Chapter 1

Basic ideas

In this chapter, we don't really answer the question 'What is probability?' Nobody has a really good answer to this question. We take a mathematical approach, writing down some basic axioms which probability must satisfy, and making deductions from these. We also look at different kinds of sampling, and examine what it means for events to be independent.

1.1 Sample space, events

The general setting is: We perform an experiment which can have a number of different outcomes. The *sample space* is the set of all possible outcomes of the experiment. We usually call it $\mathcal{S}$.

It is important to be able to list the outcomes clearly. For example, if I plant ten bean seeds and count the number that germinate, the sample space is

$$\mathcal{S} = \{0,1,2,3,4,5,6,7,8,9,10\}.$$

If I toss a coin three times and record the result, the sample space is

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

where (for example) HTH means 'heads on the first toss, then tails, then heads again'.

Sometimes we can assume that *all the outcomes are equally likely*. (Don't assume this unless either you are told to, or there is some physical reason for assuming it. In the beans example, it is most unlikely. In the coins example, the assumption will hold if the coin is 'fair': this means that there is no physical reason for it to favour one side over the other.) If all outcomes are equally likely, then each has probability $1/|\mathcal{S}|$. (Remember that $|\mathcal{S}|$ is the number of elements in the set $\mathcal{S}$).

Note the backward-sloping slash; this is not the same as either a vertical slash | or a forward slash /.

In the example, A' is the event 'more tails than heads', and $A \cap B$ is the event $\{HHH, THH, HTH\}$. Note that $P(A \cap B) = 3/8$; this is not equal to $P(A) \cdot P(B)$, despite what you read in some books!

1.2 What is probability?

There is really no answer to this question.

Some people think of it as 'limiting frequency'. That is, to say that the probability of getting heads when a coin is tossed means that, if the coin is tossed many times, it is likely to come down heads about half the time. But if you toss a coin 1000 times, you are not likely to get exactly 500 heads. You wouldn't be surprised to get only 495. But what about 450, or 100?

Some people would say that you can work out probability by physical arguments, like the one we used for a fair coin. But this argument doesn't work in all cases, and it doesn't explain what probability means.

Some people say it is subjective. You say that the probability of heads in a coin toss is $1/2$ because you have no reason for thinking either heads or tails more likely; you might change your view if you knew that the owner of the coin was a magician or a con man. But we can't build a theory on something subjective.

We regard probability as a mathematical construction satisfying some axioms (devised by the Russian mathematician A. N. Kolmogorov). We develop ways of doing calculations with probability, so that (for example) we can calculate how unlikely it is to get 480 or fewer heads in 1000 tosses of a fair coin. The answer agrees well with experiment.

1.3 Kolmogorov's Axioms

Remember that an event is a subset of the sample space $\mathcal{S}$. A number of events, say $A_1, A_2, \ldots$, are called *mutually disjoint* or *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for any two of the events A_i and A_j; that is, no two of the events overlap.

According to Kolmogorov's axioms, each event A has a probability $P(A)$, which is a number. These numbers satisfy three axioms:

Axiom 1: For any event A, we have $P(A) \geq 0$.

Axiom 2: $P(\mathcal{S}) = 1$.

On this point, Albert Einstein wrote, in his 1905 paper *On a heuristic point of view concerning the production and transformation of light* (for which he was awarded the Nobel Prize),

> In calculating entropy by molecular-theoretic methods, the word "probability" is often used in a sense differing from the way the word is defined in probability theory. In particular, "cases of equal probability" are often hypothetically stipulated when the theoretical methods employed are definite enough to permit a deduction rather than a stipulation.

In other words: Don't just assume that all outcomes are equally likely, *especially* when you are given enough information to calculate their probabilities!

An *event* is a subset of $\mathcal{S}$. We can specify an event by listing all the outcomes that make it up. In the above example, let A be the event 'more heads than tails' and B the event 'heads on last throw'. Then

$$\begin{aligned} A &= \{HHH, HHT, HTH, THH\}, \\ B &= \{HHH, HTH, THH, TTH\}. \end{aligned}$$

The probability of an event is calculated by adding up the probabilities of all the outcomes comprising that event. So, *if all outcomes are equally likely*, we have

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

In our example, both A and B have probability $4/8 = 1/2$.

An event is *simple* if it consists of just a single outcome, and is *compound* otherwise. In the example, A and B are compound events, while the event 'heads on every throw' is simple (as a set, it is $\{HHH\}$). If $A = \{a\}$ is a simple event, then the probability of A is just the probability of the outcome a, and we usually write $P(a)$, which is simpler to write than $P(\{a\})$. (Note that a is an *outcome*, while $\{a\}$ is an *event*, indeed a simple event.)

We can build new events from old ones:

- $A \cup B$ (read 'A union B') consists of all the outcomes in A or in B (or both!)
- $A \cap B$ (read 'A intersection B') consists of all the outcomes in both A and B;
- $A \setminus B$ (read 'A minus B') consists of all the outcomes in A but not in B;
- A' (read 'A complement') consists of all outcomes not in A (that is, $\mathcal{S} \setminus A$);
- $\emptyset$ (read 'empty set') for the event which doesn't contain any outcomes.

(each contains only one element which is in none of the others), and $A_1 \cup A_2 \cup \cdots \cup A_n = A$; so by Axiom 3a, we have

$$P(A) = P(a_1) + P(a_2) + \cdots + P(a_n).$$

Corollary 1.2 *If the sample space $\mathcal{S}$ is finite, say $\mathcal{S} = \{a_1, \ldots, a_n\}$, then*

$$P(a_1) + P(a_2) + \cdots + P(a_n) = 1.$$

For $P(a_1) + P(a_2) + \cdots + P(a_n) = P(\mathcal{S})$ by Proposition 1.1, and $P(\mathcal{S}) = 1$ by Axiom 2. Notice that once we have proved something, we can use it on the same basis as an axiom to prove further facts.

Now we see that, if all the n outcomes are equally likely, and their probabilities sum to 1, then each has probability $1/n$, that is, $1/|\mathcal{S}|$. Now going back to Proposition 1.1, we see that, *if all outcomes are equally likely*, then

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

for any event A, justifying the principle we used earlier.

Proposition 1.3 $P(A') = 1 - P(A)$ *for any event A.*

Let $A_1 = A$ and $A_2 = A'$ (the complement of A). Then $A_1 \cap A_2 = \emptyset$ (that is, the events A_1 and A_2 are disjoint), and $A_1 \cup A_2 = \mathcal{S}$. So

$$\begin{aligned} P(A_1) + P(A_2) &= P(A_1 \cup A_2) \quad \text{(Axiom 3)} \\ &= P(\mathcal{S}) \\ &= 1 \quad \text{(Axiom 2)}. \end{aligned}$$

So $P(A) = P(A_1) = 1 - P(A_2)$.

Corollary 1.4 $P(A) \leq 1$ *for any event A.*

For $1 - P(A) = P(A')$ by Proposition 1.3, and $P(A') \geq 0$ by Axiom 1; so $1 - P(A) \geq 0$, from which we get $P(A) \leq 1$.

Remember that if you ever calculate a probability to be less than 0 or more than 1, you have made a mistake!

Corollary 1.5 $P(\emptyset) = 0$.

For $\emptyset = \mathcal{S}'$, so $P(\emptyset) = 1 - P(\mathcal{S})$ by Proposition 1.3; and $P(\mathcal{S}) = 1$ by Axiom 2, so $P(\emptyset) = 0$.

Axiom 3: If the events $A_1, A_2, \ldots$ are pairwise disjoint, then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

Note that in Axiom 3, we have the union of events and the sum of numbers. Don't mix these up; never write $P(A_1) \cup P(A_2)$, for example. Sometimes we separate Axiom 3 into two parts: Axiom 3a if there are only finitely many events $A_1, A_2, \ldots, A_n$, so that we have

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i),$$

and Axiom 3b for infinitely many. We will only use Axiom 3a, but 3b is important later on.

Notice that we write

$$\sum_{i=1}^{n} P(A_i)$$

for

$$P(A_1) + P(A_2) + \cdots + P(A_n).$$

1.4 Proving things from the axioms

You can prove simple properties of probability from the axioms. That means, every step must be justified by appealing to an axiom. These properties seem obvious, just as obvious as the axioms; but the point of this game is that we assume only the axioms, and build everything else from that.

Here are some examples of things proved from the axioms. There is really no difference between a theorem, a proposition, and a corollary; they all have to be proved. Usually, a theorem is a big, important statement; a proposition a rather smaller statement; and a corollary is something that follows quite easily from a theorem or proposition that came before.

Proposition 1.1 *If the event A contains only a finite number of outcomes, say $A = \{a_1, a_2, \ldots, a_n\}$, then*

$$P(A) = P(a_1) + P(a_2) + \cdots + P(a_n).$$

To prove the proposition, we define a new event A_i containing only the outcome a_i, that is, $A_i = \{a_i\}$, for $i = 1, \ldots, n$. Then $A_1, \ldots, A_n$ are mutually disjoint