airline has a regular flight with s seats. It always sells t tickets. Passengers turn up for the flight with probability p, and do so independent of other passengers. What is the probability that u passengers turn up?

Solution: Now we flip the coin t times, and are interested in the probability of getting u T's. By the argument of worked example 32, there are

$$\frac{t!}{u!(t-u)!}$$

disjoint outcomes with u T's and $t-u$ N's. Each such outcome is independent, and has probability $p^u(1-p)^{t-u}$. So

$$P(u \text{ passengers turn up}) = \frac{t!}{u!(t-u)!}p^u(1-p)^{t-u}$$

Worked example 2.41 *Overbooking - 5*

An airline has a regular flight with s seats. It always sells t tickets. Passengers turn up for the flight with probability p, and do so independent of other passengers. What is the probability that the flight is oversold?

Solution: We need $P(\{s+1 \text{ turn up}\} \cup \{s+2 \text{ turn up}\} \cup \ldots \cup \{t \text{ turn up}\})$. But the events $\{i \text{ turn up}\}$ and $\{j \text{ turn up}\}$ are disjoint if $i \neq j$. So we can exploit example 40, and write

$$\begin{aligned} P(\text{overbooked}) &= P(\{s+1 \text{ turn up}\}) + P(\{s+2 \text{ turn up}\}) + \\ &\quad \ldots P(\{t \text{ turn up}\}) \\ &= \sum_{i=s+1}^{t} P(\{i \text{ turn up}\}) \\ &= \sum_{i=s+1}^{t} \frac{t!}{i!(t-i)!} p^i (1-p)^{t-i} \end{aligned}$$

2.2 CONDITIONAL PROBABILITY

If you throw a fair die twice and add the numbers, then the probability of getting a number less than six is $\frac{10}{36}$. Now imagine you know that the first die came up three. In this case, the probability that the sum will be less than six is $\frac{1}{3}$, which is slightly larger. If the first die came up four, then the probability the sum will be less than six is $\frac{1}{6}$, which is rather less than $\frac{10}{36}$. If the first die came up one, then the probability that the sum is less than six becomes $\frac{2}{3}$, which is much larger.

Each of these probabilities is an example of a **conditional probability**. We assume we have a space of outcomes and a collection of events. The conditional probability of $\mathcal{B}$, conditioned on $\mathcal{A}$, is the probability that $\mathcal{B}$ occurs given that $\mathcal{A}$ has definitely occurred. We write this as

$$P(\mathcal{B}|\mathcal{A})$$

2.2.1 Evaluating Conditional Probabilities

Now to get an expression for $P(\mathcal{B}|\mathcal{A})$, notice that, because $\mathcal{A}$ is known to have occurred, our space of outcomes or sample space is now reduced to $\mathcal{A}$. We know that our outcome lies in $\mathcal{A}$; $P(\mathcal{B}|\mathcal{A})$ is the probability that it also lies in $\mathcal{B} \cap \mathcal{A}$.

The outcome lies in $\mathcal{A}$, and so it must lie in either $P(\mathcal{B} \cap \mathcal{A})$ or in $P(\mathcal{B}^c \cap \mathcal{A})$. This means that

$$P(\mathcal{B}|\mathcal{A}) + P(\mathcal{B}^c|\mathcal{A}) = 1.$$

Now recall the idea of probabilities as relative frequencies. If $P(\mathcal{C} \cap \mathcal{A}) = kP(\mathcal{B} \cap \mathcal{A})$, this means that we will see outcomes in $\mathcal{C} \cap \mathcal{A}$ k times as often as we will see outcomes in $\mathcal{B} \cap \mathcal{A}$. But this must apply even if we know in advance that

the outcome is in $\mathcal{A}$. So we must have

$$P(\mathcal{B}|\mathcal{A}) \propto P(\mathcal{B} \cap \mathcal{A}).$$

Now we need to determine the constant of proportionality; write c for this constant, meaning

$$P(\mathcal{B}|\mathcal{A}) = cP(\mathcal{B} \cap \mathcal{A}).$$

We have that

$$P(\mathcal{B}|\mathcal{A}) + P(\mathcal{B}^c|\mathcal{A}) = cP(\mathcal{B} \cap \mathcal{A}) + cP(\mathcal{B}^c \cap \mathcal{A}) = cP(\mathcal{A}) = 1,$$

so that

$$\boxed{P(\mathcal{B}|\mathcal{A}) = \frac{P(\mathcal{B} \cap \mathcal{A})}{P(\mathcal{A})}.}$$

Using the "size" metaphor, this says the probability of that an outcome, *which is known to be in* $\mathcal{A}$, is also in $\mathcal{B}$ is the fraction of $\mathcal{A}$ that is also in $\mathcal{B}$.

Another, very useful, way to write this expression is

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

Now, since $\mathcal{B} \cap \mathcal{A} = \mathcal{A} \cap \mathcal{B}$, we must have that

$$\boxed{P(\mathcal{B}|\mathcal{A}) = \frac{P(\mathcal{A}|\mathcal{B})P(\mathcal{B})}{P(\mathcal{A})}}$$

Worked example 2.42 *Two dice*

We throw two fair dice. What is the conditional probability that the sum of spots on both dice is greater than six, conditioned on the event that the first die comes up five?

Solution: Write the event that the first die comes up 5 as $\mathcal{F}$, and the event the sum is greather than six as $\mathcal{S}$. There are five outcomes where the first die comes up 5 and the number is greater than 6, so $P(\mathcal{F} \cap \mathcal{S}) = 5/36$. $P(\mathcal{S}|\mathcal{F}) = P(\mathcal{F} \cap \mathcal{S})/P(\mathcal{F}) = (5/36)/(1/6) = 5/6$.

Notice that $\mathcal{A} \cap \mathcal{B}$ and $\mathcal{A} \cap \mathcal{B}^c$ are disjoint sets, and that $\mathcal{A} = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{B}^c)$. So, because $P(\mathcal{A}) = P(\mathcal{A} \cap \mathcal{B}) + P(\mathcal{A} \cap \mathcal{B}^c)$, we have

$$\boxed{P(\mathcal{A}) = P(\mathcal{A}|\mathcal{B})P(\mathcal{B}) + P(\mathcal{A}|\mathcal{B}^c)P(\mathcal{B}^c)}$$

a tremendously important and useful fact. Another version of this fact is also very useful. Assume we have a set of disjoint sets $\mathcal{B}_i$. These sets must have the property that (a) $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ for $i \neq j$ and (b) they cover $\mathcal{A}$, meaning that $\mathcal{A} \cap (\cup_i \mathcal{B}_i) = \mathcal{A}$. Then, because $P(\mathcal{A}) = \sum_i P(\mathcal{A} \cap \mathcal{B}_i)$, so we have

$$P(\mathcal{A}) = \sum_i P(\mathcal{A}|\mathcal{B}_i)P(\mathcal{B}_i)$$

Worked example 2.43 *Car factories*

There are two car factories, A and B. Each year, factory A produces 1000 cars, of which 10 are lemons. Factory B produces 2 cars, each of which is a lemon. All cars go to a single lot, where they are thoroughly mixed up. I buy a car.

- What is the probability it is a lemon?
- What is the probability it came from factory B?
- The car is now revealed to be a lemon. What is the probability it came from factory B, conditioned on the fact it is a lemon?

Solution:

- Write the event the car is a lemon as $\mathcal{L}$. There are 1002 cars, of which 12 are lemons. The probability that I select any given car is the same, so we have 12/1002.
- Same argument yields 2/1002.
- Write $\mathcal{B}$ for the event the car comes from factory B. I need $P(\mathcal{B}|\mathcal{L})$. This is $P(\mathcal{L}|\mathcal{B})P(\mathcal{B})/P(\mathcal{L}) = (1 \times 2/1002)/(12/1002) = 1/6$.

Worked example 2.44 *Royal flushes in poker - 1*

This exercise is after Stirzaker, p. 51.
You are playing a straightforward version of poker, where you are dealt five cards face down. A royal flush is a hand of AKQJ10 all in one suit. What is the probability that you are dealt a royal flush?

Solution: This is

$$\frac{\text{number of hands that are royal flushes, ignoring card order}}{\text{total number of different five card hands, ignoring card order}}.$$

There are four hands that are royal flushes (one for each suit). Now the total number of five card hands is

$$\left(\begin{array}{c} 52 \\ 5 \end{array} \right) = 2598960$$

so we have

$$\frac{4}{2598960} = \frac{1}{649740}.$$

Worked example 2.45 *Royal flushes in poker - 2*

This exercise is after Stirzaker, p. 51.
You are playing a straightforward version of poker, where you are dealt five cards face down. A royal flush is a hand of AKQJ10 all in one suit. The fifth card that you are dealt lands face up. It is the nine of spades. What now is the probability that your have been dealt a royal flush? (i.e. what is the conditional probability of getting a royal flush, conditioned on the event that one card is the nine of spades)

Solution: No hand containing a nine of spades is a royal flush, so this is easily zero.

Worked example 2.46 *Royal flushes in poker - 3*

This exercise is after Stirzaker, p. 51.
You are playing a straightforward version of poker, where you are dealt five cards face down. A royal flush is a hand of AKQJ10 all in one suit. The fifth card that you are dealt lands face up. It is the Ace of spades. What now is the probability that your have been dealt a royal flush? (i.e. what is the conditional probability of getting a royal flush, conditioned on the event that one card is the Ace of spades)

Solution: There are two ways to do this. The easiest is to notice that the answer is the probability that the other four cards are KQJ10 of spades, which is

$$1/\left(\begin{array}{c} 51 \\ 4 \end{array}\right) = \frac{1}{249900}.$$

Harder is to consider the events

$\mathcal{A}$ = event that you receive a royal flush and last card is the ace of spades

and

$\mathcal{B}$ = event that the last card you receive is the ace of spades,

and the expression

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

Now $P(\mathcal{A}) = \frac{1}{52}$. $P(\mathcal{A} \cap \mathcal{B})$ is given by

$$\frac{\text{number of five card royal flushes where card five is Ace of spades}}{\text{total number of different five card hands}}.$$

where we DO NOT ignore card order. This is

$$\frac{4 \times 3 \times 2 \times 1}{52 \times 51 \times 50 \times 49 \times 48}$$

yielding

$$P(\mathcal{A}|\mathcal{B}) = \frac{1}{249900}.$$

Notice the interesting part: the conditional probability is rather larger than the probability. If you see this ace, the conditional probability is $\frac{13}{5}$ times the probability that you will get a flush if you don't. Seeing this card has really made a difference.

Worked example 2.47 *False positives*

After Stirzaker, p55. You have a blood test for a rare disease that occurs by chance in 1 person in 100, 000. If you have the disease, the test will report that you do with probability 0.95 (and that you do not with probability 0.05). If you do not have the disease, the test will report a false positive with probability 1e-3. If the test says you do have the disease, what is the probability it is correct?

Solution: Write S for the event you are sick and R for the event the test reports you are sick. We need $P(S|R)$. We have

$$\begin{aligned} P(S|R) &= \frac{P(R|S)P(S)}{P(R)} \\ &= \frac{P(R|S)P(S)}{P(R|S)P(S) + P(R|S^c)P(S^c)} \\ &= \frac{0.95 \times 1e-5}{0.95 \times 1e-5 + 1e-3 \times (1 - 1e-5)} \\ &= 0.0094 \end{aligned}$$

which should strike you as being a bit alarming. The disease is so rare that the test is almost useless.

Worked example 2.48 *False positives -2*

After Stirzaker, p55. You want to *design* a blood test for a rare disease that occurs by chance in 1 person in 100, 000. If you have the disease, the test will report that you do with probability p (and that you do not with probability $(1-p)$). If you do not have the disease, the test will report a false positive with probability q. You want to choose the value of p so that if the test says you have the disease, there is at least a 50% probability that you do.

Solution: Write S for the event you are sick and R for the event the test reports you are sick. We need $P(S|R)$. We have

$$\begin{aligned} P(S|R) &= \frac{P(R|S)P(S)}{P(R)} \\ &= \frac{P(R|S)P(S)}{P(R|S)P(S) + P(R|S^c)P(S^c)} \\ &= \frac{p \times 1e-5}{p \times 1e-5 + q \times (1-1e-5)} \\ &\geq 0.5 \end{aligned}$$

which means that $p \geq 99999q$ which should strike you as being very alarming indeed, because $p \leq 1$ and $q \geq 0$. One plausible pair of values is $q = 1e-5$, $p = 1-1e-5$. The test has to be spectacularly accurate to be of any use.

2.2.2 The Prosecutors Fallacy

It is quite easy to make mistakes in conditional probability. Several such mistakes have names, because they're so common. One is the **prosecutor's fallacy**. This often occurs in the following form: A prosecutor has evidence $\mathcal{E}$ against a suspect. Write $\mathcal{I}$ for the event that the suspect is innocent. The evidence has the property that $P(\mathcal{E}|\mathcal{I})$ is extremely small. The prosecutor argues that the suspect must be guilty, because $P(\mathcal{E}|\mathcal{I})$ is so small, and this is the fallacy.

The problem here is that the conditional probability of interest is $P(\mathcal{I}|\mathcal{E})$ (rather than $P(\mathcal{E}|\mathcal{I})$). The fact that $P(\mathcal{E}|\mathcal{I})$ is small doesn't mean that $P(\mathcal{I}|\mathcal{E})$ is small, because

$$P(\mathcal{I}|\mathcal{E}) = \frac{P(\mathcal{E}|\mathcal{I})P(\mathcal{I})}{P(\mathcal{E})} = \frac{P(\mathcal{E}|\mathcal{I})P(\mathcal{I})}{(P(\mathcal{E}|\mathcal{I})P(\mathcal{I}) + P(\mathcal{E}|\mathcal{I}^c)(1-P(\mathcal{I})))}$$

Notice how, if $P(\mathcal{I})$ is large or if $P(\mathcal{E}|\mathcal{I}^c)$ is much smaller than $P(\mathcal{E}|\mathcal{I})$, then $P(\mathcal{I}|\mathcal{E})$ could be close to one. The question to look at is not how unlikely the evidence is if the subject is innocent; instead, the question is how likely the subject is to be guilty compared to some other source of the evidence. These are two very different questions.

One useful analogy may be helpful. If you buy a lottery ticket ($\mathcal{L}$), the

probability of winning ($\mathcal{W}$) is small. So $P(\mathcal{W}|\mathcal{L})$ may be very small. But $P(\mathcal{L}|\mathcal{W})$ is 1 — the winner is always someone who bought a ticket.

The prosecutor's fallacy has contributed to a variety of miscarriages of justice. One famous incident involved a mother, Sally Clark, convicted of murdering two of her children. Expert evidence by paediatrician Roy Meadow argued that the probability of both deaths resulting from Sudden Infant Death Syndrome was extremely small. Her first appeal cited, among other grounds, statistical error in the evidence (you should spot the prosecutors fallacy; others were involved, too). The appeals court rejected this appeal, calling the statistical point "a sideshow". This prompted a great deal of controversy, both in the public press and various professional journals, including a letter from the then president of the Royal Statistical Society to the Lord Chancellor, pointing out that "*statistical evidence ... (should be) ... presented only by appropriately qualified statistical experts*". A second appeal (on other grounds) followed, and was successful. The appelate judges specifically criticized the statistical evidence, although it was not a point of appeal. Clark never recovered from this horrific set of events and died in tragic circumstances shortly after the second appeal. Roy Meadow was then struck off the rolls for serious professional misconduct as an expert witness, a ruling he appealed successfully. You can find a more detailed account of this case, with pointers to important documents including the letter to the Lord Chancellor (which is well worth reading), at `http://en.wikipedia.org/wiki/Roy_Meadow`; there is further material on the prosecutors fallacy at `http://en.wikipedia.org/wiki/Prosecutor\%27s_fallacy`.

2.2.3 Independence and Conditional Probability

As we have seen, two events are **independent** if

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

If two events $\mathcal{A}$ and $\mathcal{B}$ are independent, then

$$P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A})$$

and

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

Again, this means that knowing that $\mathcal{A}$ occurred tells you nothing about $\mathcal{B}$ — the probability that $\mathcal{B}$ will occur is the same whether you know that $\mathcal{A}$ occurred or not.

Events $\mathcal{A}_1 \ldots \mathcal{A}_n$ are **pairwise independent** if each pair is independent (i.e. $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent, etc.). They are **independent** if for any collection of distinct indices $i_1 \ldots i_k$ we have

$$P(\mathcal{A}_{i_1} \cap \ldots \cap \mathcal{A}_{i_k}) = P(\mathcal{A}_{i_1}) \ldots P(\mathcal{A}_{i_k})$$

Notice that independence is a much stronger assumption than pairwise independence.

Worked example 2.49 *Cards and pairwise independence*

We draw three cards from a properly shuffled standard deck, with replacement and reshuffling (i.e., draw a card, make a note, return to deck, shuffle, draw the next, make a note, shuffle, draw the third). Let $\mathcal{A}$ be the event that "card 1 and card 2 have the same suit"; let $\mathcal{B}$ be the event that "card 2 and card 3 have the same suit"; let $\mathcal{C}$ be the event that "card 1 and card 3 have the same suit". Show these events are pairwise independent, but not independent.

Solution: By counting, you can check that $P(\mathcal{A}) = 1/4$; $P(\mathcal{B}) = 1/4$; and $P(\mathcal{A} \cap \mathcal{B}) = 1/16$, so that these two are independent. This argument works for other pairs, too. But $P(\mathcal{C} \cap \mathcal{A} \cap \mathcal{B}) = 1/16$ which is not $1/4^3$, so the events are not independent; this is because the third event is logically implied by the first two.

We usually do not have the information required to prove that events are independent. Instead, we use intuition (for example, two flips of the same coin are likely to be independent unless there is something very funny going on) or simply choose to apply models in which some variables are independent.

Some events are pretty obviously independent. On other occasions, one needs to think about whether they are independent or not. Sometimes, it is reasonable to choose to model events as being independent, even though they might not be exactly independent. In cases like this, it is good practice to state your assumptions, because it helps you to keep track of what your model means. For example, we have worked with the event that a person, selected fairly and randomly from a set of people in a room, has a birthday on a particular day of the year. We assumed that, for different people, the events are independent. This seems like a fair assumption, but one might want to be cautious if you know that the people in the room are drawn from a population where multiple births are common.

Independent events can lead very quickly to very small probabilities, as we saw in example 31. This can mislead intuition quite badly, and lead to serious problems. In particular, these small probabilities can interact with the prosecutor's fallacy in a dangerous way. In example 31, we saw how the probability of getting a chance match in a large DNA database could be quite big, even though the probability of a single match is small. One version of the prosecutors fallacy is to argue that, because the probability of a single match is small, the person who matched the DNA must have committed the crime. The fallacy is to ignore the fact that the probability of a chance match to a large database is quite high.

People quite often reason poorly about independent events. The most common problem is known as the **gambler's fallacy**. This occurs when you reason that the probability of an independent event has been changed by previous outcomes. For example, imagine I toss a coin that is known to be fair 20 times and get 20 heads. The probability that the next toss will result in a head has not changed at all — it is still 0.5 — but many people will believe that it has changed. This idea is also sometimes referred to as **antichance**.

It might in fact be sensible to behave as if you're committing some version of

the gambler's fallacy in real life, because you hardly ever know for sure that your model is right. So in the coin tossing example, if the coin wasn't known to be fair, it might be reasonable to assume that it has been weighted in some way, and so to believe that the more heads you see, the more likely you will see a head in the next toss. At time of writing, Wikipedia has some fascinating stories about the gambler's fallacy; apparently, in 1913, a roulette wheel in Monte Carlo produced black 26 times in a row, and gamblers lost an immense amount of money betting on red. Here the gambler's reasoning seems to have been that the universe should ensure that probabilities produce the right frequencies in the end, and so will adjust the outcome of the next spin of the wheel to balance the sums. This is an instance of the gambler's fallacy. However, the page also contains the story of one Joseph Jagger, who hired people to keep records of the roulette wheels, and notice that one wheel favored some numbers (presumably because of some problem with balance). He won a lot of money, until the casino started more careful maintenance on the wheels. This isn't the gambler's fallacy; instead, he noticed that the numbers implied that the wheel was not a fair randomizer. He made money because the casino's odds on the bet assumed that it was fair.

2.3 EXAMPLE: THE MONTY HALL PROBLEM

Careless thinking about probability, particularly conditional probability, can cause wonderful confusion. The Monty Hall problem is a relatively simple exercise in conditional probability. Nonetheless, it has been the subject of extensive, lively, and often quite inaccurate correspondence in various national periodicals — it seems to catch the attention, which is why we describe it in some detail. The problem works like this: There are three doors. Behind one is a car. Behind each of the others is a goat. The car and goats are placed randomly and fairly, so that the probability that there is a car behind each door is the same. You will get the object that lies behind the door you choose at the end of the game. The goats are interchangeable, and, for reasons of your own, you would prefer the car to a goat.

The game goes as follows. You select a door. The host then opens a door and shows you a goat. You must now choose to either keep your door, or switch to the other door. What should you do?

You cannot tell what to do, by the following argument. Label the door you chose at the start of the game 1; the other doors 2 and 3. Write C_i for the event that the car lies behind door i. Write G_m for the event that a goat is revealed behind door m, where m is the number of the door where the goat was revealed (which could be 1, 2, or 3). You need to know $P(C_1|G_m)$. But

$$P(C_1|G_m) = \frac{P(G_m|C_1)P(C_1)}{P(G_m|C_1)P(C_1) + P(G_m|C_2)P(C_2) + P(G_m|C_3)P(C_3)}$$

and you do not know $P(G_m|C_1)$, $P(G_m|C_2)$, $P(G_m|C_3)$, because you don't know the rule by which the host chooses which door to open to reveal a goat. Different rules lead to quite different analyses.

There are several possible rules for the host to show a goat:

- **Rule 1:** choose a door uniformly at random.

- **Rule 2:** choose from the doors with goats behind them *that are not door 1* uniformly and at random.
- **Rule 3:** if the car is at 1, then choose 2; if at 2, choose 3; if at 3, choose 1.
- **Rule 4:** choose from the doors with goats behind them uniformly and at random.

We should keep track of the rules in the conditioning, so we write $P(G_m|C_1, r_1)$ for the conditional probability that a goat was revealed behind door m when the car is behind door 1, using rule 1 (and so on). This means we are interested in

$$P(C_1|G_m, r_n) = \frac{P(G_m|C_1, r_n)P(C_1)}{P(G_m|C_1, r_n)P(C_1) + P(G_m|C_2, r_n)P(C_2) + P(G_m|C_3, r_n)P(C_3)}.$$

Worked example 2.50 *Monty Hall, rule one*

Assume the host uses rule one, and shows you a goat behind door two. What is $P(C_1|G_2, r_1)$?

Solution: To work this out, we need to know $P(G_2|C_1, r_1)$, $P(G_2|C_2, r_1)$ and $P(G_2|C_3, r_1)$. Now $P(G_2|C_2, r_1)$ must be zero, because the host could not reveal a goat behind door two if there was a car behind that door. Write O_2 for the event the host *chooses* to open door two, and B_2 for the event there happens to be a goat behind door two. These two events are independent — the host chose the door uniformly at random. We can compute

$$\begin{aligned} P(G_2|C_1, r_1) &= P(O_2 \cap B_2|C_1, r_1) \\ &= P(O_2|C_1, r_1)P(B_2|C_1, r_1) \\ &= (1/3)(1) \\ &= 1/3 \end{aligned}$$

where $P(B_2|C_1, r_1) = 0$ because we conditioned on the fact there was a car behind door one, so there is a goat behind each other door. This argument establishes $P(G_2|C_3, r_1) = 1/3$, too. So $P(C_1|G_2, r_1) = 1/2$ — the host showing you the goat does not motivate you to do anything, because if $P(C_3|G_2, r_1) = 1/2$, too.

Worked example 2.51 *Monty Hall, rule two*

Assume the host uses rule two, and shows you a goat behind door two. What is $P(C_1|G_2, r_2)$?

Solution: To work this out, we need to know $P(G_2|C_1, r_2)$, $P(G_2|C_2, r_2)$ and $P(G_2|C_3, r_2)$. Now $P(G_2|C_2, r_2) = 0$, because the host chooses from doors with goats behind them. $P(G_2|C_1, r_2) = 1/2$, because the host chooses uniformly and at random from doors with goats behind them that are not door one; if the car is behind door one, there are two such doors. $P(G_2|C_3, r_2) = 1$, because there is only one door that (a) has a goat behind it and (b) isn't door one. Plug these numbers into the formula, to get $P(C_1|G_2, r_2) = 1/3$. This is the source of all the fuss. It says that, if you know the host is using rule two, you should switch doors if the host shows you a goat behind door two (because $P(C_3|G_2, r_2) = 2/3$).

Notice what is happening: if the car is behind door three, then the *only* choice of goat for the host is the goat behind two. So by choosing a door under rule two, the host is signalling some information to you, which you can use. By using rule three, the host can tell you precisely where the car is (exercises).

Many people find the result of example 51 counterintuitive, and object (sometimes loudly, in newspaper columns, letters to the editor, etc.). One example that some people find helpful is an extreme case. Imagine that, instead of three doors, there are 1002. The host is using rule two, modified in the following way: open all but one of the doors that are not door one, choosing only doors that have goats behind them to open. You choose door one; the host opens 1000 doors — say, all but doors one and 1002. What would you do?

2.4 WHAT YOU SHOULD REMEMBER

You should be able to:

- Write out a set of outcomes for an experiment.
- Construct an event space.
- Compute the probabilities of outcomes and events.
- Determine when events are independent.
- Compute the probabilities of outcomes by counting events, when the count is straightforward.
- Compute a conditional probability.

You should remember:

- The definition of an event space.
- The properties of a probability function.

- The definition of independence.
- The definition of conditional probability.

PROBLEMS

Outcomes

2.1. You roll a four sided die. What is the space of outcomes?

2.2. King Lear decides to allocate three provinces (1, 2, and 3) to his daughters (Goneril, Regan and Cordelia - read the book) at random. Each gets one province. What is the space of outcomes?

2.3. You randomly wave a flyswatter at a fly. What is the space of outcomes?

2.4. You read the book, so you know that King Lear had family problems. As a result, he decides to allocate two provinces to one daughter, one province to another daughter, and no provinces to the third. Because he's a bad problem solver, he does so at random. What is the space of outcomes?

Probability of outcomes

2.5. You roll a fair four sided die. What is the probability of getting a 3?

2.6. You roll a fair four sided die, and then a fair six sided die. You add the numbers on the two dice. What is the probability the result is even?

2.7. You roll a fair 20 sided die. What is the probability of getting an even number?

2.8. You roll a fair five sided die. What is the probability of getting an even number?

Events

2.9. At a particular University, 1/2 of the students drink alcohol and 1/3 of the students smoke cigarettes.

(a) What is the largest possible fraction of students who do neither?

(b) It turns out that, in fact, 1/3 of the students do neither. What fraction of the students does both?

2.10. I flip two coins. What one set needs to be added to this collection of sets to form an event space?

$$\Sigma = \{\emptyset, \Omega, \{TH\}, \{HT, TH, TT\}, \{HH\}, \{HT, TT\}, \{HH, TH\}\}$$

Probability of Events

2.11. Assume each outcome in Ω has the same probability. In this case, show

$$P(\mathcal{E}) = \frac{\text{Number of outcomes in } \mathcal{E}}{\text{Total number of outcomes in } \Omega}$$

2.12. You flip a fair coin three times. What is the probability of seeing HTH? (i.e. Heads, then Tails, then Heads)

2.13. You flip a fair coin three times. What is the probability of seeing two heads and one tail?

2.14. You remove the king of hearts from a standard deck of cards, then shuffle it and draw a card.

(a) What is the probability this card is a king?

(b) What is the probability this card is a heart?

2.15. You shuffle a standard deck of cards, then draw four cards.
- **(a)** What is the probability all four are the same suit?
- **(b)** What is the probability all four are red?
- **(c)** What is the probability each has a different suit?

2.16. You roll three fair six-sided dice and add the numbers. What is the probability the result is even?

2.17. You roll three fair six-sided dice and add the numbers. What is the probability the result is even *and* not divisible by 20?

2.18. You shuffle a standard deck of cars, then draw seven cards. What is the probability that you see no aces?

2.19. Show that $P(\mathcal{A} - (\mathcal{B} \cup \mathcal{C})) = P(\mathcal{A}) - P(\mathcal{A} \cap \mathcal{B}) - P(\mathcal{A} \cap \mathcal{C}) + P(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C})$.

2.20. You draw a single card from a standard 52 card deck. What is the probability that it is red?

2.21. You remove all heart cards from a standard 52 card deck, then draw a single card from the result. What is the probability that the card you draw is red?

Conditional Probability

2.22. You roll two fair six-sided dice. What is the conditional probability the sum of numbers is greater than three, conditioned on the first die coming up even.

2.23. You take a standard deck of cards, shuffle it, and remove one card. You then draw a card.
- **(a)** What is the conditional probability that the card you draw is a red king, conditioned on the removed card being a king?
- **(b)** What is the conditional probability that the card you draw is a red king, conditioned on the removed card being a red king?
- **(c)** What is the conditional probability that the card you draw is a red king, conditioned on the removed card being a black ace?

2.24. A royal flush is a hand of five cards, consisting of Ace, King, Queen, Jack and 10 of a single suit. Poker players like this hand, but don't see it all that often.
- **(a)** You draw five cards from a standard deck of playing cards. What is the probability of getting a royal flush?
- **(b)** You draw three cards from a standard deck of playing cards. These are Ace, King, Queen of hearts. What is the probability that the next two cards you draw will result in a getting a royal flush? (this is the conditional probability of getting a royal flush, conditioned on the first three cards being AKQ of hearts).

2.25. You roll a fair five-sided die, and a fair six-sided die.
- **(a)** What is the probability that the sum of numbers is even?
- **(b)** What is the conditional probability that the sum of numbers is even, conditioned on the six-sided die producing an odd number?

Independence

2.26. You take a standard deck of cards, shuffle it, and remove both red kings. You then draw a card.
- **(a)** Is the event {card is red} independent of the event {card is a queen}?
- **(b)** Is the event {card is black} independent of the event {card is a king}?

The Monty Hall Problem

2.27. **Monty Hall, Rule 3:** If the host uses rule 3, then what is $P(C_1|G_2, r_3)$? Do this by computing conditional probabilities.

2.28. **Monty Hall, Rule 4:** If the host uses rule 4, and shows you a goat behind door 2, what is $P(C_1|G_2, r_4)$? Do this by computing conditional probabilities.

CHAPTER 3

Random Variables and Expectations

3.1 RANDOM VARIABLES

Quite commonly, we would like to deal with numbers that are random. We can do so by linking numbers to the outcome of an experiment. We define a **random variable**:

Definition 3.1 *Discrete random variable*

Given a sample space Ω, a set of events $\mathcal{F}$, and a probability function P, and a countable set of of real numbers D, a discrete random variable is a function with domain Ω and range D.

This means that for any outcome ω there is a number $X(\omega)$. P will play an important role, but first we give some examples.

Example: *Numbers from coins*

We flip a coin. Whenever the coin comes up heads, we report 1; when it comes up tails, we report 0. This is a random variable.

Example: *Numbers from coins II*

We flip a coin 32 times. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 32 bit random number, which is a random variable.

Example: *The number of pairs in a poker hand*

(from Stirzaker). We draw a hand of five cards. The number of pairs in this hand is a random variable, which takes the values $0, 1, 2$ (depending on which hand we draw)

A function of a discrete random variable is also a discrete random variable.

Example: *Parity of coin flips*

We flip a coin 32 times. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 32 bit random number, which is a random variable. The parity of this number is also a random variable.

Associated with any value x of the random variable X are a series of events. The most important is the set of outcomes such that $X = x$, which we can write

$\{\omega : X(\omega) = x\}$; it is usual to simplify to $\{X = x\}$, and we will do so. The probability that a random variable X takes the value x is given by $P(\{X = x\})$. This is sometimes written as $P(X = x)$, and rather often written as $P(x)$.

We could also be interested in the set of outcomes such that $X \leq x$ (i.e. in $\{\omega : X(\omega) \leq x\}$), which we will write $\{X \leq x\}$; The probability that X takes the value x is given by $P(\{X \leq x\})$. This is sometimes written as $P(X \leq x)$. Similarly, we could be interested in $\{X > x\}$, and so on.

Definition 3.2 *The probability distribution of a discrete random variable*

The probability distribution of a discrete random variable is the set of numbers $P(\{X = x\})$ for each value x that X can take. The distribution takes the value 0 at all other numbers. Notice that this is non-negative.

Definition 3.3 *The cumulative distribution of a discrete random variable*

The cumulative distribution of a discrete random variable is the set of numbers $P(\{X <= x\})$ for each value x that X can take. Notice that this is a non-decreasing function of x. Cumulative distributions are often written with an f, so that $f(x)$ might mean $P(\{X <= x\})$.

Worked example 3.1 *Numbers from coins III*

We flip a biased coin 2 times. The flips are independent. The coin has $P(H) = p$, $P(T) = 1 - p$. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 2 bit random number, which is a random variable taking the values 0, 1, 2, 3. What is the probability distribution and cumulative distribution of this random variable?

Solution: Probability distribution: $P(0) = (1-p)^2$; $P(1) = (1-p)p$; $P(2) = p(1-p)$; $P(3) = p^2$. Cumulative distribution: $f(0) = (1-p)^2$; $f(1) = (1-p)$; $f(2) = p(1-p) + (1-p) = (1-p^2)$; $f(3) = 1$.

Worked example 3.2 *Betting on coins*

One way to get a random variable is to think about the reward for a bet. We agree to play the following game. I flip a coin. The coin has $P(H) = p$, $P(T) = 1 - p$. If the coin comes up heads, you pay me q; if the coin comes up tails, I pay you r. The number of dollars that change hands is a random variable. What is its probability distribution?

Solution: We see this problem from my perspective. If the coin comes up heads, I get q; if it comes up tails, I get $-r$. So we have $P(X = q) = p$ and $P(X = -r) = (1 - p)$, and all other probabilities are zero.

3.1.1 Joint and Conditional Probability for Random Variables

All the concepts of probability that we described for events carry over to random variables. This is as it should be, because random variables are really just a way of getting numbers out of events. However, terminology and notation change a bit.

Assume we have two random variables X and Y. The probability that X takes the value x and Y takes the value y could be written as $P(\{X = x\} \cap \{Y = y\})$. It is more usual to write it as $P(x, y)$. You can think of this as a table of values, one for each possible pair of x and y values. This table is usually referred to as the **joint probability distribution** of the random variables. Nothing (except notation) has really changed here, but the change of notation is useful.

We will simplify notation further. Usually, we are interested in random variables, rather than potentially arbitrary outcomes or sets of outcomes. We will write $P(X)$ to denote the probability distribution of a random variable, and $P(x)$ or $P(X = x)$ to denote the probability that that random variable takes a particular value. This means that, for example, the rule we could write as

$$P(\{X = x\} \mid \{Y = y\})P(\{Y = y\}) = P(\{X = x\} \cap \{Y = y\})$$

will be written as

$$P(x|y)P(y) = P(x, y).$$

This yields **Bayes' rule**, which is important enough to appear in its own box.

Definition 3.4 *Bayes' rule*

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Random variables have another useful property. If $x_0 \neq x_1$, then the event $\{X = x_0\}$ must be disjoint from the event $\{X = x_1\}$. This means that

$$\sum_x P(x) = 1$$

and that, for any y,

$$\sum_x P(x|y) = 1$$

(if you're uncertain on either of these points, check them by writing them out in the language of events).

Now assume we have the joint probability distribution of two random variables, X and Y. Recall that we write $P(\{X = x\} \cap \{Y = y\})$ as $P(x, y)$. Now consider the sets of outcomes $\{Y = y\}$ for each different value of y. These sets must be disjoint, because y cannot take two values at the same time. Furthermore, each element of the set of outcomes $\{X = x\}$ must lie in one of the sets $\{Y = y\}$. So we have

$$\sum_y P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\})$$

which is usually written as

$$\sum_y P(x, y) = P(x)$$

and is often referred to as the **marginal probability** of X.

Definition 3.5 *Independent random variables*

The random variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent. This means that

$$P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\})P(\{Y = y\}),$$

which we can rewrite as

$$P(x, y) = P(x)P(y)$$

Worked example 3.3 *Sums and differences of dice*

You throw two dice. The number of spots on the first die is a random variable (call it X); so is the number of spots on the second die (Y). Now define $S = X + Y$ and $D = X - Y$. What is the probability distribution of S and of D?

Solution: S can have values in the range $2, \ldots, 12$. There is only one way to get a $S = 2$; two ways to get $S = 3$; and so on. Using the methods of chapter 22 for each case, the probabilities for $[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$ are $[1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1]/36$. Similarly, D can have values in the range $-5, \ldots, 5$. Again, using the methods of chapter 22, the probabilities for $[-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]$ are $[1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1]/36$.

Worked example 3.4 *Sums and differences of dice, II*

Using the terminology of example 3, what is the joint probability distribution of S and D?

Solution: This is more interesting to display, because it's an 11x11 table. Each entry of the table represents a pair of S, D values. Many pairs can't occur (for example, for $S = 2$, D can only be zero; if S is even, then D must be even; and so on). You can work out the table by checking each case; it's in Table 3.1.

Worked example 3.5 *Sums and differences of dice, III*

Using the terminology of example 3, are X and Y independent? are S and D independent?

Solution: X and Y are clearly independent. But S and D are not. There are several ways to see this. One way is to notice that, if you know $S = 2$, then you know the value of D precisely; but if you know $S = 3$, D could be either 1 or -1. This means that $P(S|D)$ depends on D, so they're not independent. Another way is to notice that the rank of the table, as a matrix, is 6, which means that it can't be the outer product of two vectors.

Worked example 3.6 *Sums and differences of dice, IV*

Using the terminology of example 3, what is $P(S|D = 0)$? what is $P(D|S = 11)$?

Solution: You could work it out either of these from the table, or by first principles. If $D = 0$, S can have values $2, 4, 6, 8, 10, 12$, and each value has conditional probability $1/6$. If $S = 11$, D can have values 1, or -1, and each value has conditional probability $1/2$.

3.1.2 Just a Little Continuous Probability

Our random variables take values from a discrete set of numbers D. This makes the underlying machinery somewhat simpler to describe, and is often, but not always, enough for model building. Some phenomena are more naturally modelled as being continuous — for example, human height; human weight; the mass of a distant star; and so on. Giving a complete formal description of probability on a continuous space is surprisingly tricky, and would involve us in issues that do not arise much in practice.

These issues are caused by two interrelated facts: real numbers have infinite precision; and you can't count real numbers. A continuous random variable is still a random variable, and comes with all the stuff that a random variable comes with. We will not speculate on what the underlying sample space is, nor on the underlying events. This can all be sorted out, but requires moderately heavy lifting that isn't

$$\frac{1}{36} \times \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

TABLE 3.1: *A table of the joint probability distribution of S (vertical axis; scale $2, \ldots, 12$) and D (horizontal axis; scale $-5, \ldots, 5$) from example 4*

particularly illuminating for us. The most interesting thing for us is specifying the probability distribution. Rather than talk about the probability that a real number takes a particular value (which we can't really do satisfactorily most of the time), we will instead talk about the probability that it lies in some interval. So we can specify a probability distribution for a continuous random variable by giving a set of (very small) intervals, and for each interval providing the probability that the random variable lies in this interval.

The easiest way to do this is to supply a **probability density function**. Let $p(x)$ be a probability density function for a continuous random variable X. We interpret this function by thinking in terms of small intervals. Assume that dx is an infinitesimally small interval. Then

$$p(x)dx = P(\{\text{event that } X \text{ takes a value in the range } [x, x + dx]\}).$$

Important properties of probability density functions follow from this definition.

Useful Facts 3.1 *Probability density functions*

- Probability density functions are non-negative. This follows from the definition; a negative value at some x would imply a negative probability.
- For $a < b$

  $$P(\{\text{event that } X \text{ takes a value in the range } [a, b]\}) = \int_a^b p(x)dx.$$

  which we obtain by summing $p(x)dx$ over all the infinitesimal intervals between a and b.
- We must have that

  $$\int_{-\infty}^{\infty} p(x)dx = 1.$$

  This is because

  $$P(\{X \text{ takes a value in the range } [-\infty, \infty]\}) = 1 = \int_{-\infty}^{\infty} p(x)dx$$

- Probability density functions are usually called pdf's.
- It is quite usual to write all pdf's as lower-case p's. If one specifically wishes to refer to probability, one writes an upper case P, as in the previous points.

One good way to think about pdf's is as the limit of a histogram. Imagine you collect an arbitrarily large dataset of data items, each of which is independent. You build a histogram of that dataset, using arbitrarily narrow boxes. You scale the histogram so that the sum of the box areas is one. The result is a probability density function.

The pdf doesn't represent the probability that a random variable takes a value. Instead, you should think of $p(x)$ as being the limit of a ratio (which is why it's called a density):

$$\frac{\text{the probability that the random variable will lie in a small interval centered on } x}{\text{the length of the small interval centered on } x}$$

Notice that, while a pdf has to be non-negative, and it has to integrate to 1, it does *not* have to be smaller than one. A ratio like this could be a lot larger than one, as long as it isn't larger than one for too many x (because the integral must be one).

Probability density functions can be moderately strange functions.

Worked example 3.7 *Strange probability density functions*

There is some (small!) voltage over the terminals of a warm resistor caused by noise (electrons moving around in the heat and banging into one another). This is a good example of a continuous random variable, and we can assume there is some probability density function for it, say $p(x)$. We assume that $p(x)$ has the property that

$$\lim_{\epsilon \to 0} \int_{v-\epsilon}^{v+\epsilon} p(x)dx = 0$$

which is what you'd expect for any function you're likely to have dealt with. Now imagine I define a new random variable by the following procedure: I flip a coin; if it comes up heads, I report 0; if tails, I report the voltage over the resistor. This random variable, u, has a probability $1/2$ of taking the value 0, and $1/2$ of taking a value from $p(x)$. Write this random variable's probability density function $q(u)$. Compute

$$\lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} q(u)du$$

Solution: We can do this from the definition. We have

$$P(\{u \in [-\epsilon, \epsilon]\}) = \int_{-\epsilon}^{\epsilon} q(u)du.$$

But u will take the value 0 with probability $1/2$, and otherwise takes the value over the resistor. So

$$P(\{u \in [-\epsilon, \epsilon]\}) = 1/2 + \int_{-\epsilon}^{\epsilon} p(x)dx.$$

and

$$\lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} q(u)du = \lim_{\epsilon \to 0} P(\{u \in [-\epsilon, \epsilon]\}) = 1/2.$$

This means $q(x)$ has the property that

$$\lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} q(u)du = 1/2,$$

which means that $q(u)$ is displaying quite unusual behavior at $u = 0$. We will not need to deal with probability density functions that behave like this; but you should be aware of the possibility.

Every probability density function $p(x)$ has the property that $\int_{-\infty}^{\infty} p(x)dx = 1$; this is useful, because when we are trying to determine a probability density function, we can ignore a constant factor. So if $g(x)$ is a non-negative function that

is proportional to the probability density function (often pdf) we are interested in, we can recover the pdf by computing

$$p(x) = \frac{1}{\int_{-\infty}^{\infty} g(x)dx} g(x).$$

This procedure is sometimes known as **normalizing**, and $\int_{-\infty}^{\infty} g(x)dx$ is the **normalizing constant**.

3.2 EXPECTATIONS AND EXPECTED VALUES

Imagine we play the game of example 2 multiple times. Our frequency definition of probability means that in N games, we expect to see about pN heads and $(1-p)N$ tails. In turn, this means that my total income from these N games should be about $(pN)q - ((1-p)N)r$. The N in this expression is inconvenient; instead, we could say that for any single game, my income is

$$pq - (1-p)r.$$

This isn't the actual income from a single game (which would be either q or $-r$, depending on what the coin did). Instead, it's an estimate of what would happen over a large number of games, on a per-game basis. This is an example of an **expected value**.

3.2.1 Expected Values of Discrete Random Variables

Definition 3.6 *Expected value*

Given a discrete random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

$$\mathbb{E}[X] = \sum_{x \in \mathcal{D}} xP(X = x).$$

This is sometimes written which is $\mathbb{E}_P[X]$, to clarify which distribution one has in mind

Notice that an expected value could take a value that the random variable doesn't take.

Example: *Betting on coins*

We agree to play the following game. I flip a fair coin (i.e. $P(H) = P(T) = 1/2$). If the coin comes up heads, you pay me 1; if the coin comes up tails, I pay you 1. The expected value of my income is 0, even though the random variable never takes that value.

Definition 3.7 *Expectation*

Assume we have a function f that maps a discrete random variable X into a set of numbers $\mathcal{D}_f$. Then $f(x)$ is a discrete random variable, too, which we write F. The expected value of this random variable is written

$$\mathbb{E}[f] = \sum_{u \in \mathcal{D}_f} uP(F = u) = \sum_{x \in \mathcal{D}} f(x)P(X = x)$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".

Expectations are linear, so that $\mathbb{E}[0] = 0$ and $\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B]$. The expectation of a constant is that constant (or, in notation, $\mathbb{E}[k] = k$), because probabilities sum to 1. Because probabilities are non-negative, the expectation of a non-negative random variable must be non-negative.

3.2.2 Expected Values of Continuous Random Variables

We can compute expectations for continuous random variables, too, though summing over all values now turns into an integral. This should be expected. Imagine you choose a set of closely spaced values for x which are x_i, and then think about x as a discrete random variable. The values are separated by steps of width Δx. Then the expected value of this discrete random variable is

$$\mathbb{E}[X] = \sum_i x_i P(X \in \text{interval centered on } x_i) = \sum_i x_i p(x_i)\Delta x$$

and, as the values get closer together and Δx gets smaller, the sum limits to an integral.

Definition 3.8 *Expected value*

Given a continuous random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

$$\mathbb{E}[X] = \int_{x \in \mathcal{D}} xp(x)dx.$$

This is sometimes written $\mathbb{E}_p[X]$, to clarify which distribution one has in mind

The expected value of a continuous random variable could be a value that the random variable doesn't take, too. Notice one attractive feature of the $\mathbb{E}[X]$ notation; we don't need to make any commitment to whether X is a discrete random

variable (where we would write a sum) or a continuous random variable (where we would write an integral). The reasoning by which we turned a sum into an integral works for functions of continuous random variables, too.

Definition 3.9 *Expectation*

Assume we have a function f that maps a continuous random variable X into a set of numbers $\mathcal{D}_f$. Then $f(x)$ is a continuous random variable, too, which we write F. The expected value of this random variable is

$$\mathbb{E}[f] = \int_{x \in \mathcal{D}} f(x)p(x)dx$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".

Again, for continuous random variables, expectations are linear, so that $\mathbb{E}[0] = 0$ and $\mathbb{E}[A + B] = \mathbb{E}[A]+\mathbb{E}[B]$. The expectation of a constant is that constant (or, in notation, $\mathbb{E}[k] = k$), because probabilities sum to 1. Because probabilities are non-negative, the expectation of a non-negative random variable must be non-negative.

3.2.3 Mean, Variance and Covariance

There are three very important expectations with special names.

Definition 3.10 *The mean or expected value*

The mean or expected value of a random variable X is

$$\mathbb{E}[X]$$

Definition 3.11 *The variance*

The variance of a random variable X is

$$\mathsf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

Notice that

$$\begin{aligned}\mathbb{E}\left[(X-\mathbb{E}[X])^2\right] &= \mathbb{E}\left[\left(X^2-2X\mathbb{E}[X]+\mathbb{E}[X]^2\right)\right] \\ &= \mathbb{E}\left[X^2\right]-2\mathbb{E}[X]\mathbb{E}[X]+\mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right]-(\mathbb{E}[X])^2\end{aligned}$$

Definition 3.12 *The covariance*

The covariance of two random variables X and Y is

$$\mathsf{cov}\,(X,Y)=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$

Notice that

$$\begin{aligned}\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] &= \mathbb{E}[(XY-Y\mathbb{E}[X]-X\mathbb{E}[Y]+\mathbb{E}[X]\mathbb{E}[Y])] \\ &= \mathbb{E}[XY]-2\mathbb{E}[Y]\mathbb{E}[X]+\mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

We also have $\mathsf{Var}[X]=\mathsf{cov}\,(X,X)$.

Now assume that we have a probability distribution $P(X)$ defined on some discrete set of numbers. There is some random variable that produced this probability distribution. This means that we could talk about the mean of a probability distribution P (rather than the mean of a random variable whose probability distribution is $P(X)$). It is quite usual to talk about the mean of a probability distribution. Furthermore, we could talk about the variance of a probability distribution P (rather than the variance of a random variable whose probability distribution is $P(X)$).

Worked example 3.8 *variance*

Can a random variable have $\mathbb{E}[X]>\sqrt{\mathbb{E}[X^2]}$?

Solution: No, because that would mean that $\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]<0$. But this is the expected value of a non-negative quantity; it must be non-negative.

Worked example 3.9 *More variance*

We just saw that a random variable can't have $\mathbb{E}[X]>\sqrt{\mathbb{E}[X^2]}$. But I can easily have a random variable with large mean and small variance - isn't this a contradiction?

Solution: No, you're confused. Your question means you think that the variance of X is given by $\mathbb{E}\left[X^2\right]$; but actually $\mathsf{Var}[X]=\mathbb{E}\left[X^2\right]-\mathbb{E}[X]^2$

Worked example 3.10 *Mean of a coin flip*

We flip a biased coin, with $P(H) = p$. The random variable X has value 1 if the coin comes up heads, 0 otherwise. What is the mean of X? (i.e. $\mathbb{E}[X]$).

Solution: $\mathbb{E}[X] = \sum_{x \in D} xP(X = x) = 1p + 0(1 - p) = p$

Useful Facts 3.2 *Expectations*

1. $\mathbb{E}[0] = 0$
2. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
3. $\mathbb{E}[kX] = k\mathbb{E}[X]$
4. $\mathbb{E}[1] = 1$
5. if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
6. if X and Y are independent, then $\mathsf{cov}\,(X, Y) = 0$.

All but 5 and 6 are obvious from the definition. If 5 is true, then 6 is obviously true. I prove 5.

Proposition: *If X and Y are independent random variables, then* $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof: Recall that $\mathbb{E}[X] = \sum_{x \in D} xP(X = x)$, so that

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{(x,y) \in D_x \times D_y} xyP(X = x, Y = y) \\
&= \sum_{x \in D_x} \sum_{y \in D_y} (xyP(X = x, Y = y)) \\
&= \sum_{x \in D_x} \sum_{y \in D_y} (xyP(X = x)P(Y = y)) \\
&\quad \text{because } X \text{ and } Y \text{ are independent} \\
&= \sum_{x \in D_x} \sum_{y \in D_y} (xP(X = x))\,(yP(Y = y)) \\
&= \left(\sum_{x \in D_x} xP(X = x) \right) \left(\sum_{y \in D_y} yP(Y = y) \right) \\
&= (\mathbb{E}[X])(\mathbb{E}[Y]).
\end{aligned}$$

This is certainly not true when X and Y are not independent (try $Y = -X$).

Useful Facts 3.3 *Variance*

It is quite usual to write $\mathsf{Var}[X]$ for the variance of the random variable X.

1. $\mathsf{Var}[0] = 0$
2. $\mathsf{Var}[1] = 0$
3. $\mathsf{Var}[X] \geq 0$
4. $\mathsf{Var}[kX] = k^2\mathsf{Var}[X]$
5. if X and Y are independent, then $\mathsf{Var}[X + Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$

1, 2, 3 are obvious. You will prove 4 and 5 in the exercises.

Worked example 3.11 *Variance of a coin flip*

We flip a biased coin, with $P(H) = p$. The random variable X has value 1 if the coin comes up heads, 0 otherwise. What is the variance of X? (i.e. $\mathsf{Var}[X]$).

Solution: $\mathsf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = (1p - 0(1 - p)) - p^2 = p(1 - p)$

The variance of a random variable is often inconvenient, because its units are the square of the units of the random variable. Instead, we could use the **standard deviation**.

Definition 3.13 *Standard deviation*

The **standard deviation** of a random variable X is defined as

$$\mathsf{std}(X) = \sqrt{\mathsf{Var}[X]}$$

You do need to be careful with standard deviations. If X and Y are independent random variables, then $\mathsf{Var}[X + Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$, but $\mathsf{std}(X + Y) = \sqrt{\mathsf{std}(X)^2 + \mathsf{std}(Y)^2}$. One way to avoid getting mixed up is to remember that variances add, and derive expressions for standard deviations from that.

3.2.4 Expectations and Statistics

You should have noticed we now have two notions each for mean, variance, covariance, and standard deviation. One, which we expounded in sections 22, describes datasets. We will call these **descriptive statistics**. The other, described above, is a property of probability distributions. We will call these **expectations**. In each case, the reason we have one name for two notions is that the notions are not really

all that different.

Imagine we have a dataset $\{\mathbf{x}\}$ of N items, where the i'th item is $\mathbf{x}_i$. We can build a probability distribution out of this dataset, by placing a probability on each data item. We will give each data item the same probability (which must be $1/N$, so all probabilities add to 1). Write $\mathbb{E}[\mathbf{x}]$ for the mean of this distribution. We have

$$\mathbb{E}[\mathbf{x}] = \sum_i \mathbf{x}_i p(\mathbf{x}_i) = \frac{1}{N} \sum_i \mathbf{x}_i = \text{mean}\left(\{\mathbf{x}\}\right).$$

The variances, standard deviations and covariance have the same property: For this particular distribution (sometimes called the **empirical distribution**), the expectations have the same value as the descriptive statistics (exercises).

In section 22, we will see a form of converse to this fact. Imagine we have a dataset that consists of independent, identically distributed samples from a probability distribution. That is, we know that each data item was obtained independently from the distribution. For example, we might have a count of heads in each of a number of coin flip experiments. Then the descriptive statistics will turn out to be accurate estimates of the expectations.

3.2.5 Indicator Functions

It is sometimes convenient when working with random variables to use **indicator functions**. This is a function that is one when some condition is true, and zero otherwise. The reason they are useful is that their expectated values have interesting properties.

Definition 3.14 *Indicator functions*

An indicator function for an event is a function that takes the value zero for values of X where the event does not occur, and one where the event occurs. For the event $\mathcal{E}$, we write

$$\mathbb{I}_{[\mathcal{E}]}(X)$$

for the relevant indicator function.

For example,

$$\mathbb{I}_{[\{|X|\}\leq a]}(X) = \begin{cases} 1 & \text{if } -a < X < a \\ 0 & \text{otherwise} \end{cases}$$

Indicator functions have one useful property.

$$\mathbb{E}\left[\mathbb{I}_{[\mathcal{E}]}\right] = P(\mathcal{E})$$

which you can establish by checking the definition of expectations.

3.2.6 Two Inequalities

Mean and variance tell us quite a lot about a random variable, as two important inequalities show.

Definition 3.15 *Markov's inequality*

Markov's inequality is

$$P(\{\| X \| \geq a\}) \leq \frac{\mathbb{E}[\| X \|]}{a}.$$

You should read this as indicating that a random variable is most unlikely to have an absolute value a lot larger than the mean of its absolute value. This should seem fairly intuitive from the definition of expectation. Recall that

$$\mathbb{E}[X] = \sum_{x \in D} xP(\{X = x\})$$

Assume that D contains only non-negative numbers (that absolute value). Then the only way to have a small value of $\mathbb{E}[X]$ is to be sure that, when x is large, $P(\{X = x\})$ is small. The proof is a rather more formal version of this observation, below.

Definition 3.16 *Chebyshev's inequality*

Chebyshev's inequality is

$$P(\{|X - \mathbb{E}[X]| \geq a\}) \leq \frac{\mathsf{Var}[X]}{a^2}.$$

It is common to see this in another form, obtained by writing σ for the standard deviation of X, substituting $k\sigma$ for a, and rearranging

$$P(\{|X - \mathbb{E}[X]| \geq k\sigma\}) \leq \frac{1}{k^2}$$

This means that the probability of a random variable taking a particular value must fall off rather fast as that value moves away from the mean, in units scaled to the variance. This probably doesn't seem intuitive from the definition of expectation. But think about it this way: values of a random variable that are many standard deviations above the mean must have low probability, otherwise the standard deviation would be bigger. The proof, again, is a rather more formal version of this observation, and appears below.

Proposition: *Markov's inequality*

$$P(\{\| X \| \geq a\}) \leq \frac{\mathbb{E}[\| X \|]}{a}.$$

Proof: (from Wikipedia). Notice that, for $a > 0$,

$$a\mathbb{I}_{[\{|X| \leq a\}]}(X) \leq |X|$$

(because if $|X| < a$, the LHS is a; otherwise it is zero). Now we have

$$\mathbb{E}\left[a\mathbb{I}_{[\{|X| \leq a\}]}\right] \leq \mathbb{E}[|X|]$$

but, because expectations are linear, we have

$$\mathbb{E}\left[a\mathbb{I}_{[\{|X| \leq a\}]}\right] = a\mathbb{E}\left[\mathbb{I}_{[\{|X| \leq a\}]}\right] = aP(\{|X| \leq a\})$$

and so we have

$$aP(\{|X| \leq a\}) \leq |X|$$

and we get the inequality by division, which we can do because $a > 0$.

Proposition: *Chebyshev's inequality*

$$P(\{|X - \mathbb{E}[X]| \geq a\}) \leq \frac{\mathsf{Var}[X]}{a^2}.$$

Proof: Write U for the random variable $(X - \mathbb{E}[X])^2$. Markov's inequality gives us

$$P(\{|U| \geq w\}) \leq \frac{\mathbb{E}[|U|]}{w}$$

Now notice that, if $a^2 = w$,

$$P(\{|U| \geq w\}) = P(\{|X - \mathbb{E}[X]|\} \geq a)$$

so we have

$$P(\{|U| \geq w\}) = P(\{|X - \mathbb{E}[X]|\} \geq a) \leq \frac{\mathbb{E}[|U|]}{w} = \frac{\mathsf{Var}[X]}{a^2}$$

3.2.7 IID Samples and the Weak Law of Large Numbers

Imagine a random variable X, obtained by flipping a fair coin and reporting 1 for an H and -1 for a T. We can talk about the probability distribution $P(X)$ of this random variable; we can talk about the expected value that the random variable takes; but the random variable itself doesn't have a value. However, if we actually

flip a coin, we get either a 1 or a -1. This number is often called a **sample** of the random variable (or of its probability distribution). Similarly, if we flipped a coin many times, we'd have a set of numbers (or samples). These numbers would be independent. Their histogram would look like $P(X)$. Collections of data items like this are important enough to have their own name.

Assume we have a set of data items x_i such that (a) they are independent; (b) each is produced from the same process; and (c) the histogram of a very large set of data items looks increasingly like the probability distribution $P(X)$ as the number of data items increases. Then we refer to these data items as **independent identically distributed samples** of $P(X)$; for short, **iid samples** or even just **samples**. For all of the cases we will deal with, it will be obvious how to get IID samples. However, it's worth knowing that obtaining IID samples from arbitrary probability distributions is very difficult.

Now assume we have a set of N IID samples x_i of a probability distribution $P(X)$. Write

$$X_N = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Now X_N is a random variable (the x_i are IID samples, and for a different set of samples you will get a different, random, X_N).

Definition 3.17 *The Weak Law of Large Numbers*

The **weak law of large numbers** states that, for any positive number ϵ

$$\lim_{N \to \infty} P(\{\| X_N - \mathbb{E}[X] \| > \epsilon\}) = 0.$$

This means that, for a large enough set of IID samples, the average of the samples (i.e. X_N) will, with high probability, be very close to the expectation $\mathbb{E}[X]$.

Proposition: *Weak law of large numbers*

$$\lim_{N\to\infty} P(\{\| X_N - \mathbb{E}[X] \| > \epsilon\}) = 0.$$

Proof: Assume that $P(X)$ has finite variance; that is, $\mathsf{var}\,(\{X\}) = \sigma^2$. Choose $\epsilon > 0$. Now we have that

$$\begin{aligned}
\mathsf{var}\,(\{X_N\}) &= \mathsf{var}\left(\left\{\frac{\sum_{i=1} Nx_i}{N}\right\}\right) \\
&= (\frac{1}{N^2})\mathsf{var}\left(\left\{\sum_{i=1} Nx_i\right\}\right) \\
&= (\frac{1}{N^2})(N\sigma^2) \\
&\quad \text{because the } x_i \text{ are independent} \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

and that

$$\mathbb{E}[X_N] = \mathbb{E}[X].$$

Now Chebyshev's inequality gives

$$P(\{\| X_N - \mathbb{E}[X] \| \geq \epsilon\}) \leq \frac{\sigma^2}{N\epsilon^2}$$

so

$$1 - P(\{\| X_N - \mathbb{E}[X] \| \geq \epsilon\}) = P(\{\| X_N - \mathbb{E}[X] \| < \epsilon\}) \geq 1 - \frac{\sigma^2}{N\epsilon^2}.$$

This means that

$$\lim_{N\to\infty} P(\{\| X_N - \mathbb{E}[X] \| > \epsilon\}) = 0$$

which is the weak law of large numbers.

3.3 USING EXPECTATIONS

The weak law of large numbers gives us a very valuable way of thinking about expectations. Assume we have a random variable X. Then the weak law says that, if you observe this random variable over a large number of trials, the mean value you observe should be very close to $\mathbb{E}[X]$. Notice that this extends to functions of random variables (because they are random variables, too). For example, I observe values x_i of a random variable X over a large number N of trials, and compute

$$\frac{1}{N}\sum_{i=1} Nf(x_i).$$

The weak law says that the value I get should be very close to $\mathbb{E}[f]$. You can show this by defining a new random variable $F = f(X)$. This has a probability

distribution $P(F)$, which might be difficult to know — but we don't need to. $\mathbb{E}[f]$, the expected value of the function f under the distribution $P(X)$. This is the same as $\mathbb{E}[F]$, and the weak law applies.

Remember: the average over repeated trials of a random variable is very close to the expectation. You can use this information to make many kinds of decision in uncertain circumstances.

3.3.1 Should you accept a bet?

We can't answer this as a moral question, but we can as a practical question, using expectations. Generally, a bet involves an agreement that amounts of money will change hands, depending on the outcome of an experiment. Mostly, you are interested in how much you get from the bet, so it is natural to give sums of money you receive a positive sign, and sums of money you pay out a negative sign. Under this convention, the practical answer is easy: accept a bet enthusiastically if its expected value is positive, otherwise decline it. It is interesting to notice how poorly this advice describes actual human behavior.

Worked example 3.12 *Red or Black?*

A roulette wheel has 36 numbers, 18 of which are red and 18 of which are black. Different wheels also have one, two, or even three zeros,which are colorless. A ball is thrown at the wheel when it is spinning, and it falls into a hole corresponding to one of the numbers (when the number is said to "come up"). The wheel is set up so that there is the same probability of each number coming up. You can bet on (among other things) whether a red number or a black number comes up. If you bet 1 on red, and a red number comes up, you keep your stake and get 1, otherwise you get -1 (i.e. the house keeps your bet).

- On a wheel with one zero, what is the expected value of a 1 bet on red?
- On a wheel with two zeros, what is the expected value of a 1 bet on red?
- On a wheel with three zeros, what is the expected value of a 1 bet on red?

Solution: Write p_r for the probability a red number comes up. The expected value is $1 \times p_r + (-1)(1 - p_r)$ which is $2p_r - 1$.

- In this case, $p_r =$ (number of red numbers)/(total number of numbers) $=$ $18/37$. So the expected value is $-1/37$ (you lose about 3 cents each time you bet).
- In this case, $p_r = 18/38$. So the expected value is $-2/38 = -1/19$ (you lose slightly more than five cents each time you bet).
- In this case, $p_r = 18/39$. So the expected value is $-3/39 = -1/13$ (you lose slightly less than 8 cents each time you bet).

Notice that in the roulette game, the money you lose will go to the house. So the expected value to the house is just the negative of the expected value to you. This is positive, which is a partial explanation of why there are lots of roulette wheels, and usually free food nearby. Not all bets are like this, though.

Worked example 3.13 *Coin game*

In this game, P1 flips a fair coin and P2 calls "H" or "T". If P2 calls right, then P1 throws the coin into the river; otherwise, P1 keeps the coin. What is the expected value of this game to P2? and to P1?

Solution: To P2, which we do first, because it's easiest: P2 gets 0 if P2 calls right, and 0 if P2 calls wrong; these are the only cases, so the expected value is 0. To P1: P1 gets -1 if P2 calls right, and 0 if P1 calls wrong. The coin is fair, so the probability P2 calls right is $1/2$. The expected value is $-1/2$. While I can't explain why people would play such a game, I've actually seen this done.

We call a bet fair when its expected value is zero. Taking a bet with a negative expected value is unwise, because, on average, you will lose money. Worse, the more times you play, the more you lose. Taking a bet with a positive expected value is likely to be profitable. However, you do need to be careful you computed the expected value right.

Worked example 3.14 *Birthdays in succession*

P1 and P2 agree to the following bet. P1 gives P2 a stake of 1. If three people, stopped at random on the street, have birthdays in succession (i.e. Mon-Tue-Wed, and so on), then P2 gives P1 100. Otherwise, P1 loses the stake. What is the expected value of this bet to P1?

Solution: Write p for the probability of winning. Then the expected value is $p \times 100 - (1-p) \times 1$. We computed p in example 22 (it was $1/49$). So the bet is worth $(52/49)$, or slightly more than a dollar, to P1. P1 should be happy to agree to this as often as possible.

The reason P2 agrees to bets like that of example 14 is most likely that P2 can't compute the probability exactly. P2 thinks the event is quite unlikely, so the expected value is negative; but it isn't as unlikely as P2 thought it was, and this is how P1 makes a profit. This is one of the reasons you should be careful accepting a bet from a stranger: they might be able to compute better than you.

3.3.2 Odds, Expectations and Bookmaking — a Cultural Diversion

Gamblers sometimes use a terminology that is a bit different from ours. In particular, the term **odds** is important. The term comes from the following idea: P1 pays a bookmaker b (the stake) to make a bet; if the bet is successful, P1 receives a, and if not, loses the original stake.

Assume the bet is fair, so that the expected value is zero. Write p for the probability of winning. The net income to P1 is $ap - b(1-p)$. If this is zero, then $p = b/(a+b)$. So you can interpret odds in terms of probability, *if* you assume the bet is fair.

A bookmaker sets adds at which to accept bets from gamblers. The bookmaker does not wish to lose money at this business, and so must set odds which are

potentially profitable. Doing so is not simple (bookmakers can, and occasionally do, lose catastrophically, and go out of business). In the simplest case, assume that the bookmaker knows the probability p that a particular bet will win. Then the bookmaker could set odds of $(1-p)/p : 1$. In this case, the expected value of the bet is zero; this is fair, but not attractive business, so the bookmaker will set odds assuming that the probability is a bit higher than it really is. There are other bookmakers out there, so there is some reason for the bookmaker to try to set odds that are close to fair.

In some cases, you can tell when you are dealing with a bookmaker who is likely to go out of business soon. For example, imagine there are two horses running in a race, both at $10 : 1$ odds — whatever happens, you could win by betting 1 on each. There is a more general version of this phenomenon. Assume the bet is placed on a horse race, and that bets pay off only for the winning horse. Assume also that exactly one horse will win (i.e. the race is never scratched, there aren't any ties, etc.), and write the probability that the i'th horse will win as p_i. Then $\sum_{i \in \text{horses}} p_i$ must be 1. Now if the bookmaker's odds yield a set of probabilities that is less than 1, their business should fail, because there is at least one horse on which they are paying out too much. Bookmakers deal with this possibility by writing odds so that $\sum_{i \in \text{horses}} p_i$ is larger than one.

But this is not the only problem a bookmaker must deal with. The bookmaker doesn't actually know the probability that a particular horse will win, and must account for errors in this estimate. One way to do so is to collect as much information as possible (talk to grooms, jockeys, etc.). Another is to look at the pattern of bets that have been placed already. If the bookmaker and the gamblers agree on the probability that each horse will win, then there should be no expected advantage to choosing one horse over another — each should pay out slightly less than zero to the gambler (otherwise the bookmaker doesn't eat). But if the bookmaker has underestimated the probability that a particular horse will win, a gambler may get a positive expected payout by betting on that horse. This means that if one particular horse attracts a lot of money from bettors, it is wise for the bookmaker to offer less generous odds on that horse. There are two reasons: first, the bettors might know something the bookmaker doesn't, and they're signalling it; second, if the bets on this horse are very large and it wins, the bookmaker may not have enough capital left to pay out or to stay in business. All this means that real bookmaking is a complex, skilled business.

3.3.3 Ending a Game Early

Imagine two people are playing a game for a stake, but must stop early — who should get what percentage of the stake? One way to do this is to give each player what they put in at the start, but this is (mildly) unfair if one has an advantage over the other. The alternative is to give each player the expected value of the game at that state for that player. Sometimes one can compute that expectation quite easily.

Worked example 3.15 *Ending a game early*

(from Durrett), two players each pay 25 to play the following game. They toss a fair coin. If it comes up heads, player H wins that toss; if tails, player T wins. The first player to reach 10 wins takes the stake of 50. But one player is called away when the state is 8-7 (H-T) — how should the stake be divided?

Solution: In this state, each player can either win — and so get 50 — or lose — and so get 0. The expectation for H is $50P(\{\text{H wins from 8-7}\}) + 0P(\{\text{T wins from 8-7}\})$, so we need to compute $P(\{\text{H wins from 8-7}\})$. Similarly, the expectation for T is $50P(\{\text{T wins from 8-7}\}) + 0P(\{\text{H wins from 8-7}\})$, so we need to compute $P(\{\text{T wins from 8-7}\})$; but $P(\{\text{T wins from 8-7}\}) = 1 - P(\{\text{H wins from 8-7}\})$. Now it is slightly easier to compute $P(\{\text{T wins from 8-7}\})$, because T can only win in two ways: 8-10 or 9-10. These are independent. For T to win 8-10, the next three flips must come up T, so that event has probability 1/8. For T to win 9-10, the next four flips must have one H in them, but the last flip may not be H (or else H wins); so the next four flips could be HTTT, THTT, or TTHT. The probability of this is 3/16. This means the total probability that T wins is 5/16. So T should get 16.625 and H should get the rest (although they might have to flip for the odd half cent).

3.3.4 Making a Decision with Decision Trees and Expectations

Imagine we have to choose an action. Once we have chosen, a sequence of random events occurs, and we get a reward with some probability. Which action should we choose? A good answer is to choose the action with the best expected outcome. In fact, choosing any other action is unwise, because if we encounter this situation repeatedly and make a choice that is even only slightly worse than the best, we could lose heavily. This is a very common recipe, and it can be applied to many situations. Usually, but not always, the reward is in money, and we will compute with money rewards for the first few examples.

For such problems, it can be useful to draw a **decision tree**. A decision tree is a drawing of possible outcomes of decisions, which makes costs, benefits and random elements explicit. Each node of the tree represents a test of an attribute (which could be either a decision, or a random variable), and each edge represents a possible outcome of a test. The final outcomes are leaves. Usually, decision nodes are drawn as squares, chance elements as circles, and leaves as triangles.

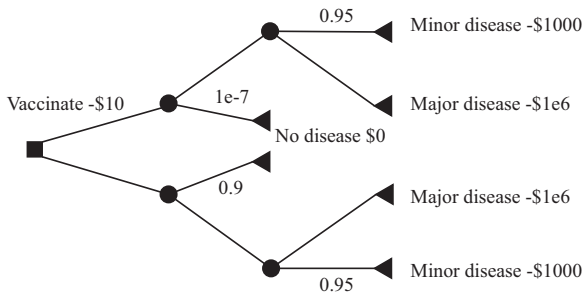

FIGURE 3.1: *A decision tree for the vaccination problem. The only decision is whether to vaccinate or not (the box at the root of the tree). I have only labelled edges where this is essential, so I did not annotate the "no vaccination" edge with zero cost. Once you decide whether to vaccinate or not, there is a random node (whether you get the disease or not), and, if you get it, another (minor or major).*

Worked example 3.16 *Vaccination*

It costs 10 to be vaccinated against a common disease. If you have the vaccination, the probability you will get the disease is $1-1e-7$. If you do not, the probability is 0.1. The disease is unpleasant; with probability 0.95, you will experience effects that cost you 1000 (eg several days in bed), but with probability 0.05, you will experience effects that cost you $1e6$. Should you be vaccinated?

Solution: Figure 3.1 shows a decision tree for this problem. I have annotated some edges with the choices represented, and some edges with probabilities; the sum of probabilities over all rightward (downgoing) edges leaving a random node is 1. It is straightforward to compute expectations. The expected cost of the disease is $0.95 \times 1000 + 0.05 \times 1e6 = 50,950$. If you are vaccinated, your expected income will be $-(10 + 1e-7 \times 50,950) = -10.01$ (rounding to the nearest cent). If you are not, your expected income is $-5,095$. You should be vaccinated.

Sometimes there is more than one decision. We can still do simple examples, though drawing a decision tree is now quite important, because it allows us to keep track of cases and avoid missing anything. For example, assume I wish to buy a cupboard. Two nearby towns have used furniture shops (usually called antique shops these days). One is further away than the other. If I go to town A, I will have time to look in two (of three) shops; if I go to town B, I will have time to look in one (of two) shops. I could lay out this sequence of decisions (which town to go to; which shop to visit when I get there) as Figure 3.2.

You should notice that this figure is missing a lot of information. What is the probability that I will find what I'm looking for in the shops? What is the value of finding it? What is the cost of going to each town? and so on. This information is

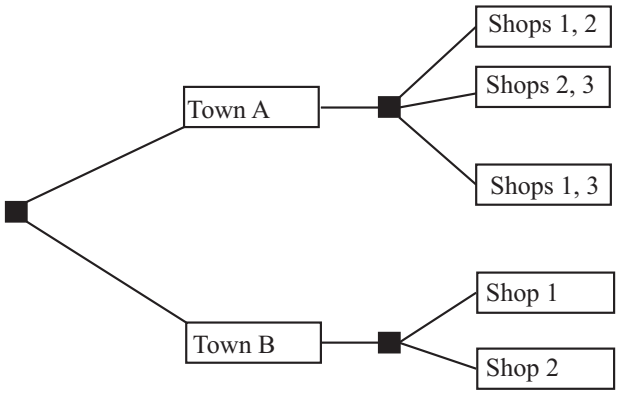

FIGURE 3.2: *The decision tree for the example of visiting furniture shops. Town A is nearer than town B, so if I go there I can choose to visit two of the three shops there; if I go to town B, I can visit only one of the two shops there. To decide what to do, I could fill in the probabilities and values of outcomes, compute the expected value of each pair of decisions, and choose the best. This could be tricky to do (where do I get the probabilities from?) but offers a rational and principled way to make the decision.*

not always easy to obtain. In fact, I might simply need to give my best subjective guess of these numbers. Furthermore, particularly if there are several decisions, computing the expected value of each possible sequence could get difficult. There are some kinds of model where one can compute expected values easily, but a good viable hypothesis about why people don't make optimal decisions is that optimal decisions are actually too hard to compute.

3.3.5 Utility

Sometimes it is hard to work with money. For example, in the case of a serious disease, choosing treatments often boils down to expected survival times, rather than money.

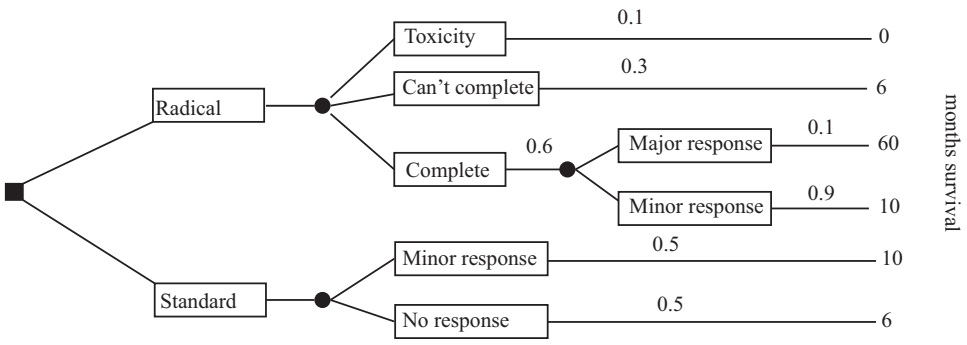

FIGURE 3.3: *A decision tree for example 17. Notations vary a bit, and here I have put boxes around the labels for the edges.*

Worked example 3.17 *Radical treatment*

(This example largely after Vickers, p97). Imagine you have a nasty disease. There are two kinds of treatment: standard, and radical. Radical treatment might kill you (with probability 0.1); might be so damaging that doctors stop (with probability 0.3); but otherwise you will complete the treatment. If you do complete radical treatment, there could be a major response (probability 0.1) or a minor response. If you follow standard treatment, there could be a major response (probability 0.5) or a minor response, but the outcomes are less good. All this is best summarized in a decision tree (Figure 3.3). What gives the longest expected survival time?

Solution: In this case, expected survival time with radical treatment is $(0.1 \times 0 + 0.3 \times 6 + 0.6 \times (0.1 \times 60 + 0.9 \times 10)) = 10.8$ months; expected survival time without radical treatment is $0.5 \times 10 + 0.5 \times 6 = 8$ months.

Working with money values is not always a good idea. For example, many people play state lotteries. The expected value of a 1 bet on a state lottery is well below 1 — why do people play? It's easy to assume that all players just can't do sums, but many players are well aware that the expected value of a bet is below the cost. It seems to be the case that people value money in a way that doesn't depend linearly on the amount of money. So, for example, people may value a million dollars rather more than a million times the value they place on one dollar. If this is true, we need some other way to keep track of value; this is sometimes called **utility**. It turns out to be quite hard to know how people value things, and there is quite good evidence that (a) human utility is complicated and (b) it is difficult to explain human decision making in terms of expected utility.

Worked example 3.18 *Human utility is not expected payoff*

Here are four games:

- **Game 1:** The player is given 1. A biased coin is flipped, and the money is taken back with probability p; otherwise, the player keeps it.
- **Game 2:** The player stakes 1, and a fair coin is flipped; if the coin comes up heads, the player gets r and the stake back, but otherwise loses the original stake.
- **Game 3:** The player bets nothing; a biased coin is flipped, and if it comes up heads (probability q), the player gets $1e6$.
- **Game 4:** The player stakes 1000; a fair coin is flipped, and if it comes up heads, the player gets s and the stake back, but otherwise loses the original stake.

In particular, what happens if $r = 3 - 2p$ and $q = (1 - p)/1e6$ and $s = 2 - 2p + 1000$?

Solution: Game 1 has expected value $(1-p)1$. Game 2 has expected value $(1/2)(r-1)$. Game 3 has expected value $q1e6$. Game 4 has expected value $(1/2)s - 500$.
In the case given, each game has the same expected value. Nonetheless, people usually have decided preferences for which game they would play. Generally, 4 is unattractive (seems expensive to play); 3 seems like free money, and so a good thing; 2 might be OK but is often seen as uninteresting; and 1 is unattractive. This should suggest to you that people's reasoning about money and utility is not what simple expectations can predict.

3.4 WHAT YOU SHOULD REMEMBER

You should be able to:

- Interpret notation for joint and conditional probability for random variables; in particular, understand notation such as: $P(\{X\})$, $P(\{X = x\})$, $p(x)$, $p(x, y)$, $p(x|y)$
- Interpret a probability density function $p(x)$ as $P(\{X \in [x, x + dx]\})$.
- Interpret the expected value of a discrete random variable.
- Interpret the expected value of a continuous random variable.
- Compute expected values of random variables for straightforward cases.
- Write down expressions for mean, variance and covariance for random variables.

- Write out a decision tree.

You should remember:

- The definition of a random variable.
- The definition of an expected value.
- The definitions of mean, variance and covariance.
- The definition of an indicator function.
- Bayes rule.
- The definition of marginalization.
- The Markov inequality.
- The Chebyshev Inequality.
- The weak law of large numbers.

PROBLEMS

Joint and Conditional Probability for Random Variables

3.1. Define a random variable X by the following procedure. Draw a card from a standard deck of playing cards. If the card is knave, queen, or king, then $X = 11$. If the card is an ace, then $X = 1$; otherwise, X is the number of the card (i.e. two through ten). Now define a second random variable Y by the following procedure. When you evaluate X, you look at the color of the card. If the card is red, then $Y = X - 1$; otherwise, $Y = X + 1$.
(a) What is $P(\{X \leq 2\})$?
(b) What is $P(\{X \geq 10\})$?
(c) What is $P(\{X \geq Y\})$?
(d) What is the probability distribution of $Y - X$?
(e) What is $P(\{Y \geq 12\})$?

3.2. Define a random variable by the following procedure. Flip a fair coin. If it comes up heads, the value is 1. If it comes up tails, roll a die: if the outcome is 2 or 3, the value of the random variable is 2. Otherwise, the value is 3.
(a) What is the probability distribution of this random variable?
(b) What is the cumulative distribution of this random variable?

3.3. Define three random variables, X, Y and Z by the following procedure. Roll a six-sided die and a four-sided die. Now flip a coin. If the coin comes up heads, then X takes the value of the six-sided die and Y takes the value of the four-sided die. Otherwise, X takes the value of the four-sided die and Y takes the value of the six-sided die. Z always takes the value of the sum of the dice.
(a) What is $P(X)$, the probability distribution of this random variable?
(b) What is $P(X, Y)$, the joint probability distribution of these two random variables?
(c) Are X and Y independent?
(d) Are X and Z independent?

3.4. Define two random variables X and Y by the following procedure. Flip a fair coin; if it comes up heads, then $X = 1$, otherwise $X = -1$. Now roll a six-sided die, and call the value U. We define $Y = U + X$.
(a) What is $P(Y|X = 1)$?
(b) What is $P(X|Y = 0)$?
(c) What is $P(X|Y = 7)$?
(d) What is $P(X|Y = 3)$?
(e) Are X and Y independent?

Expected Values

3.5. A simple coin game is as follows: we have a box, which starts empty. P1 flips a fair coin. If it comes up heads, P2 gets the contents of the box, and the game ends. If it comes up tails, P1 puts a dollar in the box and they flip again; this repeats until it comes up heads
(a) With what probability will P1 win exactly 10 units?
(b) What is the expected value of the game?
(c) How much should P1 pay to play, to make the game fair?

3.6. A simple card game is as follows. P1 pays a stake of 1 to play. P1 and P2 then each draw a card. If both cards are the same color, P2 keeps the stake and the game ends. If they are different colors, P2 pays P1 the stake and 1 extra (a total of 2).
(a) What is the expected value of the game to P1?
(b) P2 modifies the game, as follows. If both cards are court cards (that is, knave, queen, king), then P2 keeps the stake and the game ends; otherwise, the game works as before. Now what is the expected value of the game to P1?

3.7. An airline company runs a flight that has six seats. Each passenger who buys a ticket has a probability p of turning up for the flight. These events are independent.
(a) The airline sells six tickets. What is the expected number of passengers, if $p = 0.9$?
(b) How many tickets should the airline sell to ensure that the expected number of passengers is greater than six, if $p = 0.7$? **Hint:** The easiest way to do this is to write a quick program that computes the expected value of passengers that turn up for each the number of tickets sold, then search the number of tickets sold.

3.8. An airline company runs a flight that has 10 seats. Each passenger who buys a ticket has a probability p of turning up for the flight. The gender of the passengers is not known until they turn up for a flight, and women buy tickets with the same frequency that men do. The pilot is eccentric, and will not fly unless at least two women turn up.
(a) How many tickets should the airline sell to ensure that the expected number of passengers that turn up is greater than 10?
(b) The airline sells 10 tickets. What is the expected number of passengers on the aircraft, given that it flies? (i.e. that at least two women turn up). Estimate this value with a simulation.

Mean, Variance and Covariance

3.9. Show that $\mathsf{Var}[kX] = k^2\mathsf{Var}[X]$.

3.10. Show that if X and Y are independent random variables, then $\mathsf{Var}[X + Y] =$

$\mathsf{Var}[X] + \mathsf{Var}[Y]$. You will find it helpful to remember that, for X and Y independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Using Inequalities

3.11. The random variable X takes the values -2, -1, 0, 1, 2, but has an unknown probability distribution. You know that $\mathbb{E}[\|X\|] = 0.2$. Use Markov's inequality to give a *lower* bound on $P(\{X = 0\})$. *Hint:* Notice that $P(\{X = 0\}) = 1 - P(\{\|X\| = 1\}) - P(\{\|X\|\} = 2)$.

3.12. The random variable X takes the values 1, 2, 3, 4, 5, but has unknown probability distribution. You know that $\mathbb{E}[X] = 2$ and $\mathsf{var}(\{X\}) = 0.01$. Use Chebychev's inequality to give a *lower* bound on $P(\{X = 2\})$.

Using Expectations

3.13. Imagine we have a game with two players, who are playing for a stake. There are no draws, the winner gets the whole stake, and the loser gets nothing. The game must end early. We decide to give each player the expected value of the game for that player, from that state. Show that the expected values add up to the value of the stake (i.e. there won't be too little or too much money in the stake.

General Exercises

CHAPTER 4

Useful Probability Distributions

4.1 DISCRETE DISTRIBUTIONS

4.1.1 The Discrete Uniform Distribution

If every value of a discrete random variable has the same probability, then the probability distribution is the discrete uniform distribution. We have seen this distribution before, numerous times. For example, I define a random variable by the number that shows face-up on the throw of a die. This has a uniform distribution. As another example, write the numbers 1-52 on the face of each card of a standard deck of playing cards. The number on the face of the first card drawn from a well-shuffled deck is a random variable with a uniform distribution.

One can construct expressions for the mean and variance of a discrete uniform distribution, but they're not usually much use (too many terms, not often used).

4.1.2 Sums and Differences of Discrete Uniform Random Variables

Assume X and Y are discrete random variables with uniform distributions. Neither $X - Y$ nor $X + Y$ is uniform. We can see this by constructing the distribution. This is easiest to do with concrete ranges, etc. for these random variables. For concreteness, assume that both X and Y are integers in the range $1 - 100$. Write $S = X + Y$, $D = X - Y$. Then

$$P(S = k) = P(\cup_{u=1}^{u=100} \{\{X = k - u\} \cap \{Y = u\}\})$$

but the events $\{\{X = k - u\} \cap \{Y = u\}\}$ and $\{\{X = k - v\} \cap \{Y = v\}\}$ are disjoint if $u \neq v$. So we can write

$$P(S = k) = \sum_{u=1}^{u=100} P(\{\{X = k - u\} \cap \{Y = u\}\})$$

and the events $\{X = k - u\}$ and $\{Y = u\}$ are independent, so we can write

$$P(S = k) = \sum_{u=1}^{u=100} P(\{X = k - u\})P(\{Y = u\}).$$

Now, although $P(X)$ and $P(Y)$ are uniform, the shifting effect of the subtraction term in $\{X = k - u\}$ has very significant effects. For example, imagine $k = 2$; then there is only one non-zero term in the sum (i.e. $P(\{X = 1\})P(\{Y = 1\})$). But if $k = 3$, there are two (i.e. $P(\{X = 2\})P(\{Y = 1\})$ and $P(\{X = 1\})P(\{Y = 2\})$). And if $k = 100$, there are far more terms (which I'm not going to list here).

By a similar argument,

$$P(D = k) = \sum_{u=1}^{u=100} P(\{X = k + u\})P(\{Y = u\}).$$

Again, this isn't uniform; again, the shifting effect of the addition term in $\{X = k + u\}$ has very significant effects. For example, imagine $k = -99$; then there is only one non-zero term in the sum (i.e. $P(\{X = 1\})P(\{Y = 100\})$). But if $k = 98$, there are two (i.e. $P(\{X = 2\})P(\{Y = 100\})$ and $P(\{X = 1\})P(\{Y = 99\})$). And if $k = 0$, there are far more terms (which I'm not going to list here).

4.1.3 The Geometric Distribution

We have a biased coin. The probability it will land heads up, $P(\{H\})$ is given by p. We flip this coin until the first head appears. The number of flips required is a discrete random variable which takes integer values greater than or equal to one, which we shall call X. To get n flips, we must have $n - 1$ tails followed by 1 head. This event has probability $(1 - p)^{(n-1)}p$. We can now write out the probability distribution that n flips are required.

Definition 4.1 *The Geometric Distribution*

We have an experiment with a binary outcome (i.e. heads or tails; 0 or 1; and so on), with $P(H) = p$ and $P(T) = 1 - p$. We repeat this experiment until the first head occurs. The probability distribution for n, the number of repetitions, is the geometric distribution. It has the form

$$P(\{X = n\}) = (1 - p)^{(n-1)}p.$$

for $0 \leq p \leq 1$ and $n \geq 1$; for other n it is zero. p is called the **parameter** of the distribution.

Worked example 4.1 *Geometric distribution*

Show that the geometric distribution is non-negative and sums to one (and so is a probability distribution).

Solution: Recall that for $0 < r < 1$,

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}.$$

So

$$\begin{aligned}
\sum_{n=1}^{\infty} P(\{X = n\}) &= p\sum_{n=1}^{\infty}(1-p)^{(n-1)} \\
&= p\sum_{i=0}^{\infty}(1-p)^{i} \qquad \text{just reindexing the sum} \\
&= p\frac{1}{1-(1-p)} \\
&= 1
\end{aligned}$$

Useful Facts 4.1 *The geometric distribution*

1. The mean of the geometric distribution is $\frac{1}{p}$.
2. The variance of the geometric distribution is $\frac{1-p}{p^2}$.

The proof of these facts requires some work with series, and is relegated to the exercises.

4.1.4 The Binomial Probability Distribution

Assume we have a biased coin with probability p of coming up heads in any one flip. The binomial probability distribution gives the probability that it comes up heads i times in N flips.

Worked example 32 yields one way of deriving this distribution. In that example, I showed that if I flip a coin N times, then $N!/(i!(N-i)!)$ of the outcomes will have i heads. These outcomes are disjoint, and each has probability $p^i(1-p)^{(N-i)}$. As a result, we must have the probability distribution below.

Definition 4.2 *The Binomial distribution*

In N independent repetitions of an experiment with a binary outcome (ie heads or tails; 0 or 1; and so on) with $P(H) = p$ and $P(T) = 1 - p$, the probability of observing a total of i H's and $N - i$ T's is

$$P_b(i; N, p) = \left(\begin{array}{c} N \\ i \end{array} \right) p^i (1-p)^{(N-i)}$$

(as long as $0 \leq i \leq N$; in any other case, the probability is zero).

Here is another way to derive the binomial distribution. Assume we have a biased coin, so that $P(H) = p$ and $P(T) = 1-p$. Write $f(i; N, p)$ for the probability we encounter i heads in N flips. This probability distribution satisfies a recurrence relation. In particular

$$f(i; N, p) = pf(i-1; N-1, p) + (1-p)f(i; N-1, p).$$

This is equivalent to saying that you can get i heads in N flips either by having $i-1$ heads in $N-1$ flips, then flipping another, or by having i heads in N flips then flipping a tail. You can verify by induction that the binomial distribution satisfies this recurrence relation.

Worked example 4.2 *The binomial distribution*

Write $P_b(i; N, p)$ for the binomial distribution that one observes i H's in N trials. Show that

$$\sum_{i=0}^{N} P_b(i; N, p) = 1$$

Solution:

$$\sum_{i=0}^{N} P_b(i; N, p) = (p + (1-p))^N = (1)^N = 1$$

by pattern matching to the binomial theorem

Definition 4.3 *Bernoulli random variable*

A Bernoulli random variable takes the value 1 with probability p and 0 with probability $1 - p$. This is a model for a coin toss, among other things

Useful Facts 4.2 *The binomial distribution*

1. The mean of $P_b(i; N, p)$ is Np.
2. The variance of $P_b(i; N, p)$ is $Np(1-p)$

The proofs are easy and informative, and so are not banished to the exercises.

Proofs: *The binomial distribution*

Notice that the number of heads in N coin tosses is can be obtained by adding the number of heads in each toss. This means that, if X has the binomial distribution $P_b(X; N, p)$, and Y has the binomial distribution $P_b(Y; 1, p)$, so we can get the mean easily by

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{j=1}^{N} Y\right] \\ &= \sum_{j=1}^{N} \mathbb{E}[Y] \\ &= N\mathbb{E}[Y] \\ &= Np.\end{aligned}$$

The variance is easy, too. Each coin toss is independent, so the variance of the sum of coin tosses is the sum of the variances. This gives

$$\begin{aligned}\mathsf{Var}[X] &= \mathsf{Var}\left[\sum_{j=1}^{N} Y\right] \\ &= N\mathsf{Var}[Y] \\ &= Np(1-p)\end{aligned}$$

The binomial distribution can be used to demonstrate that our interpretation of probability as frequency is consistent. In particular, if a coin has probability p of coming up heads, then almost all of the probability in the binomial distribution occurs in values close to $i = pN$. This means that, in N flips, the number of heads you will see is very likely about pN. As N gets bigger, this effect becomes more pronounced. Showing this requires some extra machinery we haven't seen yet; section 22 and the exercises do that.

4.1.5 Multinomial probabilities

The binomial distribution describes what happens when a coin is flipped multiple times. But we could toss a die multiple times too. Assume this die has k sides, and we toss it N times. The distribution of outcomes is known as the **multinomial**

distribution.

We can guess the form of the multinomial distribution in rather a straightforward way. The die has k sides. We toss the die N times. This gives us a sequence of N numbers. Each toss of the die is independent. Assume that side 1 appears n_1 times, side 2 appears n_2 times, ... side k appears n_k times. Any single sequence with this property will appear with probability $p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$, because the tosses are independent. However, there are

$$\frac{N!}{n_1! n_2! ... n_k!}$$

such sequences. Using this reasoning, we arrive at the distribution below

Definition 4.4 *Multinomial Distribution*

I perform N independent repetitions of an experiment with k possible outcomes. The i'th such outcome has probability p_i. I see outcome 1 n_1 times, outcome 2 n_2 times, etc. Notice that $n_1 + n_2 + n_3 + \ldots + n_k = N$. The probability of observing this set of outcomes is

$$P_m(n_1, \ldots, n_k; N, p_1, \ldots, p_k) = \frac{N!}{n_1! n_2! ... n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}.$$

Worked example 4.3 *Dice*

I throw five fair dice. What is the probability of getting two 2's and three 3's?

Solution: $\frac{5!}{2!3!}(\frac{1}{6})^2(\frac{1}{6})^3$

4.1.6 The Poisson Distribution

Assume we are interested in counts that occur in an interval of space (e.g. along some ruler) or of time (e.g. within a particular hour). Because they are counts, they are non-negative and integer valued. We know these counts have two important properties. First, they occur with some fixed average rate. Second, an observation occurs independent of the interval since the last observation. Then the Poisson distribution is an appropriate model.

There are numerous such cases. For example, the marketing phone calls you receive during the day time are likely to be well modelled by a Poisson distribution. They come at some average rate — perhaps 5 a day as I write, during the last phases of an election year — and the probability of getting one clearly doesn't depend on the time since the last one arrived. As another example, mark the height of each dead insect on your car windscreen on a ruler; these heights should be well-modelled by a Poisson distribution. Classic examples include the number of Prussian soldiers killed by horse-kicks each year; the number of calls arriving at a call center each

minute; the number of insurance claims occurring in a given time interval (outside of a special event like a hurricane, etc.).

Definition 4.5 *The Poisson Distribution*

A non-negative, integer valued random variable X has a Poisson distribution when its probability distribution takes the form

$$P(\{X = k\}) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where $\lambda > 0$ is a parameter often known as the **intensity** of the distribution.

Notice that the Poisson distribution is a probability distribution, because it is non-negative and because

$$\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda}$$

so that

$$\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = 1$$

Useful Facts 4.3 *The Poisson Distribution*

1. The mean of a Poisson distribution with intensity λ is λ.
2. The variance of a Poisson distribution with intensity λ is λ (no, that's not an accidentally repeated line or typo).

The proof of these facts requires some work with series, and is relegated to the exercises.

4.2 CONTINUOUS DISTRIBUTIONS

4.2.1 The Continuous Uniform Distribution

Some continuous random variables have a natural upper bound and a natural lower bound but otherwise we know nothing about them. For example, imagine we are given a coin of unknown properties by someone who is known to be a skillful maker of unfair coins. The manufacturer makes no representations as to the behavior of the coin. The probability that this coin will come up heads is a random variable, about which we know nothing except that it has a lower bound of zero and an upper bound of one.

If we know nothing about a random variable apart from the fact that it has a lower and an upper bound, then a **uniform distribution** is a natural model. Write l for the lower bound and u for the upper bound. The probability density

function for the uniform distribution is

$$p(x) = \begin{cases} 0 & x < l \\ 1/(u-l) & l \leq x \leq u \\ 0 & x > u \end{cases}$$

A continuous random variable whose probability distribution is the uniform distribution is often called a **uniform random variable**. The matlab function `rand` will produce independent samples from a continuous uniform distribution, with lower bound 0 and upper bound 1.

4.2.2 Sums of Continuous Random Variables

You can guess the expression for a sum of two continuous random variables from the expression for discrete random variables above. Write p_x for the probability density function of X and p_y for the probability density function of Y. Then the probability density function of $S = X + Y$ is

$$p(s) = \int_{-\infty}^{\infty} p_x(s-u)p_y(u)du.$$

Notice that the procedure we have applied to add two random variables could be applied to three, as well. So if we need to know the probability density function for $S_3 = X + Y + Z$, we have

$$p(s) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} p_x((s-v)-u)p_y(u)du \right) p_z(v)dv$$

and so on.

4.2.3 The Normal Distribution

Definition 4.6 *The Standard Normal Distribution*

The probability density function

$$p(x) = \left(\frac{1}{\sqrt{2\pi}} \right) \exp\left(\frac{-x^2}{2} \right).$$

is known as the **standard normal distribution**

The first step is to plot this probability density function (Figure 4.1). You should notice it is quite familiar from work on histograms, etc. in Chapter 22. It has the shape of the histogram of standard normal data, or at least the shape that the histogram of standard normal data aspires to.

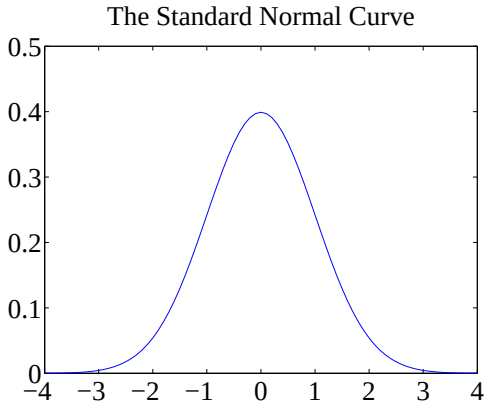

FIGURE 4.1: *A plot of the probability density function of the standard normal distribution. Notice how probability is concentrated around zero, and how there is relatively little probability density for numbers with large absolute values.*

Useful Facts 4.4 *The standard normal distribution*

1. The mean of the standard normal distribution is 0.
2. The variance of the standard normal distribution is 1.

These results are easily established by looking up (or doing!) the relevant integrals; they are relegated to the exercises.

A continuous random variable is a **standard normal random variable** if its probability density function is a standard normal distribution.

Any probability density function that is a standard normal distribution *in standard coordinates* is a **normal distribution**. Now write μ for the mean of a random variable and σ for its standard deviation; we are saying that, if

$$\frac{x-\mu}{\sigma}$$

has a standard normal distribution, then $p(x)$ is a normal distribution. We can work out the form of the probability density function of a general normal distribution in two steps: first, we notice that for any normal distribution, we must have

$$p(x) \propto \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

But, for this to be a probability density function, we must have $\int_{-\infty}^{\infty} p(x)dx = 1$. This yields the constant of proportionality, and we get

Definition 4.7 *The Normal Distribution*

The probability density function

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

is a normal distribution.

Useful Facts 4.5 *The normal distribution*

The probability density function

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

has

1. mean μ
2. and variance σ.

These results are easily established by looking up (or doing!) the relevant integrals; they are relegated to the exercises.

This argument explains why standard normal data is quite common. A continuous random variable is a **normal random variable** if its probability density function is a **normal distribution**.

Normal distributions are important for two reasons. It turns out that anything that behaves like a binomial distribution with a lot of trials — for example, the number of heads in many coin tosses; as another example, the percentage of times you get the outcome of interest in a simulation in many runs — should produce a normal distribution (Appendix 22). For this reason, pretty much any experiment where you perform a simulation, then count to estimate a probability or an expectation, should give you an answer that has a normal distribution.

The second reason I've hinted at above, but not shown in detail because it's a nuisance to prove. If you add together many random variables, each of pretty much any distribution, then the answer has a distribution close to the normal distribution. For these two reasons, we see normal distributions often. Notice that it is quite usual to call normal distributions **gaussian distributions**.

A normal random variable tends to take values that are quite close to the mean, measured in standard deviation units. We can demonstrate this important fact by computing the probability that a standard normal random variable lies between u and v. We form

$$\int_u^v \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

It turns out that this integral can be evaluated relatively easily using a special function. The **error function** is defined by

$$\mathsf{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-t^2\right) dt$$

so that

$$\frac{1}{2}\mathsf{erf}\left((\frac{x}{\sqrt{2}})\right) = \int_0^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

Notice that $\mathsf{erf}(x)$ is an odd function (i.e. $\mathsf{erf}(-x) = \mathsf{erf}(x)$). From this (and tables for the error function, or Matlab) we get that, for a standard normal random variable

$$\frac{1}{\sqrt{2\pi}} \int_{-1}^1 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.68$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-2}^2 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.95$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-2}^2 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.99.$$

These are very strong statements. They measure how often a standard normal random variable has values that are in the range $-1, 1$, $-2, 2$, and $-3, 3$ respectively. But these measurements apply to normal random variables if we recognize that they now measure how often the normal random variable is some number of standard deviations away from the mean. In particular, it is worth remembering that:

Useful Facts 4.6 *Normal Random Variables*

- About 68% of the time, a normal random variable takes a value within one standard deviation of the mean.
- About 95% of the time, a normal random variable takes a value within one standard deviation of the mean.
- About 99% of the time, a normal random variable takes a value within one standard deviation of the mean.

4.3 PROBABILITY AS FREQUENCY

Assume we flip a coin N times, where N is a very large number. The coin has probability p of coming up heads, and so probability $q = 1 - p$ of coming up tails. The number of heads h follows the binomial distribution, so

$$P(h) = \frac{N!}{h!(N-h)!} p^h q^{(N-h)}$$

The mean of this distribution is Np, the variance is Npq, and the standard deviation is $\sqrt{Npq}$.

Worked example 4.4 *Binomial Probabilities - I*

Assume that $N = 4$, $p = q = 1/2$. What is the probability that you would see 1, 2, or 3 heads?

Solution: To compute this probability, you compute

$$\sum_{h=1}^{3} \frac{4!}{h!(4-h)!} \frac{1}{2^4}$$

which you can do with whatever programming environment takes your fancy. I got 0.875.

Worked example 4.5 *Binomial Probabilities - II*

Assume that $N = 16$, $p = q = 1/2$. What is the probability that you would see 6, 7, 8, 9 or 10 heads?

Solution: To compute this probability, you compute

$$\sum_{h=6}^{10} \frac{16!}{h!(16-h)!} \frac{1}{2^{16}}$$

which you can do with whatever programming environment takes your fancy. I got 0.790.

Worked example 4.6 *Binomial Probabilities - III*

Assume that $N = 36$, $p = q = 1/2$. What is the probability that you would see 15, 16, 17, 18, 19, 20, or 21 heads?

Solution: To compute this probability, you compute

$$\sum_{h=15}^{21} \frac{36!}{h!(36-h)!} \frac{1}{2^{36}}$$

which you can do with whatever programming environment takes your fancy. You may find it a bit difficult to actually do the sum, however, because 36! is irritatingly large. I got 0.75 — there were more digits, but they're hard to trust.

Worked example 4.7 *Binomial Probabilities - IV*

Assume that $N = 64$, $p = q = 1/2$. What is the probability that you would see a number of heads from 28 to 36, inclusive?

Solution: To compute this probability, you compute

$$\sum_{h=24}^{40} \frac{64!}{h!(64-h)!} \frac{1}{2^{64}}$$

which you can do with whatever programming environment takes your fancy. You may find it a bit difficult to actually do the sum, however, because 64! is irritatingly large. I got 0.74 — there were more digits, but they're hard to trust.

You should notice that, in each of these examples, I am computing the probability that the number of heads you see lies within one standard deviation of the mean. Doing so gets difficult when N is large (as you should have noticed). We need a way to approximate the binomial distribution for very large N. When we do this approximation — which requires some manipulation — we will see that about 68% of the time the number of heads I see will be within one standard deviation of the mean.

Now the crucial point is this. As N gets bigger, the size of that interval, *relative to the total number of flips*, gets smaller, because the standard deviation is $\sqrt{Npq}$. If I flip a coin N times, in principle I could see a number of heads that ranges from 0 to N. But about 68% of the time, the fraction of those numbers that I actually see is within one standard deviation of the mean. This means the fraction of those numbers I will see is

$$2\sqrt{\frac{pq}{N}}$$

which limits to zero as $N \to \infty$. Figure 4.2 illustrates this effect.

The main difficulty with Figure 4.2 (and with the argument above) is that the mean and standard deviation of the binomial distribution tends to infinity as the number of coin flips tends to infinity. This can confuse issues. For example, the plots of Figure 4.2 show narrowing probability distributions — but is this because the scale is compacted, or is there a real effect? It turns out there is a real effect, and a good way to see it is to consider the normalized number of heads.

4.3.1 Large N

Recall that to normalize a dataset, you subtract the mean and divide the result by the standard deviation. We can do the same for a random variable. We now consider

$$x = \frac{h - Np}{\sqrt{Npq}}.$$

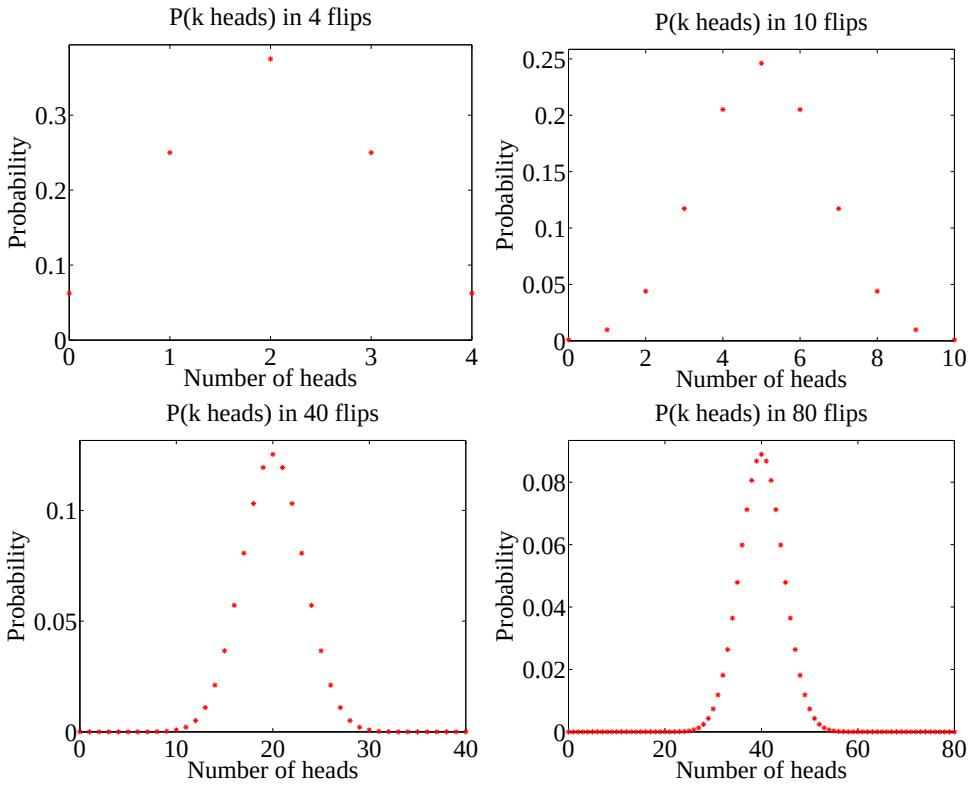

FIGURE 4.2: *Plots of the binomial distribution for $p = q = 0.5$ for different values of N. You should notice that the set of values of h (the number of heads) that have substantial probability is quite narrow compared to the range of possible values. This set gets narrower as the number of flips increases. This is because the mean is pN and the standard deviation is $\sqrt{N}pq$ — so the fraction of values that is within one standard deviation of the mean is $1/\sqrt{N}$.*

The probability distribution of x can be obtained from the probability distribution for h, because $h = Np + x\sqrt{Npq}$, so

$$P(x) = \left(\frac{N!}{(Np + x\sqrt{Npq})!(Nq - x\sqrt{Npq})!} \right) p^{(Np + x\sqrt{Npq})} q^{(Nq - x\sqrt{Npq})}.$$

I have plotted this probability distribution for various values of N in Figure 4.3.

But it is hard to work with this distribution for very large N. The factorials become very difficult to evaluate. Second, it is a discrete distribution on N points, spaced $1/\sqrt{Npq}$ apart. As N becomes very large, the number of points that have non-zero probability becomes very large, and x can be very large, or very small. For example, there is some probability, though there may be very little indeed, on the point where $h = N$, or, equivalently, $x = N(p + \sqrt{Npq})$. For sufficiently large N, we think of this probability distribution as a probability density function. We can

do so, for example, by spreading the probability for x_i (the i'th value of x) evenly over the interval between x_i and x_{i+1}. We then have a probability density function that looks like a histogram, with bars that become narrower as N increases. But what is the limit?

4.3.2 Getting Normal

To proceed, we need Stirling's approximation, which says that, for large N,

$$N! \approx \sqrt{2\pi}\sqrt{N}\left(\frac{N}{e}\right)^N.$$

This yields

$$P(h) \approx \left(\frac{Np}{h}\right)^h \left(\frac{Nq}{N-h}\right)^{(N-h)} \sqrt{\frac{N}{2\pi h(N-h)}}$$

Recall we used the normalized variable

$$x = \frac{h - Np}{\sqrt{Npq}}.$$

We will encounter the term $\sqrt{Npq}$ often, and we use $\sigma = \sqrt{Npq}$ as a shorthand. We can compute h and $N-h$ from x by the equalities

$$h = Np + \sigma x \qquad\qquad N - h = Nq - \sigma x.$$

So the probability distribution written in this new variable x is

$$P(x) \approx \left(\frac{Np}{(Np+\sigma x)}\right)^{(Np+\sigma x)} \left(\frac{Nq}{(Nq-\sigma x)}\right)^{(Nq-\sigma x)} \sqrt{\frac{N}{2\pi(Np+\sigma x)(Nq-\sigma x)}}$$

There are three terms to deal with here. It is easiest to work with $\log P$. Now

$$\log(1+x) = x - \frac{1}{2}x^2 + O(x^3)$$

so we have

$$\begin{aligned} \log\left(\frac{Np}{(Np+\sigma x)}\right) &= -\log\left(1 + \frac{\sigma x}{Np}\right) \\ &\approx -\frac{\sigma x}{Np} + (\frac{1}{2})(\frac{\sigma x}{Np})^2 \end{aligned}$$

and

$$\log\left(\frac{Nq}{(Nq-\sigma x)}\right) \approx \frac{\sigma x}{Nq} + (\frac{1}{2})(\frac{\sigma x}{Nq})^2.$$

From this, we have that

$$\begin{aligned} \log\left[\left(\frac{Np}{Np+\sigma x}\right)^{(Np+\sigma x)} \left(\frac{Nq}{Nq-\sigma x}\right)^{(Nq-\sigma x)}\right] &\approx [Np+\sigma x]\left[-\frac{\sigma x}{Np} + \left(\frac{1}{2}\right)\left(\frac{\sigma x}{Np}\right)^2\right] + \\ &\quad [Nq-\sigma x]\left[\frac{\sigma x}{Nq} + \left(\frac{1}{2}\right)\left(\frac{\sigma x}{Nq}\right)^2\right] \\ &= -\left(\frac{1}{2}\right)x^2 + O((\sigma x)^3) \end{aligned}$$

(recall $\sigma = \sqrt{Npq}$ if you're having trouble with the last step). Now we look at the square-root term. We have

$$\begin{aligned}\log\sqrt{\frac{N}{2\pi(Np+\sigma x)(Nq-\sigma x)}} &= -\frac{1}{2}\left(\log\left[Np+\sigma x\right]+\log\left[Nq-\sigma x\right]-\log N+\log 2\pi\right)\\ &= -\frac{1}{2}\left(\begin{array}{l}\log Np+O\left(\left(\frac{\sigma x}{Np}\right)\right)\\ +\log Nq-O\left(\left(\frac{\sigma x}{Nq}\right)\right)\\ -\log N+\log 2\pi\end{array}\right)\end{aligned}$$

but, since N is very large compared to σx, we can ignore the $O(\left(\frac{\sigma x}{Np}\right))$ terms. Then this term is not a function of x. So we have

$$\log P(x) \approx \frac{-x^2}{2} + \text{constant}.$$

Now because N is very large, our probability distribution P limits to a probability density function p, with

$$p(x) \propto \exp\left(\frac{-x^2}{2}\right).$$

We can get the constant of proportionality from integrating, to

$$p(x) = \left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(\frac{-x^2}{2}\right).$$

This constant of proportionality deals with the effect in figure 4.3, where the mode of the distribution gets smaller as N gets bigger. It does so because there are more points with non-zero probability to be accounted for. But we are interested in the limit where N tends to infinity. This must be a probability density function, so it must integrate to one.

Review this blizzard of terms. We started with a binomial distribution, but standardized the variables so that the mean was zero and the standard deviation was one. We then assumed there was a very large number of coin tosses, so large that that the distribution started to look like a continuous function. The function we get is the standard normal distribution.

4.3.3 So What?

I have proven an extremely useful fact, which I shall now put in a box.

Useful Fact 4.7 *The Binomial Distribution for Large N*

Assume h follows the binomial distribution with parameters p and q. Write $x = \frac{h-Np}{\sqrt{Npq}}$. Then, for sufficiently large N, the probability distribution $P(x)$ can be approximated by the probability density function

$$\left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(\frac{-x^2}{2}\right)$$

in the sense that

$$P(\{x \in [a,b]\}) \approx \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(\frac{-u^2}{2}\right)du$$

This justifies our model of probability as frequency. I interpreted an event having probability p to mean that, if I had a large number N of independent repetitions of the experiment, the number that produced the event would be close to Np, and would get closer as N got larger. We know that, for example, 68% of the time a standard normal random variable takes a value between 1 and -1. In this case, the standard normal random variable is

$$\frac{h-(Np)}{\sqrt{Npq}}$$

so that 68% of the time, h must take a value in the range $[Np-\sqrt{Npq}, Np+\sqrt{Npq}]$. Equivalently, the relative frequency h/N must take a value in the range

$$[p-\frac{pq}{\sqrt{N}}, p+\frac{pq}{\sqrt{N}}]$$

but as $N \to \infty$ this range gets smaller and smaller, and h/N limits to p. So our view of probability as a frequency is consistent.

To obtain h, we added N independent binomial random variables. So you can interpret the box as saying that the sum of many independent binomial random variables has a probability distribution that limits to the normal distribution as the number added together gets larger. It is a remarkable and deep fact, known as the **central limit theorem**, that adding many independent random variables produces a normal distribution *whatever* the distributions of those random variables.

4.4 WHAT YOU SHOULD REMEMBER

You should remember:

- The form of the uniform distribution.
- The form of the geometric distribution, and its mean and variance.
- The form of the binomial distribution, and its mean and variance.

- The form of the Poisson distribution, and its mean and variance.
- The form of the Normal distribution, and its mean and variance.
- The fact that a sum of a large number of IID binomial random variables is normally distributed, and the mean and variance of that distribution.
- The fact that a sum of a large number of IID random variables is normally distributed for most cases you will encounter.

PROBLEMS

The Geometric Distribution

4.1. Write $S_\infty = \sum_{i=0}^{\infty} r^i$. Show that $(1-r)S_\infty = 1$, so that

$$S_\infty = \frac{1}{1-r}$$

4.2. Show that

$$\sum_{i=0}^{\infty} ir^i = (\sum_{i=1}^{\infty} r^i) + r(\sum_{i=1}^{\infty} r^i) + r^2(\sum_{i=1}^{\infty} r^i) + \ldots$$

(look carefully at the limits of the sums!) and so show that

$$\sum_{i=0}^{\infty} ir^i = \frac{r}{(1-r)^2}.$$

4.3. Write $S_\infty = \sum_{i=0}^{\infty} r^i$. Show that

$$\sum_{i=0}^{\infty} i^2 r^i = (\gamma - 1) + 3r(\gamma - 1) + 5r^2(\gamma - 1) + \ldots$$

and so that

$$\sum_{i=0}^{\infty} i^2 r^i = \frac{r(1+r)}{(1-r)^3}$$

4.4. Show that, for a geometric distribution with parameter p, the mean is

$$\sum_{i=1}^{\infty} i(1-p)^{(i-1)}p = \sum_{i=0}^{\infty}(i+1)(1-p)^i p.$$

Now by rearranging and using the previous results, show that the mean is

$$\sum_{i=1}^{\infty} i(1-p)^{(i-1)}p = \frac{1}{p}$$

4.5. Show that, for a geometric distribution with parameter p, the variance is $\frac{1-p}{p^2}$. To do this, note the variance is $\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. Now use the results of the previous exercises to show that

$$\mathbb{E}\left[X^2\right] = \sum_{i=1}^{\infty} i^2(1-p)^{(i-1)}p = \frac{p}{1-p}\frac{(1-p)(2-p)}{p^3},$$

then rearrange to get the expression for variance.

The Binomial Distribution

4.6. Show that $P_b(N-i;N,p) = P_b(i;N,p)$ for all i.

4.7. Write h_r for the number of heads obtained in r flips of a coin which has probability p of coming up heads. Compare the following two ways to compute the probability of getting i heads in five coin flips:

- Flip the coin three times, count h_3, then flip the coin twice, count h_2, then form $w = h_3 + h_2$.
- Flip the coin five times, and count h_5.

Show that the probability distribution for w is the same as the probability distribution for h_5. Do this by showing that

$$P(\{w=i\}) = \sum_{j=0}^{5} P(\{h_3 = j\} \cap \{h_2 = i-j\}) = P(\{h_5 = i\}).$$

4.8. Now we will do the previous exercise in a more general form. Again, write h_r for the number of heads obtained in r flips of a coin which has probability p of coming up heads. Compare the following two ways to compute the probability of getting i heads in N coin flips:

- Flip the coin t times, count h_t, then flip the coin $N-t$ times, count h_{N-t}, then form $w = h_t + h_{N-t}$.
- Flip the coin N times, and count h_N.

Show that the probability distribution for w is the same as the probability distribution for h_N. Do this by showing that

$$P(\{w=i\}) = \sum_{j=0}^{N} P(\{h_t = j\} \cap \{h_{N-t} = i-j\}) = P(\{h_N = i\}).$$

You will likely find the recurrence relation

$$P_b(i;N,p) = pP_b(i-1;N-1,p) + (1-p)P_b(i;N-1,p).$$

is useful.

4.9. An airline runs a regular flight with six seats on it. The airline sells six tickets. The gender of the passengers is unknown at time of sale, but women are as common as men in the population. All passengers always turn up for the flight. The pilot is eccentric, and will not fly a plane unless at least one passenger is female. What is the probability that the pilot flies?

4.10. An airline runs a regular flight with s seats on it. The airline always sells t tickets for this flight. The probability a passenger turns up for departure is p, and passengers do this independently. What is the probability that the plane travels with exactly 3 empty seats?

4.11. An airline runs a regular flight with s seats on it. The airline always sells t tickets for this flight. The probability a passenger turns up for departure is p, and passengers do this independently. What is the probability that the plane travels with 1 or more empty seats?

4.12. An airline runs a regular flight with 10 seats on it. The probability that a passenger turns up for the flight is 0.95. What is the smallest number of seats the airline should sell to ensure that the probability the flight is full (i.e. 10 or more passengers turn up) is bigger than 0.99? (you'll probably need to use a calculator or write a program for this).

The Multinomial Distribution

4.13. Show that the multinomial distribution

$$P_m(n_1,\ldots,n_k;N,p_1,\ldots,n_k) = \frac{N!}{n_1!n_2!\ldots n_k!}p_1^{n_1}p_2^{n_2}\ldots p_k^{n_k}$$

must satisfy the recurrence relation

$$\begin{aligned} P_m(n_1,\ldots,n_k;N,p_1,\ldots,p_k) &= p_1P_m(n_1-1,\ldots,n_k;N-1,p_1,\ldots,p_k)+ \\ & \quad p_2P_m(n_1,n_2-1,\ldots,n_k;N-1,p_1,\ldots,p_k)+\ldots \\ & \quad p_kP_m(n_1,n_2,\ldots,n_k-1;N-1,p_1,\ldots,p_k) \end{aligned}$$

The Poisson Distribution

4.14. Compute the Taylor series for xe^x around $x=0$. Use this and pattern matching to show that the mean of the Poisson distribution with intensity parameter λ is λ.

4.15. Compute the Taylor series for $(x^2+x)e^x$ around $x=0$. Use this and pattern matching to show that the variance of the Poisson distribution with intensity parameter λ is λ.

The Normal Distribution

4.16. Write

$$f(x) = \left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(\frac{-x^2}{2}\right).$$

(a) Show that $f(x)$ is non-negative for all x.

(b) By integration, show that

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

so that $f(x)$ is a probability density function.

(c) Show that

$$\int_{-\infty}^{\infty} xf(x)dx = 0.$$

The easiest way to do this is to notice that $f(x) = f(-x)$

(d) Show that

$$\int_{-\infty}^{\infty} xf(x-\mu)dx = \mu.$$

The easiest way to do this is to change variables, and use the previous two exercises.

(e) Show that

$$\int_{-\infty}^{\infty} x^2f(x)dx = 1.$$

You'll need to either do, or look up, the integral to do this exercise.

4.17. Write

$$g(x) = \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Show that

$$\int_{-\infty}^{\infty} g(x)dx = \sqrt{2\pi}\sigma.$$

You can do this by a change of variable, and the results of the previous exercises.

4.18. Write

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

(a) Show that

$$\int_{-\infty}^{\infty} xp(x)dx = \mu$$

using the results of the previous exercises.

(b) Show that

$$\int_{-\infty}^{\infty} x^2 p(x)dx = \sigma$$

using the results of the previous exercises.

The Binomial Distribution for Large N

4.19. I flip a fair coin N times and count heads. We consider the probability that h, the fraction of heads, is in some range of numbers. *Hint:* If you know the range of numbers for h, you know the range for h/N.

(a) For $N = 1e6$, what is $P(\{h \in [49500, 50500]\})$?

(b) For $N = 1e4$, what is $P(\{h > 9000\})$?

(c) For $N = 1e2$, what is $P(\{h > 60\} \cup \{h < 40\})$?

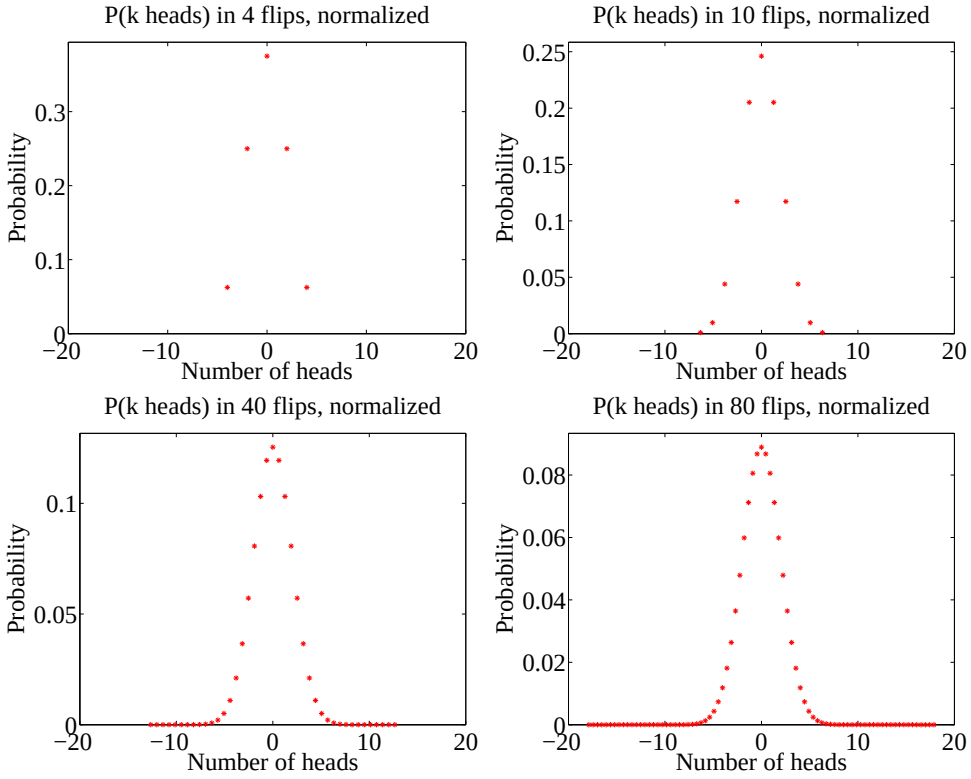

FIGURE 4.3: *Plots of the distribution for the normalized variable x, with $P(x)$ given in the text, obtained from the binomial distribution with $p = q = 0.5$ for different values of N. Because these distributions are normalized, we have that the mean of each is 0 and the standard deviation of each is 1. These distributions look increasingly like a standard normal distribution EXCEPT that the value at their mode gets smaller as N gets bigger. This occurs because there are more possible outcomes; in the text, we will establish that the standard normal distribution is a limit, in a useful sense.*

CHAPTER 5

Markov Chains and Simulation

There are many situations where one must work with a sequence of random variables. For example, consider a bus queue; people arrive at random, and so do buses. What is the probability that the queue gets to a particular length? and what is the expected length of the queue? As another example, we could have a random model of purchases from a shop — under any particular rule for restoring inventory, what is the largest (resp. smallest) amount of stock on the shelf? what is the expected amount of stock on the shelf? It turns out that there is a simple and powerful model that applies to many sequences of random variables. One can often use this model to make closed-form predictions. In cases where closed-form predictions aren't available, one can use simulation methods to estimate probabilities and expectations.

5.1 MARKOV CHAINS

A Markov chain is a sequence of random variables which has important independence properties. We will describe these properties in some detail below; first, we give some examples. Markov chains are easily represented with the language of finite state machines. For some Markov chains, it is easy to determine probabilities and expectations of interest in closed form using simple methods. For others, it is tricky (straightforward, but unreasonable amounts of straightforward work). In these cases, we can estimate the relevant probabilities and expectations by simulating the finite state machine.

5.1.1 Motivating Example: Multiple Coin Flips

We start with three examples, each of which is easy to work. Each suggests a much harder question, however, which we need new machinery to handle.

Worked example 5.1 *Multiple Coin Flips - 1*

You choose to flip a fair coin until you see two heads in a row, and then stop. What is the probability that you flip the coin twice?

Solution: Because you stopped after two flips, you must have seen two heads. So $P(2 \text{ flips}) = P(\{HH\}) = P(\{H\})^2 = 1/4$.

Worked example 5.2 *Multiple Coin Flips - 2*

You choose to flip a fair coin until you see two heads in a row, and then stop. What is the probability that you flip the coin three times?

Solution: Because you stopped after three flips, you must have seen T, then H, then H; any other sequence either doesn't stop, or stops too early. We write this with the last flip last, so THH. So $P(3 \text{ flips}) = P(\{THH\}) = P(\{T\})P(\{H\})^2 = 1/8$.

Worked example 5.3 *Multiple Coin Flips - 3*

You choose to flip a fair coin until you see two heads in a row, and then stop. What is the probability that you flip the coin four times?

Solution: This is more interesting. The last three flips must have been THH (otherwise you'd go on too long, or end too early). But, because the second flip must be a T, the first could be either H or T. This means there are two sequences that work: $HTHH$ and $TTHH$. So $P(4 \text{ flips}) = 2/8 = 1/4$.

The harder question here is to ask what is $P(N)$, for N some number (larger than 4, because we know the answers in the other cases). It is unattractive to work this case by case (you could try this, if you don't believe me). One very helpful way to think about the coin flipping experiment is as a finite state machine (Figure 5.1). If you think of this machine as a conventional finite state machine, it accepts any string of T, H, that (a) ends with HH, and (b) has no other HH in it. Alternatively, you could think of this machine as encoding a probabilistic process. At each state, an event occurs with some probability (in this case, a coin comes up H or T, with probability $(1/2)$). The event causes the machine to follow the appropriate edge. If we take this view, this machine stops when we have flipped two heads in succession. It encodes our problem of computing the probability of flipping a coin N times then stopping — this is just the probability that the machine hits the end state after N transitions.

It turns out to be straightforward to construct a recurrence relation for $P(N)$ (i.e. an equation for $P(N)$ in terms of $P(N-1)$, $P(N-2)$, and so on). This property is quite characteristic of repeated experiments. For this example, first, notice that $P(1) = 0$. Now imagine you have flipped the coin N times and stopped. We can assume that $N > 3$, because we know what happens for the other cases. The *last* three flips must have been THH. Write Σ_N for a sequence that is: (a) N flips long; (b) ends in HH; and (c) contains no other subsequence HH. Equivalently, this is a string accepted by the finite state machine of figure 5.1. We must have that any Σ_N has either the form $T\Sigma_{N-1}$ or the form $HT\Sigma_{N-2}$. But this means that

$$\begin{aligned} P(N) &= P(T)P(N-1) + P(HT)P(N-2) \\ &= (1/2)P(N-1) + (1/4)P(N-2) \end{aligned}$$

It is possible to solve this recurrence relation to get an explicit formula, but doing

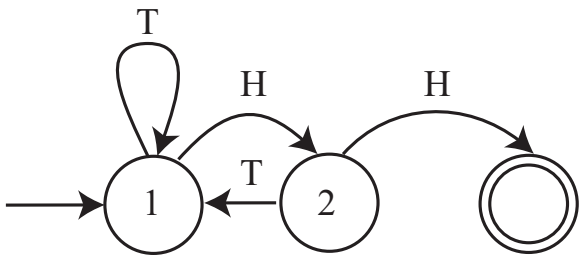

FIGURE 5.1: *A finite state machine representing the coin flip example. By convention, the end state is a double circle, and the start state has an incoming arrow. I've labelled the arrows with the event that leads to the transition, but haven't bothered to put in the probabilities, because each is 0.5.*

so would take us out of our way. You will check this recurrence relation in an exercise.

One really useful way to think of this recurrence relation is that represents an exercise in counting. We want to count all sequences that are of length N, and are accepted by the finite state machine of figure 5.1. This means they: (a) are N flips long; (b) end in HH; and (c) contain no other subsequence HH. Now work backward along the FSM. The only way to arrive at the final state is to be in state 1, then see HH. So you can obtain an acceptable N element sequence by (a) prepending a T to an acceptable $N-1$ element sequence or (b) prepending TH (which takes you to 2, then back to 1) to an acceptable $N-2$ element sequence. This line of reasoning can be made much more elaborate. There are a few examples in the exercises.

5.1.2 Motivating Example: The Gambler's Ruin

Another useful example is known as the **gambler's ruin**. Assume you bet \$1 a tossed coin will come up heads. If you win, you get \$1 and your original stake back. If you lose, you lose your stake. But this coin has the property that $P(H) = p < 1/2$. We will study what happens when you bet repeatedly.

Assume you have $\$s$ when you start. You will keep betting until either (a) you have \$0 (you are ruined; you can't borrow money) or (b) the amount of money you have accumulated is $\$j$, where $j > s$. The coin tosses are independent. We will compute $P(\text{ruined, starting with s}|p)$ the probability that you leave the table with nothing, when you start with $\$s$. For brevity, we write $P(\text{ruined, starting with s}|p) = p_s$. You can represent the gambler's ruin problem with a finite state machine as well. Figure 5.2 shows a representation of the gambler's ruin problem.

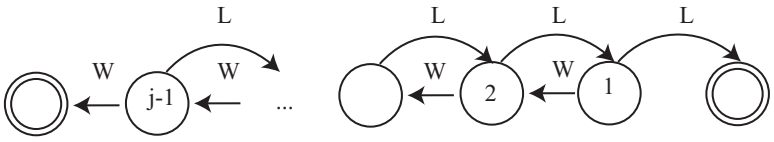

FIGURE 5.2: *A finite state machine representing the gambler's ruin example. I have labelled each state with the amount of money the gambler has at that state. There are two end states, where the gambler has zero (is ruined), or has j and decides to leave the table. The problem we discuss is to compute the probability of being ruined, given the start state is s. This means that any state except the end states could be a start state. I have labelled the state transitions with "W" (for win) and "L" for lose, but have omitted the probabilities.*

Worked example 5.4 *The gambler's ruin - 1*

Using the notation above, determine p_0 and p_j

Solution: We must have $p_0 = 1$, because if you have \$0, you leave the table. Similarly, if you have $\$j$, you leave the table with $\$j$, so you don't leave the table with nothing, so $p_j = 0$.

Worked example 5.5 *The gambler's ruin - 2*

Using the notation above, write a recurrence relation for p_s (the probability that you leave the table with nothing when you started with $\$s$.

Solution: Assume that you win the first bet. Then you have $\$s + 1$, so your probability of leaving the table with nothing now becomes p_{s+1}. If you lose the first bet, then you have $\$s - 1$, so your probability of leaving the table with nothing now becomes p_{s-1}. The coin tosses are independent, so we can write

$$p_s = pp_{s+1} + (1-p)p_{s-1}.$$

Some fairly lively work with series, relegated to the end of the chapter as exercises, yields

$$p_s = \frac{\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^s}{\left(\frac{1-p}{p}\right)^j - 1}.$$

This expression is quite informative. Notice that, if $p < 1/2$, then $(1-p)/p > 1$. This means that as $j \to \infty$, we have $p_s \to 1$. If you gamble repeatedly on an unfair coin, the probability that you run out of money before you hit some threshold ($\$j$ in this case) tends to one.

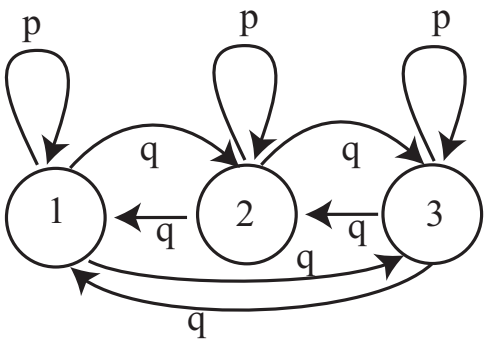

FIGURE 5.3: *A virus can exist in one of* 3 *strains. At the end of each year, the virus mutates. With probability* α*, it chooses uniformly and at random from one of the* 2 *other strains, and turns into that; with probability* $1 - \alpha$*, it stays in the strain it is in. For this figure, we have transition probabilities* $p = (1 - \alpha)$ *and* $q = (\alpha/2)$*.*

5.1.3 Motivating Example: A Virus

Problems represented with a finite state machine don't need to have an end state. As one example, (which I got from ACC Coolen's lecture notes), we have a virus that can exist in one of k strains. At the end of each year, the virus mutates. With probability α, it chooses uniformly and at random from one of the $k - 1$ other strains, and turns into that; with probability $1 - \alpha$, it stays in the strain it is in (Figure 5.3 shows an example with three strains). For that figure, we have $p = (1 - \alpha)$ and $q = (\alpha/2)$. The virus just keeps on changing, and there is no end state. But there are a variety of very interesting questions we can try to answer. We would like to know, for example, the expected time to see a strain a second time, i.e. if the virus is in strain 1 at time 1, what is the expected time before it is in strain 1 again? If the virus has mutated many times, what is the probability that it is in each strain? and do these probabilities depend on the start strain?

5.1.4 Markov Chains

In Figure 5.1 and Figure 5.2, I showed the event that caused the state transitions. It is more usual to write a probability on the figure, as I did in Figure 5.3, because the probability of a state transition (rather than what caused it) is what really matters. The underlying object is now a weighted directed graph, because we have removed the events and replaced them with probabilities. These probabilities are known as **transition probabilities**; notice that the sum of transition probabilities over *outgoing* arrows must be 1.

We can now think of our process as a **biased random walk** on a weighted directed graph. A bug (or any other small object you prefer) sits on one of the graph's nodes. At each time step, the bug chooses one of the outgoing edges at random. The probability of choosing an edge is given by the probabilities on the drawing of the graph (equivalently, the transition probabilities). The bug then follows that edge. The bug keeps doing this until it hits an end state.

This bug produces a sequence of random variables. If there are k states in the finite state machine, we can label each state with a number, $1 \dots k$. At the n'th time step, the state of the process — the node that the bug is sitting on — is a random variable, which we write X_n. These random variables have an important property. The probability that X_n takes some particular value depends only on X_{n-1}, and not on any other previous state. If we know where the bug is at step $n-1$ in our model, we know where it could go, and the probability of each transition. Where it was at previous times does not affect this, *as long as we know its state at step* $n-1$.

A sequence of random variables X_n is a **Markov chain** if it has the property that, $P(X_n = j|\text{values of all previous states}) = P(X_n = j|X_{n-1})$, or, equivalently, only the last state matters in determining the probability of the current state. The probabilities $P(X_n = j|X_{n-1} = i)$ are the **transition probabilities**. Any model built by taking the transitions of a finite state machine and labelling them with probabilities must be a Markov chain. However, this is not the only way to build or represent a Markov chain.

One representation of a Markov chain uses a matrix of transition probabilities. We define the matrix $\mathcal{P}$ with $p_{ij} = P(X_n = j|X_{n-1} = i)$. Notice that this matrix has the properties that $p_{ij} \geq 0$ and

$$\sum_j p_{ij} = 1$$

because at the end of each time step the model must be in some state. Equivalently, the sum of transition probabilities for outgoing arrows is one. Non-negative matrices with this property are **stochastic matrices**. By the way, you should look very carefully at the i's and j's here — Markov chains are usually written in terms of *row* vectors, and this choice makes sense in that context.

Worked example 5.6 *Viruses*

Write out the transition probability matrix for the virus of Figure 5.3, assuming that $\alpha = 0.2$.

Solution: We have $P(X_n = 1|X_{n-1} = 1) = (1-\alpha) = 0.8$, and $P(X_n = 2|X_{n-1} = 1) = \alpha/2 = P(X_n = 3|X_{n-1} = 1)$; so we get

$$\begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

Now imagine we do not know the initial state of the chain, but instead have a probability distribution. This gives $P(X_0 = i)$ for each state i. It is usual to take these k probabilities and place them in a k-dimensional row vector, which is usually written π. For example, we might not know what the initial strain of the virus is, but just that each strain is equally likely. So for a 3-strain virus, we would have $\pi = [1/3, 1/3, 1/3]$. From this information, we can compute the probability

distribution over the states at time 1 by

$$\begin{aligned} P(X_1 = j) &= \sum_i P(X_1 = j, X_0 = i) \\ &= \sum_i P(X_1 = j | X_0 = i) P(X_0 = i) \\ &= \sum_i p_{ij} \pi_i. \end{aligned}$$

If we write $\mathbf{p}^{(n)}$ for the row vector representing the probability distribution of the state at step n, we can write this expression as

$$\mathbf{p}^{(1)} = \pi \mathcal{P}.$$

Now notice that

$$\begin{aligned} P(X_2 = j) &= \sum_i P(X_2 = j, X_1 = i) \\ &= \sum_i P(X_2 = j | X_1 = i) P(X_1 = i) \\ &= \sum_i p_{ij} \left(\sum_{ki} p_{ki} \pi_k \right). \end{aligned}$$

so that

$$\mathbf{p}^{(n)} = \pi \mathcal{P}^n.$$

This expression is useful for simulation, and also allows us to deduce a variety of interesting properties of Markov chains.

Worked example 5.7 *Viruses*

We know that the virus of Figure 5.3 started in strain 1. After two state transitions, what is the distribution of states when $\alpha = 0.2$? when $\alpha = 0.9$? What happens after 20 state transitions? If the virus starts in strain 2, what happens after 20 state transitions?

Solution: If the virus started in strain 1, then $\pi = [1, 0, 0]$. We must compute $\pi(\mathcal{P}(\alpha))^2$. This yields $[0.66, 0.17, 0.17]$ for the case $\alpha = 0.2$ and $[0.4150, 0.2925, 0.2925]$ for the case $\alpha = 0.9$. Notice that, because the virus with small α tends to stay in whatever state it is in, the distribution of states after two years is still quite peaked; when α is large, the distribution of states is quite uniform. After 20 transitions, we have $[0.3339, 0.3331, 0.3331]$ for the case $\alpha = 0.2$ and $[0.3333, 0.3333, 0.3333]$ for the case $\alpha = 0.9$; you will get similar numbers even if the virus starts in strain 2. After 20 transitions, the virus has largely "forgotten" what the initial state was.

In example 7, the distribution of virus strains after a long interval appeared not to depend much on the initial strain. This property is true of many Markov chains. Assume that any state can be reached from any other state, by some

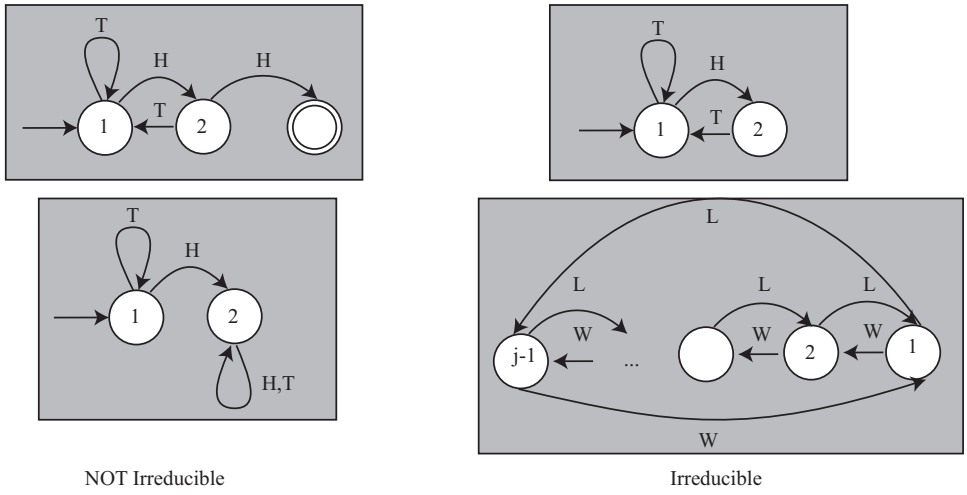

FIGURE 5.4: *Examples of finite state machines that give rise to Markov chains that are NOT irreducible (***left***) and irreducible (***right***). To obtain Markov chains from these drawings, we would have to give the probabilities of the events that lead to state transitions. We'll assume that none of the probabilities are zero; after that, the values don't matter for irreducibility analysis. The* **top left** *FSM is not irreducible, because it has an end state; once the bug reaches this point, it can't go anywhere else. The* **bottom left** *FSM is not irreducible, because the bug can get stuck in state 2.*

sequence of transitions. Such chains are called **irreducible**; notice this means there is no end state, like the virus example. Irreducibility also means that the chain cannot get "stuck" in a state or a collection of states (Figure 5.4). Then there is a unique vector $\mathbf{s}$, usually referred to as the **stationary distribution**, such that for *any* initial state distribution π,

$$\lim_{n \to \infty} \pi \mathcal{P}^{(n)} = \mathbf{s}.$$

Equivalently, if the chain has run through many steps, it no longer matters what the initial distribution is. You expect that the probability distribution over states is $\mathbf{s}$.

5.1.5 Example: Particle Motion as a Markov Chain

One can find Markov chains in quite unexpected places, often with useful consequences. In this example, I will obtain a Markov chain without reasoning about graphs or finite state machines. We will investigate a particle moving under gravity, in 3 dimensions. Write the position of the particle as $\mathbf{p}(t)$, its velocity as $\mathbf{v}(t)$, its acceleration as $\mathbf{a}(t)$, its mass as m, and the gravitational force as $\mathbf{g}$. Then we know

that

$$\begin{aligned}\mathbf{v}(t) &= \frac{d\mathbf{p}}{dt} \\ \mathbf{a}(t) &= \frac{d\mathbf{v}}{dt} \\ &= \mathbf{g}.\end{aligned}$$

Now stack the position and the acceleration into a single vector $X(t) = (\mathbf{p}(t), \mathbf{v}(t))^T$. We could write these equations as

$$\frac{dX}{dt} = \mathcal{A}X + \mathbf{b}$$

where

$$\mathcal{A} = \begin{pmatrix} 0 & \mathcal{I} \\ 0 & 0 \end{pmatrix}$$

and

$$\mathbf{b} = \begin{pmatrix} 0 \\ \mathbf{g} \end{pmatrix}.$$

Now imagine that we look at the position, velocity and acceleration of the particle at fixed time instants, so we are really interested in $X_i = X(t_0 + i\Delta t)$. In this case, we can approximate $\frac{dX}{dt}$ by $(X_{i+1} - X_i)/\Delta t$. We then have

$$X_{i+1} = X_i + (\Delta t)\left(\mathcal{A}X_i + \mathbf{b}\right).$$

This is beginning to look like a Markov chain, because X_{i+1} depends only on X_i but not on any previous X. But there isn't any randomness here. We can fix that by assuming that the particle is moving in, say, a light turbulent wind. Then the acceleration at any time consists of (a) the acceleration of gravity and (b) a small, random force produced by the wind. Write $\mathbf{w}_i$ for this small, random force at time i, and $P(\mathbf{w}_i|i)$ for its probability distribution, which could reasonably depend on time. We can then rewrite our equation by writing

$$X_{i+1} = X_i + (\Delta t)\left(\mathcal{A}X_i + \mathbf{b}_r\right).$$

where

$$\mathbf{b}_r = \begin{pmatrix} 0 \\ \mathbf{g} + \mathbf{w}_i \end{pmatrix}.$$

Now the X_i are clearly random variables. We could get $P(X_{i+1}|X_i)$ by rearranging terms. If I know X_{i+1} and X_i, I know the value of $\mathbf{w}_i$; I could plug this into $P(\mathbf{w}_i|i)$ to yield $P(X_{i+1}|X_i)$.

Using a finite state machine in this example would be a bit unnatural. This example turns out to be surprisingly useful in applications, because (with other algorithmic machinery we can't go into here) it offers the basis for algorithms that can track moving objects by predicting where they will go next. Applications are widespread. Military applications include tracking aircraft, UFO's, missiles, etc. Civilian applications include surveillance in public places; games console interfaces that track moving people; and methods that can follow experimental mice as they move around their cages (useful for testing medicines).

5.2 SIMULATION

Many problems in probability can be worked out in closed form if one knows enough combinatorial mathematics, or can come up with the right trick. Textbooks are full of these, and we've seen some. Explicit formulas for probabilities are often extremely useful. But it isn't always easy or possible to find a formula for the probability of an event in a model. An alternative strategy is to build a simulation, run it many times, and count the fraction of outcomes where the event occurs. This is a simulation experiment.

5.2.1 Obtaining Uniform Random Numbers

Simulation is at its most useful when we try to estimate probabilities that are hard to calculate. Usually we have processes with some random component, and we want to estimate the probability of a particular outcome. To do so, we will need a supply of random numbers to simulate the random component. Here I will describe methods to get uniformly distributed random numbers, and Section 5.2.5 describes a variety of methods to get random numbers from other distributions.

I will describe features in Matlab, because I'm used to Matlab. It's a good programming environment for numerical work, particularly for working with matrices and vectors. You can expect pretty much any programming environment to provide a random number generator that returns a number uniformly distributed in the range $[0-1]$. If you know anything about representing numbers, you'll already have spotted something funny. The number that comes out of the random number generator must be represented by a small set of bits, but almost all numbers in the interval $[0-1]$ require an infinite number of bits to represent. We'll sweep this issue under the general carpet of floating point representations; it doesn't matter for anything we need to do. The Matlab function function to do this is called `rand`. It can also return matrices; for example, `rand(10, 20)` will get you a 10×20 table of independent uniformly distributed random numbers in the range $[0-1]$.

It is useful to get a uniformly distributed random integer in a particular range (for example, you might want to choose a random element in an array). You can do so with a random number generator and a function (like Matlab's `floor`) that returns the largest integer smaller than its argument. If I want a discrete random variable with uniform distribution, maximum value 100 and minimum value 7, I habitually choose a very tiny number (for this example, say $1e-7$) and do `floor((100-7-1e-7)*rand()+7)`. I use the $1e-7$ because I can never remember whether `rand` produces a number no larger than one, or one that is guaranteed to be smaller than one, and I never need to care about the very tiny differences in probability caused by the 1e-7. You might be able to do something cleaner if you bothered checking this point.

5.2.2 Computing Expectations with Simulations

Simulation is also a very good way to estimate expectations. Imagine we have a random variable X with probability distribution $P(X)$ that takes values in some domain D. Assume that we can easily produce independent simulations, and that we wish to know $\mathbb{E}[f]$, the expected value of the function f under the distribution

$P(X)$.

The weak law of large numbers tells us how to proceed. Define a new random variable $F = f(X)$. This has a probability distribution $P(F)$, which might be difficult to know. We want to estimate $\mathbb{E}[f]$, the expected value of the function f under the distribution $P(X)$. This is the same as $\mathbb{E}[F]$. Now if we have a set of IID samples of X, which we write x_i, then we can form a set of IID samples of F by forming $f(x_i) = f_i$. Write

$$F_N = \frac{\sum_{i=1}^{N} f_i}{N}.$$

This is a random variable, and the weak law of large numbers gives that, for any positive number ϵ

$$\lim_{N \to \infty} P(\{\| F_N - \mathbb{E}[F] \| > \epsilon\}) = 0.$$

You can interpret this as saying that, that for a set of IID random samples x_i, the probability that

$$\frac{\sum_{i=1}^{N} f(x_i)}{N}$$

is very close to $\mathbb{E}[f]$ is high for large N

Worked example 5.8 *Computing an Expectation*

Assume the random variable X is uniformly distributed in the range $[0-1]$, and the random variable Y is uniformly distributed in the range $[0-10]$. X and Z are independent. Write $Z = (Y - 5X)^3 - X^2$. What is $\mathsf{var}(\{Z\})$?

Solution: With enough work, one could probably work this out in closed form. An easy program will get a good estimate. We have that $\mathsf{var}(\{Z\}) = \mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2$. My program computed 1000 values of Z (by drawing X and Y from the appropriate random number generator, then evaluating the function). I then computed $\mathbb{E}[Z]$ by averaging those values, and $\mathbb{E}[Z]^2$ by averaging their squares. For a run of my program, I got $\mathsf{var}(\{Z\}) = 2.76 \times 10^4$.

5.2.3 Computing Probabilities with Simulations

You can compute a probability using a simulation, too, because a probability can be computed by taking an expectation. Recall the property of indicator functions that

$$\mathbb{E}\left[\mathbb{I}_{[\mathcal{E}]}\right] = P(\mathcal{E})$$

(Section 3.2.5). This means that computing the probability of an event $\mathcal{E}$ involves writing a function that is 1 when the event occurs, and 0 otherwise; we then estimate the expected value of that function.

The weak law of large numers justifies this procedure. An experiment involves drawing a sample from the relevant probability distribution $P(X)$, then determining

whether event $\mathcal{E}$ occurred or not. We define a new random variable E, which takes the value 1 when the event $\mathcal{E}$ occurs during our experiment (i.e. with probability $P(\mathcal{E})$) and 0 when it doesn't (i.e. with probability $P(\mathcal{E}^c)$). Our experiment yields a sample of this random variable. We perform N experiments yielding a set of IID samples of this distribution e_i and compute $E_N = \frac{\sum_{i=1}^{N} e_i}{N}$. From the weak law of large numbers, we have that for any $\epsilon > 0$,

$$\lim_{N \to \infty} P(\{\| E_N - \mathbb{E}[E] \| \geq \epsilon\}) = 0$$

meaning that for large enough N, E_N will be very close to $\mathbb{E}[E] = P(\mathcal{E})$.

Worked example 5.9 *Computing a Probability for Multiple Coin Flips*

You flip a fair coin three times. Use a simulation to estimate the probability that you see three H's.

Solution: You really should be able to work this out in closed form. But it's amusing to check with a simulation. I wrote a simple program that obtained a 1000x3 table of uniformly distributed random numbers in the range $[0-1]$. For each number, if it was greater than 0.5 I recorded an H and if it was smaller, I recorded a T. Then I counted the number of rows that had 3 H's (i.e. the expected value of the relevant indicator function). This yielded the estimate 0.127, which compares well to the right answer.

Worked example 5.10 *Computing a Probability*

Assume the random variable X is uniformly distributed in the range $[0-1]$, and the random variable Y is uniformly distributed in the range $[0-10]$. Write $Z = (Y - 5X)^3 - X^2$. What is $P(\{Z > 3\})$?

Solution: With enough work, one could probably work this out in closed form. An easy program will get a good estimate. My program computed 1000 values of Z (by drawing X and Y from the appropriate random number generator, then evaluating the function) and counted the fraction of Z values that was greater than 3 (which is the relevant indicator function). For a run of my program, I got $P(\{Z > 3\}) \approx 0.619$

5.2.4 Simulation Results as Random Variables

The estimate of a probability or of an expectation that comes out of a simulation experiment is a random variable, because it is a function of random numbers. If you run the simulation again, you'll get a different value (unless you did something silly with the random number generator). Generally, you should expect this random variable to behave like a normal random variable. You can check this by constructing a histogram over a large number of runs. The mean of this random variable is the parameter you are trying to estimate. It is useful to know that this random variable tends to be normal, because it means the standard deviation of the random variable tells you a lot about the likely values you will observe.

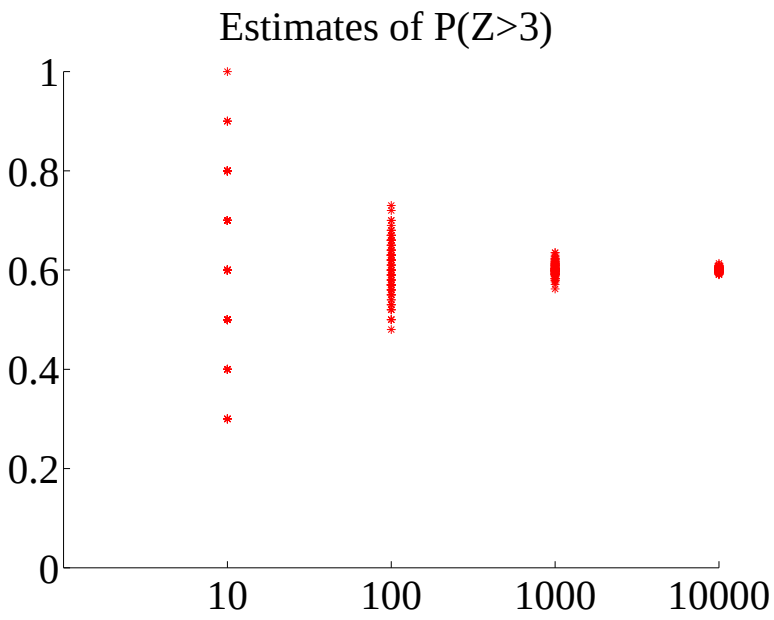

FIGURE 5.5: *Estimates of the probability from example 10, obtained from different runs of my simulator using different numbers of samples. In each case, I used 100 runs; the number of samples is shown on the horizontal axis. You should notice that the estimate varies pretty widely when there are only 10 samples, but the variance (equivalently, the size of the spread) goes down sharply as the number of samples increases to 1000. Because we expect these estimates to be roughly normally distributed, the variance gives a good idea of how accurate the original probability estimate is.*

Another helpful rule of thumb, which is almost always right, is that the standard deviation of this random variable behaves like

$$\frac{C}{\sqrt{N}}$$

where C is a constant that depends on the problem and can be very hard to evaluate, and N is the number of runs of the simulation. What this means is that if you want to (say) double the accuracy of your estimate of the probability or the expectation, you have to run four times as many simulations. Very accurate estimates are tough to get, because they require immense numbers of simulation runs.

Figure 5.5 shows how the result of a simulation behaves when the number of runs changes. I used the simulation of example 10, and ran multiple experiments for each of a number of different samples (i.e. 100 experiments using 10 samples; 100 using 100 samples; and so on).

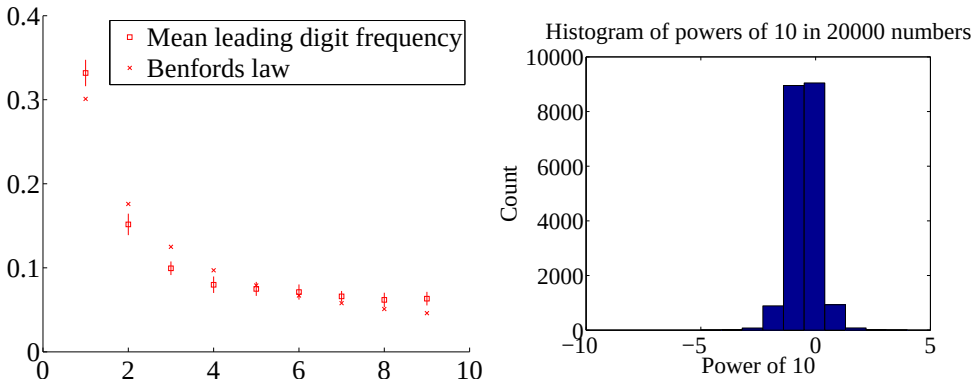

FIGURE 5.6: **Left** *summarizes results of a simulation where, for 20 runs, I generated 1000 pairs of independent random numbers uniformly distributed in* $[0-1]$ *and divided one by the other. The boxes show the mean over 20 runs, and the vertical bars show one standard deviation up and down. The crosses are predictions from Benford's law. Numbers generated in this way have a wide range of orders magnitude, so a conventional histogram isn't easy to plot. On the* **Right**, *a histogram of the first digit of the logarithm (i.e. the power of 10, or order of magnitude) of these numbers. Notice the smallest number is eight orders of magnitude smaller than the biggest.*

Worked example 5.11 *The First Digit of a Random Number*

We have seen (section 4.2.2) that adding random variables tends to produce a normal random variable. What happens if one divides one random number by another? Solving this in closed form is tricky, but simulation gives some insight.

Solution: I wrote a short program that produced 20 experiments, each consisting of 1000 numbers. I obtained the numbers by generating two uniform random numbers in the range $[0-1]$, then dividing one by the other. Notice that doing so can produce both very big and very small numbers, so that plotting a histogram is tricky — most boxes will be empty unless we produce an immense quantity of data. Instead, I plot a histogram of the leading digit of the number in Figure 5.7.

Figure 5.7 also shows the mean over all 20 experiments of the fraction of leading digits that is one, etc. together with one standard deviation error bars. The figure also shows a histogram of the rounded log (i.e. the power of 10) to give some idea of the range of values one obtains. It turns out that there is a widely applicable law, called **Benford's law**, that predicts the frequency of the first digit of many types of random number. The main condition for Benford's law to apply is that the numbers cover a wide range of scales. The figure shows predictions from Benford's law as well.

Small probabilities can be rather hard to estimate, as we would expect. In the case of example 10, let us estimate $P(\{Z > 950\})$. A few moments with a computer will show that this probability is of the order of 1e-3 to 1e-4. I obtained a million different simulated values of Z from my program, and saw 310 where $Z > 950$. This means that to know this probability to, say, three digits of numerical accuracy might involve a daunting number of samples. Notice that this does not contradict the rule of thumb that the standard deviation of the random variable defined by a simulation estimate behaves like $\frac{C}{\sqrt{N}}$; it's just that in this case, C is very large indeed.

5.2.5 Obtaining Random Samples

Building a successful simulation requires appropriate random numbers. Recall that anything we compute from a simulation can be thought of as an expectation (we used indicator functions to estimate probabilities). So we have some random variable X with distribution $P(X)$, a function f, and we wish to compute $\mathbb{E}[f]$, where the expectation is with respect to $P(X)$.

Most programming environments provide random number generators for uniform random numbers (I described the one for MATLAB briefly in Section 5.2.1) and for normally distributed random numbers. Building a really good, really fast random number generator is a major exercise. All sorts of tricks are involved, because it really matters to produce executable code that is as fast as possible on the target machine. This means that you don't have to build your own (and shouldn't, unless you can spend a lot of time and trouble on doing so).

Normal Random Variables

In pretty much any programming environment, you would also expect to find a random number generator that returns a normal random variable, with mean zero and standard deviation one. In Matlab, this function is called `randn`. Conveniently, `randn(3, 4)` will give you a 3×4 table of such numbers, which are independent of each other. As you would expect from section 22, to change the mean of this random number, you add a constant; to change the variance, you multiply by a constant. So in Matlab, if you want a normal random variable with mean 3 and standard deviation 4, you use `4*randn()+3`.

Rejection Sampling

Imagine you know a probability distribution describing a discrete random variable. In particular, you have one probability value for each of the numbers $1, 2, \ldots, N$ (all the others are zero). Write $p(i)$ for the probability of i. You can generate a sample of this distribution by the following procedure: first, generate a sample x of a uniform discrete random variable in the range $1, \ldots, N$; now generate a sample t of a uniformly distributed continuous random variable in the range $[0, 1]$; finally, if $t < p(x)$ report x, otherwise generate a new x and repeat. This process is known as **rejection sampling** (Algorithm 5.1).

To understand this procedure, it is best to think of it as a loop around a basic process. The basic process generates an x, then decides whether it is acceptable or not. The loop keeps invoking the basic process until it gets an acceptable x. Now

Listing 5.1: Matlab code for simple rejection sampling; this is inefficient, but simple to follow.

```
function rnum=rejectsample(pvec)
%
% pvec is a probability distribution over numbers
%  1, ..., size(pvec, 1)
%
nv=size(pvec, 1);
done=0;
while done==0
    ptr=floor(1+(nv-1e-10)*rand);
    % this gives me a uniform random
    % number in the range 1, .. nv
    pval=pvec(ptr);
    if rand<pval
        done=1;
        % i.e. accept
    end
end
rnum=ptr;
```

the probability that the basic process produces the value x_i, decides it is acceptable, and reports it is:

$$\begin{aligned} P(\{\text{report } x_i \text{ acceptable}\}) &= \left(\begin{array}{c} P(\{\text{accept } x_i\} \,|\, \{\text{generate } x_i\}) \\ \times \\ P(\{\text{generate } x_i\}) \end{array} \right) \\ &= p(x_i) \times \frac{1}{N}. \end{aligned}$$

Now the loop keeps going until the basic process obtains an x_i that it decides is acceptable. This means the probability that the *loop* reports x_i is proportional to $p(x_i)$. But since $\sum_i p(x_i) = 1$, the probability that the *loop* reports x_i is *equal* to $p(x_i)$.

The problem here is that the basic process may not come up with an acceptable value x_i the first time; the loop might have to go on multiple times. If there are many x with small values of $p(x)$, we may find that it takes a very long time indeed to come up with a single sample. In the worst case, all values of $p(x)$ will be small. For example, for a uniform distribution on the range $1, \ldots, N$, $p(x)$ is $1/N$.

We can make things more efficient by noticing that multiplying the probability distribution by a constant doesn't change the relative frequencies with which numbers are selected. So this means that we can find the largest value $\hat{p}$ of $p(x)$, and form $q(x) = (1/\hat{p})p(x)$. Our process then becomes that shown in algorithm 5.2.

This process is not the most efficient available.

***** tree and point location algorithm

Listing 5.2: Matlab code for simple rejection sampling; this is somewhat more efficient.

```
function rnum=fasterrejectsample(pvec)
%
% pvec is a probability distribution over numbers
%  1, ..., size(pvec, 1)
%
nv=size(pvec, 1);
wv=max(pvec);
pv2=pvec/wv;  % we rescale
done=0;
while done==0
    ptr=floor(1+(nv-1e-10)*rand);
    % this gives me a uniform random
    % number in the range 1, .. nv
    pval=pv2(ptr); % work with rescaled probs
    if rand<pval
        done=1;
        % i.e. accept
    end
end
rnum=ptr;
```

5.3 SIMULATION EXAMPLES

Computing probabilities and expectations from simulations should be so natural to a computer science student that it can be hard to see the magic. You estimate the probability of an event $\mathcal{E}$ by writing a program that runs N independent simulations of an experiment, counts how many times the event occurs (which we write $\#(\mathcal{E})$), and reports

$$\frac{\#(\mathcal{E})}{N}$$

as an estimate of $P(\mathcal{E})$. This estimate will not be exact, and may be different for different runs of the program. It's often a good, simple estimate.

Choosing N depends on a lot of considerations. Generally, a larger N means a more accurate answer, and also a slower program. If N is too small, you may find that you report 1 or 0 for the probability (think of what would happen if you measured $P(H)$ for a coin with one flip). One strategy is to run several simulations, report the mean, and use the standard deviation as some guide to the accuracy.

5.3.1 Simulating Experiments

Worked example 5.12 *Getting 14's with 20-sided dice*

You throw 3 fair 20-sided dice. Estimate the probability that the sum of the faces is 14 using a simulation. Use $N = [1e1, 1e2, 1e3, 1e4, 1e5, 1e6]$. Which estimate is likely to be more accurate, and why?

Solution: You need a fairly fast computer, or this will take a long time. I ran ten versions of each experiment for $N = [1e1, 1e2, 1e3, 1e4, 1e5, 1e6]$, yielding ten probability estimates for each N. These were different for each version of the experiment, because the simulations are random. I got means of $[0, 0.0030, 0.0096, 0.0100, 0.0096, 0.0098]$, and standard deviations of [00.00670.00330.00090.00020.0001]. This suggests the true value is around 0.0098, and the estimate from $N = 1e6$ is best. The reason that the estimate with $N = 1e1$ is 0 is that the probability is very small, so you don't usually observe this case at all in only ten trials.

Worked example 5.13 *Comparing simulation with computation*

You throw 3 fair six-sided dice. You wish to know the probability the sum is 3. Compare the true value of this probability with estimates from six runs of a simulation using $N = 10000$. What conclusions do you draw?

Solution: I ran six simulations with $N = 10000$, and got $[0.0038, 0.0038, 0.0053, 0.0041, 0.0056, 0.0049]$. The mean is 0.00458, and the standard deviation is 0.0007, which suggests the estimate isn't that great, but the right answer should be in the range $[0.00388, 0.00528]$ with high probability. The true value is $1/216 \approx 0.00463$. The estimate is tolerable, but not super accurate.

Worked example 5.14 *The First Digit of a Random Number*

We have seen (section 4.2.2) that adding random variables tends to produce a normal random variable. What happens if one divides one random number by another? Solving this in closed form is tricky, but simulation gives some insight.

Solution: I wrote a short program that produced 20 experiments, each consisting of 1000 numbers. I obtained the numbers by generating two uniform random numbers in the range $[0 - 1]$, then dividing one by the other. Notice that doing so can produce both very big and very small numbers, so that plotting a histogram is tricky — most boxes will be empty unless we produce an immense quantity of data. Instead, I plot a histogram of the leading digit of the number in Figure 5.7.

Figure 5.7 also shows the mean over all 20 experiments of the fraction of

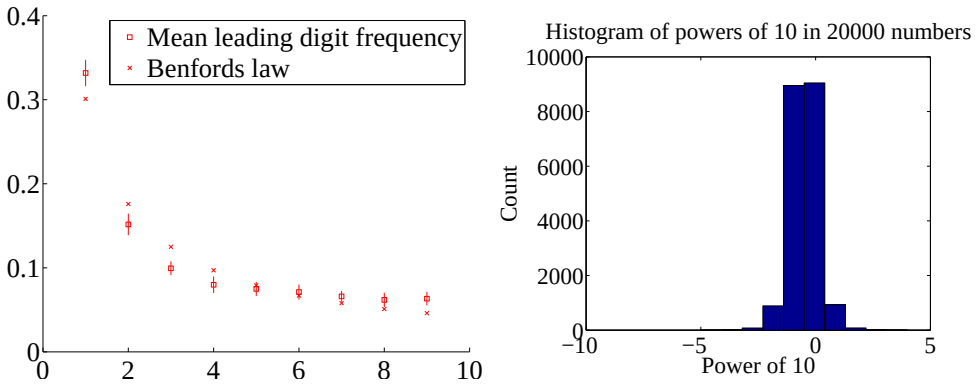

FIGURE 5.7: **Left** *summarizes results of a simulation where, for 20 runs, I generated 1000 pairs of independent random numbers uniformly distributed in* $[0-1]$ *and divided one by the other. The boxes show the mean over 20 runs, and the vertical bars show one standard deviation up and down. The crosses are predictions from Benford's law. Numbers generated in this way have a wide range of orders magnitude, so a conventional histogram isn't easy to plot. On the* **Right**, *a histogram of the first digit of the logarithm (i.e. the power of 10, or order of magnitude) of these numbers. Notice the smallest number is eight orders of magnitude smaller than the biggest.*

leading digits that is one, etc. together with one standard deviation error bars. The figure also shows a histogram of the rounded log (i.e. the power of 10) to give some idea of the range of values one obtains. It turns out that there is a widely applicable law, called **Benford's law**, that predicts the frequency of the first digit of many types of random number. The main condition for Benford's law to apply is that the numbers cover a wide range of scales. The figure shows predictions from Benford's law as well.

5.3.2 Simulating Markov Chains

Worked example 5.15 *Coin Flips with End Conditions*

I flip a coin repeatedly until I encounter a sequence HTHT, at which point I stop. What is the probability that I flip the coin nine times?

Solution: You might well be able to construct a closed form solution to this if you follow the details of example 22 and do quite a lot of extra work. A simulation is really straightforward to write; notice you can save time by not continuing to simulate coin flips once you've flipped past nine times. I got 0.0411 as the mean probability over 10 runs of a simulation of 1000 experiments each, with a standard deviation of 0.0056.

Worked example 5.16 *A Queue*

A bus is supposed to arrive at a bus stop every hour for 10 hours each day. The number of people who arrive to queue at the bus stop each hour has a Poisson distribution, with intensity 4. If the bus stops, everyone gets on the bus and the number of people in the queue becomes zero. However, with probability 0.1 the bus driver decides not to stop, in which case people decide to wait. If the queue is ever longer than 15, the waiting passengers will riot (and then immediately get dragged off by the police, so the queue length goes down to zero). What is the expected time between riots?

Solution: I'm not sure whether one could come up with a closed form solution to this problem. A simulation is completely straightforward to write. I get a mean time of 441 hours between riots, with a standard deviation of 391. It's interesting to play around with the parameters of this problem; a less conscientious bus driver, or a higher intensity arrival distribution, lead to much more regular riots.

Worked example 5.17 *Inventory*

A store needs to control its stock of an item. It can order stocks on Friday evenings, which will be delivered on Monday mornings. The store is old-fashioned, and open only on weekdays. On each weekday, a random number of customers comes in to buy the item. This number has a Poisson distribution, with intensity 4. If the item is present, the customer buys it, and the store makes $100; otherwise, the customer leaves. Each evening at closing, the store loses $10 for each unsold item on its shelves. The store's supplier insists that it order a fixed number k of items (i.e. the store must order k items each week). The store opens on a Monday with 20 items on the shelf. What k should the store use to maximise profits?

Solution: I'm not sure whether one could come up with a closed form solution to this problem, either. A simulation is completely straightforward to write. To choose k, you run the simulation with different k values to see what happens. I computed accumulated profits over 100 weeks for different k values, then ran the simulation five times to see which k was predicted. Results were $21, 19, 23, 20, 21$. I'd choose 21 based on this information.

For example 17, you should plot accumulated profits. If k is small, the store doesn't lose money by storing items, but it doesn't sell as much stuff as it could; if k is large, then it can fill any order but it loses money by having stock on the shelves. A little thought will convince you that k should be near 20, because that is the expected number of customers each week, so $k = 20$ means the store can expect to sell all its new stock. It may not be exactly 20, because it must depend a little on the balance between the profit in selling an item and the cost of storing it. For

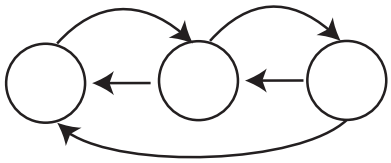

FIGURE 5.8: *A very small fraction of the web, drawn to suggest a finite state machine; each state represents a page, and each directed edge represents an outgoing link. A random web surfer could either (a) follow an outgoing link, chosen at random or (b) type in the URL of any page, chosen at random. Such a surfer would see lots of pages that have many incoming links from pages that have lots of incoming links, and so on. Pages like this are likely important, so that finding important pages is analogous to simulating a random web surfer.*

example, if the cost of storing items is very small compared to the profit, an very large k might be a good choice. If the cost of storage is sufficiently high, it might be better to never have anything on the shelves; this point explains the absence of small stores selling PC's.

5.3.3 Example: Ranking the Web by Simulating a Markov Chain

Perhaps the most valuable technical question of the last thirty years has been: Which web pages are interesting? Some idea of the importance of this question is that it was only really asked about 20 years ago, and at least one gigantic technology company has been spawned by a partial answer. This answer, due to Larry Page and Sergey Brin, and widely known as PageRank, starts with a Markov chain.

Figure 5.8 shows a picture of (a very small fraction of) the world wide web. I have drawn the web using a visual metaphor that should strongly suggest a finite state machine, and so a Markov chain. Each page is a state. Directed edges from page to page represent links. I count only the first link from a page to another page. Some pages are linked, others are not. I want to know how important each page is.

One way to think about importance is to think about what a random web surfer would do. The surfer can either (a) choose one of the outgoing links on a page at random, and follow it or (b) type in the URL of a new page, and go to that instead. As Figure 5.8 suggests, it is useful to think of this as a random walk on a finite state machine. We expect that this random surfer should see a lot of pages that have lots of incoming links from other pages that have lots of incoming links that (and so on). These pages are important, because lots of pages have linked to them.

For the moment, ignore the surfer's option to type in a URL. Write $r(i)$ for the importance of the i'th page. We model importance as leaking from page to page across outgoing links (the same way the surfer jumps). Page i receives importance down each incoming link. The amount of importance is proportional to the amount of importance at the other end of the link, and inversely proportional to the number of links leaving that page. So a page with only one outgoing link transfers all its importance down that link; and the way for a page to receive a lot of importance

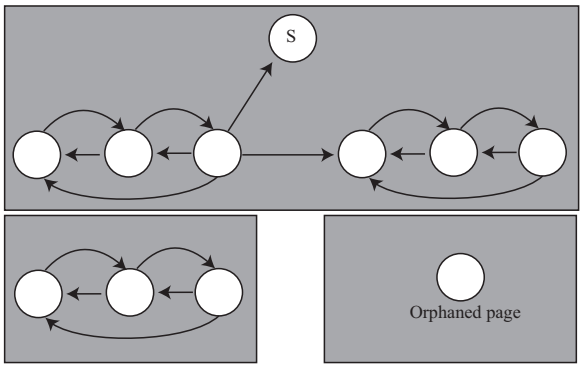

FIGURE 5.9: *The web isn't really like Figure 5.8. It's more like this figure (only bigger). In this figure, we see sections of the web that can't be reached by following links from other sections (the gray boxes); orphaned pages; and pages where a random walker would get stuck (S). Any algorithm for assigning importance must be able to deal with these effects.*

is for it to have a lot of important pages link to it alone. We write

$$r(j) = \sum_{i \to j} \frac{r(i)}{|i|}$$

where $|i|$ means the total number of links pointing *out* of page i. We can stack the $r(j)$ values into a *row* vector $\mathbf{r}$, and construct a matrix $\mathcal{P}$, where

$$p_{ij} = \begin{cases} \frac{1}{|i|} & \text{if } i \text{ points to } j \\ 0 & \text{otherwise} \end{cases}$$

With this notation, the importance vector has the property

$$\mathbf{r} = \mathbf{r}\mathcal{P}$$

and should look a bit like the stationary distribution of a random walk to you, except that $\mathcal{P}$ isn't stochastic — there may be some rows where the row sum of $\mathcal{P}$ is zero, because there are *no* outgoing links from that page. We can fix this easily by replacing each row that sums to zero with $(1/n)\mathbf{1}$, where n is the total number of pages. Call the resulting matrix $\mathcal{G}$ (it's quite often called the **raw Google matrix**). Notice that doing this doesn't really change anything significant; pages with no outgoing links leak a tiny fraction of their importance to every other page.

Figure 5.8 isn't a particularly good model of the web (and not because it's tiny, either). Figure 5.9 shows some of the problems that occur. There are pages with no outgoing links (which we've dealt with), pages with no incoming links, and even pages with no links at all. Worse, a random walk can get trapped (in one of the gray boxes). One way to fix all this would be to construct a process that wandered around the web inserting links that clean up the structure of the graph. This strategy is completely infeasible, because the real web is much too

big. Allowing the surfer to randomly enter a URL sorts out all of these problems, because it inserts an edge of small weight from every node to every other node. Now the random walk cannot get trapped.

There are a variety of possible choices for the weight of these inserted edges. The original choice was to make each inserted edge have the same weight. Write $\mathbf{1}$ for the n dimensional column vector containing a 1 in each component, and let $0 < \alpha < 1$. We can write the matrix of transition probabilities as

$$\mathcal{G}(\alpha) = \alpha\frac{(\mathbf{1}\mathbf{1}^T)}{n} + (1-\alpha)\mathcal{G}$$

where $\mathcal{G}$ is the original Google matrix. An alternative choice is to choose a weight for each web page, using anything from advertising revenues to page visit statistics to thaumaturgy (Google keeps quiet about the details). Write this weight vector $\mathbf{v}$, and require that $\mathbf{1}^T\mathbf{v} = 1$ (i.e. the coefficients sum to one). Then we could have

$$\mathcal{G}(\alpha, \mathbf{v}) = \alpha\frac{(\mathbf{1}\mathbf{v}^T)}{n} + (1-\alpha)\mathcal{G}.$$

Now the importance vector $\mathbf{r}$ is the (unique, though I won't prove this) *row* vector $\mathbf{r}$ such that

$$\mathbf{r} = \mathbf{r}\mathcal{G}(\alpha, \mathbf{v}).$$

How do we compute this vector? One natural algorithm is to start with some initial estimate, and propagate it. We write $\mathbf{r}^{(k)}$ for the estimated importance after the k'th step. We define updates by

$$(\mathbf{r}^{(k)}) = (\mathbf{r}^{(k-1)})\mathcal{G}(\alpha, \mathbf{v}).$$

We can't compute this directly, either, because $\mathcal{G}(\alpha, \mathbf{v})$ is unreasonably big so (a) we can't form or store $\mathcal{G}(\alpha, \mathbf{v})$ and (b) we can't multiply by it either. But we could estimate $\mathbf{r}$ with a random walk, because $\mathbf{r}$ is the stationary distribution of a Markov chain. If we simulate this walk for many steps, the probability that the simulation is in state j should be $r(j)$, at least approximately.

This simulation is easy to build. Imagine our random walking bug sits on a web page. At each time step, it transitions to a new page by either (a) picking from all existing pages at random, using $\mathbf{v}$ as a probability distribution on the pages (which it does with probability α); or (b) chooses one of the outgoing links uniformly and at random, and follows it (which it does with probability $1-\alpha$). The stationary distribution of this random walk is $\mathbf{r}$. Another fact that I shall not prove is that, when α is sufficiently large, this random walk very quickly "forgets" it's initial distribution. As a result, you can estimate the importance of web pages by starting this random walk in a random location; letting it run for a bit; then stopping it, and collecting the page you stopped on. The pages you see like this are independent, identically distributed samples from $\mathbf{r}$; so the ones you see more often are more important, and the ones you see less often are less important.

5.3.4 Example: Simulating a Complicated Game

I will build several examples around a highly simplified version of a real card game. This game is Magic: The Gathering, and is protected by a variety of trademarks,

etc. My version — MTGDAF — isn't very interesting as a game, but is simple enough to be studied, and interesting enough it casts some light on the real game. The game is played with decks of 60 cards. There are two types of card: Lands, and Spells. Lands can be placed on the play table and stay there permanently; Spells are played and then disappear. A Land on the table can be "tapped" or "untapped". Players take turns. Each player draws a hand of seven cards from a shuffled deck. In each turn, a player first untaps any Lands on the table, then draws a card, then plays a land onto the table (if the player has one in hand to play), then finally can play one or more spells. Each spell has a fixed cost (of $1, \ldots, 10$), and this cost is played by "tapping" a land (which is not untapped until the start of the next turn). This means that the player can cast only cheap spells in the early turns of the game, and expensive spells in the later turns.

Worked example 5.18 *MTGDAF — The number of lands*

Assume a deck of 60 cards has 24 Lands. It is properly shuffled, and you draw seven cards. You could draw $0, \ldots, 7$ Lands. Estimate the probability for each, using a simulation. Furthermore, estimate the error in your estimates.

Solution: The matlab function `randperm` produces a random permutation of given length. This means you can use it to simulate a shuffle of a deck, as in listing 5.3. I then drew 10, 000 random hands of seven cards, and counted how many times I got each number. Finally, to get an estimate of the error, I repeated this experiment 10 times and computed the standard deviation of each estimate of probability. This produced

0.0218 0.1215 0.2706 0.3082 0.1956 0.0686 0.0125 0.0012

for the probabilities (for 0 to 7, increasing number of lands to the right) and

0.0015 0.0037 0.0039 0.0058 0.0027 0.0032 0.0005 0.0004

for the standard deviations of these estimates.

Listing 5.3: Matlab code used to simulate the number of lands

```
simcards=[ones(24, 1); zeros(36, 1)]
% 1 if land, 0 otherwise
ninsim=10000;
nsims=10;
counts=zeros(nsims, 8);
for i=1:10
    for j=1:10000
        shuffle=randperm(60);
        hand=simcards(shuffle(1:7));
        %useful matlab trick here
        nlands=sum(hand);
        %ie number of lands
        counts(i, 1+nlands)=...
            counts(i, 1+nlands)+1;
        % number of lands could be zero
    end
end
probs=counts/ninsim;
mean(probs)
std(probs)
%%
```

Worked example 5.19 *MTGDAF — The number of lands*

What happens to the probability of getting different numbers of lands if you put only 15 Lands in a deck of 60? It is properly shuffled, and you draw seven cards. You could draw $0, \ldots, 7$ Lands. Estimate the probability for each, using a simulation. Furthermore, estimate the error in your estimates.

Solution: You can change one line in the listing to get

0.1159 0.3215 0.3308 0.1749 0.0489 0.0075 0.0006 0.0000

for the probabilities (for 0 to 7, increasing number of lands to the right) and

0.0034 0.0050 0.0054 0.0047 0.0019 0.0006 0.0003 0.0000

for the standard deviations of these estimates.

Worked example 5.20 *MTGDAF — Playing spells*

Assume you have a deck of 24 Lands, 10 Spells of cost 1, 10 Spells of cost 2, 10 Spells of cost 3, 2 Spells of cost 4, 2 Spells of cost 5, and 2 Spells of cost 6. Assume you always only play the cheapest spell in your hand (i.e. you never play two spells). What is the probability you will be able to play at least one spell on each of the first four turns?

Solution: This simulation requires just a little more care. You draw the hand, then simulate the first four turns. In each turn, you can only play a spell whose cost you can pay, and only if you have it. I used the matlab of listing 5.4 and listing 5.5; I found the probability to be 0.64 with standard deviation 0.01. Of course, my code might be wrong....

Worked example 5.21 *MTGDAF — Playing spells*

Now we use a different distribution of cards. Assume you have a deck of 20 Lands, 9 Spells of cost 1, 5 Spells of cost 2, 5 Spells of cost 3, 5 Spells of cost 4, 5 Spells of cost 5, and 11 Spells of cost 6. Assume you always only play the cheapest spell in your hand (i.e. you never play two spells). What is the probability you will be able to play at least one spell on each of the first four turns?

Solution: This simulation requires just a little more care. You draw the hand, then simulate the first four turns. In each turn, you can only play a spell whose cost you can pay, and only if you have it. I found the probability to be 0.33 with standard deviation 0.05. Of course, my code might be wrong....

One engaging feature of the real game that is revealed by these very simple simulations is the tension between a players goals. The player would like to have few lands — so as to have lots of spells — but doing so means that there's a bigger chance of not being able to play a spell. Similarly, a player would like to have lots of powerful (=expensive) spells, but doing so means there's a bigger chance of not being able to play a spell. Players of the real game spend baffling amounts of time arguing with one another about the appropriate choices for a good set of cards.

Worked example 5.22 *MTGDAF — How long before you can play a spell of cost 3?*

Assume you have a deck of 15 Lands, 15 Spells of cost 1, 14 Spells of cost 2, 10 Spells of cost 3, 2 Spells of cost 4, 2 Spells of cost 5, and 2 Spells of cost 6. What is the expected number of turns before you can play a spell of cost 3? Assume you always play a land if you can.

Solution: I get 6.3, with a standard deviation of 0.1. The problem is it can take quite a large number of turns to get three lands out. I used the code of listings 5.6 and 5.7

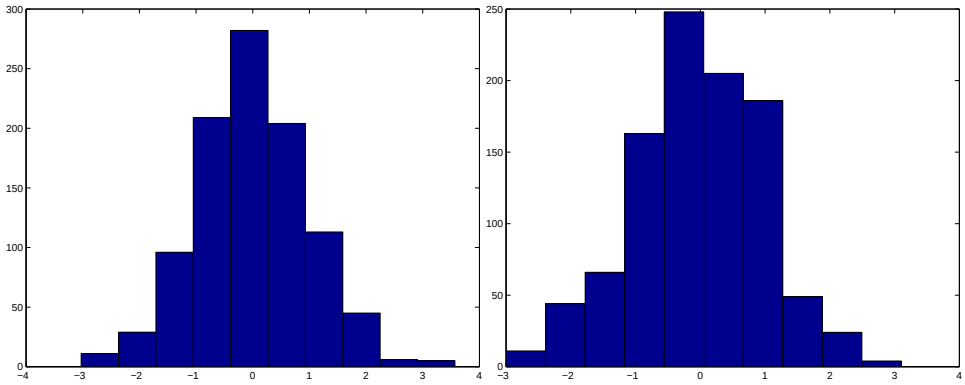

FIGURE 5.10: *Estimates of probabilities produced by simulation typically behave like normal data. On the* **left**, *I show a histogram of probabilities of having a hand of 3 Lands in the simulation of example 18; these are plotted in standard coordinates. On the* **right**, *I show a histogram of probability of playing a spell in each of the first four turns (example 21), from 1000 simulation experiments; again, these are plotted in standard coordinates. Compare these to the standard normal histograms of the previous chapter.*

Recall that you can reasonably expect that the probability you compute from a simulation behaves like normal data. Run a simulation experiment a large number of times and construct a data set whose entries are the probability from each experiment. This data should be normal. This means that, if you subtract the mean and divide by the standard deviation, you should get a histogram that looks like the standard normal curve. Figure 5.10 shows some examples of this effect. This effect is important, because it means that the right answer should be very few standard deviations away from the mean you compute — the standard deviation gives you quite a good idea of the accuracy of your estimate.

PROBLEMS

5.1. Multiple coin flips:
- **(a)** For example **??**, show that $P(5) = (3/32)$ by directly writing out the sequences of flips, and summing their probabilities.
- **(b)** Now check the recurrence relation $P(N) = (1/2)P(N-1)+(1/4)P(N-2)$ for the case $N = 5$.
- **(c)** What is $P(7)$?

5.2. Multiple die rolls: You roll a fair die until you see a 5, then a 6; after that, you stop. Write $P(N)$ for the probability that you roll the die N times.
- **(a)** What is $P(1)$?
- **(b)** Show that $P(2) = (1/36)$.
- **(c)** Draw a finite state machine encoding all the sequences of die rolls that you could encounter. Don't write the events on the edges; instead, write their probabilities. There are 5 ways not to get a 5, but only one probability, so this simplifies the drawing.

(d) Show that $P(3) = (1/36)$.

(e) Now use your finite state machine to argue that $P(N) = (5/6)P(N-1) + (25/36)P(N-2)$.

5.3. More complicated multiple coin flips: You flip a fair coin until you see either HTH or THT, and then you stop. We will compute a recurrence relation for $P(N)$.

(a) Figure **??** shows a finite state machine. Check that this finite state machine represents all coin sequences that you will encounter.

(b) Write Σ_N for some string of length N accepted by this finite state machine. Use this finite state machine to argue that $Sigma_N$ has one of four forms:

1. $TT\Sigma_{N-2}$
2. $HH\Sigma_{N-3}$
3. $THH\Sigma_{N-2}$
4. $HTT\Sigma_{N-3}$

(c) Now use this argument to show that $P(N) = (1/2)P(N-2)+(1/4)P(N-3)$.

5.4. The gambler's ruin:

(a) Show that you can rearrange the recurrence relation of example 22 to get

$$p_{s+1} - p_s = \frac{(1-p)}{p}\left(p_s - p_{s-1}\right).$$

Now show that this means that

$$p_{s+1} - p_s = \left(\frac{(1-p)}{p}\right)^2 (p_{s-1} - p_{s-2})$$

so that

$$\begin{aligned} p_{s+1} - p_s &= \left(\frac{(1-p)}{p}\right)^s (p_1 - p_0) \\ &= \left(\frac{(1-p)}{p}\right)^s (p_1 - 1). \end{aligned}$$

(b) Now we need a simple result about series. Assume I have a series u_k, $k \geq 0$, with the property that

$$u_k - u_{k-1} = cr^{k-1}.$$

Show that

$$u_k - u_0 = c\left(\frac{r^k - 1}{r - 1}\right).$$

(c) Use the results of the last two steps to show that

$$p_s - 1 = (p_1 - 1)\left(\frac{\left(\frac{1-p}{p}\right)^s - 1}{\left(\frac{1-p}{p}\right) - 1}\right)$$

(d) Use the fact that $p_j = 0$ and the result of the last exercise to show

$$(p_1 - 1) = \frac{-1}{\left(\frac{\left(\frac{1-p}{p}\right)^j - 1}{\left(\frac{1-p}{p}\right) - 1}\right)}.$$

(e) Use the results of the previous exercises to show that

$$p_s = \frac{\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^s}{\left(\frac{1-p}{p}\right)^j - 1}.$$

Listing 5.4: Matlab code used to simulate the four turns

```
simcards=[zeros(24, 1); ones(10, 1);...
    2*ones(10, 1);3*ones(10, 1); ...
    4*ones(2, 1); 5*ones(2, 1); 6*ones(2, 1)];
nsims=10;
ninsim=1000;
counts=zeros(nsims, 1);
for i=1:nsims
    for j=1:ninsim
        % draw a hand
        shuffle=randperm(60);
        hand=simcards(shuffle(1:7));
        %reorganize the hand
        cleanhand=zeros(7, 1);
        for k=1:7
            cleanhand(hand(k)+1)=cleanhand(hand(k)+1)+1;
            % ie count of lands, spells, by cost
        end
        landsontable=0;
        [playedspell1, landsontable, cleanhand]=...
            playround(landsontable, cleanhand, shuffle, ...
            simcards, 1);
        [playedspell2, landsontable, cleanhand]=...
            playround(landsontable, cleanhand, shuffle, ...
            simcards, 2);
        [playedspell3, landsontable, cleanhand]=...
            playround(landsontable, cleanhand, shuffle, ...
            simcards, 3);
        [playedspell4, landsontable, cleanhand]=...
            playround(landsontable, cleanhand, shuffle, ...
            simcards, 4);
        counts(i)=counts(i)+...
            playedspell1*playedspell2*...
            playedspell3*playedspell4;
    end
end
```

Listing 5.5: Matlab code used to simulate playing a turn

```
function [playedspell, landsontable, cleanhand]=...
    playround(landsontable, cleanhand, shuffle, simcards, ...
    turn)
        % draw
ncard=simcards(shuffle(7+turn));
cleanhand(ncard+1)=cleanhand(ncard+1)+1;
% play land
if cleanhand(1)>0
     landsontable=landsontable+1;
     cleanhand(1)=cleanhand(1)-1;
end
playedspell=0;
if landsontable>0
    i=1; done=0;
    while done==0
        if cleanhand(i)>0
            cleanhand(i)=cleanhand(i)-1;
            playedspell=1;
            done=1;
        else
            i=i+1;
            if i>landsontable
                done=1;
            end
        end
    end
end
```

Listing 5.6: Matlab code used to estimate number of turns before you can play a spell of cost 3

```
simcards=[zeros(15, 1); ones(15, 1);...
    2*ones(14, 1);3*ones(10, 1); ...
    4*ones(2, 1); 5*ones(2, 1); 6*ones(2, 1)];
nsims=10;
ninsim=1000;
counts=zeros(nsims, 1);
for i=1:nsims
    for j=1:ninsim
        % draw a hand
        shuffle=randperm(60);
        hand=simcards(shuffle(1:7));
        %reorganize the hand
        cleanhand=zeros(7, 1);
        for k=1:7
            cleanhand(hand(k)+1)=cleanhand(hand(k)+1)+1;
            % ie count of lands, spells, by cost
        end
        landsontable=0;
        k=0; played3spell=0;
        while played3spell==0;
            [played3spell, landsontable, cleanhand]=...
                play3round(landsontable, cleanhand, shuffle, ...
                simcards, k+1);
            k=k+1;
        end
        counts(i)=counts(i)+k;
    end
   counts(i)=counts(i)/ninsim;
end
```

Listing 5.7: Matlab code used to simulate a turn to estimate the number of turns before you can play a spell of cost 3

```
function [played3spell, landsontable, cleanhand]=...
    play3round(landsontable, cleanhand, shuffle, simcards, ...
    turn)
        % draw
ncard=simcards(shuffle(7+turn));
cleanhand(ncard+1)=cleanhand(ncard+1)+1;
% play land
if cleanhand(1)>0
     landsontable=landsontable+1;
     cleanhand(1)=cleanhand(1)-1;
end
played3spell=0;
if (landsontable>=3)&&(cleanhand(4)>0)
    played3spell=1;
end
```