Chapter 2

Probability

The ***probability*** of something occurring is the quantification of the chance of observing a particular outcome given a single event. 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, and the way we refer to data implies if data is singular or plural. If we quantify the probability of a repeatable experiment, then this understanding can be used to make predictions of the outcomes of future experiments. We cannot predict the outcome of a given experiment with certainty, however we can assign a level of confidence to our predictions that incorporates the uncertainty from our previous knowledge and any information of the limitations of the experiment to be performed.

Consider the following: A scientist builds an experiment with two distinct outputs A and B. Having prepared the experiment, the apparatus is configured to always return the result A, and never return the result B. If the experiment is performed over and over again one will always obtain the result A with certainty. The probability of obtaining this result is 1.0 (100%). The result B will never be observed, and so the probability of obtaining that result is 0.0 (0%). Some time later the scientist decides it is interesting to see what happens when the experiment is 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 are often faced with quantifying probabilities like the second case above. This is discussed in more detail in the coin flipping example in section 2.7.1, and in the context of a more general approach in section 4.2. We often work with an experiment that will yield a number of possible results, only a sub-set of which we consider interesting enough to analyse further.

The following sections illustrate the concept of probability and how one can formally compute probabilities. Having introduced probability, the related concepts of probability density functions and likelihood are introduced. Following the formal discussion, several examples are reviewed. The first example discussed in section 2.7 is the case of repeatedly flipping a coin. This is an extension of the idealised experiment discussed above.

2.1 Elementary rules

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.

1. The probability of any x_i is not negative: $P(x_i) \geq 0$ for all i.
2. The probability of any x_i has an upper bound of unity. $P(x_i) = 1$ corresponds to an event x_i being the certain outcome of a measurement.

3. Given two ***independent events*** x_i and x_j, where $i \neq j$, the probability of x_i occurring is independent of the probability of x_j happening, and the probability for one or other event to occur is given by $P(x_i \text{ or } x_j) = P(x_i \vee x_j) = P(x_i) + P(x_j)$.

4. The ***total probability*** for all of the possible events in Ω to occur is normalised to unity, so $P(\Omega) = \sum_i P(x_i) = 1$.

5. Given two independent events x_i and x_j, the probability of x_i occurring is independent of the probability of x_j happening, and the probability for both events to occur is given by: $P(x_i \text{ and } x_j) = P(x_i \wedge x_j) = P(x_i) \cdot P(x_j)$.

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

$$\begin{aligned} P(A \cup B) &= P(A \vee B) - P(A \wedge B), & (2.1.1) \\ &= P(A) + P(B) - P(A \cap B), & (2.1.2) \end{aligned}$$

where $P(A \cap B)$ is the probability of intersection between the outcomes A and B. If $P(A \cap B) = \emptyset$, then A and B are independent outcomes, and we recover feature (3).

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

$$P(\overline{A}) = 1 - P(A), \tag{2.1.3}$$

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

2.2 Bayesian probability

The Bayesian school of probability stems from the work of Rev. Bayes published posthumously in 1763, reprinted in Ref. (Bayes, 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 often 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

$$P(B|A) = \frac{P(A|B)}{P(A)} P(B). \tag{2.2.1}$$

The terms used in this theorem have the following meanings

$P(B|A)$ 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. This is sometimes referred to as the ***a priori probability***.

$P(B)$ 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.

$P(A)$ the probability of obtaining the data A, this is a normalisation 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. $P(A) = \sum_i P(A|B_i)P(B_i)$.

Sometimes it is not possible to compute the normalisation constant $P(A)$, in which case it is possible to work with the following proportionality derived from Eq. (2.2.1)

$$P(B|A) \propto P(A|B)P(B). \tag{2.2.2}$$

The use of this proportionality is discussed further in section 7.4.

The notation introduced here naturally lends itself to discrete problems, however one can extend the notation to accommodate continuous variables. In this case the quantities $P(B|A)$, $P(A|B)$, $P(B)$, and hence $P(A)$ become functions of the continuous variables, and may also become functions of some parameters required to define the shapes of the underlying hypotheses being tested. Several examples of the use of Bayes theorem to compute posterior probabilities can be found in section 2.7.

2.2.1 Priors

The subjective part of the Bayesian prescription of probability is the choice of the prior probability. What form should this take? If the possible outcomes are discrete then the prior probability $P(B_i)$ will correspond to some number. If however the prior probability depends on the value of some continuous variable x, then the prior probability will take the form of a distribution $P(B_i) = \theta(x)$. Instead of assigning a single number for the prior one assumes the functional form $\theta(x)$ of a prior probability distribution. Often it is convenient to choose a ***uniform prior*** probability distribution (see appendix B.2.3), so effectively taking $P(B) = constant$. Such a choice is a representation of our ignorance of a problem where all values of x are taken to be equally likely however this is not always well motivated. One constraint on the choice of prior is that the result should be stable, and independent of this choice. Therefore one can check to see how a result depends on a given prior in order to establish the reliability of an estimate.

When dealing with a discrete problem, for example "Will it rain tomorrow?" from section 2.7.4, the prior probability that it will be rain is a discrete quantity associated with the hypothesis that it will rain, or the complement, that it will not rain. In order to compute a value to associate with the prior, one needs additional information, for example that it typically rains 131 days of the year in England, thus the prior probability that it will rain can be computed for this geographical location: $P(rain) = 131/365$. The posterior probability obtained for this example is an 88.7% chance of rain.

Without that prior information, one might assume that it would be equally likely to rain or not, and the choice of $P(rain)$ could be different. It would not be unreasonable to assume $P(rain) = 0.5$ in such a scenario. If one repeated the calculation in section 2.7.4 with $P(rain) = 0.5$, and $P(no\,rain) = 0.5$, then the probability obtained for rain tomorrow is 93.3%. While this numerically differs from the original calculation, one can see that the overall conclusion remains unchanged. It is more likely to rain than not, given that rain has been forecast.

When dealing with a continuous problem there are an infinite number of possible priors $\theta(x)$ to choose from, however in practice one may prefer to investigate results using a limited number of priors based on any available information. As mentioned above a common choice is a uniform prior

$$\theta(x) = \text{constant}, \tag{2.2.3}$$

where the prior is normalised appropriately to preserve the total probability. There is no variable dependence for a uniform prior, and the use of this can simplify computation (for example see the upper limit calculation discussed in section **??**). In the total absence of information a uniform prior over all allowed physical values

of the problem space may be reasonable. However if results already exist that may provide a weak constraint on the value of an observable then one could incorporate that information in the prior.

A detailed discussion of priors is beyond the scope of this book, and the interested reader is encouraged to refer to the references listed at the end of this book, in particular (James, 2007; Silvia and Skilling, 2008). The important issue to note has already been mentioned above: *any physical result reported should be independent of the choice of prior*, hence in general one should check a given result by validating that it is stable with respect to the choice of prior. If a result or conclusion depends strongly on the prior then this may indicate insufficient data is available to draw a conclusion, or poor choice of prior, or both. In general one should be careful to include sufficient information used to derive the result so that a third party can understand exactly what has been done.

2.3 Classic approach

The classic approach to probability is restricted to situations where it is possible to definitely assign a probability based on existing information. For example if one rolls an unbiased six-sided die, then the probability of having the die land face up on any particular side is equal to 1/6. The classic approach to probability can be used to calculate the probability of many different outcomes including the results of throwing dice, card games and lotteries. However this approach is limited to situations where the outcomes are finite and well defined.

If one considers the scenario of rolling two fair dice, it is possible to construct a matrix of possible outcomes for the first die and for the second. The probabilities for the set of outcomes is illustrated in Table 2.1, where each specific outcome has equal probability of occurring, namely $1/6 \times 1/6 = 1/36$. Possible outcomes of interest can be grouped together. For example the probability of rolling two sixes is 1/36 as this is one distinct outcome, which is the same as that for any other single outcome. If one only cares about rolling a double (both die giving the same value), then the probability is the sum of probabilities for the combinations 11, 22, 33, 44, 55, and 66, which is 1/6. If one only cares that a particular configuration of two different numbers is the result, for example that a one and a three is given, then there are two possible combinations to obtain that solution, hence the probability is 2/36.

Table 2.1: The probabilities of rolling two dice and obtaining the outcomes shown for the first die (rows) and the second die (columns).

	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

2.4 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

$$P(I) = \lim_{N\to\infty} \left(\frac{n}{N}\right). \tag{2.4.1}$$

At first glance it may appear that it is impractical to compute a frequentist probability as we 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 using the classic approach. 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. This approach is not without problems, and to illustrate this the important issue of checking the coverage quoted for an interval or limit computed using a frequentist approach is examined in section 6.8.3.

2.4.1 Which approach should I use?

Sometimes individuals become very attached to one of the main schools of thought and loose focus on the problem at hand in order to concentrate on which statistical philosophy is correct. While the classic or frequentist approach can lead to a well defined probability for a given situation, it is not always usable. In such circumstances one is then left with only one option: use Bayesian statistics[1]. In practice these approaches can give somewhat different predictions when the data are scarce to come by, however given sufficiently large samples of data 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 arbitrary as long as both approaches remain valid. In fact it can be very useful to consider the results of using both approaches when trying to interpret data, where the results obtained can be considered as cross-checks of each other. The more complex an analysis, the more desirable it may be to perform such an independent cross check to gain confidence that there are no significant problems in the analysis[2].

2.5 Probability density functions

A ***probability density function*** (PDF) is a distribution where the total area is unity and the variation of the PDF is related to the probability of something occurring at that point in the parameter space. If we consider the simple case of a one-dimensional PDF that describes a uniform distribution of some variable x between the values -1 and $+1$, then the PDF $f(x)$ is simply a straight line with zero gradient. The function itself is given by

$$f(x) = \frac{1}{A}, \tag{2.5.1}$$

where A is a ***normalisation constant*** determined by

$$A = \int_{-1}^{1} dx, \tag{2.5.2}$$

$$= 2. \tag{2.5.3}$$

[1] For example if it is not possible to repeat an experiment, then the computation of probability lends itself to the Bayesian approach.

[2] This is especially true when it comes to algorithms dependent on computer programs that may easily incorporate coding errors. If two independent methods provide consistent results one can be confident that either no such errors exist or are small, or alternatively complementary mistakes have been made. The latter possibility, while an option, is usually very unlikely.

So in this instance, the PDF is given by

$$f(x) = \frac{1}{2}. \tag{2.5.4}$$

If a set of data is described by a PDF, then we can use the PDF to tell us if we are more likely to find a data point at $x = x_1$ than at a neighbouring point $x = x_2$.

For example if we consider the uniform PDF above, then it follows that we are equally likely to find data points anywhere between the minimum and maximum values of x. We have already encountered this function above in the context of a uniform prior, where here the valid domain is $x \in [-1, +1]$.

More generally a PDF will be described by some function $f(x)$, where

$$\int_a^b f(x)dx = 1, \tag{2.5.5}$$

and a and b represent the limits of the valid domain for the function. The probability of obtaining a result between x and $x + dx$ is $f(x)dx$.

PDFs are useful concepts in many branches of statistical data analysis, and in particular they are useful in defining models that can be used to extract detailed information about data through optimisation techniques such as those described in chapter 8. Many commonly used PDFs can be found in chapter 4 and appendix B. While we often consider PDFs to be continuous distributions, it can also be useful to use discrete distributions to represent a PDF.

2.6 Likelihood

The ***likelihood*** $\mathcal{L}$ of something to occur is proportional to the probability and has the property

$$P = C \cdot \mathcal{L}, \tag{2.6.1}$$

where C is some constant of proportionality. The function $\mathcal{L}$ is typically normalised such that the maximum value is unity. For example, if one considers Eq. (2.5.4), then the corresponding likelihood function is given by

$$f(x) = 1, \tag{2.6.2}$$

where the constant of proportionality between the PDF and likelihood is $C = 1/2$. In addition to using the likelihood function $\mathcal{L}$, one often uses $-\ln \mathcal{L}$ in order to take advantage of the computational benefit associated with logarithms converting products into summations.

Whereas probability gives one the ability to discuss the absolute expectation of some experiment, the use of likelihood provides a natural way to investigate the relative expectations of an outcome. A number of practical applications of likelihood include the constructs based on ratios of likelihoods or minimising $-\ln \mathcal{L}$ in the case of the maximum likelihood fit approach discussed in chapter 8. A detailed treatment of likelihood can be found in Edwards (1992).

2.7 Case studies

This section discusses the outcomes of some familiar events that depend on chance, including the outcomes of tossing a coin, a lottery, and a popular card game. The final two examples demonstrate the use of a

Bayesian approach. These are examples of what is more generally referred to as Game Theory.

2.7.1 Tossing a coin (classical)

We can perform the repeatable experiment of tossing a coin. Each coin toss 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 toss 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 $1/2$, hence the probability of obtaining T is also $1/2$. So the probability of obtaining $20H$ is $(1/2)^{20} = 1/1,048,576$. Now consider the particular result we obtained above: $THHTHHTTTTTTHTT\ HHHHH$. The probability for this to happen is also given by $(1/2)^{20} = 1/1,048,576$. We can express this result in the following way: if you do something a million times, then expect an event with probability of a million to one to happen. 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 toss a coin in order to obtain a $10H10T$ outcome where each particular outcome has a probability of $(1/2)^{20}$. In fact the number of distinct $10H+10T$ outcomes from flipping a coin 20 times is given by 20!/10!10!. There are 184,756 combinations that result in a $10T + 10H$ result, each with a probability of $(1/2)^{20}$. So the probability of tossing a coin 20 times and obtaining one of these solutions is 17.6%. A formal treatment of this problem is discussed in the context of the binomial distribution section 4.2.

2.7.2 The national lottery (classical)

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 tossing 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 calculate the number of possible outcomes, and the number of possible successful outcomes for each lottery draw. There are 49! combinations 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. As the total number of winning combinations is given by the product of the combinations of the winning six numbers and the complementary set, the ratio is:

$$\frac{6!43!}{49!} = \frac{6!}{49 \times 48 \times 47 \times 46 \times 45 \times 44} = \frac{1}{13,983,816}. \tag{2.7.1}$$

So the probability of winning the jackpot is 1 in 13,983,816. 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 numbers, leaving a set of 43 complementary numbers from a total set of 49 numbers. As successive lottery draws are independent, the probability calculated here is the probability of winning the lottery any given draw. Given this, we can see that one needs to play the national lottery 13,983,816 times using unique combinations of numbers in any given draw in order to be certain of winning the jackpot.

2.7.3 Blackjack (classical)

Blackjack is a card game, where the aim is to obtain a card score of 21 or less, being as close to that number 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 points 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 of being dealt an ace and a 10 point card. 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% (assuming that there are no other players). So the combined probability of being dealt an ace followed by a 10 point card is the product of these two probabilities: 2.4%. 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.

2.7.4 Will it rain tomorrow? (Bayesian)

A weather forecast predicts that it will rain tomorrow. When it rains, 70% of the time, the forecast was correctly predicting rain, and when there was no rain forecast, only 5% of the time would it rain. It rains on average 131 days of the year in England. So given this information, what is the probability that it will rain tomorrow given the forecast? Starting from Bayes theorem,

$$P(B|A) = \frac{P(A|B)}{P(A)}P(B), \tag{2.7.2}$$

where the hypothesis that it will rain is given by $B = rain\ forecast$, and the data of it actually raining tomorrow is A. We know that the prior probability of rain, $P(rain)$, is $131/365 = 0.3589$, and the corresponding prior probability that it will not rain is $P(no\,rain) = 234/365 = 0.6411$. We are told above that $P(rain|rain\ forecast) = 0.7$, and that $P(rain|not\,forecast) = 0.05$, so the normalisation $P(A) = P(rain)$ is given by

$$\begin{aligned} P(rain) &= P(rain|rain\,forecast)P(rain) + \\ &\qquad P(rain|not\,forecast)P(no\,rain) &(2.7.3)\\ &= 0.7 \times 0.3589 + 0.05 \times 0.6411, &(2.7.4)\end{aligned}$$

and hence the posterior probability is

$$P(rain\,forecast|rain) = \frac{0.7 \times 0.3589}{0.7 \times 0.3589 + 0.05 \times 0.6411}. \tag{2.7.5}$$

Thus the probability that it will rain tomorrow given the forecast is 88.7%, so it may be a good idea to pick up an umbrella given this forecast.

2.7.5 The three cups problem (Bayesian)

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:

$$P(B_1|A) = \frac{P(A|B_1)}{P(A)}P(B_1). \tag{2.7.6}$$

Here hypothesis B_1 corresponds to the ball being being found under cup C_1. Intuitively we can consider that all cups are equal, and so one can 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. This gives us the prior probability of $P(B_1)$ based on our belief that the cups are all equivalent or unbiased. The prior probabilities that the ball will be found under C_2 or C_3, given by $P(B_2)$ and $P(B_3)$, respectively, are also 1/3.

So we have determined the prior probabilities, we now need to determine the probability of obtaining the data $P(A)$, and the probability of observing the outcome A given the prior B_1. The probability of observing the ball under cup C_1 is 1/3 which is $P(A|B_1)$, and similarly one finds the probabilities of 1/3 each for observing the ball under C_2 or C_3. The only thing that is still to be calculated is $P(A)$ which in general is given by

$$\begin{aligned} P(A) &= \sum_i P(A|B_i)P(B_i), &(2.7.7)\\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3), &(2.7.8)\end{aligned}$$

where the probability $P(B_i)$ is 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

$$\begin{aligned} P(A) &= \left(\frac{1}{3}\times\frac{1}{3}\right) + \left(\frac{1}{3}\times\frac{1}{3}\right) + \left(\frac{1}{3}\times\frac{1}{3}\right), &(2.7.9)\\ &= \frac{1}{3}, &(2.7.10)\end{aligned}$$

as we consider all cups to be equal and thus $P(B_i) = 1/3$ for all i. So in this example the normalisation term given by $P(A) = 1/3$, the prior probability $P(B_1) = 1/3$, thus the final quantity to be determined is $P(A|B_1)$. This is the probability of obtaining an outcome A given the hypothesis B_1, which again is 1/3. Now returning to Bayes theorem we find

$$P(B|A) = \frac{1/3}{1/3}\times 1/3 = 1/3. \tag{2.7.11}$$

This is the same result as one would have determined using a classical approach.