Chapter 3

Probability

The probability of something occurring is the quantification of the chance of observing a particular outcome given a single event or belief. The event itself may be the result of a single experiment, or one single data point collected by an un-repeatable experiment. We refer to a single event or an ensemble of events as data. The way we refer to data implies if data is singular or plural. If we quantify the probability of a repeatable experiment, then our understanding of past experiments can be used to make predictions of the outcomes of future experiments. We cannot predict the outcome of an experiment with certainty, however we can assign a level of confidence to our predictions that incorporates the uncertainty from our previous information and any knowledge of the limitations of the experiment to be performed.

Consider the following thought experiment: A scientist builds an experiment with two distinct outputs A and B. Having prepared the experiment, the scientist configures the apparatus to always return the result A, and never return the result B. If we perform this experiment over and over again we will always obtain the result A with certainty. The probability of obtaining this result is 100%. We will never observe the result B, and so the probability of obtaining that result is 0%. Some time later the scientist decides it is interesting to see what happens when the experiment is run run, however this time half of the time the result produced will be A, and the other half of the time the result B will be the outcome. In practice we often are faced with quantifying non-trivial probabilities like the second case of the above thought experiment. We normally work with an experiment that will yield a number of possible results, only a sub-set of which we consider interesting enough to analyze further.

The following sections illustrate the concept of probability and how one can formally consider how to compute the probability. Following the formal discussion, several popular examples are reviewed. These examples all involve discrete outcomes of something happening or not happening in a given sequence of events or repeatable experiments. The first example discussed in Section 3.2 is the case of repeatedly flip a coin. This is an extension of our idealized thought experiment discussed above.

3.1 Probability: A coin with two faces

When considering how to solve a problem in terms of probability there is a philosophical choice to be made. Sometimes the solution to the choice is straight forward, and sometimes it is simply an arbitrary one. The choice at hand is 'which philosophical approach shall I use to solve my problem?'. There are two schools of thought on this subject: Bayesian and Frequentist probability. These are summarised in brief in the following, and later material will give a more detailed digression on the subject.

If we consider a data set $\Omega = \{x|x_i\}$ that contains all possible elementary events x_i, then we can define the probability of obtaining a certain outcome or event as $P(x_i)$. We are able to write down several features of $P(x_i)$ that are relevant: (i) $P(x_i) \geq 0$ for all i; (ii) $P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$ where $i \neq j$; (iii) the total probability $P(\Omega) = 1$. We can put these features in words as:

- The probability of any x_i is not negative: $P(x_i) \geq 0$.
- Given two independent events x_i and x_j, the probability of x_i occurring is independent of the probability of x_j event from happening: $P(x_i \text{ or } x_i) = P(x_i) + P(x_j)$.
- The total probability for all of the possible events in Ω to happen is normalized to unity, so $P(\Omega) = 1$.

These features of probability can be used to determine the law of addition of probability, for two outcomes A and B we have:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

where $P(A \cap B)$ is the probability of intersection between the outcomes A and B, which is equivalent to saying that $i \neq j$. If $P(A \cap B) = \emptyset$, then A and B are independent outcomes, and we recover feature (ii).

It is also often useful to recall that if $P(A)$ is the probability for some outcome A to happen, then

$$P(\bar{A}) = 1 - P(A),$$

that is to say that the probability for some outcome A to happen or not is unity, and $P(\bar{A})$ is the compliment of $P(A)$.

3.1.1 Bayesian probability

The Bayesian school of probability stems from the work of Rev. Bayes in 1763. This approach interprets probability in terms of some degree of belief that a given event will happen.

The definition of Bayesian probability depends on a subjective input based on theory. In many circumstances, the theory can be relatively advanced, based on previous information that is relevant for a particular problem, however the opposite is also often the case and the theory expectations can be extremely crude. Bayes theorem states that for data A given some theory B:

$$P(B|A) = \frac{P(A|B)}{P(A)} P(B). \tag{3.1.1}$$

The terms used in this theorem have the following meanings:

$P(A)$	the probability of obtaining the data A, this is a normalization constant to ensure that the total probability for anything happening is unity. More generally $P(A)$ is a sum over possible outcomes or hypotheses B_i, i.e. $P(A) = \sum\limits_i P(A\|B_i)P(B_i)$.
$P(B)$	is called the prior probability. This is the subjective part of the Bayesian approach and it represents our degree of belief in a given theory or hypothesis B before any measurements are made.
$P(B\|A)$	is called the posterior probability and it represents the probability of the theory or hypothesis B given the data A.
$P(A\|B)$	is the probability of observing the data A given the theory or hypothesis B.

Sometimes it is not possible to compute the normalisation constant, in which case it is possible to work with the following proportionality derived from Eq. (3.1.1):

$$P(B|A) \propto P(A|B)P(B). \tag{3.1.2}$$

3.1.2 Frequentist probability

The frequentist approach interprets probability in terms of the relative frequency of a particular repeatable event to occur. A consequence of this is that one has to be able to discuss a repeatable event or experiment in order to apply frequentist techniques.

Consider a repeatable experiment where the set of outcomes is described completely in Ω. If there is a subset of n interesting events I, so $I \subset \Omega$, and a total of N elements in Ω, then the frequentist probability of obtaining an interesting event is given by

$$P(I) = \lim_{N\to\infty} \left(\frac{n}{N}\right). \tag{3.1.3}$$

At first glance it may appear that it is impractical to compute a frequentist probability as we clearly need to have an infinite set in order to produce an exact result. In practice however it turns out that in many cases one can compute an exact value of $P(I)$ without resorting to an infinite set. In other cases it is possible to compute $P(I)$ to sufficient precision with finite N. At all steps, frequentist probability is computed by following a defined set of logical steps. This in itself is an attractive feature of the prescription for many people.

3.1.3 Which approach should I use?

Sometimes individuals become very attached to one of the two schools of thought and loose focus on the problem at hand in order to concentrate on which statistical philosophy is correct. Such discussions detract from the task at hand: that of furthering scientific endeavor, and so it is best to remain impartial and get on with work rather than to engage in this subjective debate. In practice the Bayesian and Frequentist approaches on result in somewhat different predictions when the data are scarce to come by. When large data samples are available the probabilities computed using one method are much the same as those computed using the other. For this reason the choice between approaches should be considered subjective or arbitrary as long as both approaches remain valid. If it is simply not possible to repeat an experiment, then the computation of probability lends itself to the Bayesian approach.

3.2 Some Examples

This section discusses the outcomes of some familiar events that depend on chance, including the outcomes of flipping a coin, a lottery, and a popular card game. The final example demonstrates the use of a Bayesian approach.

3.2.1 Flipping a Coin

We can perform the repeatable experiment of flipping a coin. Each coin flip is an independent '*event*' and has two possible outcomes. We denote these outcomes corresponding to the coin landing heads-up and tails-up with H and T, respectively. If the coin is unbiased there is an equal probability of the coin toss resulting in H or T. If we run an ensemble of 20 experiments with an unbiased coin, then we will obtain a random sequence of results. One such set of data is the sequence $THHTHHTTTTTTHTTHHHHH$. There are $10H$ and $10T$ in this outcome, which is consistent with our naive expectations that on average half of the time we will obtain the result H and half of the time we will obtain T.

Is it possible to flip the coin 20 times and obtain the result $20H$? Yes of course, this is just one of the valid outcomes. The probability of obtaining H for any particular event is $1/2$. So the probability of obtaining $20H$ is $(1/2)^{20} = 1/1,048,576$. Now consider the particular result we obtained above:

$THHTHHTTTTTTHTTHHHHH$. The probability for this to happen is also given by $(1/2)^{20} = 1/1,048,576$[1]. So our particular outcome of $10H10T$ has a probability of a million to one, but so does $20H$ (and for that matter $20T$). If we don't care about the ordering of H and T, then there are many ways to flip a coin in order to obtain a $10H10T$ outcome where each particular outcome has a probability of $(1/2)^{20}$.

If we consider any combination of $10T + 10H$, there are $10! \times 10!$ ways to produce this combination out of a total of 20! combinations of a total of 20 H and T in any combination. This corresponds to a probability of obtaining any combination of $10T + 10H$ from 20 coin flips of 1 in 184,756, and the probability of obtaining any $10T + 10H$ combination from 20 coin tosses is 17.6%.

3.2.2 The National Lottery

Many countries run national or regional lotteries. The UK's National Lottery is an example of a unbiased experiment with many possible outcomes. This is more complicated than flipping a coin as six numbers are picked at random from a total of 49 numbers in the sequence 1 through 49. Each number can only be picked once, so after the first number is picked, there are 48 remaining numbers to choose from and so on. If the six numbers chosen by the lottery machine match the numbers on a lottery ticket purchased by a member of the public they will win the jackpot. But what is the probability of this event occurring?

In order to determine this we need to be able to easily calculate the number of possible outcomes, and the number of possible successful outcomes for each lottery draw. There are 49! combinations of numbers for the whole set of numbers. The order with which the winning numbers are selected does not matter, so there are 6! ways of selecting these six numbers. The remaining problem to solve is to determine how many combinations will result in the wrong set of numbers being selected. In selecting six numbers, we leave a further 43 numbers unselected in a complementary set. The total number of possible combinations of this complementary set is 43!. We can compute the probability of selecting the six winning numbers as the ratio of possible winning combinations divided by the total number of combinations. The total number of winning combinations is given by the product of the combinations of the winning six numbers and the complementary set. This ratio is

$$\frac{6!43!}{49!} = \frac{6!}{49 \times 48 \times 47 \times 46 \times 45 \times 44} = \frac{1}{13,983,816}.$$

So the probability of winning the jackpot is about 1 in 14 million. The notation 49!/(6!43!) is often written in short hand as ${}^{49}C_6$. The quantity ${}^{49}C_6$ corresponds to the number of unique combinations of selecting six good numbers, leaving a set of 43 complementary numbers from a total set of 49 numbers. There are 13,983,816 such combinations, only one of which is the winning combination for any given lottery. As successive lottery draws are independent, the probability calculated here is the probability of winning the lottery any given week. Given this, we can see that on average one needs to play the national lottery 13,983,816 times in any given week in order to be certain of winning the jackpot.

3.2.3 Blackjack

Blackjack is a card game, where the aim of the game is to obtain a card score of 21 or as close to that as possible with as few cards as possible. In this game the value of numbered cards is given by the number on the card, picture cards count for 10 points, and aces can count for either 11 or 1 point. So in order to win the game with 21 points and two cards dealt the player must have a picture card or 10, and an ace. There are 16 cards with a value of 10 points, and 4 cards with a value of 11 points out of a pack of 52 cards. Lets consider this simplest case. The probability of being dealt one ace is four out of 52 times, or 7.7%. The probability of subsequently being dealt a 10 point card is 16 out of 51 times, or 31.4%. So the combined probability of being dealt an ace followed by a 10 point card is the product of these two probabilities: 2.4%.

[1] Terry Pratchet would note at this point that *"It's a million to one chance, but it might just work"*. A slightly different viewpoint is that if you do something a million times, then expect an event with probability of a million to one to happen!

We don't care if we are dealt the ace as the first or second card, so we can also consider the possibility that we are dealt the 10 point card before the ace. This probability is also 2.4%. So the total probability of being dealt an ace and a 10 point card is the sum of these two probabilities, namely 4.8%. In reality the game is played by more than one person, so the probabilities are slightly modified to account for the fact that more than two cards are dealt from the pack before any of the players can reach a total point score of 21. With a little thought it is not difficult to understand how likely you are to be dealt any given combination of cards at a game such as this one.

3.2.4 The three cups problem

Consider the case when someone presents three cups C_1, C_2, and C_3, only one of which contains a ball. You're asked to guess which cup contains the ball, and on guessing you will make a measurement to identify if you have found a ball or not. What is the probability of finding a ball under cup C_1? In this example we will use Bayes theorem to compute the probability, so we start from:

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

Intuitively we can consider that all cups are equal, and so can simply assume that the probability of a cup containing a ball would be 1/3. That means that there is a probability of 1/3 that cup C_1 will contain a ball, and this gives us the prior probability of $P(B)$ based on our belief that the cups are all equivalent or unbiased.

So we have determined the prior, we now need to determine the probability of obtaining the data $P(A)$, and the probability of observing the outcome A given the prior. In order to compute the posterior probability we have to make a measurement, and in this case it means that we have to pick one of the cups and see if indeed that there is a ball under C_1. On doing this we find that you were lucky and found the ball under cup C_1. The probability of observing the ball under cup C_1 is 1/3 which is simply $P(A|B)$. The only thing that remains is for us to determine $P(A)$ which in general is given by:

$$\begin{aligned} P(A) &= \sum_i P(A|B_i)P(B_i) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3), \end{aligned}$$

where the probability $P(B_i)$ is simply the prior that we find the ball under the i^{th} cup, and B_i is the corresponding hypothesis (i.e. under which cup we look). So

$$\begin{aligned} P(A) &= \left(\frac{1}{3}\times\frac{1}{3}\right) + \left(\frac{1}{3}\times\frac{1}{3}\right) + \left(\frac{1}{3}\times\frac{1}{3}\right) \\ &= \frac{1}{3}. \end{aligned}$$

So in this example the normalisation term is simply given by $P(A) = 1/3$, where the probability $P(A|B_i)$ is the probability of obtaining an outcome A given the hypothesis B_i, which again is 1/3. Now if we return to Bayes theorem we find:

$$P(B|A) = \frac{1/3}{1/3} \times 1/3 = 1/3,$$

as we had originally expected.

In computing the probability of finding a ball under one particular cup, we have gone through the recipe prescribed by Bayes theorem. Given such a simple problem, the use of Bayes theorem was not the most sensible solution as it required a lot more work than the frequentist approach would have told us intuitively. What has been useful is to see how the theorem performs in a simple case so that we can apply Bayes theorem to more complicated problems in the future.