Bayesian book

Ben Lambert

June 15, 2015

Contents

Chapter 1

How to best use this book

1.1 Why don't more people use Bayesian statistics?

Many are discouraged from using Bayesian approaches to analysis due to its supposed *difficulty*, and dependence on mathematics. However, we would argue that this is, in part, a weakness of the existent literature on the subject, which this book looks to address. It also highlights how many books on classical statistics sweep their inherent complexity and assumptions under the carpet, resulting in texts which are easy to digest; meaning that for many the path of least resistance is to forge ahead with classical tools.

By its dependence on the logic of probability, this means on first glances, Bayesian statistics appears more mathematically-complex. However, what is often lost in introductory texts on Bayesian theory, is the intuitive explanations behind the mathematical formulae. In this text instead, we shift the emphasis towards the latter; choosing to focus on graphical and illustrative explanations rather than getting lost in the details of the mathematics, which to be honest, is not necessary for much of modern Bayesian analysis. We hope that by doing so, we shall lose fewer casualties to mathematical complexity, and redress the imbalance between classical and Bayesian analysis applications.

Again, on first appearances, the concept of the *prior* no doubt leads many to 'abandon ship' early on the path to understanding better Bayesian methodologies. However, as discussed in section 2.12.2, and will be covered in detail in chapter 5 which is fully-devoted to this subject, we hope to banish

this particular thorn in the side of would-be Bayesian statisticians.

The reliance on computing, in particular simulation, is also seen to add to the complexity of Bayesian approaches. Whilst, this is true, we argue that the modern algorithms used for simulation are straightforward to understand, and with modern software, easy to implement. Furthermore, the added complexity of simulation methods is more than compensated by the straightforward extension of Bayesian models to handle arbitrarily complex situations. Like most things worth learning, there is a slight learning curve to become acquainted with the languages used to write modern Bayesian simulations. However, we hope to make this curve sufficiently shallow by incremental introduction of elements used in these computational applications.

Part I

Understanding the Bayesian formula

Chapter 2

The subjective worlds of Frequentist and Bayesian statistics

2.1 Chapter mission statement

At the end of this chapter the reader will understand the similarities and differences between Frequentist and Bayesian statistics. Furthermore, the reader will gain knowledge about probability distributions, as well as how they can be manipulated to calculate quantities of interest. Finally, we discuss the benefits of using Bayesian statistics for a particular analysis.

2.2 Chapter goals

In life, we are often tasked with building predictive models to understand complex phenomena. As a first approximation, we often disregard parts of the system, which are not directly of interest; making the models *statistical* rather than deterministic. There are two distinct approaches to statistical modelling: Frequentist or Classical inference, and Bayesian. This chapter will explain the similarities between these two approaches, and importantly, indicate where they differ substantively. Bayesian statistics formulates models in terms of entities called *probability distributions*. It is thus

paramount to understand how to interpret, and manipulate these objects, and a significant proportion of this chapter will be devoted to this cause. Finally, the central part of Bayesian inference - the *Bayesian formula* - will be introduced.

2.3 The purpose of statistical inference

How much does a particular drug contribute to treatment success? What can an average student earn after obtaining a college education? Will the Democrats win the next US Presidential election? In life we often want to test theories, then go on to draw conclusions.

However, it is often impossible to exactly isolate the parts of a system which we want to examine. The outcome of history is hence determined by a nexus of interacting elements; each of which contributes to the reality that we witness. In the case of a drug trial, we may not be able to control the diets of participants, and are certainly unable to control for their idiosyncratic metabolisms, both of which could impact the results we see. Evaluating the return to a college education - there are a range of factors which affect the wage which an individual ultimately commands, of which education is only one element. The outcome of the next US Presidential election depends on party politics, the performance of the incumbent government, as well as the media's portrayal of the candidates.

In life noise obfuscates the signal. The wind, rain, snow and sleet make it difficult to forge a path to where we want to go.

Statistical inference allows us to draw conclusions in this blustery landscape; separating the signal from the noise. It is the logical framework which we can use to trial our beliefs against *data*.

In statistics, we formalise our beliefs in models of *probability*. The models are probabilistic because we assume we are ignorant to many of the multitude of interacting parts of a system, meaning we cannot say with certainty whether something will, or will not, occur.

Suppose we are evaluating the efficacy of a drug in a trial. We might suppose that *on average*, the drug might have a given probability of working as desired. However, before we carry out any trials, we are unaware of its exact treatment success rate. Fortunately, statistical inference allows us to

estimate this unknown characteristic, or *parameter*, from the data we are given.

There are two schools of thought for carrying out this process of inference: *Frequentist* and *Bayesian*. Although this book is devoted to the latter, we now spend some time comparing the two approaches, so that a reader is aware of the different paths taken to their shared goal.

2.4 The world according to Frequentists

In Frequentist or classical statistics, we suppose that our sample of data is the result of one of an infinite number of exactly-repeated experiments. The sample we see in this context is hence assumed to be the outcome of some probabilistic process. Any conclusions that we draw from this approach are based on the supposition that events occur with probabilities, which represent long-run frequencies.

For example, if we flip a coin, we might assume that any sequence of outcomes we obtain is indicative of the results that we might obtain if we were to conduct the experiment an infinite number of times. Further, we might take the proportion of heads observed in this infinite set of throws as defining the probability of obtaining a 'heads'. We suppose that this probability actually exists, and is fixed for each set of coin-throws that we carry out (see figure 2.1).

In general, in Frequentist statistics assume that the data is *random* and results from *sampling*, from a fixed and defined *population* distribution. For a Frequentist the noise that obscures the true signal of the population relationship in which we are interested is due to *sampling variation*; the fact that the sample that we pick will each time be slightly different, and not exactly representative of the population.

We may flip our coin 10 times, obtaining 7 heads even if the long-run proportion of heads is $\frac{1}{2}$. To a Frequentist, this is because we have picked a slightly odd sample from the population of infinitely-many repeated throws. Further, if we flip the coin another 10 times, we will likely get a different result, because we have picked a different sample.

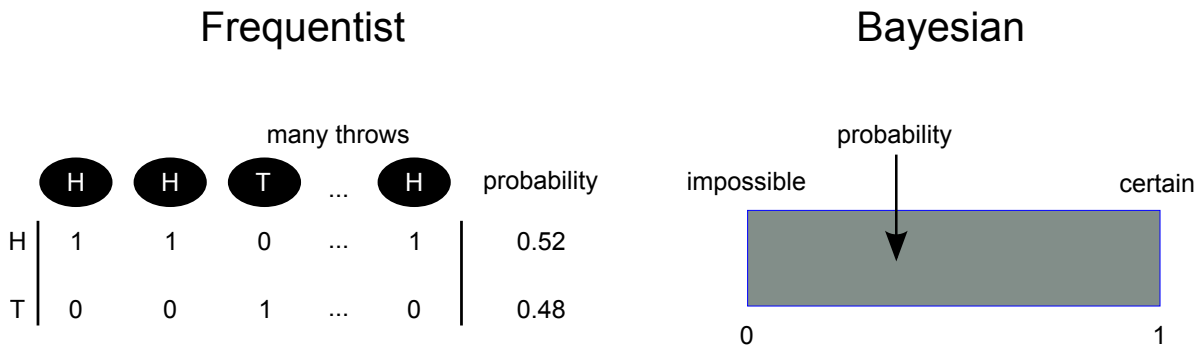

Figure 2.1: The Frequentist and Bayesian approaches to probability.

2.5 The world according to Bayesians

Bayesians do not imagine repetitions of an experiment in order to define and specify a probability. It is merely taken as a measure of certainty of a particular belief. From this viewpoint, the probability of us throwing a 'heads' measures and quantifies our underlying belief, that before we flip the coin, it will land this way.

In this sense, Bayesians do not view probabilities as concrete entities that actually exist. They are merely abstractions which we can use to help express our uncertainty. In this frame of reference there is no necessity for events to be repeatable in order to define a probability. We are thus equally able to say, 'The probability of a heads is 0.5', or, 'The probability of the democrats winning the 2020 US Presidential election is 0.75'. Probability is merely seen as a scale from: 0 where we are certain an event will not happen, to 1 where we are certain it will (see figure 2.1).

A statement such as 'The probability of the democrats winning the 2020 US Presidential election is 0.75' is hard to explain using the Frequentist definition of a probability. There is only ever one possible sample - the history that we witness - and what would we actually mean by a 'population of all possible US elections which happen in the year 2020'?

Probabilities are therefore seen as an expression of subjective beliefs, meaning they can be updated in light of new data. The formula invented by the Reverend Thomas Bayes provides the *only* logical manner in which to carry out this process, and is central to Bayesian inference, where we aim to express probabilistically our uncertainty in parameters after we have seen

the *data*.

Bayesians assume, since we are witness to the data, that it is *fixed*, and therefore does not vary. We do not need to imagine that there are an infinite number of possible samples. We 'see' our data, and hence do not need to view it as the outcome of some random process.

In contrast, we do not ever learn exactly the value of an unknown parameter. This epistemic uncertainty means that in Bayesian inference we choose to view the parameter as a quantity that is probabilistic in nature. We can view this in one of two perspectives. Either we view the unknown parameter as truly being *fixed* in some absolute sense, but our beliefs are uncertain, and thus probabilistic. Alternatively, we can take the view that there isn't some definitive *population* process, and for each sample we take, we get a slightly different parameter.

In the latter perspective we get different results from the coin flipping because each time we are subjecting our system to a slightly different probability of it landing 'heads' up. This could be because we mildly altered our throwing technique, or started with the coin in a different position. In the former perspective, we view the sample as a noisy representation of the signal, and hence we get different results for each set of throws. Although these two descriptions are different philosophically, they are not mathematically, meaning we can apply the same analysis to both.

2.6 Frequentist and Bayesian inference

The Bayesian inference process is the only logical way to modify our beliefs to take into account new data. We start out before we collect data with a probabilistic description of our beliefs, which we call a *prior*. We then collect data, and together with a model describing our theory, Bayes' formula for probability allows us to calculate our post-data or *posterior* belief:

$$prior + data \xrightarrow{model} posterior \tag{2.1}$$

In Bayesian inference, we want to draw conclusions based on purely probabilistic descriptions of phenomena. If we wish to summarise our evidence for a particular hypothesis, we describe this probabilistically, as the 'probability of the hypothesis *given* the data obtained'.

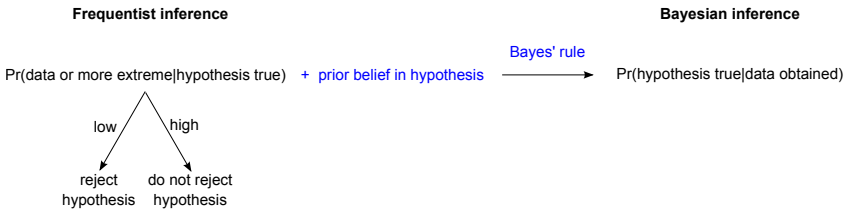

Figure 2.2: Frequentist and Bayesian inference.

The difficulty in obtaining this type of conclusion is that when we write down a probability model describing our process of interest, we can only use it to compute the 'probability of obtaining our data *given* our hypothesis being true'; the inverse of that which we desire. This probability is calculated by taking into account all the possible samples that we could have obtained from the population, if we assume the hypothesis is true.

Frequentists stop here; using this probability as evidence for a particular hypothesis in question. If the probability of obtaining the data, or a data sample more extreme, given a hypothesis is small, then it is assumed that it is unlikely that the hypothesis is true, and is rejected. Note however if reject a hypothesis, we have no way of telling whether it is true, and we just witnessed a weird sample, or that it is actually false. Also, note that in this inference process, we have also had to imagine obtaining a sample more extreme than our result, in order to get a usable probability.

Bayes' formula allows us to circumvent these difficulties, by inverting the Frequentist probability to get the 'probability of the hypothesis given the *actual* data we obtained'. There is no need for an arbitrary cut-off in the probability in order to validate the hypothesis; all information is summarised in this probability, and hence can be viewed as the end point of an analysis in itself.

The next few, albeit silly, and (more than) somewhat contrived, illustrate a difference both in methodology, but perhaps more significantly, in philosophy, between the two different approaches.

2.6.1 The Frequentist and Bayesian murder trials

Assume you find yourself in the unfortunate situation where you are (hopefully falsely) accused of murder, and face a trial by jury. A variation in the

usual tale is that you personally have a choice over the method used by the jury to assign guilt: either Frequentist or Bayesian. Another unfortunate twist is that the legal system of the country starts by presuming *guilt* rather than *innocence*.

Let's assume that you have been shown by a security camera to have definitely been in the same house as the victim - Sally - on the night of her demise.

If you choose the Frequentist trial, your jurors start by coming up with a model based on previous trials, which assigns a probability of you being seen by the security camera if you were guilty. They then use this to make the statement that, 'If you did commit the murder, then 30% of the time you would have been seen by the security camera', based on a hypothetical infinity of repetitions of the same conditions. Since this is not sufficiently unlikely (the p value is not below 5%), the jurors cannot reject the null hypothesis of *guilt*, and you are sentenced to life in prison.

In a Bayesian trial, the jury are first introduced to an array of evidence, that suggests that you neither knew Sally, nor had any previous record of violent conduct; being otherwise a perfectly respectable citizen. Furthermore, the ex-boyfriend of Sally is a multiple-violent-offending convict on the run from prison after being sentenced by a judge on the basis of witness testimony by Sally. On this basis, the jury determine a *prior* probability in the hypothesis that you are guilty that is $\frac{1}{1000}$. They then use the same model as the Frequentists, to determine that the probability of you being seen by the security camera given your guilt is 30%. However, they then coolly use Bayes' rule, and conclude that the probability of you committing the crime is $\frac{1}{1000}$ (see section 2.15.1 for a full description of this calculation). Based on this evidence, the jury acquits you, and you go home to your family.

2.6.2 Radio control towers: example

In a hypothetical war two radio control workers sit side-by-side, Mr Pearson (from frequentland), and Mr Laplace (from the county of Bayesdom), and are tasked with finding an enemy plane that has been spotted over the country's borders. They will each feed this information to the nearest airforce base(s) which will respond by sending up aircraft of their own. There are however, administratively two different airforces, which correspond to the two different counties. Although the airforces of frequentland and

Bayesdom share airbases, they are distinct, and only respond to Mr Pearson and Mr Laplace's advice respectively. The war, although short, has been costly to both allies, and they each want to avoid needless expenditure, as well as the unwarranted scaring of the local populace by sending up jets.

Mr Pearson starts by inputting the radar information into a computer program which uses a model of a plane's position which has been calibrated against a dataset of historical plane data in this short war. The result comes out instantly.

"...The plane is most likely 5 miles from the town of Tunbridge Wells."

Without another moment's thought, Mr Pearson radios the base of Tunbridge Wells, telling them to scramble all 10 available Frequentist fighter jets immediately. He then gets up to get himself a well-earned coffee.

Laplace knows from experience there are three different flight paths that the enemy has used to attack previously. Accordingly, he gives these regions a high probability density in his prior for the plane's current location, and feeds this into the same computer program that Pearson used. The output this time is different. By using the optional input, the program now outputs a map with the most likely regions shown via a colour shading. There is the highest posterior density over the region near Tunbridge Wells, where Pearson radioed, although the map suggests there are two other towns which might be likely victims of the plane's bombing. Accordingly, Laplace radios to Tunbridge Wells, asking them to send up four jets, and to the other two towns, asking them to send up two jets each. At the end of this all, Laplace remains seated, tired but contented that he has done his best for his own.

The enemy bomber turned out to be approaching berkstad, one of the towns which Laplace radioed. The Bayesdom jets intercept the encroaching aircraft, and escort it out of allied airspace. Laplace is awarded a medal in honour of his efforts. Pearson looks on jealously.

2.7 Probability distributions: helping us explicitly state our ignorance

Before we look out the window in the morning, before we get our exam results, before the cards are dealt, we are uncertain of the world that lies in

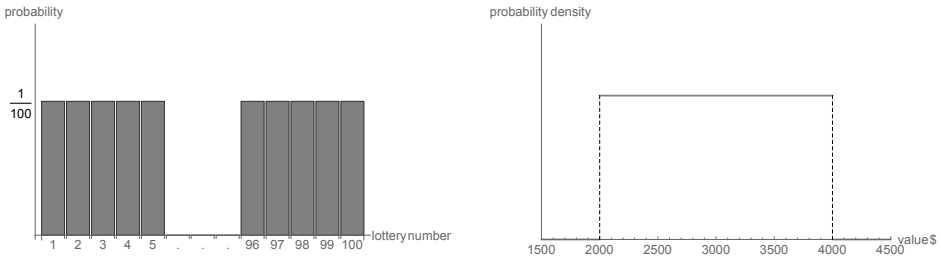

Figure 2.3: Probability distributions representing **left:** the chance of winning a lottery, and **right:** the value of a second-hand car.

wait of us. In order to plan, as well as make sense of things, we often predict the relative likelihood of different outcomes. However, in order to allow interrogation of thought, with a view to transparency and self-improvement, we sometimes would like to state our pre-conceptions *explicitly*, using a suitable framework.

The mathematical theory of probability provides a logic and language which is suitable to describe the majority of cases in which we are uncertain. Imagine that we enter a lottery, where we select a number from 1-100, to have a chance of winning $1000. We suppose that in the lottery only one ball is drawn, and it is fair with all numbers being equally likely to win. Although we haven't stated this world-view in mathematical notation, we have without realising it, formulated a valid probability distribution[1] for the number drawn in the lottery (see 2.3).

2.7.1 What make a probability distribution *valid*?

The lottery example given in section 2.7 refers to discrete probability distribution, since the variable we were measuring - the winning number - is confined to take on finite set of values. However, we could similarly define a probability distribution where our variable is able to take on an infinity of values across a spectrum. Imagine that before test drive a second-hand car we are uncertain about its value. We might think that from seeing pictures of the car, that it could be worth anywhere from $2000 to $4000, with all

[1] This is technically a probability mass function, since we are describing a discrete random variable, but I prefer to not differentiate terminology.

values being equally likely (see 2.3).

The aforementioned examples are both examples of valid/proper probability distributions. So, what are their defining properties?

- All values of the distribution must be real, and non-negative.
- The sum (integral) across all possible values of the discrete (continuous) random variable must be 1.

In the lottery case, this is satisfied since $p(X) = \frac{1}{100} \geq 0$, and:

$$\frac{1}{100} + \frac{1}{100} + ... + \frac{1}{100} = \sum_{i=1}^{100} \frac{1}{100} = 1 \tag{2.2}$$

For the continuous case of the probability of a heads when flipping a coin, the probability density function is always $\frac{1}{2000} \geq 0$, and when we do the continuous analogue of summing - integrating - we find that:

$$\begin{aligned} \int_{2000}^{4000} p(v)\mathrm{d}v &= \int_{0}^{1} \frac{1}{2000}\mathrm{d}v \\ &= 1 \end{aligned} \tag{2.3}$$

Although, it may seem that this definition is relatively arbitrary, and perhaps well-trodden-territory for some readers, it is of *central* importance to Bayesian statistics. This is because Bayesians like to work with, and produce *valid* probability distributions. The pursuit of this ideal underlies the majority of *all* methods in applied Bayesian statistics - analytic and computational - and hence its importance cannot be overstated!

2.7.2 Interpreting discrete and continuous probability distributions

The discrete probability distribution for the lottery shown on the left hand side in figure 2.3, is straightforward to interpret. To calculate the probability that the winning number is 3, we simply read off the probability from the graph corresponding to the height of the leftmost bar, and find that:

$$p(X = 3) = \frac{1}{100} \tag{2.4}$$

In the discrete case, if we want to calculate the probability that a random variable takes on a range of values, then we simply need to sum the individual probabilities corresponding to each specific event. In the die example, if we want to calculate the probability that the winning number is 10 or less, we just add together the probabilities of it being {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}:

$$\begin{aligned} p(X \leq 10) &= p(X = 1) + p(X = 2) + p(X = 3) + ... + p(X = 9) + p(X = 10) \\ &= \frac{1}{100} + \frac{1}{100} + \frac{1}{100} + ... + \frac{1}{100} + \frac{1}{100} \\ &= \frac{1}{10} \end{aligned} \tag{2.5}$$

How can we use the continuous probability distribution such as the one shown on the right hand side of figure 2.3? If we want to calculate the probability that the value of the second-hand car is $2,500, then we could simply draw a vertical line from this point on the *value* axis up to the line of the distribution; concluding that $p(value = \$2,500) = \frac{1}{2000}$! However, under this logic, we could also deduce that the probability of the value of the car being {$2,500,$2,500.10,$2,500.01,$2,500.001} are all $\frac{1}{2000}$. Furthermore, we could generate an infinity of these test values of *value*, meaning that if we summed them all together we could get a total probability of ∞.

There is evidently something wrong with our method for interpreting continuous densities. If we reconsider the test values $\{\$2,500, \$2,500.10, \$2,500.01, \$2,500.001\}$, we reason that these are all equally unlikely, and part of a set of an infinity of potential values we could draw. This means for a continuous random variable, we always have $p(\theta = number) = 0$. Hence, when we write $p(\theta)$ for a continuous random variable, we should be careful to interpret the value of it at a particular value as a probability *density*, *not* a probability.

However, we can use a continuous probability distribution to calculate the probability that a random variable lies between two bounds. To do this we use the continuous analogue of a sum, an *integral*. For the car example, we can calculate, $\$2,500 \leq \theta \leq \$3,000$:

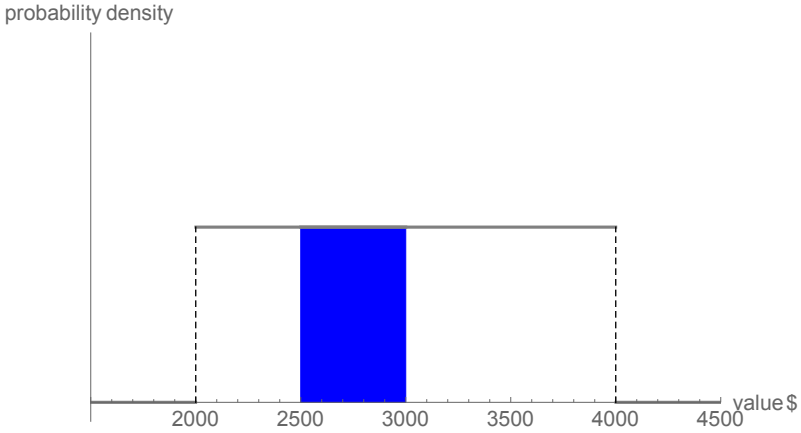

Figure 2.4: The probability that a second-hand car's value lies between $2,500 and $3,000.

$$
\begin{aligned}
Pr(2500 \leq value \leq 3000) &= \int_{2500}^{3000} p(v)\mathrm{d}v \\
&= \int_{2500}^{3000} \frac{1}{2000}\mathrm{d}v \\
&= \left[\frac{1}{2000}\, v\right]_{3000}^{2500} = \frac{1}{2000}(3000 - 2500) = 0.25
\end{aligned}
\tag{2.6}
$$

In (2.6), we have used Pr to explicitly state that the result is a *probability*, whereas $p(\theta)$ is a probability density. Of course, the calculation carried out in (2.6), is equivalent to working out the area under the graph within those limits (see figure 2.4).

2.7.3 Mean and variance of distributions

A popular way of summarising a distribution is via its *mean*, which is one measure of central tendency of a distribution. More intuitively, a mean, or *expected value*, of a distribution represents the long-run average value that would be obtained if we sampled from that particular distribution in question, an infinite number of times.

The way in which we calculate the *mean* of a distribution depends on whether it is *discrete* or *continuous* in nature. However, the concept is essentially the same in both cases. The mean is calculated as a weighted sum of the values taken on by the random variable in question, where the weights are provided by the probability distribution. This results in the following forms for the mean of a discrete and continuous variable respectively:

$$\mathbb{E}(X) = \sum_{All\ \alpha} \alpha Pr(X = \alpha) \tag{2.7}$$

$$\mathbb{E}(X) = \int_{All\ \alpha} \alpha p(\alpha) \mathrm{d}\alpha \tag{2.8}$$

In (2.7) and (2.8), α represents the multitude, or continuum of *values* taken on by the random variable X respectively. We have chosen to use Pr in (2.7), and p in (2.8), to illustrate that these represent probabilities and probability *densities* respectively.

We can now apply (2.7) to allow us to calculate the mean winning number from the lottery:

$$\begin{aligned} \mathbb{E}(X) &= \sum_{\alpha=1}^{100} \alpha Pr(X = \alpha) \\ &= 1 \times \frac{1}{100} + 2 \times \frac{1}{100} + 3 \times \frac{1}{100} + ... + 99 \times \frac{1}{100} + 100 \times \frac{1}{100} \\ &= 50\tfrac{1}{2} \end{aligned} \tag{2.9}$$

We can also demonstrate the *long-run* nature of the mean value of $50\frac{1}{2}$ found in (2.9) by simulating a number of rolls of a fair die computationally (see figure 2.5). As the number of rolls increases, the running mean tends towards this value.

We can also apply (2.8) to calculate our expectation of the second-hand car value:

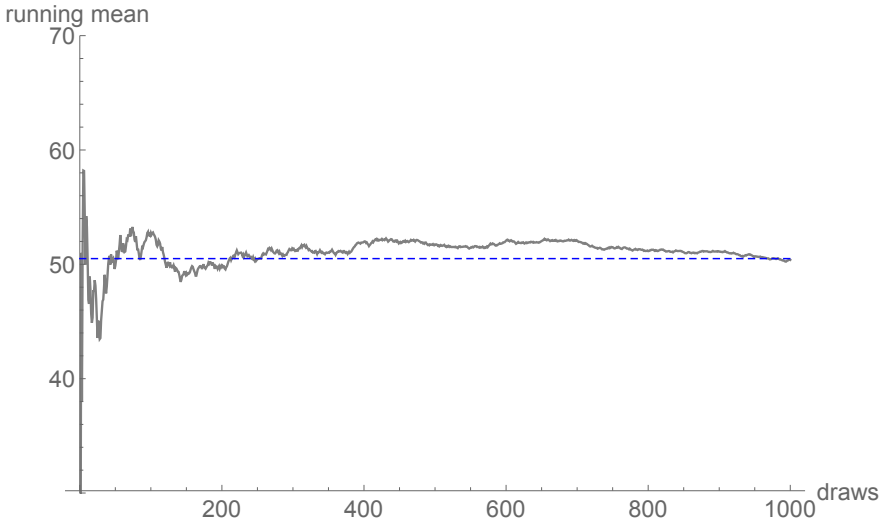

Figure 2.5: Playing a computational lottery. We see the approach of the running mean of repeatedly playing the lottery to the long-run mean of $50\frac{1}{2}$, as the number of plays increases.

$$\begin{aligned}\mathbb{E}(value) &= \int_{2000}^{4000} vp(v)v\mathrm{d}v \\ &= \frac{1}{2000}\left[\frac{v^2}{2}\right]_{2000}^{4000} \\ &= \$3,000\end{aligned} \tag{2.10}$$

If we were to have a business buying (and selling) second-hand cars, we can imagine keeping tabs on the values of cars we buy over time. If all cars came from the same uniform distribution that we have proposed, then we would see the sample average value approaching the above long-run mean of \$3,000, as the number of cars we buy gets large[2] (see figure 2.6).

If you can grasp the process undertaken to produce figures 2.5 and 2.6 respectively, then you already understand the basis behind modern computational Bayesian statistics! If you need a bit more explanation of the theory of Bayesian computational, then fear not, we devote an entire Part

[2]Technically, tends to infinity.

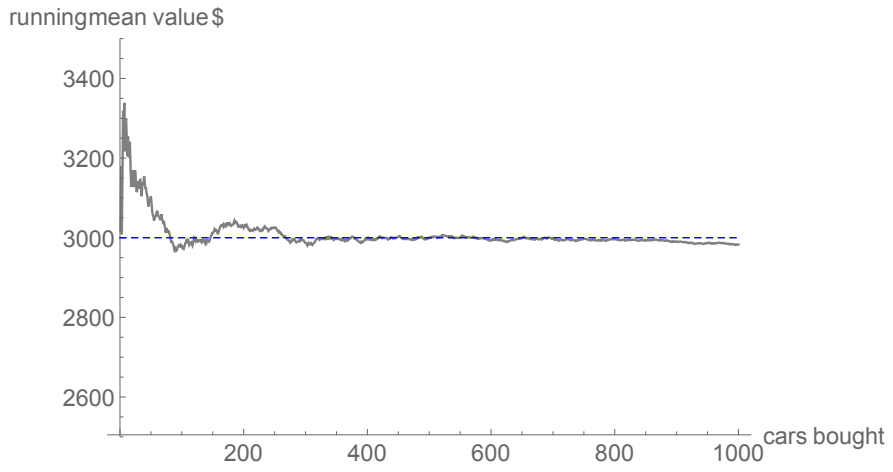

Figure 2.6: Career second-hand car sales. We can see the approach of the sample mean towards the long-run mean of $3,500.

of the book for this purpose (see Part III).

Whilst the *mean* of a distribution is a measure of central tendency for a particular distribution, we do not yet have a way of summarising the width of the range of the values of the random variable which are most likely. This motivates the introduction of the concept of a *variance* of a distribution:

$$var(X) = \mathbb{E}\left([X - \mathbb{E}(X)]^2\right) \tag{2.11}$$

To apply this to discrete and continuous distributions respectively, we straightforwardly replace the α on the right-hand side of (2.7) and (2.8) respectively, by $(\alpha - \mathbb{E}(X))^2$ in each case[3]:

$$var(X) = \sum_{All\ \alpha} (\alpha - \mathbb{E}(X))^2 Pr(X = \alpha) \tag{2.13}$$

[3]This is a specific example of the general rule, that to calculate the mean value of some function $f(X)$, where X is governed by a particular distribution, we do:

$$\mathbb{E}(f(X)) = \sum_{All\ \alpha} f(\alpha) Pr(X = \alpha) \tag{2.12}$$

for a discrete distribution, and analogously for the continuous case, but using an integral opposed to a sum.

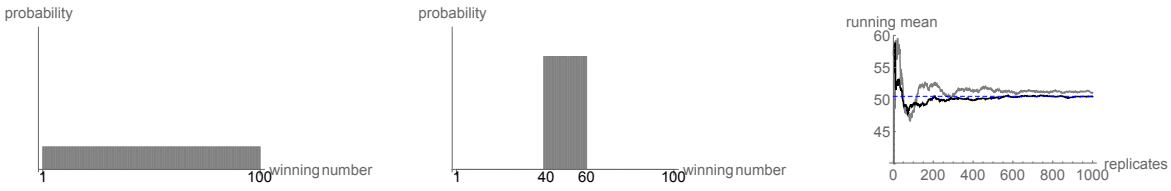

Figure 2.7: Comparing the variance of two lotteries, left: a lottery where all values between 0 and 100 are equally likely. Middle: a lottery where only values between 40 and 60 have a positive probability. Right: comparing the variability of these distributions about their common mean.

$$var(X) = \int_{All\ \alpha} (\alpha - \mathbb{E}(X))^2 p(\alpha) \mathrm{d}\alpha \tag{2.14}$$

If the equations are starting to overwhelm, then fret not, we really only wanted to include them for completeness. What is more important is their significance. Essentially, a *variance* measures the width of the distribution of values obtained around its mean. A wider variance therefore signifies a greater variety of values away from the mean. In figure **??** we compare the variability of the fair lottery, with one heavily biased to take on the values between 40 and 60. We see that the variability of the running mean for the biased lottery is smaller than that of the fair one, particularly as the number of rolls increases. This is due to the fact that the fair lottery has a variance of:

$$\begin{aligned} var(X) &= \sum_{\alpha=1}^{100} (\alpha - 50\tfrac{1}{2})^2 \times Pr(X = \alpha) \\ &= 833\tfrac{1}{4} \end{aligned} \tag{2.15}$$

whereas similar calculations for the loaded die distribution shown in the middle of figure 2.7, yield a value of approximately 265. To be concrete, the variance of a distribution is an indicator of the long-run average square distance of values away from the mean.

2.7.4 Generalising probability distributions to two dimensions

Life is often more complex than the examples of section 2.7. Often we are tasked with formulating opinions on a range of different outcomes; each of which may influence or shed light on the other results. We begin by considering the outcome of two measurements, in order to introduce the reader to the mechanics of probability. The great thing is that these rules do not become any more complex when we generalise to higher dimensional problems, meaning that if the reader is comfortable with the following examples, then they should be able to handle the vast majority of probability distribution operations encountered. In Bayesian statistics, being comfortable manipulating probability distributions is essential, since the output of the Bayesian formula - the posterior probability distribution - is used to derive all post-experiment quantities of interest. As such, it is important to devote some time to introduce two examples which we will use to describe and explain the manipulations of 2-dimensional probability distributions.

Horses for courses: a 2-dimensional discrete probability example

Imagine that you are a horse racing aficionado, and are interested in quantifying the uncertainty regarding the outcome of two (fictitious) races for two thoroughbreds in a particular stable. From their historical performance you notice that both horses tend to react in the same way to the racing conditions. When horse A wins, it is more likely that, later in the day, horse B will win, and vice versa. Similarly regarding the losses; when horse A finds the going tough, so does horse B. Wanting to flex your statistical muscle, you choose to represent this information by the two-dimensional probability distribution shown in table 2.1

		horse A	
		0	**1**
horse B	**0**	0.3	0.1
	1	0.1	0.5

Table 2.1: A probability distribution regarding the performance of two horses, A and B, in two separate races. $\{0, 1\}$ refers to each horse losing or winning in their respective races.

How can we check whether this distribution satisfies the requirements for

a valid probability distribution? We simply apply the rules described in section 2.7.1. Firstly, all the values of the distribution are real and non-negative; satisfying our first requirement. For the second rule rather than summing over the values of one random variable, we now have to sum over the outcome of two:

$$\sum_{X_A=0}^{1}\sum_{X_B=0}^{1} Pr(X_A, X_B) = 0.3 + 0.1 + 0.1 + 0.5 = 1 \tag{2.16}$$

In (2.16), X_A and X_B are random variables[4] which refer to the outcome of the races for horse A and horse B respectively. Notice that since we are now considering a situation with the outcome of two random variables, we are now required to index the probability, $Pr(X_A, X_B)$, by both. Due to the probability now being a function of two variables, we say that the probability distribution is 2-dimensional.

How can we interpret the probability distribution shown in table 2.1? The probability that both horses lose (and hence both their random variables take on the value of 0), is simply read off from the top-left entry in the table, meaning $Pr(X_A = 0, X_B = 0) = 0.3$. We ascribe a smaller likelihood to heterogeneous outcomes, $Pr(X_A = 0, X_B = 1) = 0.1$ or $Pr(X_A = 1, X_B = 0) = 0.1$, since we believe that the horses are more likely to react similarly to the racing conditions. We believe that the most likely outcome is that both horses win, since historically the horses have done well on this particular racing course, and hence ascribe the highest probability to this result, with $Pr(X_A, X_B) = 0.5$.

Foot length and intelligence: a 2-dimensional continuous probability example

We suppose that we have a sample of individuals, and we measure their foot size, as well as how well they score on an IQ test. Both of these variables can reasonably be assumed to be continuous, meaning that we are now required to represent our strength of belief, by specifying a probability distribution across a continuum of values (see figure 2.8).

[4] A function which associates a unique numerical value with each outcome of an experiment. In this case the function gives value 0 if the result is a tails, and 1 if it is heads.

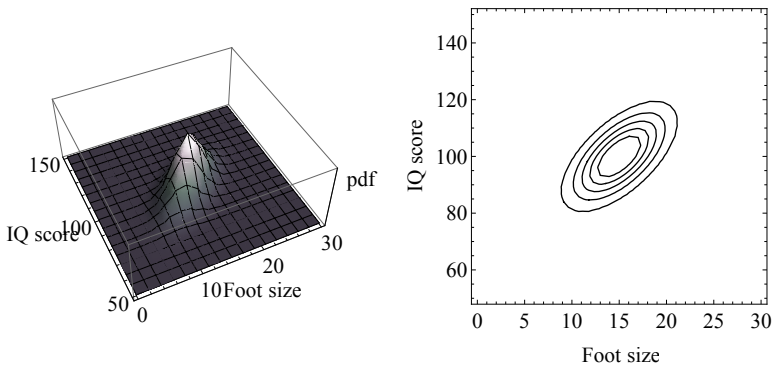

Figure 2.8: A probability distribution describing the foot size and IQ for an individual within our sample. Left) Represented as a 3-dimensional plot, and Right) Contour lines specify isolines of probability.

We could verify that the distribution shown in figure 2.8 is in fact valid, by showing that the volume underneath the left hand plot is 1, via integration. However, since we don't want to overcomplicate things now, you will have to take our word for it.

Notice that we have chosen to allow there to be a degree of correlation between foot size and IQ. Why might we choose to do this[5]?

2.7.5 Marginal distributions

We may be interested in simplifying the preceding analysis, by stating the distribution of one variable, completely *unconditional* of the other. In our horses example, we might be interested in say, only the result of the first horse race, which involves horse A. Alternatively, we might want to remove the dependence on foot size, in our IQ example, and what remains would then be an *unconditional* probability distribution for IQ.

In order to do this, we essentially need to *average* out the dependence of the other variable. In our horses example, if we are only interested in the result of horse A, we can sum down the column values for horse B, obtaining the *marginal* distribution of horse A, shown at the bottom of table 2.2.

[5]Our sample of individuals here is a sample of children of various ages. Age is correlated with shoe size and intelligence.

		horse A		
		0	**1**	$Pr(X_B)$
horse B	**0**	0.3	0.1	**0.4**
	1	0.1	0.5	**0.6**
	$Pr(X_A)$	**0.4**	**0.6**	

Table 2.2: The marginal distribution of horses A and B, achieved by summing the values in each column or row respectively.

Hence, we have that the *marginal* probability of horse A winning is 0.6. This value is composed out of the two possible ways in which this *single* event can occur:

$$Pr(X_A = 1) = Pr(X_A = 1, X_B = 0) + Pr(X_A = 1, X_B = 1) \tag{2.17}$$

In (2.17), we see that A can win with B losing, or alternatively both horses can win.

Thus, in order to calculate the probability of a single event, we simply need to sum across all possible occurrences of it, allowing the other variable to take on its possible values. Mathematically, we can summarise this rule by the following for the case of two discrete random variables:

$$Pr(A = \alpha) = \sum_{\beta} Pr(A = \alpha, B = \beta) \tag{2.18}$$

In (2.18), α and β refer to the specific values taken on by the random variables A and B.

We can use (2.18) for the horses example to calculate the probability that horse B loses:

$$\begin{aligned} Pr(X_B = 0) &= \sum_{\alpha=0}^{1} Pr(X_B = 0, X_A = \alpha) \\ &= Pr(X_B = 0, X_A = 0) + Pr(X_B = 0, X_A = 1) \\ &= 0.3 + 0.1 = 0.4 \end{aligned} \tag{2.19}$$

For continuous random variables we need the continuous analogue of a

sum, an *integral*, in order to calculate the marginal distribution. Intuitively, this is because the other variable is now able to take on an continuum of values:

$$p_A(\alpha) = \int_{All\ \beta} p_{AB}(\alpha, \beta) \mathrm{d}\beta \tag{2.20}$$

In (2.20), $p_{AB}(\alpha, \beta)$ corresponds to the joint probability distribution of random variables A and B evaluated at $(A = \alpha, B = \beta)$. Similarly, $p_A(\alpha)$ refers to the marginal distribution of random variable A, evaluated at $A = \alpha$. Although it is somewhat of an abuse of notation, for simplicity, from now on we will now write $p_{AB}(\alpha, \beta)$ as $p(A, B)$, and $p_A(\alpha)$ as $p(A)$.

In the foot size/IQ example, we may not be interested in foot size; wanting only the distribution of IQ in our sample. We can obtain this by simply integrating out the dependence on foot size:

$$p(IQ) = \int_0^{30} p(IQ, FS) \mathrm{d}FS \tag{2.21}$$

The result of carrying out the step in (2.21) is that we are left with the distribution shown on the right of figure 2.9. We have rotated this graph to emphasise that it is the result of essentially summing[6] across the joint density at each particular value of IQ.

Another way to think about marginal densities, is imagine that you are walking along the landscape of the joint density. The total distance walked - horizontally and vertically - from $FS = 0$ to $FS = 30$ along a line of constant IQ, gives the height of the marginal density for IQ at that point. If the path is relatively flat, indicating a low value of joint density, then the corresponding marginal density is low. However, if the path encompasses a large hill, indicating a high value of joint density, then the marginal density will be relatively high.

Add a 3D version of the figure with the contours traced out on the landscape, and leading to the height of the marginals, perhaps with stick figures walking along lines of iso-IQ.

[6] We really mean integrating, but it is more intuitive to think about this in terms of discrete summing.

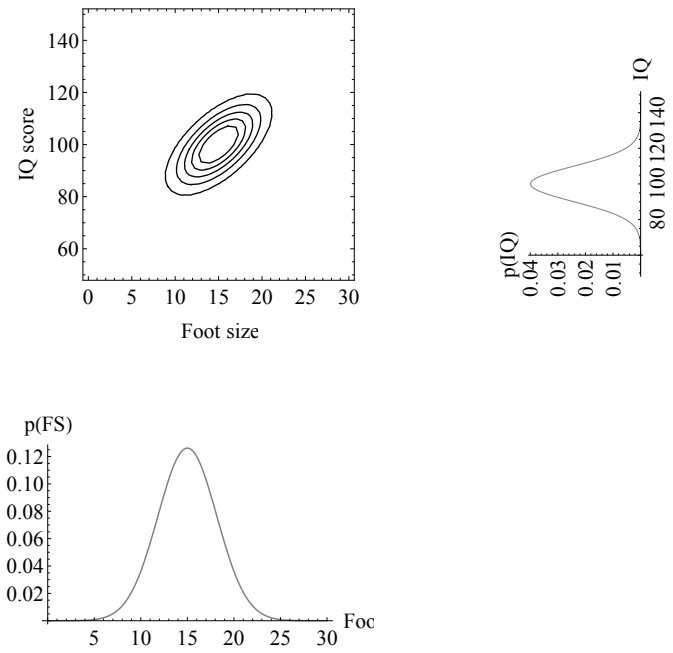

Figure 2.9: Top-left: the joint density of foot size and intelligence. Right: the marginal density of IQ. Bottom: the marginal density of foot size. **I want to add a line at a particular value of IQ, and at a particular value of FS, to illustrate the horizontal and vertical summing.**

Venn diagrams

An alternative way of thinking about marginal distributions is provided by the Venn diagram shown in figure 2.10. In a Venn diagram, the area of a particular event indicates its probability, and the rectangular area represents all the events that can possibly happen, and so has an area of 1. We have chosen to specify the events of horse A winning, and B winning A as sub-areas the diagram, which overlap indicating a region of joint probability, $p(X_A = 1, X_B = 1)$. In this set-up it is straightforward to calculate the marginal probability of horse A winning, or horse B; we find the area of the elliptic shapes A or B respectively. Considering horse A, when we calculate the area of the entire ellipse, we are implicitly carrying out the sum of the form indicated in (2.18):

$$p(A) = p(A, B) + p(A, not\ B) \tag{2.22}$$

and are working out the probability that A wins unconditionally[7].

In (2.22), the terms on the right hand side correspond to the overlap region, and the remaining part of A (where B does not occur) respectively.

2.7.6 Conditional distributions

We sometimes only receive partial information by observing part of the system in which we are interested. In horses example, we might only see the result of one horse race, and on this basis update our probabilities of the other horse winning. Alternatively, in the foot size - IQ example described before, we might measure an individual's shoe size, and then want to obtain the updated probability distribution for IQ scores.

In probability, when we observe one variable, and reformulate the probability distribution for the other variable, we say that we are deriving the *conditional* distribution of the latter. *Conditional* refers to the fact that we are deriving the probability distribution of one variable, *conditional* on the value of the other(s).

In each case, we have reduced some of the uncertainty in the system, by observing one of its characteristics. Hence in the two-dimensional exam-

[7] Irrespective of what happens to B.

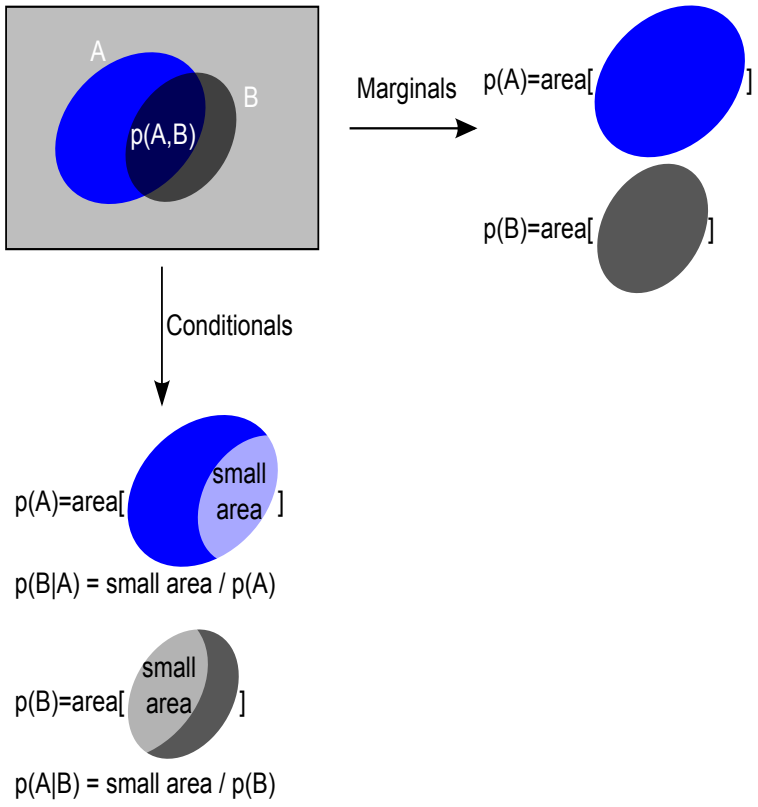

Figure 2.10: A Venn diagram showing one way of interpreting marginal and conditional distributions for the horse racing example.

ples described above the conditional distribution is only one-dimensional, because we are only now uncertain about one variable.

Luckily, there is a simple rule that we can use to obtain the probability of one variable, conditional on the value of the other:

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{2.23}$$

In (2.23), $p(A|B)$ refers to the probability of A occurring, given that B has occurred. In the right hand side of (2.23), $p(B)$ is the *marginal* distribution of B occurring, and $p(A,B)$ is the joint probability of A and B occurring.

We can use (2.23) for the horses example to calculate the probability that *given* that horse A wins, what is the probability of horse B also winning?

$$\begin{aligned} Pr(X_B = 1|X_A = 1) &= \frac{Pr(X_A = 1, X_B = 1)}{Pr(X_A = 1)} \\ &= \frac{Pr(X_A = 1, X_B = 1)}{Pr(X_A = 1, X_B = 0) + Pr(X_A = 1, X_B = 1)} \\ &= \frac{0.5}{0.1 + 0.5} \\ &= \frac{5}{6} \end{aligned} \tag{2.24}$$

In (2.24), we have used the rule we discussed earlier for calculating marginal probabilities, shown in (2.18), to calculate the denominator, $Pr(X_A = 1)$ (allowing us to move from line 1 to 2).

Another way to see the workings of this calculation is shown in table 2.3. When we see that horse A wins, we essentially reduce our solution space to only the central column (highlighted in blue). Therefore we need to renormalise the solution space such that it has a probability of 1, by dividing each of its entries through by its original total of probabilities, 0.6; yielding the conditional probabilities shown in the right hand column of table 2.3.

The Venn diagram in figure 2.10 shows another way of interpreting conditional distributions. If we are told that horse B wins, then our event space collapses to only the area specified by B. The conditional probability, $p(X_A = 1|X_B = 1)$ is then simply given by the ratio of the area of overlap

		horse A		
		0	**1**	$Pr(X_B\|X_A = 1)$
horse B	**0**	0.3	0.1	**=0.1/0.6 = 1/6**
	1	0.1	0.5	**=0.5/0.6 = 5/6**
	$Pr(X_A)$	**0.4**	**0.6**	

Table 2.3: The highlighted region indicates the new solution space, since we know that horse A has won.

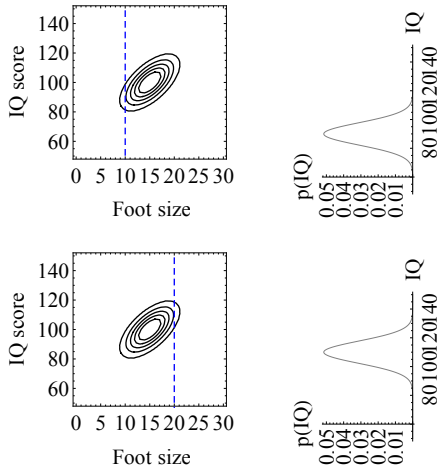

Figure 2.11: The dashed blue lines indicate the new event space in each case. The height walked following these lines is related to the magnitude of the conditional distributions shown on the right.

between A and B to the total area of B. This makes intuitive sense, since this is the only way that horse A can win, given that B has already won.

We can also use (2.23) to allow us to calculate the conditional distribution of IQ for individuals after we have measured their shoe size. The only difference with the discrete example is that we now have to use an integral to work out the marginal probability for foot size; the denominator of (2.23). Figure 2.11 shows the conditional distributions traced out when we measure an individual's foot size to be 10cm and 20cm respectively. The blue dashed lines show the new event space, since we have lost our uncertainty over foot size in each of the cases. Therefore the heights traversed on the walk along these lines of constant foot size indicate the relative likelihood of different

values of IQ.

Add a stick man diagram.

2.8 Higher dimensional probability densities: no harder than 2-D, just looks it!

Now that we are equipped with the tools to calculate marginal and conditional distributions in two dimensions, we can use these to work with probability distributions that depend on arbitrary many variables. Although formulae appear more complex, this is really just a result of having to keep track of each individual variable.

Imagine that we are told that there is another horse C, which comes from the same stable as horses A and B, which will compete in a third race, on the same day as the other two. This probability density could be represented by a 3-dimensional array, or alternatively as two separate tables of the same form as table 2.1, one for each outcome for horse C's race. We could write this probability density as before, but with a third random variable $p(X_A, X_B, X_C)$. If we were only interested in the result of horses A and B, we can define a vector $\boldsymbol{Z} = (X_A, X_B)$, meaning that we could write our density as $p(\boldsymbol{Z}, X_C)$. We can then just apply the same rule as before (in (2.18)) to get the marginal density, because our notation makes it look '2-dimensional':

$$\begin{aligned} p(X_A, X_B) = p(\boldsymbol{Z}) &= \sum_{X_C=0}^{1} p(\boldsymbol{Z}, X_C) \\ &= p(\boldsymbol{Z}, 0) + (\boldsymbol{Z}, 1) \end{aligned} \tag{2.25}$$

If we wanted to work out the new probabilities of the two horses winning *conditional* on the fact that horse C has won, we can simply use our 'two-dimensional' notation, and equation (2.23):

$$\begin{aligned} p(X_A, X_B | X_C = 1) &= \frac{p(X_A, X_B, X_C = 1)}{p(X_C = 1)} \\ &= \frac{p(X_A, X_B, X_C = 1)}{p(1, 1, X_C = 1) + p(1, 0, X_C = 1) + p(0, 1, X_C = 1) + p(0, 0, X_C = 1)} \end{aligned} \tag{2.26}$$

Notice now however, that the denominator $p(X_C = 1)$; representing the marginal probability of horse C winning is actually obtained by a sum over all the four possible ways this can occur: A and B winning, A winning and B not, B winning and A not, and finally both A and B losing[8].

For another example, suppose that we have a continuous posterior density that is defined in terms of three person-specific variables, $p(IQ, FS, W)$; where W represents the weight of an individual. If we wish to determine the posterior solely as a function of IQ and FS, then we start by defining the parameter vector $\boldsymbol{\theta} = (IQ, FS)$, meaning we are left with $p(\boldsymbol{\theta}, W)$. Note that all we have done is defined a new composite variable, $\boldsymbol{\theta}$. Now our density is of the '2-dimensional' form shown in (2.20), and we can apply this relation:

$$\begin{aligned} p(IQ, FS) &= p(\boldsymbol{\theta}) \\ &= \int_{All\ W} p(\boldsymbol{\theta}, W) \mathrm{d}W \end{aligned} \tag{2.27}$$

If we wish to find the marginal distribution for *IQ only*, then all we do is integrate the resultant distribution in (2.27) with respect to FS.

We can use exactly the same trick to calculate the *conditional* density of IQ and FS, conditional on observing weight:

$$\begin{aligned} p(IQ, FS|W) &= p(\boldsymbol{\theta}|W) \\ &= \frac{p(\boldsymbol{\theta}, W)}{p(W)} \end{aligned} \tag{2.28}$$

In (2.28), we have used (2.23) to arrive at the second line.

Finally, note that combining all parameters of interest into a single parameter vector $\boldsymbol{\theta}$ allows us to calculate *marginal* and *conditional* distributions for probability distributions that depend on an arbitrary number of parameters.

[8] I have shortened the notation to allow all the outcomes to be shown on a single line, so $p(X_A = 1, X_B = 1, X_C) = p(1, 1, X_C)$, for example.

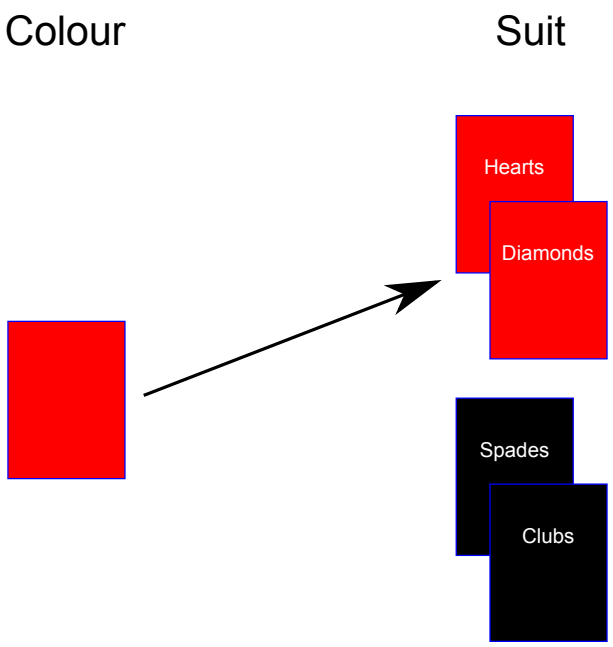

Figure 2.12: Knowledge of the colour of a card provides information about the suit of the card. The colour and suit of a card are *dependent*.

2.9 Independence

If we think that there is a relationship between two random variables, then we say that they are *dependent*. This does not necessarily mean *causal* dependence, as it is sometimes supposed, in that the behaviour of variable A affects the outcome of variable B. What it really means is that the value taken on by A is informative for predicting B.

An example of dependence might be the *colour* and *suit* of a playing card. If we are told that the colour of a playing card is *red*, this means that our other variable *suit* is constrained to be either *hearts* or *diamonds*. In this case, knowing with certainty the value of the first variable, *colour*, helps us to narrow down the list of outcomes of the second variable, *suit* (see figure 2.12).

Another example of dependent variables, is the weather outside and the suntan of a particular, vein, individual[9]. If it is sunny, then we assume it is

[9]Discounting sunbeds.

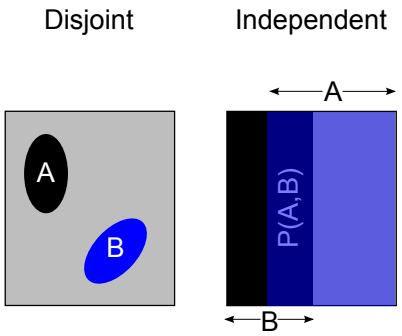

Figure 2.13: Venn diagram depictions of left: disjoint, and right: independent, events A and B.

more likely that an individual is tanned. Whereas, if the weather is cloudy, it is less so.

If two variables, A and B are *disjoint*, then if one occurs, then the other cannot. In this case, it is often mistakenly believed that the variables are *independent*, although this is very much not the case (see the left hand panel of figure 2.13). In this case, knowledge of variable A, provides significant information about variable B. If A occurs, then we know for *certain* that B cannot!

By contrast, if two events are *independent*, then knowledge of B provides no additional information on A. Mathematically, this means that the conditional probability of A is equal to the marginal:

$$p(A|B) = p(A) \tag{2.29}$$

Using our conditional probability rule given in (2.23), we can then use this to rewrite the above as:

$$\frac{p(A,B)}{p(B)} = p(A) \tag{2.30}$$

In words, the ratio of the area of overlap between A and B to the area of B, is the same as the overall probability of A (see the right hand panel of figure 2.13). This makes intuitive sense, since uncovering that B has occurred (being in B) should result in no change to the probability of A occurring (now $p(A|B)$).

Another way of stating independence that is commonly used, is obtained by multiplying (2.29) by its denominator:

$$p(A,B) = p(A) \times p(B) \tag{2.31}$$

To make this idea more concrete, we can think again of our horses example. Imagine that now are are considering two horses C and D, that come from separate stables, and race on different days. Using historical race results we have come up with the probability distribution shown in table 2.4.

We can use this table to test whether or not the results for the two horses are independent using (2.31). We should be able to get the joint probabilities of both C and D winning from multiplying together the individual marginal densities:

$$\begin{aligned} p(X_C = 1, X_D = 1) &= 0.3 \\ &= p(X_C = 1) \times p(X_D = 1) \quad = 0.6 \times 0.5 = 0.3 \end{aligned} \tag{2.32}$$

which we see is true. We could similarly also validate this by checking the other three joint outcomes in the table, and find that this is the case.

2.10 Central Limit Theorems

Life is rarely simple, and in stating statistical models of phenomena we are attempting to at best, *approximate* what is actually happening. In this environment, anything that brings a degree of certainty and coherence to our models is most welcome.

		horse C		
		0	**1**	$Pr(X_D)$
horse D	**0**	0.2	0.3	**0.5**
	1	0.2	0.3	**0.5**
	$Pr(X_C)$	**0.4**	**0.6**	

Table 2.4: The probability distribution for horses C and D. The marginal distribution of horses C and D are achieved by summing the values in each column or row respectively.

Imagine that we are tasked with coming up with a model of probability for the *mean* IQ test score in a particular school. Furthermore, as a simplification, we imagine that IQ is constrained to lie in the range $[0, 300]$, and we believe that a reasonable probability distribution for an individual's test score is uniform on this range[10] (see figure 2.14). We also suppose that individuals' scores are independent of one another.

Before we consider a sample size of N, we imagine that we only have a sample of two individuals, and we are asked to describe the distribution for the mean of their scores. If we use our assumption of uniformity, we might then ask, whether any mean scores are more likely than others, or are *all* values equally likely, as per the individual case? We start by considering the extremes: there is only one way to achieve a mean test score of 300; both individuals would have to had scored 300. Similarly, to obtain a mean of 0, both individuals must score 0. However, consider obtaining a mean of 150. This could have been obtained with a number[11] of combinations of scores, for example $(score_A, score_B) =: (150, 150)\,,\ (100, 200)\,,\ (125, 175)$. Intuitively, there are many more ways of obtaining moderate values for the sample mean, that there are for the extremes.

This tendency towards the centre increases the more values we sum over, since in order to get extreme values we would require *more* individual scores to be simultaneously extreme; which is less likely. We can see this increasing central-tendency in figure 2.14, as the number of points being summed over increases.

However, we also see another tendency of the probability densities: *an*

[10] Albeit this isn't a particularly good model, but suspend your disbelief for this thought experiment.

[11] Technically an infinity.

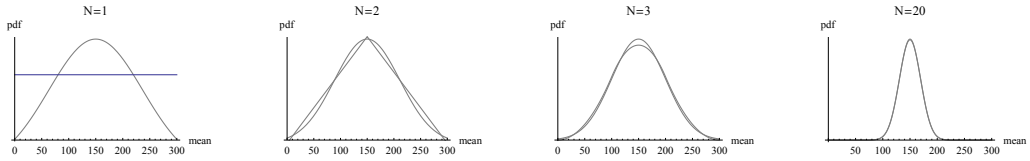

Figure 2.14: The convergence to a normal distribution for the mean of a sum of uniform distributions for IQ. The pdf for the average is shown in blue, with a normal distribution of the same mean and variance indicated in grey.

increasingly good fit of the normal distribution to these sample averages. This approximation, it turns out, becomes *exact* in the limit that we have an infinite sample, and is known as the *Central Limit Theorem*. Although for our purposes, it will often be practically-exact if the sample size is sufficiently large.

There are in fact a number of central limit theorems, of which we have only exposed you to the one for the average of independent, identically-distributed random variables[12]. However, what is important for you to note is that, whenever we have a number of factors, which additively result in an outcome, then an assumption of normality may be reasonable.

For example, there is an argument which would suggest that an individual's intelligence is the result of a number of factors including: un-bringing, genetics, life-experience, and health amongst others. Hence, we might tentatively propose that an individual's test score picked at random from the population would actually be normally-distributed! This is why I earlier discussed that the assumption that individuals' test scores were uniform might not be reasonable.

2.11 The Bayesian formula

We first of all rewrite the conditional probability formula (2.23), regarding the probability of event A occurring, *given* that event B has occurred:

[12]Known as the Lindeberg-Levy CLT. **Note: need an accent on the e.**

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{2.33}$$

However, we could also swap A and B around leading to the following for the probability of B *given* that A has already occurred:

$$p(B|A) = \frac{p(B,A)}{p(A)} \tag{2.34}$$

We however reason from the Venn diagram in figure 2.10, that the overlap region of $p(A,B)$ really translated means, the probability of A *and* B occurring. This means that this is exactly the same as the reverse; the probability of B *and* A coinciding, $p(B,A)$. We can therefore rearrange (2.34) for this joint probability:

$$p(A,B) = p(B|A) \times p(A) \tag{2.35}$$

We can use (2.35) to break down the probability of both A and B occurring into two steps. Firstly, for this to happen we require that A *must* happen, with its corresponding probability $p(A)$. Then for both to occur, we straightforwardly require the probability of B occurring, *given* that A has already occurred, which is given by $p(B|A)$. This reasoning provides a little intuition as to the workings of the conditional probability law that we wrote down in (2.23).

We can finally substitute (2.35) into the numerator of the fraction in (2.33), to yield the famous Bayesian formula!

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \tag{2.36}$$

The Bayesian formula importantly tells us how to correctly convert from $p(B|A)$ to its inverse $p(A|B)$, which is central to Bayesian statistics.

2.11.1 The intuition behind the formula

If we multiply both sides of (2.36) by $p(B)$, we arrive at the following alternative statement of Bayes' rule:

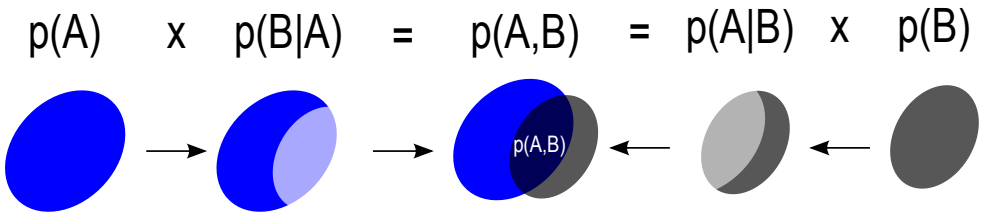

Figure 2.15: The two ways of arriving at the joint probability $p(A, B)$; providing some intuition behind Bayes' rule.

$$p(A|B) \times p(B) = p(B|A) \times p(A) \quad [= p(A, B)] \tag{2.37}$$

In (2.37), we have added the final part in square parentheses due to the reasoning of section 2.11, that both sides are equivalent to the joint probability of A and B.

What the relation (2.37) tells us however, is that there are two ways of arriving at this joint probability (see figure 2.15). The first way is given by the left hand side, and is due to B occurring, with probability $p(B)$, followed by A *given* that B has occurred, with probability $p(A|B)$. An exactly equivalent way route to both A and B occurring is given by the right hand side of the leftmost equals sign. Here we require that A occurs first, with probability $p(A)$, followed by B *given* that A has occurred, with probability $p(B|A)$.

Let's now make this discussion more concrete by means of an example. Let's imagine that we are an oncology doctor, working in breast cancer diagnosis. We suppose that out of all women aged forty who participate in screenings, about 1% of them will have breast cancer. We suppose that the screening process is relatively robust, and it is known that for women who have breast cancer, the tests will indicate a positive result 80% of the time. However, there is also the risk of false-positives, with 10% of women without breast cancer also testing positive[13].

We now suppose that we are in the position where a woman has tested positive. What we would like to do, is work out the probability that she has breast cancer.

In the language of probability we would like to work out the conditional

[13]I am not necessarily indicating clinically-up-to-date values, more I am using these example values to indicate the importance of reducing false-positives of any medical-test.

probability: $p(cancer| + ve)$. In words, the probability that she has breast cancer, *given* that she has tested positive. However, summarising the information we currently have in probability language: $p(cancer) = 0.01$, $p(+ve|cancer) = 0.8$, and finally $p(+ve|no\ cancer) = 0.1$. How can we proceed? Bayes' formula to the rescue:

$$\begin{aligned} p(cancer| + ve) &= \frac{p(+ve|cancer) \times p(cancer)}{p(+ve)} \\ &= \frac{p(+ve|cancer) \times p(cancer)}{p(+ve, cancer) + p(+ve, no\ cancer)} \\ &= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99} \approx 0.08 \end{aligned} \tag{2.38}$$

In (2.38), we got from the first line to the second by using (2.18) to work out the marginal probability of an individual testing positive. The fact that the probability is so small is due to the risk of false positives (see video XXX for a fully-intuitive explanation of this), which produce many more positive diagnoses than the real-positives (because the risk of cancer is much lower than the probability of not having cancer).

2.12 The Bayesian inference process from the Bayesian formula

In Bayesian statistics we aim to use probability distributions to describe all components of our system. Our starting point is Bayes' rule:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \tag{2.39}$$

In statistics we are typically looking to estimate a number of parameters, which we will from now on call θ, which represent a component of a statistical model which we build to represent a particular phenomena. These parameters can be real (such as the proportion of individuals within a given population that have a disease), or mere abstractions (for example the scale parameter of a hyper-distribution).

In Bayesian statistics, we want to update our beliefs about values of a parameter *given* that we have obtained a particular sample of data. Being

Bayesians, we would like to represent these beliefs via a probability distribution, which we write as $p(\theta|data)$. However, we can use (2.39), (if we associate A with θ, and B with the *data*) to write:

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{2.40}$$

Although we have straightforwardly made two substitutions to arrive at (2.40) from (2.39), what have we gained by doing so? Also, what exactly do the terms on the right hand side actually mean? I will explain these in detail in the first part of this book, although provide some short examples next.

2.12.1 Likelihoods

Starting with the numerator on the right hand side of (2.39), we come across the term $p(data|\theta)$, which we call the *likelihood*. This tells us the probability of generating the particular sample of *data*, if the parameters in our statistical model were equal to θ. When we write down a statistical model, we can generally calculate the probability of particular outcomes, so this is easily obtained. Imagine that we have a coin that we believe to be fair. By *fair*, we typically mean that the probability of the coin falling 'heads-up' is $\theta = \frac{1}{2}$. If we flip the coin twice, we might suppose that it is reasonable to model the outcomes as independent (see section 2.9), and hence we can calculate the probabilities of the four possible outcomes, by multiplying the probabilities of the individual outcomes together:

$$\begin{aligned} p(HH) &= p(H) \times p(H) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\ p(HT) &= p(H) \times p(T) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\ p(TH) &= p(T) \times p(H) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\ p(TT) &= p(H) \times p(H) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \end{aligned} \tag{2.41}$$

Hence, we obtained a *sample* of two heads, we could write down the corresponding likelihood, $p(HH|\theta = \frac{1}{2}) = \frac{1}{4}$.

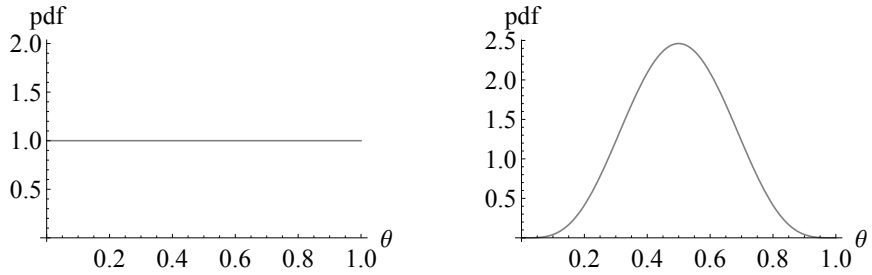

Figure 2.16: Left: All values of a the bias of a coin are equally likely. Right: It is believed that the coin is most likely fair.

Do not worry if you do not fully understand this concept, as we will be devoting an entire chapter to likelihoods in chapter 4.

2.12.2 Priors

The next term in the numerator of the right hand side, $p(\theta)$ is the most controversial[14] part of the Bayesian formula, which we call the *prior* distribution of θ. It is a probability distribution which represents our pre-data beliefs as to the likely values of the parameters in our model, θ. This appears at first to be slightly counter-intuitive, particularly if you are used to the world of classical statistics, which does not require us to state our preferences *explicitly*[15]. Continuing the coin example, we might assume that we do not know whether the coin is fair or biased beforehand, so we might think that all possible values of $\theta \in [0, 1]$ - which represents the probability of the coin falling 'heads-up' - are equally likely. We might represent these beliefs by a continuous uniform probability density on this interval (see the left-hand graph of figure 2.16). Normally however, we might think that coins are manufactured such that the weight distribution is fairly symmetrical on either face; meaning that we expect that the majority of coins are reasonably fair. These latter beliefs of unbiasedness might be fairly well represented by the right-hand graph of figure 2.16.

The concept of *priors* will be covered in detail in chapter 5.

[14] Although this controversy is unwarranted, as we explain in section 2.13.
[15] Although, we always do *implicitly*, as we explain in section 2.13.

2.12.3 The denominator

The final term on the right hand side, on the denominator is $p(data)$. This represents the probability of obtaining our particular sample of data, if we assume a particular model and prior. We will mostly postpone further discussion of this term until chapter 6, when we understand better the significance of *likelihoods* and *priors* respectively. However, note that this denominator is actually a marginal probability density (see section 2.7.5), which can be obtained from summing/integrating the joint density $p(\theta, data)$ with respect to the parameter(s) of the model.

The concept of *the denominator* will be covered in detail in chapter 6.

2.12.4 Posteriors: the goal of Bayesian inference

The posterior probability distribution $p(\theta|data)$ is often the main goal of Bayesian inference. For example, in the coin example described before, we might want to derive a probability distribution representing our post-experimental beliefs of the inherent bias, θ, of a coin, *given* that we flipped it 10 times, and it came up 'heads' 7 times. If we use (2.40), assuming the likelihood model specified in section 2.12.1 and the flat uniform prior shown in figure 2.16, then we would end up with a posterior distribution shown in figure 2.17. Notice that the peak of the distribution occurs at $\theta = 0.7$, which corresponds exactly with the percentage of 'heads' seen in the experiment[16].

The posterior distribution summaries our uncertainty in parameter values. If the distribution is more peaked, then this emphasises that there is a greater degree of certainty with a particular value for a parameter. This increased certainty over a parameter value is frequently obtained by collecting more data. In figure 2.18, we compare the posterior distribution for the previous case of 7/10 times a coin appearing 'heads' up, with a new, larger, sample where 70/100 times the same coin comes up 'heads'. In both cases the same ratio of 'heads' to 'tails' appeared, resulting in the same peak value of $\theta = 0.7$. However, in the latter case, since we have more evidence to support our claim, we end up with greater certainty over the parameter value after the experiment.

The posterior distribution is also used as a starting point for prediction of

[16]Note that if we chose a non-uniform prior, this peak would most likely shift.

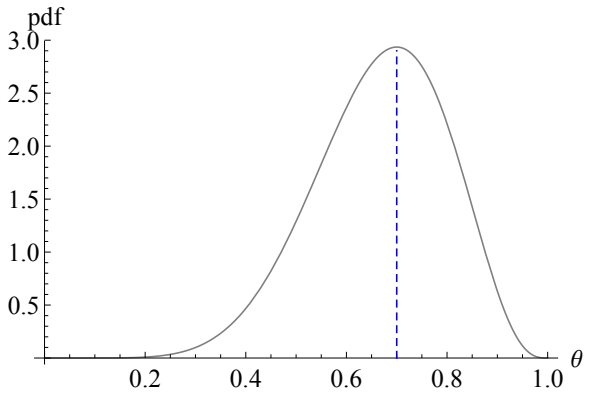

Figure 2.17: The posterior distribution for, θ, the bias of a coin when flipped, assuming a flat uniform prior and Bernoulli likelihood. We assume that 7/10 times the coin came up 'heads'.

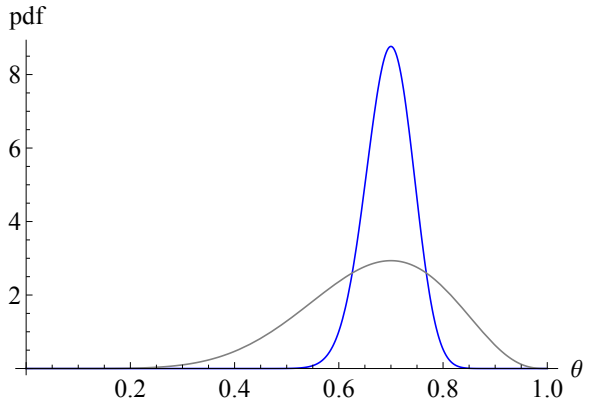

Figure 2.18: Posterior distributions for, θ, the bias of a coin when flipped, assuming a flat uniform prior and Bernoulli likelihood. The grey line assumes that 7/10 times the coin came up 'heads'. The blue line is for the case where 70/100 times the coin came up 'heads'.

future outcomes of an experiment, as well as for model testing. However, we will leave discussion of these until chapter 3.

2.13 Implicit vs Explicit subjectivity

One of the major arguments levied against Bayesian statistics is that it is by its nature *subjective*, due to its dependence on the analyst specifying their pre-experimental beliefs through *priors*. This experimenter prejudice towards certain outcomes is said to bias the results away from the types of fair, objective outcomes resultant from a classical analysis.

We argue that *all* analyses involve a degree of subjectivity, which is *implicitly* assumed. In a classical analysis, the statistician typically states a model for probability which depends on a range of assumptions, should be justified explicitly. This process of justification is indicative of the subjective nature of the assumptions on which most analyses rest. For example, the choice to use a simple *linear regression model* in many applied classical analyses assumes that the response of a dependent variable is linear in the model's parameters. This choice of model architecture is generally arbitrary, and used mostly to simplify the analysis.

In science, there is a tendency amongst scientists to use data to suit one's needs, although this practice should really be discouraged (see [**?**]). This choice as to which data points to include is subjective, and will remain independent of the type of analysis applied.

A further source of subjectivity is in the way in which models are checked and tested. In analyses, both classical and Bayesian, there is a need to exercise (subjective) judgement in suggesting a methodology which will be used in this process. We would argue that a Bayesian analysis allows greater flexibility, and suitable methodologies for these processes, since the prior- and posterior- predictive distributions are straightforwardly manipulated to suit most situations. A Bayesian methodology also allows different models to be compared in a logically-coherent manner, whilst classical analysis relies on fairly arbitrary criteria [17] to do so.

In contrast to the examples of *subjectivity* which we have mentioned above, Bayesian *priors* are *explicitly* stated. This makes this part of the analysis

[17] $\bar{R}^2$, AIC and BIC are examples

openly available to the reader, allowing it to be as thoroughly interrogated and debated, as any part of an argument. This transparent nature of Bayesian statistics has lead many to suggest that it is *honest*; whilst classical analyses hide behind a fake veil of *objectivity*, Bayesian equivalents explicitly acknowledge the subjective nature knowledge.

Furthermore, the more data that is collected, the less impact the prior exerts on posterior distributions. In any case, if slight modification to priors results in a different conclusion being reached, it is the job of the researcher to report this sensitivity. In fact, in contrast to classical analyses, a Bayesian analysis allows for a range of models/priors to be stated, which can then be used to test the sensitivity of conclusions to any subjective assumptions made.

Finally, comparing the classical and Bayesian approach to pursuit of knowledge, we find two different solutions; both of which require a subjective judgement to be made. In both cases we would like to have access $p(\theta|data)$ - the probability of the parameter/hypothesis of interest after we have obtained a given data set. In classical hypothesis testing we do not calculate this quantity directly, but use a rule of thumb: we calculate the probability that the data would have been more extreme than that which we obtained under a 'null hypothesis'. If the probability is sufficiently small, typically less than a cut-off of 5% or 1%, then we reject the null. Note that this choice of threshold probability - known as a statistical test's *size* - is completely *arbitrary*, and subjective. In Bayesian statistics, we instead use a prior to invert the likelihood from $p(data|\theta) \rightarrow p(\theta|data)$. There is no need to have a null hypothesis and an alternative, since all information is summarised neatly in the posterior. In this way we see a symmetry in the choice of classical *size* and Bayesian priors; they are both attempts to invert the likelihood to get a posterior.

2.14 What are the tangible (non-academic) benefits of Bayesian statistics?

In Bayesian textbooks much discourse is devoted to advocating the academic reasons for choosing to use a Bayesian analysis over classical approaches. However, often authors neglect to promote the more tangible, everyday benefits of the former. Here, we list the following *real* benefits of a Bayesian approach:

- **Simple and intuitive model testing and comparison**. The prior- and posterior-predictive distributions allow for in-depth testing of any particular aspect of a model, by comparing it with the same aspects from the data collected. The Bayesian approach also provides a logical framework in which to compare different models.

- **Straightforward interpretation of results.** In classical analyses, the *confidence interval* is often taken to be a measure of uncertainty for a particular parameter. As we shall see in section 3.3.4, this is not the case, and interpretation of this concept is not straightforward. By contrast Bayesian *credible intervals*, can be taken to be a measure of uncertainty in a parameter, as they are obtained directly from probability distributions.

- **Full model flexibility**. Modern Bayesian analyses use computational simulation in order to carry out analyses. Whilst this might appear excessive when compared to classical approaches, an additional benefit is the straightforward extension to almost arbitrarily complex models when using Bayesian approaches. This means that Bayesian models can be extended to encompass any complexity of data process. This is in contrast to classical approaches, where the intrinsic difficulty of analysis scales with the complexity of the model chosen.

- **The best predictions.** Leading figures both inside and outside of academia use Bayesian approaches for prediction. An example being Nate Silver's correct prediction of the 2008 US Presidential election results [**?**].

2.15 Appendix

2.15.1 The Frequentist and Bayesian murder trial

In the Bayesian trial the probability that you are guilty given being seen by the security camera on the night of the murder is:

$$\begin{aligned} p(guilt|security\ camera footage) &= \frac{p(security\ camera footage|guilt) \times p(guilt)}{p(security\ camera footage)} \\ &= \frac{\frac{30}{100} \times \frac{1}{1000}}{\frac{30}{100} \times \frac{999}{1000} + \frac{30}{100} \times \frac{1}{1000}} \\ &= \frac{1}{1000} \end{aligned} \tag{2.42}$$

In (2.42) we have implicitly assumed that the security camera is hidden, and hence the murderer does not alter his behaviour to avoid being seen; meaning that the probability of being seen by the security camera in each case is 30%.

Chapter 3

The posterior - the goal of Bayesian inference

3.1 Chapter Mission statement

At the end of this chapter the reader will understand the central importance, and use of the posterior probability distribution in Bayesian statistics.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)} \tag{3.1}$$

3.2 Chapter goals

Calculating the posterior distribution for a model's parameters is the focus of Bayesian analysis. This *probability distribution* which results from the application of Bayes' rule (see 3.1) can be used to infer the effects of given variables, to forecast, compare different models of phenomena, as well as test its own foundations! In order to do justice to the multitude of uses of the posterior distribution, it is necessary that the reader is familiar with the basics of probability distributions explained in chapter 2.

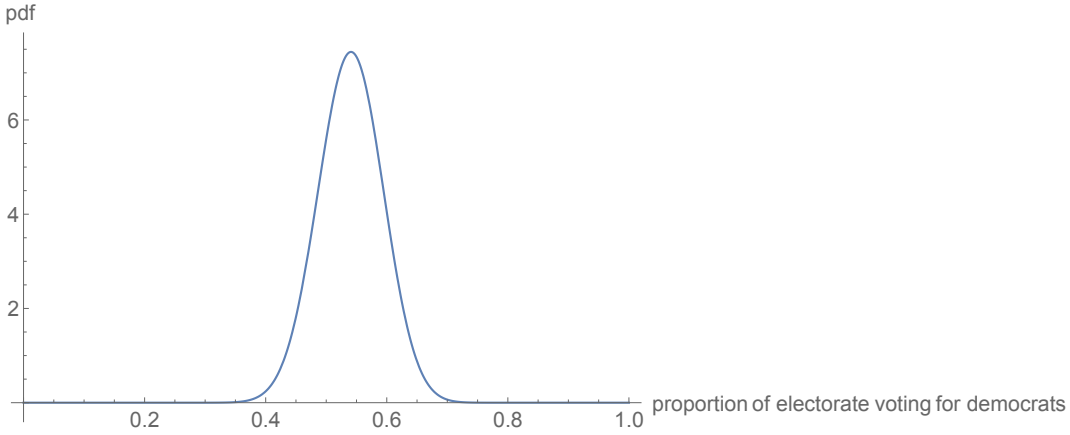

Figure 3.1: A probability distribution representing uncertainty over the proportion of the electorate that will vote for the Democrats in an upcoming election.

3.3 Expressing uncertainty through the posterior probability distribution

Unlike looking out the window, getting exam results, or playing a hand at blackjack, we frequently in inference never learn the *true*[1] state of nature. The uncertainty here is both in the future and *present*; the latter meaning we are unable to perfectly measure the state of the world today, and hence cannot hope to perfectly know the former.

A way of representing our ignorance, or uncertainty in a parameter's value is through probability distributions. For example, suppose that we wanted to know the proportion of individuals that would vote for the democrats in an upcoming election. We might, on the basis of past exit poll surveys, calculate a posterior[2] uncertainty which is represented by the probability distribution which is shown in figure 3.1.

How can we interpret the probability distribution shown in figure 3.1? And further, how can we use it to express our uncertainty to a non-mathematician?

[1] Whether a true value for a parameter actually exists we leave until section 3.3.3.

[2] Via Bayes' rule - don't worry if you don't know how to do this, that is the goal of this whole first part!

Often we describe a distribution by its summary characteristics. These are aspects of the distribution that we commonly want to know. For example, we normally want to know the *mean* value of a parameter. This is a measure of central tendency of our estimates, that is essentially a weighted mean (where the weights are provided by the values of the probability density function). If we have the mathematical description of the distribution shown in figure 3.1, we can calculate this by simply finding its mean, by taking the expectation:

$$\mathbb{E}[\theta] = \int_0^1 p(\theta)\theta \mathrm{d}\theta = 54\% \tag{3.2}$$

This provides a point estimate of the proportion of individuals - 54% - that we expect to vote for the Democrats, which may be a useful piece of information to pass on to an interested party.

A point estimate of the proportion of individuals that we expect to vote for the Democrats is not useful in itself, (and quite dangerous to pass on), without some measure of our inherent confidence/uncertainty in this particular value. One measure of uncertainty in a parameter's value is its variance:

$$var(\theta) = \int_0^1 p(\theta)(\theta - (E[\theta])^2 \mathrm{d}\theta \tag{3.3}$$

In many cases, it is easier to understand the meaning of uncertainty if it is expressed in the same units as the mean, which we do by taking the square root of the variance, yielding 5.3% for the case of figure 3.1. A larger variance indicates that we view a wider range of outcomes as being feasible. In this case, a wider variance would mean that we would be less surprised if the electorate voted in the Republican party.

In sections 3.3.5 and 3.3.5 we will introduce other summary measures of distributions, that are also often presented in research articles, and books. However, the important thing to note is that all of these are derived from the posterior distribution for our parameters.

Another example of the use of posteriors can be illustrated by a regression example. Suppose that we are investigating the effect of military partic-

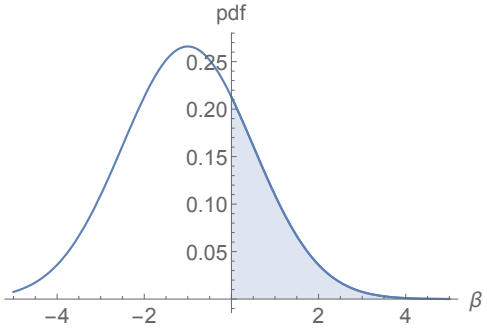

Figure 3.2: An estimated posterior probability distribution for the parameter β in (3.4). The shaded region represents the posterior probability that the parameter is positive.

ipation on lifetime earnings[3]. We suppose that the effect is, on average, negative due in part to the psychological stresses of warfare. We suppose that the effect is can be modelled as being linear, and hence our relationship of interest can be represented as:

$$LI_i = \alpha + \beta MP_i + \epsilon_i \tag{3.4}$$

where we expect that β is negative. Here ϵ_i represents the myriad of other factors that are important for determination of an individual's income.

If we formulate a Bayesian model, with a given prior and likelihood, then we might end up with a posterior density for β, that is shown in figure 3.2. We can use this density to help us to calculate the posterior probability that the given parameter is in fact non-negative, by finding the area under the curve corresponding to $\beta \geq 0$ (shown as the shaded region in figure 3.2), which we find in this case to be approximately 25%. This suggests that we are not all that confident in the fact that the parameter is negative (at 75%), and cannot reasonably go along with our hypothesis.

[3]See Angrist's fantastic 1990 article for a detailed study of this effect for Vietnam war veterans [**?**].

3.3.1 Bayesian coastguard: introducing the prior and the posterior

Imagine you are stationed in a radio control tower at the top of a steep cliff, in the midst of a stormy night. The tower received a distress call over radio from a ship - The Frequentasy - which has got engine trouble, somewhere out in the bay. It is your job to direct a search helicopter to rescue the poor sailors.

When you first receive the weak crackled radio signal, you are not made aware of their location. However, you know that the boat must be somewhere in the bay, less that 25km away from the tower, since this is the maximum possible range of the radio. Accordingly, you represent these views via the prior shown in the left hand panel of figure 1 (fairly flat prior in a circle from the tower, perhaps going down to zero gradually at 25km). The search area represented by this prior is currently far too wide for a rescue crew to reach the flagging ship in time!

In an attempt to improve the odds, you radio to the ship, and ask that they switch on their emergency transmitter. After radioing a number of times, you receive a weak signal, which you feed into the computer, resulting in a posterior probability density for the ship's location shown in the central panel of figure 1. (This panel shows a broadly peaked density along a line from the ship to the station, near a particular location 15km away).

The trouble is, the search area inferred from the aforementioned posterior is still wide. Luckily for the crew however, another nearby radio station has also picked up the signal, and they share this information with you. Finally, feeding this information into the computer, you obtain a final estimation of the ship's location, shown in the right hand panel of figure 1.

Since there is only a small area of high density, you direct the rescue helicopter to search this area, and they find the find the captain and crew in time.

3.3.2 Bayesian statistics: updating our pre-analysis uncertainty

In Bayesian statistics the posterior distribution summarises the combination of our pre-study, and post-analysis knowledge about a given situation, and is used as the starting point for any further analysis or descriptions of our results. In order to calculate it, we need to choose a probability model,

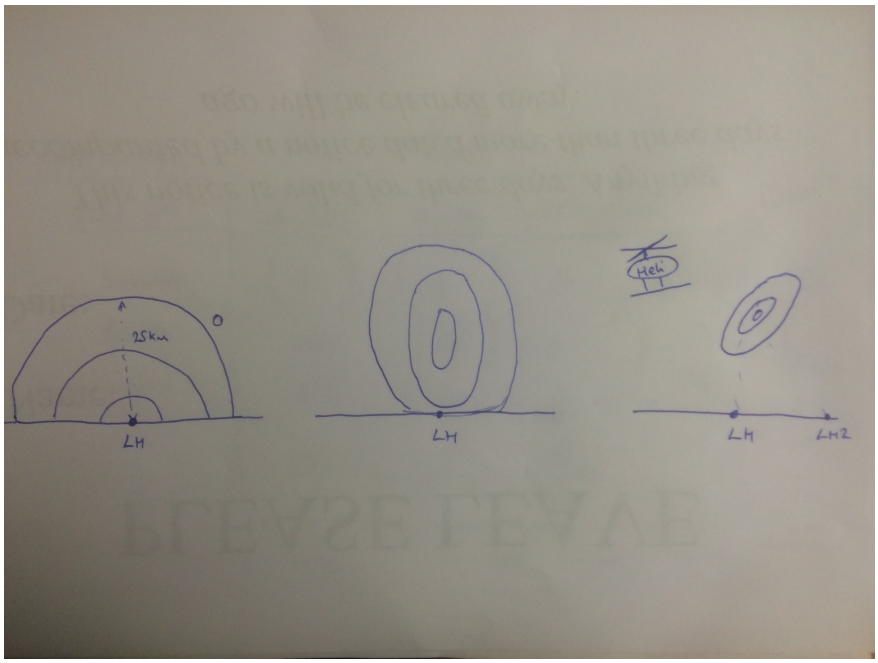

Figure 3.3: Three plots. Left hand plot is a contour plot of probability radiated symmetrically in a semi-circle away from the lighthouse, with the density declining to zero at 25km. The middle plot shows a contour plot of probability, with a higher density towards the centre of the diagram (here the densities are still relatively smooth, indicating high uncertainty). The final plot shows a definite peak in intensity around a particular point about 10km away from the coast, just right of centre. The different contours will be increasing shades of a particular colour.

which in turn defines uncertain parameters, over which we place priors. The model also provides us with a *likelihood* of the data we obtained. The priors and likelihoods are then combined in a certain way - using Bayes' rule - to yield the posterior distribution.

$$prior + data \rightarrow posterior \tag{3.5}$$

In the Bayesian lighthouse example, we started off with a fairly wide prior, since we were quite uncertain of the boat's location. We then fed the data from the ships' emergency transmitter, along with the prior, into the computer - which uses Bayes' rule - to provide an updated estimate of the ship's location. We actually went through this process twice, to emphasise the fact that Bayes' rule can be used iteratively to update knowledge about an uncertain situation.

Statistical inference is useful whenever there is uncertainty regarding a parameter of interest. Bayesians use the posterior distribution, and various summaries of it, in order to describe the degree of uncertainty regarding a parameter. Before we delve too deep though, it is useful to take a step back, and ask the somewhat philosophical question, 'Do parameters actually exist?'

3.3.3 Do parameters actually exist and have a point value?

For Bayesians, the parameters of the system are taken to vary, whereas the known part of the system - the data - is taken as given. Whereas Frequentist statisticians view the unseen part of the system - the parameters of the probability model - as being fixed, whereas the known parts of the system - the data - as varying. Whether you agree with one of these views more than the other mainly comes down to how you want to interpret the parameters of a given statistical model.

The Bayesian perspective on parameters can be viewed as having a duality of meaning. Either we view the parameters as truly *varying*, or we view our knowledge about the parameters as imperfect. The fact that we will get different estimates of parameters from different studies can be taken to reflect either of these two views. Either we view the parameters of interest as varying - taking on different values in each of the samples we pick (see the bottom panel of figure 3.4). Alternatively, we can view our uncertainty

over a parameter's value as the reason we will estimate slightly different values in different samples. This uncertainty is thought of as decreasing as we collect more data (see the middle panel of figure 3.4). Bayesians are more at ease with using parameters as a means to an end - taking them not as real immutable constants, but as tools from which to make inferences about a given situation.

The Frequentist perspective is less flexible, and assumes that these parameters are constant; alternatively representing the average of a long run - typically infinite number - of identical experiments. There are occasions when we might think that this is a reasonable assumption. For example, if our parameter represented the proportion of the electorate that voted for the Democrat party in the last election, or the probability that an individual taken at random from the UK population will have dyslexia. In both these examples, it is reasonable to assume that there is a *true*, or fixed *population* value of the parameter of question. Whilst the Frequentist view might be in some ways reasonable here, the Bayesian view easily extends here to encompass these two situations by assuming that we are uncertain about the value of these fixed parameters before we measure them, and using a probability distribution to represent this lack of perfect knowledge.

However, there are circumstances when the Frequentist view runs into trouble. When we are estimating parameters of an arbitrarily complicated distribution, we normally do not view them as actually existing. Unless you view the universe as being built from mathematical building blocks[4], then it seems incorrect to assert that a given parameter has any deeper existence than that with which we endow it. The less restrictive Bayesian perspective here seems more reasonable.

The Frequentist view of parameters as a limiting value of an average across an infinity of identically repeated experiments (see the top panel of figure 3.4) also runs into difficulty when we think about one-off events, such as the 2016 US Presidential Election result. Any parameter we wish to estimate about such an event cannot easily be existentially justified on these grounds, since elections cannot be rerun.

[4]See [**?**] for an interesting argument for this hypothesis

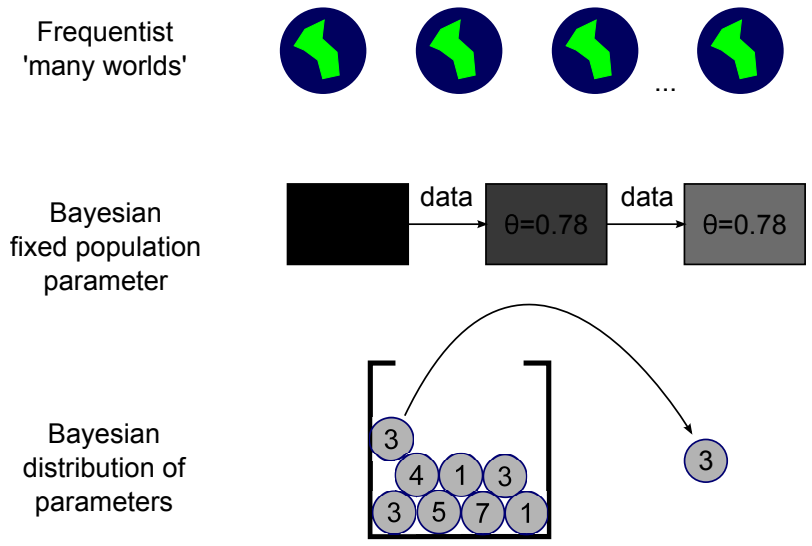

Figure 3.4: The Frequentist and Bayesian perspectives on parameters.

3.3.4 Failings of the Frequentist confidence interval

A mainstay of the Frequentist estimation procedure is the *confidence interval*. In empirical research we often see these intervals as stated for a given parameter (where for now we assume that the parameters are unknown and fixed). For example,

'From our research, we concluded that the percentage of penguins with red tails, RT**, has a 95% confidence interval of** $1\% \leq RT \leq 5\%$**.'**

This is often incorrectly taken as having an implicit meaning, 'We are 95% sure that the true percentage of penguins with red tails lies in the range of 1% to 5%.' However, what it actually captures is not uncertainty about the parameter in question, but about the interval we calculate.

In the Frequentist paradigm we imagine that we are taking repeated samples from a population of interest, and for each of the samples, we estimate a confidence interval (see figure 3.5). A 95% confidence interval means that across all of the intervals we calculate, the true value of the parameter will lie in this range 95% of the time.

However, what is important to note here is that in reality, we only draw one sample from the population, and we have no way of knowing whether the

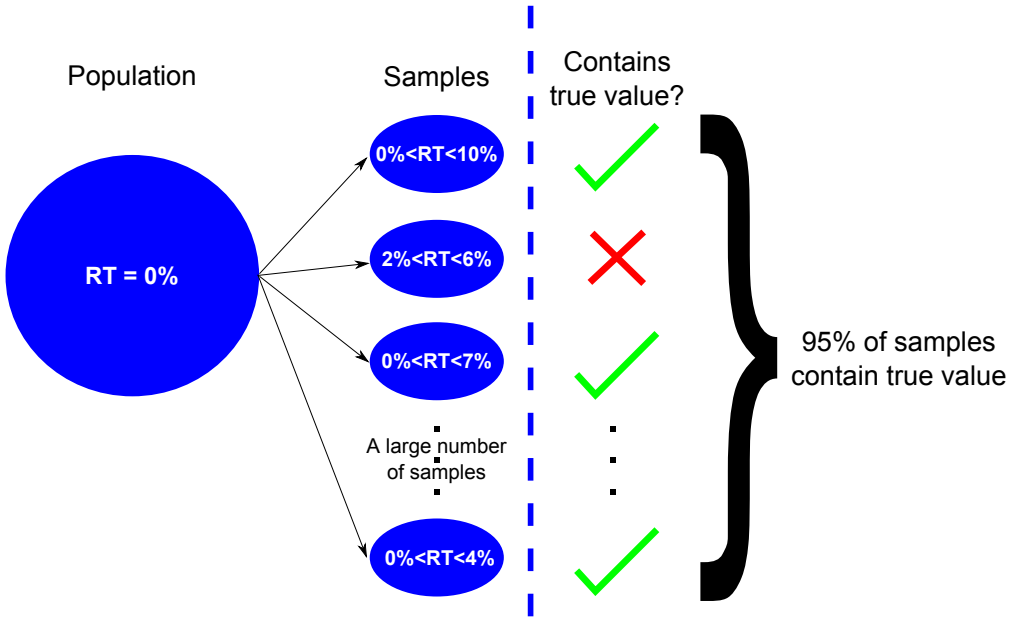

Figure 3.5: The classical confidence interval. In each sample, we can calculate a 95% confidence interval. Across repeated samples from a given population distribution, the classical confidence interval will contain the true parameter value 95% of time.

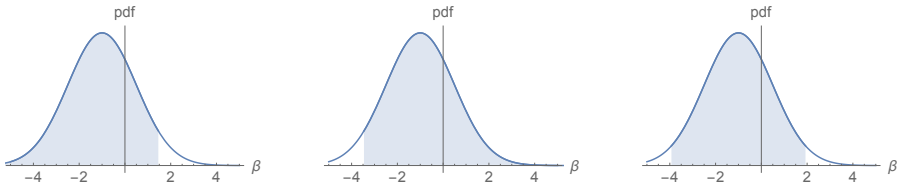

Figure 3.6: Three examples of a 95% credible interval for the regression parameter β of the example described in section 3.3.

confidence interval we calculate contains the true parameter value. This means that although 95% of the time the confidence intervals we calculate will contain the true value of the parameter, and 5% of confidence intervals will be nonsense!

In general, a confidence interval represents the uncertainty about the interval we obtained, rather than a statement of probability about the parameter of interest. The uncertainty is quantified in terms of all the samples we *could* have taken, not only the one we have in our hands.

3.3.5 Credible intervals

Credible intervals, in contrast to confidence intervals, describe our uncertainty in the location of the parameter values and thus can be interpreted as a probabilistic statement about the parameters. They are a Bayesian concept, that is calculated from the posterior density.

In particular, a 95% credible region satisfies the condition that 95% of the posterior density's area lies in this parameter range. The statement below,

'From our research, we concluded that the percentage of penguins with red tails, RT, has a 95% credible interval of $0\% \leq RT \leq 4\%$**.'**

can be interpreted as, 'From our research, we conclude that there is a 95% probability that the percentage of penguins with red tails, lies in the range $0\% \leq RT \leq 4\%$'.

In general an arbitrary credible interval of $X\%$ can be constructed from the posterior density, by finding a region whose area is equal to $\frac{X}{100}$.

In contrast to the classical confidence interval, a credible interval is more straightforward to understand. It is a probability statement of confidence

in the location of a parameter. Also, in contrast to the classical confidence intervals, the uncertainty here refers to our inherent uncertainty in the value of the parameter, rather than counter-factual samples.

There are usually an infinite number of regions which satisfy this condition, as figure 3.6 indicates for the regression example used in section 3.3. All three of the examples shown in figure 3.6 satisfy the condition, that given our choice of model and prior, we conclude that there is a 95% probability that the parameter lies in this range.

In order to reduce the number of credible intervals down to one, there are 'industry standards' that are followed in most applied research. We introduce two of the most frequently used metrics now.

Treasure hunting: The central posterior and highest density intervals

Imagine you (as a pirate) were told by a fortune-teller that treasure of \$1000 is buried somewhere along the seashore of an imagined island. Further, imagine that the mystic has gone to the trouble of using their past experience, and intuition, to arrive at a posterior density for the location of the treasure, along the x-axis seashore, that is shown in figure 3.7. The cost to hire a digger to dig up 1km of coastline is \$100.

Suppose you want to find the gold with 95% certainty, and maximise your profit in doing so[5]. In order to reach this level of confidence in plundering the gold, you have the choice of the two 95% credible intervals shown in figure 3.7: the left-hand *central posterior interval*, and the right-hand *highest density interval*.

Both of these intervals have the same area, so we are equally likely to find the gold in either. So which one would be best to choose?

The central posterior interval spans a range of $0.25km - 9.75km$ along the beach. This would entail a cost of $9.5 \times \$100 = \$950$.

The highest density interval by contrast spans two non-contiguous regions, given by $0km - 2.5km$ and $7.5km - 10km$. Each has a cost of $2.5 \times \$100 = \$250$, meaning a total cost of \$500. We pick the right-hand strategy, and cross our fingers!

[5]This is really a Bayesian Decision Theory question, where you are choosing an *action*, which corresponds to an interval; specifying a constant cost function across the domain.

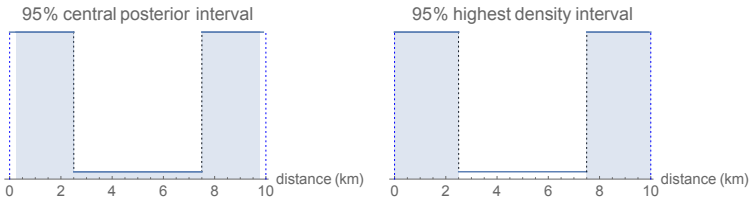

Figure 3.7: The posterior probability for treasure being found along the seashore (represented by a linear x-axis).

Intuitively, we should avoid the left-hand central posterior interval since it this involves digging up more coastline which has a very low probability of containing the gold. The right-hand highest density interval avoids this misspent effort by directing our search efforts towards the areas most likely to contain the treasure.

If it was costly to drive a digger a given distance, without digging, then we might change our minds, and favour the contiguous region of the *central posterior interval* over the *highest density interval*. However, in most practical (non-pirate) situations, the most sensible thing to do is to report the highest density interval.

To work out the upper and lower bounds of an X% central posterior interval, we find the $\frac{X}{2}$% and $(100 - \frac{X}{2})$% quantiles of the posterior distribution. This will result in an interval that is centred on the median parameter value.

To find the X% highest posterior interval, we find the set of values which contains this percentage of the posterior density area, with the property that the probability density in this set is never lower than outside.

For a unimodal, symmetric distribution the central posterior density and highest density intervals will be the same. However, for more complex distributions, this is no longer the case (compare the left and right hand panels of figure 3.7).

3.3.6 Reconciling the difference between confidence and credible intervals

It is easy to jump on the bandwagon, and dismiss classical confidence intervals as misleading and misleading; favouring the Bayesian alternative

definitively. However, in doing so we are somewhat guilty of zealotry[6]. The two concepts really just represent different measures of uncertainty. As we explained in section 2.4, Frequentists view data sets as one of an infinity of exactly repeated experiments, and hence design an interval which contains the true value X% of the time across all these repetitions. The Frequentist confidence interval states uncertainty in terms of the interval itself. By contrast, Bayesians view the data as fixed, and the parameter as coming from an over-arching distribution. They correspondingly calculate an interval where X% of the distribution of the parameter which has been drawn from the overall sits.

The problem with the classical confidence interval is more that it is often interpreted *incorrectly*, as a *credible* interval. It is not a problem with the concept itself. It just depends on your personal preference, and situation, as to which you find more useful.

The following (slightly silly) example hopefully makes this difference of viewpoint clearer.

The interval ENIGMA

Suppose that at the outbreak of war, you are employed as a code breaker in hut 8 at Bletchley Park. By monitoring enemy communications we are able to identify the *source* of the message, although the message *contents* itself, is not. The source of the message is either submarine, boat, tank or aircraft. Messages on the frequencies being monitored contain details of the next domestic target of the enemy, and can either be dams, ports, towns or airfields.

Fortunately, previous code-breakers have managed to decode a significant proportion of messages, and have tallied up the proportions of communications from each source, which resulted in a particular attack destination (see figure 3.1). We also know from experience that the proportion of attacks on each destination are roughly similar.

Our job is to predict the next attack destination *given* that we have received the mode of communication used. Since there is uncertainty regarding the attack destination, we shall be making confidence intervals which consist

[6]Fanaticism.

	Attack destination			
Communication method	Dam	Port	Town	Airfield
Submarine	73%	50%	50%	13%
Boat	9%	25%	25%	16%
Tank	0%	25%	25%	66%
Aircraft	18%	0%	0%	5%
Total	100%	100%	100%	100%

Table 3.1: Historical communication frequencies resulting in an attack on a given location.

of groups of these entities. We are told to use the most narrow[7] intervals of width greater than or equal to 75% in all cases.

From these historical evidence we first put the data into a 'statistics-machine', turning the knob that says 'classical confidence intervals'. The result are the confidence intervals shown in table 3.2. In words, this is because the sum of the interval values contained in each column exceeds the threshold. So, for every attack destination, our intervals ensure that the true attack destination lies within these bounds at least 75% of the time.

	Attack destination				
Communications method	Dam	Port	Town	Airfield	**Credibility**
Submarine	[73%	50%	50%]	13%	**93%**
Boat	[9%	25%	25%	16%]	**100%**
Tank	0%	25%	25%	[66%]	**57%**
Aircraft	18%	0%	0%	5%	**0%**
Coverage	**82%**	**75%**	**75%**	**82%**	

Table 3.2: Classical confidence intervals calculated from data shown in table 3.1. Confidence intervals greater than or equal to 75% are indicated in red, surrounded by parentheses.

We next turn the dial to 'Bayesian credible intervals', and obtain the results shown in table 3.3 (see section 3.8.1 for a full explanation). In this case, we are implicitly assuming that the choice of attack mode is a random variable, and that the enemy chooses amongst them uniformly[8]. In this case, since the

[7] We suppose there is a cost to readying a destination against attack.

[8] Which isn't unreasonable given that we know from experience that the enemy attacks

sum of interval values in each row exceeds 75%, we have credible intervals. With these intervals, for each mode of communication, we will ensure that the true attack destination is contained within these destinations at least 75% of the time.

	Attack destination				
Communication method	Dam	Port	Town	Airfield	**Credibility**
Submarine	[73%	50%	50%]	13%	**93**%
Boat	[9%	25%	25%]	16%	**79**%
Tank	0%	25%	25%	[66%]	**78**%
Aircraft	[18%]	0%	0%	5%	**78%**
Coverage	**100%**	**75%**	**100%**	**66%**	

Table 3.3: Bayesian credible intervals calculated from data shown in table 3.1. Credible intervals greater than or equal to 75% are indicated in red, surrounded by parentheses. Note: 'credibility' is calculated by dividing the sum of interval values in each row by the row's total (see section 3.8.1 for a full explanation.)

The difference between these two measures is subtle. In fact, as is often the case, the intervals are actually similar, and overlap considerably. But which should we choose? Using the classical confidence intervals, we are assured that whatever mode of attack the enemy chooses, our interval will contain the true attack mode at least 75% of the time. A Bayesian would criticise the confidence interval for the case of the *Aircraft* communication method, since this is the empty interval! This clearly is nonsensical, since we know that one of the locations is about to be attacked. This error could be particularly costly if attacks coordinated via Aircraft are particularly costly. A Frequentist would argue that since, *at most*, Aircraft communications happen 18% of the time (for dams), this isn't something to worry about.

A Bayesian would also criticise a classical confidence interval, since for a given communication mode, what is the use in worrying about all the other communication modes? We aren't uncertain about the communication mode!

A Frequentist would argue that for attacks on airfields, the Bayesian confidence intervals only correctly predict this as the attack destination 66% of the time. Again, if these types of attack are particularly costly, then this

each of these locations in similar proportions.

interval might not be ideal[9]. A Bayesian would argue that, assuming a uniform prior, this type of attack only happens 25% of the time, and so is not something to worry about. Further, for every mode of communication, our credible intervals are guaranteed to never be nonsense, in contrast to the classical confidence interval.

3.4 Prediction using predictive distributions

Parameters of a statistical distribution are typically only of interest insomuch as they influence *real variables*, for which we collected data in the first place. Be it income levels, disease cases, or electoral votes; the data is what drove us to conduct a statistical analysis. As such, it is frequently useful to compare models (which may have very different statistical formulations) using this common currency, of *data*.

Before we have run our model, we only have our prior views of the likely values of parameters. It is often informative to convert from the currency of *parameters*, to that of *data*, in order to evaluate the tangible implications of the chosen priors[10].

Also, after we have fed our priors and likelihoods into the Bayesian formula, we are outputted with the posterior distribution for our parameters of interest. Fortunately, both of these are simple, due to the manipulable nature of probabilities.

3.4.1 Example: number of Republican voters within a sample

You find yourself working for a polling organisation, ahead of the next US Presidential election. Your job is to try to predict, out of a sample of 100 people, what will be the number voting for the Republican party. Based on previous work, you expect that the proportion of Republican voters in a sample, θ, can be represented by the prior distribution, $p(\theta)$, shown in the top-left of figure 3.8. To evaluate the implications of this prior, we would like to know what this means in terms of number, x, of people out of our sample of 100, who will vote for the Republicans; in other words, the *prior predictive*

[9] If we were to assign costs to each of these attack events, then we could work out optimal intervals, although this is the realm of Bayesian Decision Theory; not covered in this text.

[10] See chapter 5 for a much more in-detail discussion.

distribution. Fortunately, we can obtain this by manipulating probabilities, although we need to specify a likelihood function, $p(x|theta)$; in this case we pick a binomial distribution[11]. The prior predictive distribution here is essentially the marginal distribution of x, which we know from section 2.7.5 can be obtained by integrating[12] out the dependence of the parameter θ from the joint distribution $p(x, \theta)$:

$$\begin{aligned} p(x) &= \int_0^1 p(x, \theta) \mathrm{d}\theta \\ &= \int_0^1 p(x|\theta) p(\theta) \mathrm{d}\theta \end{aligned} \tag{3.6}$$

where to get to the second line from the first, we have used the conditional law of probability (2.23), to decompose the joint distribution into a conditional and prior. Notice that the prior predictive distribution is essentially a sum of the probability of x conditional on θ, weighted by our prior probability that we assign to that particular proportion of the sample voting Republican, θ. In this case, this results in the intuitive result, where the prior density and the prior predictive distribution[13] exactly line up (albeit on different scales).

After our first sample of 100 people from that particular location, we find that 32 of them would vote Republican. We use this data to calculate our likelihood, then combine this with our prior via Bayes' rule; obtaining the posterior density shown in the top-right of figure 3.8. Our job is now to estimate the number of people who will vote republican in a new sample, of the same size. Ideally, we would like to have a probability distribution to describe all the possible outcomes. This distribution is known as the *posterior predictive distribution*, because it is that which we would predict *after* obtaining our data sample. It is written as $p(x'|x)$ where x' represents the number of people voting Republican in the new sample, and x is the number voting Republicans in the previous sample ($x = 32$). We can use a

[11]The relevance and nice fit of this distribution is explained in full in chapter 4, so do not worry if you don't follow its use here.

[12]Since the parameter is continuous here.

[13]This is predictably called a Beta-binomial distribution, since it is made from combining a Beta prior, with a Binomial likelihood. Don't worry though, we will discuss this in detail in chapter 8.

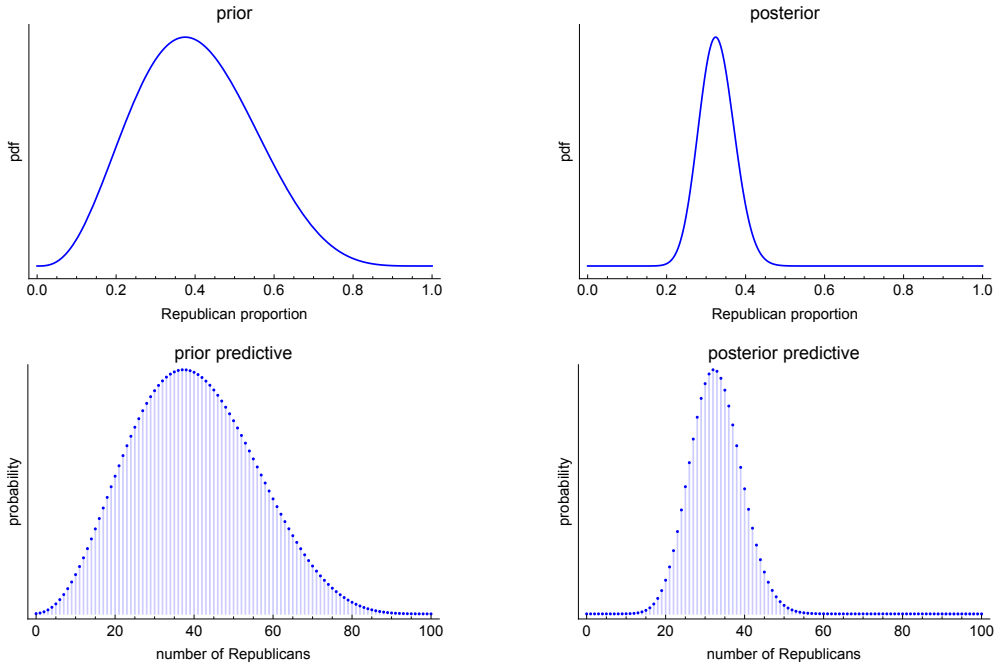

Figure 3.8: Top-left: the prior proportion of people voting for the Republican party in a sample of 100, resulting in the prior predictive distribution shown in the bottom-left. Top-right: the posterior proportion of people voting Republican, resulting the bottom-right posterior predictive distribution.

method exactly analogous to that which we used previously, by integrating out θ dependence of the joint distribution of x' and θ, although now we have to condition on x:

$$\begin{aligned} p(x'|x) &= \int_0^1 p(x', \theta|x) \mathrm{d}\theta \\ &= \int_0^1 p(x'|\theta, x) p(\theta|x) \mathrm{d}\theta \\ &= \int_0^1 p(x'|\theta) p(\theta|x) \mathrm{d}\theta \end{aligned} \tag{3.7}$$

Note that to get from the second line from the first, we have used the same conditional probability law (2.23), as we used for the prior predictive case. The only difference here is that we are additionally conditioning on x. To get from the second to the third line we have used a typical assumption, which is that once θ is known, the likelihood for our new sample does not depend on the previous data x. Another way to think about this is that all the information in x has been used to estimate θ; meaning that it does not confer any further information which is helpful for predicting x'. Again, we note in figure 3.8 how the posterior predictive distribution lines up exactly with the posterior distribution of θ. This makes intuitive sense, since if we predict the most likely *proportion* to vote Republicans is 0.32, we should expect that this will translate into the most likely *number* of $0.32 \times 100 = 32$, in a sample of 100.

3.4.2 Example: interest rate hedging

Suppose that you work as an analyst in an investment bank, focussing on predicting the actions of the central bank, so that your investments can be sufficiently hedged. The return of a certain investment, x, is probabilistically dictated by the rate of interest, and is also discrete; following a poisson distribution with rate parameter given by the chosen rate. Imagine that the current rate of interest is 1%, and the bank sets rates at 0.5% intervals; yielding a discrete distribution of possible values: $i \in$

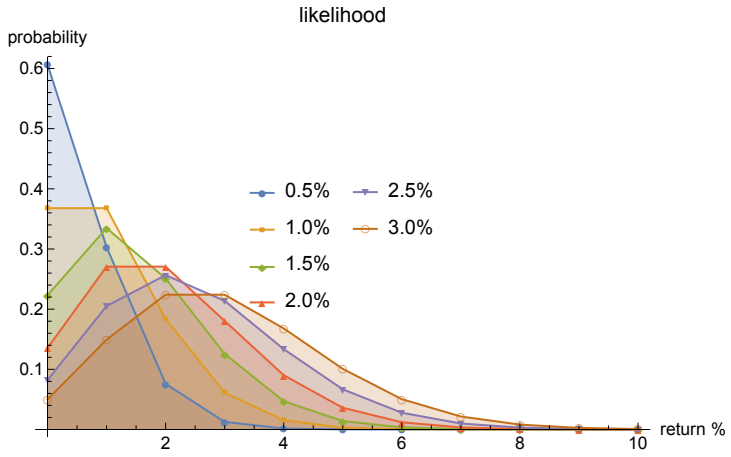

Figure 3.9: The likelihood of different rates of return across different central bank interest rates.

[0.5%, 1.0%, 1.5%, 2.0%, 2.5%, 3.0%]; resulting in the likelihood shown in figure 3.9.

The pre-data-analysis probabilities that your in-house economist gives to each different central bank rate are shown in figure 3.10. After combining these expert views with a probabilistic model, which looks at historical rate decisions, this results in the posterior distribution of probabilities shown in figure 3.10; illustrating considerable discordance between the data's view, and those of your economist[14].

From the prior distribution, $p(i)$, we would like to estimate the probability of particular rates of investment return. Mathematically, what we want is the marginal distribution of investment return, $p(x)$. Fortunately, we know from section 2.7.5, that we can get a marginal distribution from a joint distribution by summing (for a discrete parameter) across values of the parameter:

[14]This might indicate that your economist doesn't know his head from his hands, *or* he believes there is reason to suggest that this rate decision will be different to those historically.

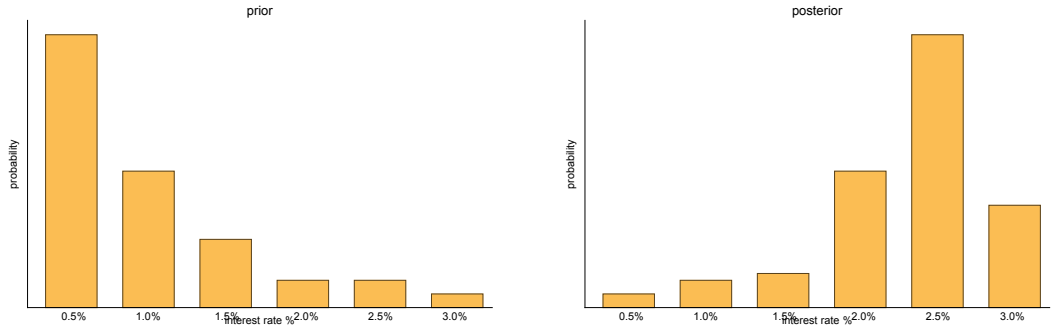

Figure 3.10: The prior and posterior probabilities of different central ban interest rates.

$$
\begin{aligned}
p(x) &= \sum_{i=0.5\%}^{3.0\%} p(x, i) \\
&= \sum_{i=0.5\%}^{3.0\%} p(x|i)p(i)
\end{aligned}
\tag{3.8}
$$

This distribution essentially weights each of the conditional probabilities, $p(x|i)$, by its corresponding prior probability $p(i)$ (see the left-hand side of figure 3.11), then sums them to yield the marginal probability of the data (see the right hand side of figure 3.11).

To calculate the posterior predictive distribution, we proceed in an analogous way to the prior case. What we would like to obtain is $p(x'|x)$, where x' indicates the predicted return on the investment, and x represents the historical data that has been used to produce the posterior. We can again achieve this by summing out all i dependence from the joint-conditional distribution $p(x', i|x)$:

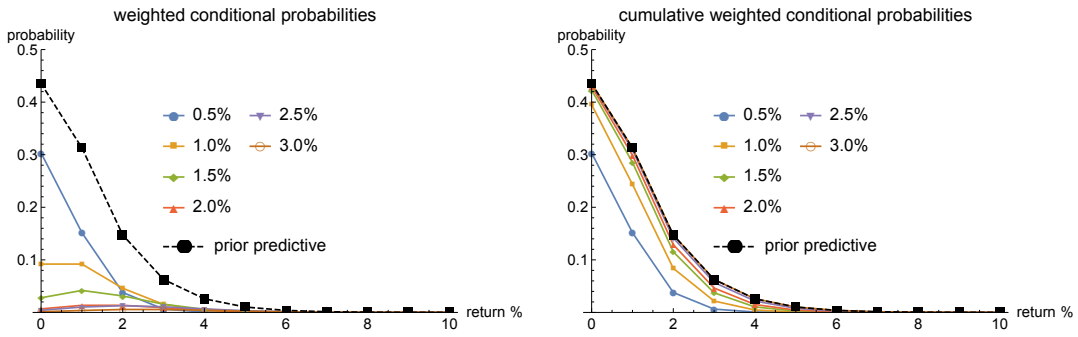

Figure 3.11: Building the prior predictive distribution, as a sum of weighted conditional probabilities. Left: the weighted conditional probabilities (compare with figure 3.9). Right: the cumulative weighted conditional probabilities, which converge on the prior predictive distribution.

$$
\begin{aligned}
p(x'|x) &= \sum_{i=0.5\%}^{3.0\%} p(x', i|x) \\
&= \sum_{i=0.5\%}^{3.0\%} p(x'|i, x)p(i|x) \\
&= \sum_{i=0.5\%}^{3.0\%} p(x'|i)p(i|x)
\end{aligned} \tag{3.9}
$$

Note that we have gone from the second line to the third line, because we suppose that, once i is known, the likelihood is independent of past values of x, and thus $p(x'|i, x) = p(x'|i)$. Also, note the similarity between this derivation and that for the prior predictive distribution in 3.8; the only difference in both second lines, is that in the posterior case, everything is conditioned on the historical values of x. We suppose that the likelihood function remains constant, so that like the prior predictive distribution, the posterior predictive distribution is obtained by weighting each of the conditional probability lines $p(x'|i)$ by a probability - in this case, the posterior probability for that value of i.

The resultant prior and posterior distributions are shown in figure 3.12. We notice that the posterior predictive distribution is shifted rightwards, towards higher investment returns, because of the fact that the posterior expected interest rate is higher than for the prior case.

3.5 Model comparison using the posterior

Suppose we have two competing models ($M1$ and $M2$) which we could use to explain a given dataset, and would like a way of evaluating their respective worth. The Bayesian framework can be used to incorporate the choice between two (or more) models, in a straightforward way. Suppose we denote a discrete parameter, $M \in [M1, M2]$, which indexes the choice of model. A way of gauging a model's performance is to calculate the probability of the model *given* the data obtained, $p(M|data)$. Since we have introduced this new parameter M, we can use Bayes' rule to calculate this quantity:

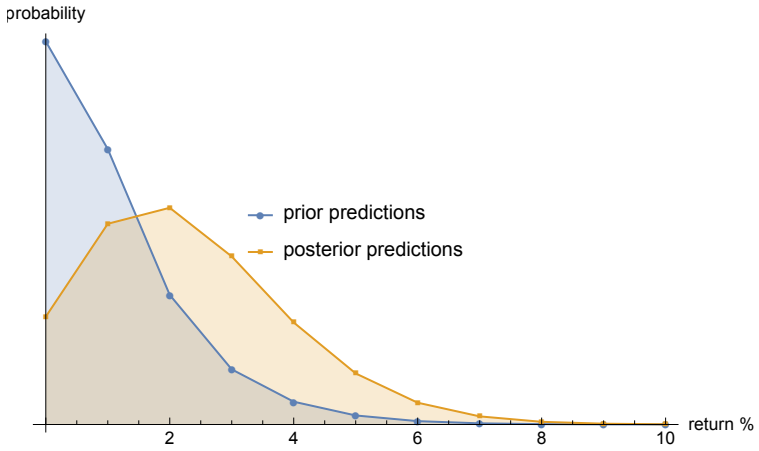

Figure 3.12: The prior and posterior predictive distributions for investment returns.

$$p(M|data) = \frac{p(data|M)p(M)}{p(data)} \tag{3.10}$$

Looking at (3.10) in detail, we examine the following quantities: $p(data|M)$, which represents what we, in the single model set-up, simply call $p(data)$; the $p(M)$ which represents our *a priori* probability of each model being correct; and $p(data)$, which in this case represents the marginal probability of the model when we have summed out our dependence on the model choice:

$$\begin{aligned} p(data) &= \sum_{M=M1}^{M2} p(data, M) \\ &= \sum_{M=M1}^{M2} p(data|M)p(M) \\ &= p(data|M1)p(M1) + p(data|M2)p(M2) \end{aligned} \tag{3.11}$$

We can then calculate the ratio of the probabilities of each model, given the data obtained:

$$\begin{aligned}\frac{p(M1|data)}{p(M2|data)} &= \frac{p(data|M1)p(M1)}{p(data|M2)p(M2)} \\ &= \frac{p(data|M1)}{p(data|M1)} \times \frac{p(M1)}{p(M2)} \\ &= BF \times \frac{p(M1)}{p(M2)}\end{aligned} \tag{3.12}$$

If this ratio is greater than 1, then this test suggests that we should favour model 1, and vice versa if the ratio is less than 1. We have also defined a quantity, BF, called the *Bayes Factor*:

$$BF = \frac{p(data|M1)}{p(data|M1)} \tag{3.13}$$

This factor represents in a narrow sense, the strength of support of model 1, over model 2, provided by the data. Note that if our prior probabilities of each model are identical, $p(M1) = p(M2) = \frac{1}{2}$, then:

$$\frac{p(M1|data)}{p(M2|data)} = BF \tag{3.14}$$

And the choice of model is determined solely through the Bayes Factor.

However, there is uncertainty over when is the BF enough to prefer one model over another. Whereas a BF of 100 may be enough to prefer a given model, how about 1.01? Jeffrey's scale (introduced in chapter 12), is an arbitrary, albeit industry-standard, meter to determine when a given model should be preferred over another.

I do importantly want to state here, that this is *not* the correct way to choose between competing models. We will learn a much more nuanced, and reasonable way of choosing between models when we introduce the concept of *Posterior Predictive Checks* in chapter 12. That is not to say that these methods introduced above can't be used as additional evidence for or against a given model, it is just that they should not be used as *sole*, or even *primary*, method.

3.5.1 Example: epidemiologist comparison

Suppose you find yourself in the (bizarre) situation, where you would like to compare two epidemiologists in terms of their ability to predict the underlying proportion of individuals having colds. The first epidemiologist, named *optimist*, gives his estimation of the proportion of individuals with colds, θ, via the (prior) distribution shown in figure 3.13. The second, named *pessimist*, gives a slightly more conservative estimate of the underlying proportion of individuals with colds (also shown in figure 3.13)[15].

The data we have is 10 samples of 100 people, where survey respondents have indicated whether or not they are currently suffering from a cold, $\boldsymbol{x} = \{22, 18, 18, 12, 16, 15, 21, 19, 14, 15\}$.

We can then go through and calculate the $p(data|M)$ for each of the two different priors (see figure 3.13 for a graphical depiction of this), by simply integrating the joint density with respect to θ:

$$\begin{aligned} p(data) &= \int_0^1 p(data, \theta) \mathrm{d}\theta \\ &= \int_0^1 p(data|\theta) p(\theta) \mathrm{d}\theta \end{aligned} \tag{3.15}$$

where $p(data|\theta)$ is the likelihood function[16]. Carrying out this integrand for each of the different priors results in the probabilities of data given in the bottom panel of figure 3.13. From this, we note two things: firstly both of the probabilities are very small. This is typical, and is unsurprising when you consider the vast array of possible outcomes that could have occurred. Secondly, we see that the probability of the data for the pessimist is higher than for the optimist. Indeed, we can calculate the Bayes factor via (3.13), and find here $BF \approx 6$ for the pessimist vs the optimist model; which, if we assign equal priors to each model, results in us favouring the pessimistic perspective.

I should add that this use of Bayes factors is atypical; it is not usually used

[15] Here I have actually generated these priors by assuming, for the optimist, $\theta \sim Beta[3, 40]$ and $\theta \sim Beta[4, 15]$ for the optimist and pessimist respectively.

[16] Which here is taken to be a binomial density.

as a means of testing between differing priors. However, since this forms part of a model, we can test its support with the data like any other part.

3.5.2 Example: customer footfall

Suppose that your job at a consultancy is to try to develop a statistical model which describes the footfall of customers into a particular store location at 1pm-2pm, over a span of 2 months. We have collected the data shown in the histogram in the left hand panel of figure 3.14.

Firstly, we fit a poisson likelihood to the data, since we know that the data were collected over a fixed period of time, and we might initially posit that the entry of an individual into the store is independent of others' entry[17]. However, we notice the point to the far right of the histogram. A poisson distribution is not reasonably able to cope with this degree of extremity, since it has a variance which is given by its mean - in this case this is estimated to be 10.5. For a small dataset, it seems hard to believe that a data point at 26, could have been generated from such a model[18].

An alternative model, which allows for a variance which exceeds its mean is the negative binomial. This choice of distribution would make most sense if we believed that one consumers' entry into a store was not independent of another's. This would make sense if people tend to shop in groups, and if the shop is outside, with people tending to enter *en masse* when it rains. The expense of this extra freedom, compared to the poisson, is that it is a two-parameter distribution, opposed to a single one.

Again, we can go through and calculate the probability of the data in each case (see figure 3.14), and then take the ratio to find $BF \approx 3$ for the negative binomial vs the poisson, which is unsurprising given the extreme observation. Here though, we might be tempted to *a priori* favour the poisson, due to its relative parsimony; setting a prior for this model that more than accounted for its loss in explanatory power. This might mean that overall we end up using the poisson model, acknowledging its shortcomings in predicting extreme footfall. However this depends on the nature of these extreme events. If they account for a disproportionate part of sales, then it may be worth the extra flexibility of the negative binomial. If by contrast, the extreme footfall is often due to rain, where consumers simply enter to

[17] See chapter 7 for a more complete examination of this distribution.
[18] This would be more easily seen by posterior predictive checks.

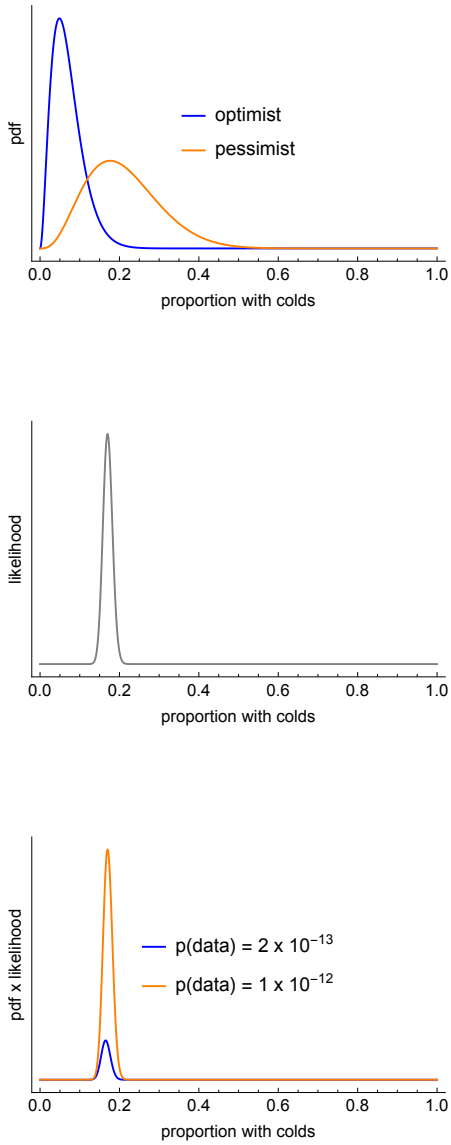

Figure 3.13: Calculating the probability of the data for the two epidemiologists' opinions on colds.

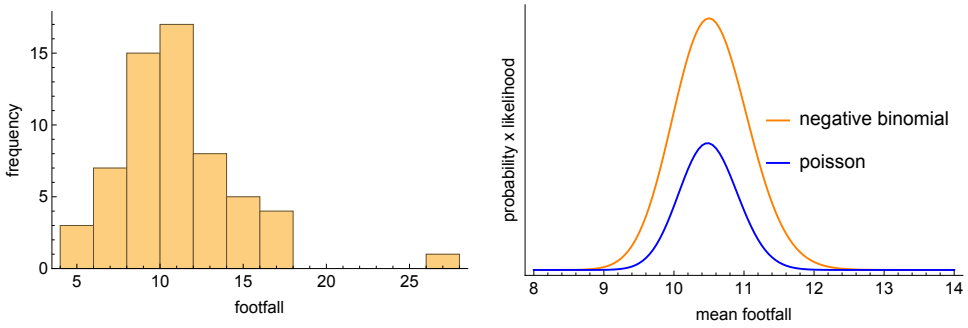

Figure 3.14: Left: the footfall data. Right: the area under the curves represents the probability of the data from each of the two models.

avoid getting wet, and fail to buy, we would perhaps not be so worried about the modelling of these events.

3.6 Model comparison through posterior predictive checks

The posterior is also the predominant means for testing the worth of a model, and comparing different models. The idea behind this method is that a reasonable model should be able to *simulate* data which is in some correspondence with the *real* data.

The idea behind this methodology is to simulate data, of the same dimension, and in the case of regression, with the same covariates, then compare this to the actual data. To go about simulating the data, we typically go through the following two steps:

1. Sample parameter values from the posterior: $\theta \sim p(\theta|x)$
2. Use these parameters in the likelihood function, then use this distribution to sample new data: $x' \sim p(x'|\theta)$

We shall see a full description of this methodology in chapter 12, so I only mention it briefly as a sign of things to come, and illustrate its basic mechanism through the following example.

3.6.1 Example: stock returns

Suppose we are tasked with modelling the daily-stock returns for a particular company over a one year period (see the leftmost panel of figure 3.15).

We firstly notice the symmetry of the returns, and reason that a normal likelihood may be a reasonable fit. We then use Bayes' rule, with appropriate priors (see chapter 8 for a good guide as to appropriate priors for the parameters of a normal distribution), to calculate posteriors for the mean and variance of this distribution. We then use these posterior distributions, which are fairly narrow, to firstly sample the mean, and variance, then use a normal likelihood to simulate a number of samples of the same size as the original data. For each of the simulated series, we compare the histogram of returns to that of the actual dataset. A typical plot of this form is shown in the middle panel of figure 3.15. We notice that the simulated data is a poor fit to the actual in a number of ways: it under-predicts the number of days with little stock movement; there are an over-abundance of days with moderate returns; and an under-weighting given to those days with more extreme stock movements. Overall, the dataset is not well represented by our model.

Instead of throwing in the towel, we decide to go for a distribution with fatter tails, which allows a greater degree of flexibility than that of a normal - a student's t distribution. We then go through the rigmarole of setting priors on its three parameters, and finally use Bayes' rule to find the posteriors of these. We then use the posterior distributions to sample these parameters, and then a t-likelihood to simulate the data a number of times. What we now typically see is a better fit (see the rightmost panel of figure 3.15), with a similar abundance of observations near zero, and a similar distribution of extreme daily movements.

Whilst these visual checks may become cumbersome with more complex data series, it is always a good idea to at least start with them, since the eyes can often be less misleading that, for example, p-values. However, we postpone a more complete discussion of posterior predictive checks until chapter 12.

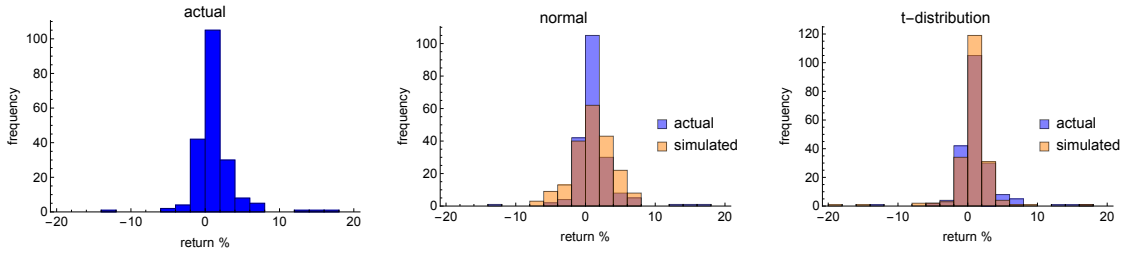

Figure 3.15: Left: the actual data. Middle: actual returns vs normal-simulated returns. Right: actual returns vs t-distribution-simulated returns.

3.7 Chapter summary

In this chapter we have seen how the posterior distribution can be used to produce summaries of parameters. In particular, we have used these distributions to produce Bayesian credible intervals. We then compared and contrasted these with the classically-equivalent confidence interval, and have reasoned that in many cases the Bayesian formulation is more straightforward and intuitive than the former. We then move from the realm of the parameter, to that of the data, in order to produce prior and posterior data probabilities; which can be used to sense-check the implications of statistical models. We finally discussed methods of model comparison, introducing firstly the Bayes Factor method, followed by a short introduction to posterior predictive checks. Although the former method is common in the literature, we advocated the more nuanced approach of posterior predictive checks, albeit postponing more in-depth discussion until chapter 12. Now that we have seen the utility of posterior probability distributions, we need to know how we can obtain them via Bayes' rule. In order to use the latter, we need to understand its constituent parts: the likelihood, prior and denominator. It is these three parts to which I devote the rest of the first part of this book.

3.8 Appendix

3.8.1 The interval ENIGMA - explained in full

Chapter 4

Likelihoods

The world is everything that is the case. **Wittgenstein**

4.1 Chapter Mission statement

At the end of this chapter a reader will know how to go about selecting a likelihood which is appropriate to a given situation. Further the reader will understand the basis behind maximum likelihood estimation.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{4.1}$$

4.2 Chapter goals

The starting point of the right hand side of the Bayesian formula is the likelihood function. This chapter will explain what is meant by a likelihood function, and why it is incorrect to view it as a probability in Bayesian analyses. The choice over which likelihood to use for a given situation is often difficult; especially to those unfamiliar with statistics. This chapter will

provide practical guidance on likelihood choice, describing a framework that can be used to select a model in a systematic way. As an important stepping stone to Bayesian estimation, this chapter will also explain how classical maximum likelihood estimation works.

4.3 What is a likelihood?

In all statistical inference, we use an idealised, simplified, model to try to mimic relationships between real variables of interest. This model is then used to test hypotheses about the nature of the relationships between these variables. In Bayesian statistics the evidence for a particular hypothesis is summarised in posterior probability distributions. Bayes' magic rule tells us how we can compute this posterior probability distribution for a given parameter within a model, θ:

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{4.2}$$

The first step to understanding this formula (so that we can ultimately use it!) is to understand what is meant by the numerator term, $p(data|\theta)$, which Bayesians call a *Likelihood*. Firstly, it's important to say that what we really mean by the numerator is:

$$p(data|\theta) = Probability(data|\theta, Model\ Choice) \tag{4.3}$$

(4.3) represents the probability that we would have obtained the 'data', given (this is represented by the | symbol) a particular value of θ and a particular choice of model. In other words, if our statistical model were true, and the value of the model's parameter were θ, (4.3) tells us the probability that we would have obtained our data.

But what does this mean in simple, everyday language? Imagine that we flip a *fair* coin. The most simple statistical model for coin flipping we can pick is to disregard the angle it was thrown at, as well as its height above the surface, along with any other details, and just pick the probability of the coin coming heads to be $\theta = \frac{1}{2}$. Furthermore, if a coin is thrown twice, we might

choose to model the situation by assuming that the throwing technique is sufficiently similar between the two throws such that we can model each throw as independently having a probability of $\frac{1}{2}$. It's important to note that it is an assumption to forget about the throwing angle, as well as height of throw for each throw, and this forms part of our model of the situation.

We can use our simple model[1] to calculate the probability that we obtain two heads in a row:

$$\begin{aligned} Pr(HH|\theta, Model) &= Pr(H|\theta, Model) \times Pr(H|\theta, Model) \\ &= \theta \times \theta = \theta^2 \\ &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \end{aligned} \tag{4.4}$$

The last row of (4.3) is obtained by assuming the probability of a head, $\theta = \frac{1}{2}$. If we continue to use this *same* value of θ, we can calculate the corresponding probabilities for all outcomes of throwing the coin twice. The most heads that can show up is 2, and the least being zero (if both flips come up tails). Figure 4.1 displays the probabilities for this model of the situation. The most likely number of heads to occur is 1, since this can occur in two different ways - either the first coin comes up heads, and the second is tails, or vice versa - whereas the other possibilities (all heads, or no heads) can each only occur in one way. However, the important thing to note about figure 4.1 isn't the individual probabilities, it is that it is a *valid* probability distribution[2], because:

- The individual event probabilities are all non-negative.
- The sum of the individual probabilities is 1.

So it appears when we assume a particular value of θ, and vary the data (in this case the number of heads obtained), the collection of resultant probabilities form a probability distribution. So, why do Bayesians call $p(data|\theta)$ a 'likelihood', and eschew the name 'probability'?

[1] Albeit in practicality, this is a pretty reasonable representation of the situation for most purposes.

[2] See section 2.7.1 for a refresher if you are unsure what is meant by a valid probability distribution.

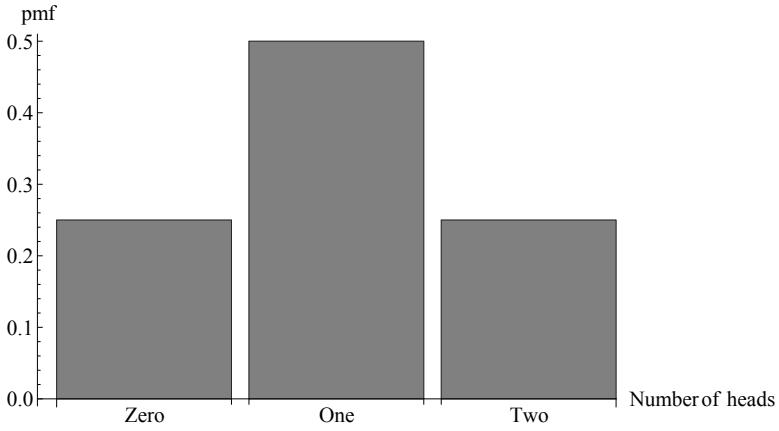

Figure 4.1: The probabilities of all possible numbers of heads for a fair coin.

4.4 Why use 'likelihood' rather than 'probability'?

When we hold the parameters of our model fixed, as when we held the probability of an individual throw turning up heads, $\theta = \frac{1}{2}$, we've reasoned that the first term of the numerator of Bayes' rule in (4.3) is a probability. So why don't we just keep calling it that, instead of renaming it a *likelihood*?

The reason is that in Bayesian inference, we *don't* keep the parameters of our model fixed! In Bayesian analysis, it is the *data* that is fixed, and the parameters that vary. This is because a posterior distribution shows the probability a parameter in a model lies in a particular range, assuming that we have obtained our particular data sample. For the case of a coin, where we don't know the probability of a head beforehand, what we hope to get out is a probability distribution of the kind shown in figure 4.2, where the x-axis is the value of θ. In order to get $p(\theta|data)$ however, we must calculate $p(data|\theta)$ from the numerator of Bayes' rule in (4.3) for each *possible* value of θ. If we assume we obtained one head and one tail, then we can calculate the probability of this occurring for a fixed θ:

$$Pr(HT|\theta) + Pr(TH|\theta) = \theta(1-\theta) + \theta(1-\theta) \quad = 2\theta(1-\theta) \tag{4.5}$$

Since we are unsure as to the 'correct' value of θ, we can graph this expression as a function of this parameter, to try to understand which values of the parameter are more or less likely, given our data (see figure 4.3).

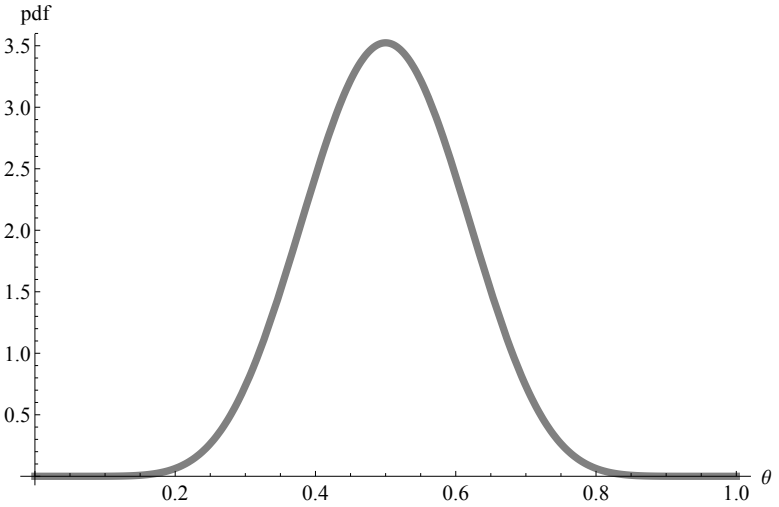

Figure 4.2: An example posterior distribution for the probability of obtaining a heads in a coin toss.

On first glances it appears that 4.3 could be a probability distribution, but first looks can be deceiving.

Checking off our necessary components of a probability distribution, we first note that all the values of the distribution in figure 4.3 are non-negative; which is what we require. However, if we calculate the area underneath the curve in figure 4.3:

$$\mathrm{I} = \int_0^1 2\theta(1-\theta)\mathrm{d}\theta = \frac{1}{3} \neq 1 \tag{4.6}$$

we find that it does not integrate to 1. Thus we have a violation of the second condition for a valid probability distribution. Hence, when we vary θ we find that, $p(data|\theta)$ is not a valid probability distribution! We thus introduce the term 'likelihood' to represent $p(data|\theta)$ when we vary the parameter, θ. Often the following notation is used to emphasise that likelihood is a function of the parameter θ with the data held fixed:

$$\mathcal{L}(\theta|data) = p(data|\theta) \tag{4.7}$$

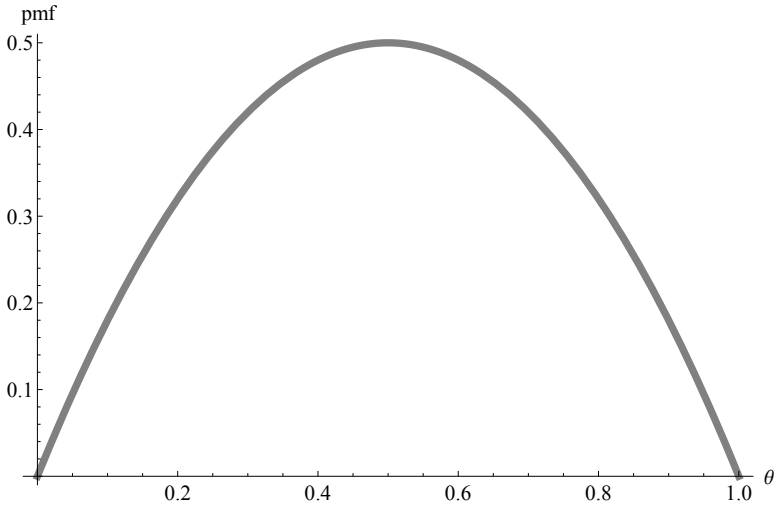

Figure 4.3: The likelihood function for obtaining a single head from two throws. The area under the curve is $\frac{1}{3}$.

However, in this book, we will persist with the original notation as this is most typical in the literature, under the implicit assumption that when we vary the parameters in question, the term is not strictly a probability.

To provide further justification for this argument, consider the following (albeit contrived) example. Suppose that, we throw a coin twice, and we are told beforehand that the probability of obtaining a head on a particular throw is one of six discrete values: $\theta \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. We can then use our model to calculate the probability of obtaining a number of heads, X:

$$Pr(X = 0|\theta) = Pr(TT|\theta) = Pr(T|\theta) \times Pr(T|\theta) = (1 - \theta)^2 \tag{4.8}$$

$$Pr(X = 1|\theta) = Pr(HT|\theta) + Pr(TH|\theta) = 2 \times Pr(T|\theta) \times Pr(H|\theta) = 2\theta(1 - \theta) \tag{4.9}$$

$$Pr(X = 2|\theta) = Pr(HH|\theta) = Pr(H|\theta) \times Pr(H|\theta) = \theta^2 \tag{4.10}$$

In (4.8), the probability is simply given by the product of the probabilities of not obtaining a head on the first throw, $(1 - \theta)$, by the probability of not obtaining a head in the second[3], which is also $(1 - \theta)$. The factor of two

[3]Since we have assumed a model whereby the results of the first and second throws are

arises in (4.10) since there are two ways of getting one head: {HT,TH}.

We can represent the corresponding values of likelihood/probability as is shown in table 4.1. In this form we can see the impact of varying the data (moving along each row), and contrast it with the effect of varying θ (moving down each column). Note that if we hold the parameter fixed - regardless of this initial choice of θ - and move along each row summing the entries, we find that the values sum to 1; meaning that this is a valid probability distribution. By contrast, when we hold the number of heads fixed, and vary the parameter θ, moving down each column, summing the entries, we find that the values do not sum to 1. Hence, when we vary θ, we are not dealing with a proper probability distribution, meriting the use of the term 'likelihood'.

In Bayesian inference, we always vary the parameter, and implicitly hold the data fixed. Thus, from a Bayesian perspective it is important to use the term *likelihood* to indicate that we recognise we are not dealing with a probability distribution.

	Number of heads			
θ	**0**	**1**	**2**	**Total**
0.0	1.00	0.00	0.00	**1.00**
0.2	0.64	0.32	0.04	**1.00**
0.4	0.36	0.48	0.16	**1.00**
0.6	0.16	0.48	0.36	**1.00**
0.8	0.04	0.32	0.64	**1.00**
1.0	0.00	0.00	1.00	**1.00**
Total	**1.20**	**1.60**	**2.20**	

Table 4.1: The values of likelihood for the case of tossing a coin twice, where the probability of heads is constrained to take on a discrete value: {0.0,0.2,0.4,0.6,0.8,1.0}.

independent, conditional on θ. In other words, all the similarity between the two throws is captured in the parameter θ.

4.5 What are models and why do we need them?

All models are wrong. They are idealised representations of reality resultant from making assumptions, which if reasonable, may emulate some of the behaviour of a system of interest. Joshua Epstein in an article titled, 'Why model?' emphasises that we perennially build *implicit* mental models for various phenomena [**?**]. Before we go to bed at night we set our alarms for the next morning on the basis of a model. We imagine an idealised - model - morning when it takes us 15 minutes to wake up as a result of an alarm. We use this model to predict how long it will take us to rise from bed, shower, and get changed into clothes in sufficient time to get to work. Whenever we go to the Doctor, they use an internalised biological model of the human body to advise on the best course of treatment for a particular ailment. Whenever we hear expert opinions on TV about the outcome of an upcoming election, the pundits are using mental models of society to explain the results of current polls, as well as make forecasts. As is the case with all models, some are better than others. Hopefully, the models a Doctor uses to prescribe medicine are subject to less error than the opinions of pundits seen on TV[4]!

Epstein goes on to emphasise that the question, 'Why model?' really means why should we build an *explicit* - written down - model of phenomena? The point being that *implicit* models are by their very nature, opaque, and not subject to the sort of interrogation that can be obtained by writing the model on paper.

We can also ask more narrowly, what are we hoping to gain by building an *explicit* model of a situation? Epstein suggests the following motivations:

- Prediction
- Explanation
- Guide data collection
- Discover new questions
- Bound outcomes to plausible ranges
- Illuminate uncertainties

[4]For a great discussion of the performance of TV pundits, read Thinking Fast. Insert reference

- Challenge the robustness of prevailing theory through perturbations
- Reveal the apparently simple (complex) to be complex (simple)

There are of course other reasons to build models, but we believe that this list is a reasonable starting point. However, we should not think of this list as static. Whenever we build a model, whether it is statistical, biological or sociological, we should ask, 'What are we hoping to gain by building this model, and how can I judge its success?'. Only when we have a grasp on the answers to these basic questions should we proceed to model building.

4.6 How to choose an appropriate likelihood?

Bayesians are acutely aware that their models are wrong. At best the abstraction from reality allows us to explain some aspect of real behaviour; at worst they can be very misleading. Before we use a model for prediction, we require that it can explain some reasonable proportion of the system's behaviour for the past and present. With this in mind we introduce the following model selection framework:

1. Write down the real life behaviour/data patterns that the model should be capable of explaining.
2. Write down the assumptions that it is believed are reasonable in order to achieve the above point.
3. Search the literature for models which utilise these assumptions; extracting only the relevant components.
4. Test your model's ability to explain said behaviour/data patterns. If unsuccessful go back to the second step and re-evaluate the appropriateness of your assumptions.

Whilst this methodology is useful for building a statistical model in general, it is more applicable for use with a full Bayesian model, resulting in a posterior distribution. In which case how do we go about specifying a likelihood for a given situation? To answer this we will start with going through a simple example.

4.6.1 A likelihood model for an individual's disease status

Suppose we work for the State as a healthcare analyst, and we want to build a statistical model to explain the prevalence of a certain disease within a sample, which can then be used to make inferences about the population incidence. Also, (unrealistically) let's imagine that we start off with a sample of only one person, for whom we have no prior information. Let the disease status of that individual be denoted by the variable X which takes on the following binary outcome values dependent on the disease status the individual:

$$X = \begin{cases} O & , \textit{No disease} \\ 1 & , \textit{Positive diagnosis} \end{cases} \tag{4.11}$$

The goal of our model is to output a probability that this individual has the disease. We might assume that a fraction θ of the population has the disease, and that this individual has come from that population. For each possible outcome, we can use this simple model to calculate the probability of each outcome:

$$Pr(X = 0|\theta) = (1 - \theta) \tag{4.12}$$
$$Pr(X = 1|\theta) = \theta \tag{4.13}$$

Note the similarity between these probabilities and those of the coin flipping example in the previous section. Often, a given model can be reused in a multitude of different settings.

However, we would like to write down a single rule which yields (4.12) or (4.13) respectively, dependent on whether $X = 0$ or $X = 1$. This can be achieved with the following:

$$Pr(X = \alpha|\theta) = \theta^{\alpha}(1 - \theta)^{1-\alpha} \tag{4.14}$$

Note that in (4.14) that $\alpha \in \{0, 1\}$ refers to the numeric value taken by the variable X. The function (4.14) is known as a *Bernoulli* probability density.

Although this rule for calculating a probability of a particular disease status, α, looks complex, we see that it reduces to (4.12) and (4.13) if the individual

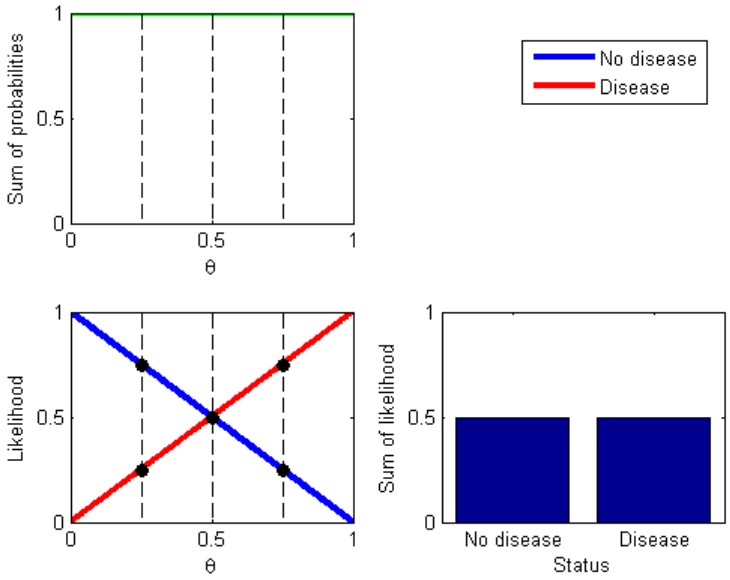

Figure 4.4: The likelihood function as theta varies for the case of the two possible data. The sum of likelihoods is found by the area under each line, whereas the sum of probabilities is a discrete sum.

is disease -negative/-positive respectively:

$$Pr(X = 0|\theta) = \theta^0(1-\theta)^1 = (1-\theta) \tag{4.15}$$

$$Pr(X = 1|\theta) = \theta^1(1-\theta)^0 = \theta \tag{4.16}$$

When we hold the datum X fixed and vary θ, (4.14) represents a likelihood. However, figure 4.4 shows that for a fixed value of *theta* the sum (here we mean the vertical sum) of the two probability densities is always equal to 1; demonstrating that in this case (4.14) is a valid probability density. Notice also in figure 4.4 that the sum of probability density is defined continuously on $\{0, 1\}$, whereas the sum of likelihoods is discrete.

4.6.2 A likelihood model for disease prevalence of a group

Now we imagine that instead of this solitary individual, we have a group of N individuals. What we would like to do is to calculate the develop a model which will tell us the probability of obtaining Z disease cases within our sample. We would also like to be able to use our model to predict the most likely number of individuals who have the disease in a sample, for a given value of the parameters[5].

In order to write down a model we first need to make some simplifying assumptions. We might assume that one individual's disease status tells us nothing about the probability of another individual in the sample having the disease[6]. This would not be a reasonable assumption if the disease were contagious, and if the individuals in the sample came from the same neighbourhood or household. It also would not be a good assumption if (as is often the case with volunteer-dependent studies) the individuals who volunteered for the experiment, self-selected on the basis of some common pre-existing ailment/underlying-factor. If an advert for participants reads, 'Psychological experiment on sleep disorders: participants wanted', we might suspect that there would be an over-presence of insomniacs than is found in the population as a whole. This first assumption is that which in statistical language we call 'independence'. We also suppose that all individuals in our sample come from the same population - the one we are trying to draw conclusions about. If we knew beforehand that some individuals came from different populations, with significantly different prevalence rates, then we might abandon this assumption. Combining these two assumptions we say in statistical language that our data sample is *independent* and *identically-distributed*.

With our two assumptions in hand, we can begin to formulate a model for the probability of obtaining Z disease-positive individuals out of a total of N individuals. We start by considering each person's disease status individually, meaning we can reuse (4.14):

$$Pr(X = \alpha|\theta) = \theta^{\alpha}(1-\theta)^{1-\alpha} \tag{4.17}$$

[5]We are starting off by assuming that we know the parameters. Later in this chapter we will obtain a point estimate of the parameters using *Maximum likelihood* estimation.

[6]Other than, if the disease prevalence were unknown, through our ability to estimate overall disease prevalence from their individual statuses.

Note that in (4.17) the $\alpha \in \{0, 1\}$ refers to a particular numeric value taken by the variable X. The assumption of *independence* means that we can get the overall probability by multiplying together the individual probabilities[7]. In words, we obtain the probability that the first person has disease status X_1 *and* the second person has status X_2:

$$\begin{aligned} Pr(X_1 = \alpha_1, X_2 = \alpha_2|\theta_1, \theta_2) &= Pr(X_1 = \alpha_1|\theta_1) \times Pr(X_2 = \alpha_2|\theta_2) \\ &= \theta_1^{\alpha_1}(1-\theta_1)^{1-\alpha_1} \times \theta_2^{\alpha_2}(1-\theta_2)^{1-\alpha_2} \end{aligned} \tag{4.18}$$

In (4.18) we have assumed that each individual has a different predisposition to having the disease, denoted by θ_1 and θ_2 respectively.

The second assumption of *identically-distributed* individuals means that we can set $\theta_1 = \theta_2$:

$$\begin{aligned} Pr(X_1 = \alpha_1, X_2 = \alpha_2|\theta) &= \theta^{\alpha_1}(1-\theta)^{1-\alpha_1} \times \theta^{\alpha_2}(1-\theta)^{1-\alpha_2} \\ &= \theta^{\alpha_1+\alpha_2}(1-\theta)^{2-\alpha_1-\alpha_2} \end{aligned} \tag{4.19}$$

In (4.19) we have obtained the second line by using the simple exponent rule: $a^b \times a^c = a^{b+c}$, for the components θ and $(1-\theta)$ respectively.

For our sample of 2 we are now in a position to calculate the probability that we obtain Z cases of the disease. We first realise that we can get from X_1 and X_2 to Z by:

$$Z = X_1 + X_2 \tag{4.20}$$

We can then use (4.19) to generate the respective probabilities.

$$\begin{aligned} Pr(Z = 0|\theta) &= Pr(X_1 = 0, X_2 = 0|\theta) = \theta^{0+0}(1-\theta)^{2-0-0} = (1-\theta)^2 \\ Pr(Z = 1|\theta) &= Pr(X_1 = 1, X_2 = 0|\theta) + Pr(X_1 = 0, X_2 = 1|\theta) = 2\theta(1-\theta) \\ Pr(Z = 2|\theta) &= Pr(X_1 = 1, X_2 = 1|\theta) = \theta^{1+1}(1-\theta)^{2-1-1} = \theta^2 \end{aligned} \tag{4.21}$$

To complete our probability model we want to write out a single rule for calculating the probability of any value taken on by Z. To do this we note that we could rewrite (4.21) as:

[7] See section 2.9 for an explanation of this.

$$\begin{aligned} Pr(Z = 0|\theta) &= \theta^0(1-\theta)^2 \\ Pr(Z = 1|\theta) &= 2\theta^1(1-\theta)^1 \\ Pr(Z = 2|\theta) &= \theta^2(1-\theta)^0 \end{aligned} \tag{4.22}$$

In (4.22) we notice the common term $\theta^\beta(1-\theta)^{2-\beta}$ in each of the expressions, where $\beta \in \{0, 1, 2\}$ represents the number of disease cases found. Therefore this suggests that we may be able to write down a single rule as something similar to:

$$Pr(Z = \beta|\theta) \sim \theta^\beta(1-\theta)^{2-\beta} \tag{4.23}$$

The only problem with matching (4.23) with the previously obtained result is the factor of 2 on the middle line of (4.22). However, as a complete aside we note that when we expand a quadratic factor we get the following:

$$(x+1)^2 = x^2 + 2x + 1 \tag{4.24}$$

The numbers $\{1, 2, 1\}$ correspond here to the non-b-dependent coefficients of $\{x^2, x^1, x^0\}$ respectively. This sequence of numbers normally appears in early secondary school maths classes, and is either known as the binomial expansion coefficients or simply ${}^nC_r$. The expansion coefficients are normally written in compact form:

$$\binom{2}{\beta} = \frac{2!}{(2-\beta)!\beta!} \tag{4.25}$$

In (4.25) the ! has its usual meaning of factorial, and $\beta \in \{0, 1, 2\}$. We can therefore use this notation to help us to write down a single model for the probability of obtaining Z disease cases out of a total of 2 individuals using our model:

$$Pr(Z = \beta|\theta) = \binom{2}{\beta}\theta^\beta(1-\theta)^{2-\beta} \tag{4.26}$$

This likelihood function is illustrated for the three possible numbers of disease cases in figure 4.5.

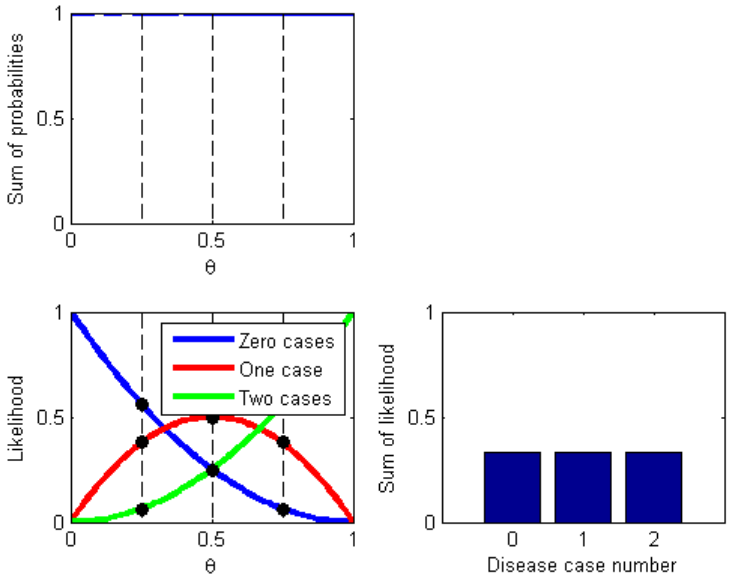

Figure 4.5: The likelihood function as theta varies for a sample of 2 individuals.

We will now extend the analysis to cover the case when we have groups of N individuals. Firstly, consider the case when we have a group size of 3. If we assume that the individuals are identically distributed, then the 4 probabilities are of the form:

$$\begin{aligned} Pr(Z=0|\theta) &= Pr(X_1=0|\theta)Pr(X_2=0|\theta)Pr(X_3=0|\theta) \\ Pr(Z=1|\theta) &= 3Pr(X_1=1|\theta)Pr(X_2=0|\theta)Pr(X_3=0|\theta) \\ Pr(Z=2|\theta) &= 3Pr(X_1=1|\theta)Pr(X_2=1|\theta)Pr(X_3=0|\theta) \\ Pr(Z=3|\theta) &= Pr(X_1=1|\theta)Pr(X_2=1|\theta)Pr(X_3=1|\theta) \end{aligned} \tag{4.27}$$

Again, we notice a numeric pattern in terms of the first part of each expression $\{1,3,3,1\}$, which happens to correspond exactly to the coefficients on terms for the expansion of $(x+1)^3$. Hence, we can again rewrite the likelihood using the binomial expansion notation:

$$Pr(Z=\beta|\theta) = \binom{3}{\beta}\theta^{\beta}(1-\theta)^{3-\beta} \tag{4.28}$$

We recognise a pattern in the likelihoods of (4.26) and (4.28) which allows us to deduce that, for a sample size of N, the likelihood is given by:

$$Pr(Z=\beta|\theta) = \binom{N}{\beta}\theta^{\beta}(1-\theta)^{N-\beta} \tag{4.29}$$

(4.29) is known as the *binomial* probability distribution.

If we had data, then we could test whether the assumptions made were appropriate by calculating the model-implied-probability of this outcome. For example, if we had a sample of 100 people of which 10 were disease-positive, and we assumed beforehand that the proportion of the population who have the disease is $\theta = 1\%$, then we could calculate the probability that we would have achieved a number of cases as bad, or worse than this using (4.29):

$$Pr(Z \geq 10|\theta=0.01) = \sum_{Z=10}^{100}\binom{100}{Z}0.01^{Z}(1-0.01)^{100-Z} = 7.63\times10^{-8} \tag{4.30}$$

We have summed over all the disease cases from 10 to 100 here, because we wanted the probability that we would have obtained a result as bad, or worse, than the one which we actually achieved. This is a particular way of carrying out classical hypothesis tests, which we will dispense with later on, but for now it seems a reasonable way of testing our model.

The probability found in this case is extremely small. What does this tell us? Well, it basically says that there is something wrong with our model which we have chosen here. It could be that the actual disease incidence in the population is much higher than the 1% which we have assumed beforehand. It could also be that our assumption of *independence* is violated in this case, for example if we sampled whole households rather than individuals. This could mean that in a particular household, the chance of having the disease, if another member of your family has the disease, is substantially higher than for the population as a whole.

It is difficult to gauge what in particular is wrong with our model without knowing further details of data collection, as well as how the estimate of 1% incidence was estimated for the population. However, it does suggest that we need to do adjust one or more of our assumptions, and reformulate the model to take these into account. We should never simply accept that our model is *correct*. A model is only as good as its capability to reproduce the data which we see in real life. In this case we find it is not a good representation, and we should readjust appropriately.

4.6.3 The intelligence of a group of people

We are now tasked with formulating a model of intelligence test scores for a group of individuals for whom we have data. We are told that the test score is on a continuous scale from 0-200. We do not have any information on individual characteristics which might help us to predict scores, although we are going to, for this simplified example, assume that we do know the mean test score $\mu = 70$, and its variance $\sigma^2 = 81$ in the population (although we will relax this assumption in section 4.9). We might assume that there are a range of factors which overall result in an individual's performance on this test. For example, these might include their schooling, parental education, 'innate' ability, as well as how tired they were feeling on the day of the test. If we assume that there are a large range of such factors and the score which results is an average of all these, then we might assume that the Central Limit Theorem might be appropriate for determining the distribution of

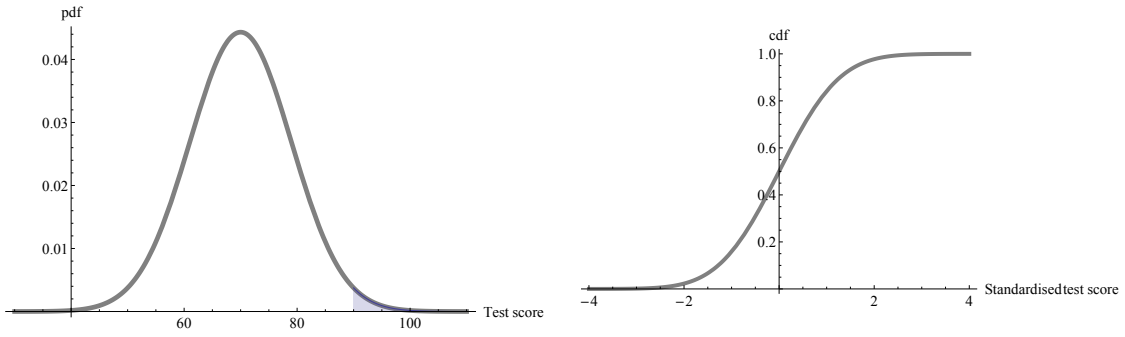

Figure 4.6: Left panel shows a normal with $\mu = 70$ and $\sigma^2 = 81$, with the area corresponding to a result as extreme as 90 indicated. This translates into a standard normal cdf shown in the right panel, which can be used to calculate this area from the first figure. This translation to the standard normal is done by taking away μ, and dividing through by σ. This is done since usually only standard normal cdf tables are available.

test scores[8]. In which case, we assume that a normal distribution for our likelihood function for an individual's test score, X:

$$p(X = \alpha|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha-\mu)^2}{2\sigma^2}} \tag{4.31}$$

Note that since this distribution is continuous, we have written p rather than Pr. The first p represents a density, whereas Pr represents a probability, which is only found in the continuous case by integrating over some bounds.

If we obtain an individual within our sample who achieved a test score of 90, we ask what's the probability of achieving a result as extreme as this? Using our idealised model, we just integrate the probability density (this is the continuous analogue to the discrete summing that we did in (4.6.2)):

[8]See section 2.10 for an introduction to the Central Limit Theorem.

$$\begin{aligned} Pr(X \geq 90|\mu = 70, \sigma^2 = 81) &= \int_{90}^{\infty} \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(\alpha-70)^2}{2\times 10}} \, \mathrm{d}\alpha \\ &= 1 - \Phi\left(\frac{90-70}{9}\right) \approx 0.0131 \end{aligned} \tag{4.32}$$

In (4.32), Φ stands for the value of the *standard* normal cumulative distribution function[9] at the value of 90 (see figure 4.6 for an explanation). Since we find that the probability of obtaining this data point under our current model is extremely small, we conclude that it is likely that there is something wrong with our model, and go back to examine the various assumptions that were made in deriving it.

If we also assume that information regarding one individual's test score tells us nothing about another's[10], then we might assume *independence* for our data. We might also assume that all individuals come from the same population; resulting in a *random sample*[11]. We calculate the joint probability density for a sample of N individuals by multiplying together the individual densities:

$$P(X_1 = \alpha_1, X_2 = \alpha_2, ..., X_N = \alpha_N|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha_i-\mu)^2}{2\sigma^2}} \tag{4.33}$$

We could then use (4.6.3) to calculate the probability of obtaining a given sample of observations as extreme as the values obtained, again by integrating. However, here it would be slightly more complicated than that of (4.32) since we would have to integrate across all individuals' variables.

4.7 Exchangeability vs random sampling

We have already been introduced to the concept of a *random sample*, in developing a probability model for the disease status of patients (section

[9] A standard normal has mean 0, and a variance of 1. By taking away the mean of 70, and dividing through by the standard deviation, we transform from an arbitrary mean- and variance-normal, to a *standard* one.

[10] Apart from their joint reliance on μ and σ^2.

[11] See section 4.7 for further discussion of random samples.

4.6.2), and the intelligence of a group of people (see section 4.6.3). The use of this term is really just a shorthand for an *independent*, and *identically-distributed* sample of data. Often however, Bayesians eschew this term in want of a (slightly) weaker condition that still allows us to write down an overall likelihood as a product of individual likelihoods in many situations.

Suppose we have a sequence of random variables representing the height of individuals in a sample of size 3: $\{H_1, H_2, H_3\}$. If this sequence is as likely as the reordered sequence: $\{H_2, H_1, H_3\}$, or any other re-ordering, then the sequence is said to be *exchangeable*[12].

Since the assumption of random sampling is stronger than that of exchangeability, it turns out that any random sample is automatically exchangeable. However, the converse is not necessarily true. A particular example of this is for the case of an urn containing 3 red and 3 blue balls, which are drawn at random without replacement. The probability of obtaining the sequence *RBR* is given by:

$$Pr(RBR) = \frac{3}{6} \times \frac{3}{5} \times \frac{2}{4} = \frac{3}{20} \tag{4.34}$$

The sequence of random variables representing the outcome of this sampling *is* exchangeable, since we have that any permutation of this sequence is equally likely:

$$\begin{aligned} Pr(BRR) &= \frac{3}{6} \times \frac{3}{5} \times \frac{2}{4} &= \frac{3}{20} \\ Pr(RRB) &= \frac{3}{6} \times \frac{2}{5} \times \frac{3}{4} &= \frac{3}{20} \end{aligned} \tag{4.35}$$

However, this sequence of random variables is *not* a random sample. The probability distribution for the first ball drawn is different to that when the second is drawn. In the first case there are 6 balls in total, with equal numbers of each. However, for the second case there are only 5 balls, and *dependent* on the first draw, there may be either more red balls or blue balls remaining.

In general we may not be able to assume we have a conditional[13] random sample of observations for reasons similar to that of the urn example.

[12]Formally, a sequence which is exchangeable requires that the joint probability distribution is invariant under any permutation of the order.

[13]Conditional on a distribution of a vector of parameters θ which sits above all the

However, a brilliant theory originally by Bruno de Finetti allows us to assume that a sequence behaves as if it is a random sample, *so long as it is exchangeable.* Technically this requires that we need an infinite sample of observations, but for a reasonably large sample making this approximation is reasonable.

Much of the time we will have a random sample, and so do not need to worry about any of this. However, due to this theorem, we are often free to write down an overall likelihood as the product of individual likelihoods, so long as the observations are exchangeable.

4.8 The subjectivity of model choice

It is hoped that the analysis in the preceding sections has given us a taste of how we can go about specifying a likelihood for a hitherto unknown circumstance. We start by writing down the behaviours that we want to emulate, then make simplifying assumptions, which we then use to look for an appropriate model in the literature. This model is then used to test the validity of the assumptions with the sample data. If the model struggles to explain the data, then we should go back and iteratively modify, then test our model, until it adequately explains the range of behaviours.

However, it should be re-emphasised that by its nature, a model is always a simplification of reality. As such, no one model is *correct*. There are often many models that could be used to explain the data which we have to hand. We should always take care to test each of these against its ability to explain the aspect of the data in which we are interested, and only proceed with it if it is adequate in this regard. Real life is complicated, and thus with each of the assumptions that were used to justify a particular model, there will inevitably be a degree of *subjectivity*. As such, no analysis - whether Frequentist or Bayesian - can be thought to be purely *objective*. Hence, the human analyst cannot, and should not, be replaced by automata for statistical analysis. A degree of subjective judgement is always necessary in statistics, as in all other walks of life.

---

observations.

4.9 Maximum likelihood - a short introduction

The analysis in section 4.6 assumes that we know beforehand the fraction, θ, of the populous that are predisposed to having the disease. In reality we rarely know such a thing. Often the main focus of building a statistical model is to try to estimate such parameters from our sample of data to which we have access. A popular Frequentist method for achieving this goal is the estimation strategy known as *Maximum Likelihood*. In this section we will examine how this estimation strategy yields estimates of parameters.

The principle of Maximum Likelihood estimation is simple. Firstly, we assume a model which we use to approximate the data generating process which resulted in our sample, based on the various assumptions about the real life process which we make. We then calculate what is known as the joint probability of obtaining the sample of observations, assuming that we do not know the parameters which specify completely those distributions. We then choose the parameters which *maximise* the likelihood of obtaining that particular sample of observations. We will go through some simple examples to illustrate this process.

4.9.1 Estimating disease prevalence

In section 4.6.2 we assumed that we knew beforehand the fraction of individuals who are disease-positive within the population. As mentioned previously, it is uncommon that such a thing be known before carrying out an analysis. If in a sample of 100 individuals, 10 test positively[14], and we make the same assumptions as in section 4.6.2 - that of a random sample - then we can write down the overall likelihood function using (4.29) as:

$$L(\theta|data) = \binom{100}{10}\theta^{10}(1-\theta)^{90} \tag{4.36}$$

Remember, that since we are varying θ and holding the data constant here, that (4.36) is a *likelihood*, not a probability. We then need to simply choose θ so that we can maximise the likelihood. We could simply differentiate (4.36) as it stands, and set the derivative equal to 0; rearranging the resultant equation for θ. However, to make life a little easier for us, we are first going

[14] Assuming for simplicity that there are no false-positives.

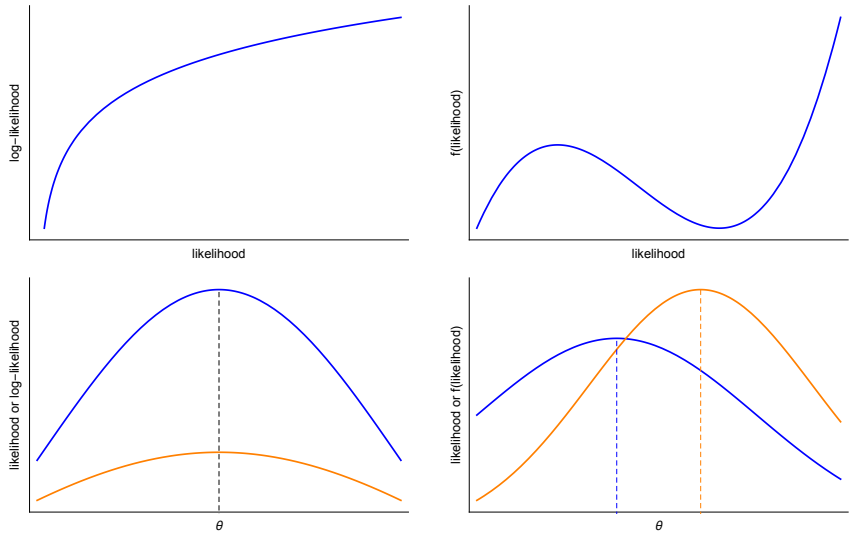

Figure 4.7: The monotonicity of log-likelihood (top-left), means that the peaks of likelihood and log-likelihood coincide (bottom-left). However, this is not the case for an arbitrary function (top-right and bottom-right).**Add legends to the bottom two graphs.**

to take the *log* of this expression, then differentiate it, setting the derivative to 0; resulting in the same value of θ. We are able to do this because of the simple properties of the log transformation (see figure 4.9.1):

$$l(\theta|data) = Log\left(L(\theta|data)\right) = log\binom{100}{10} + 10log(\theta) + 90log(1-\theta) \tag{4.37}$$

Where to get the result (4.37), we have used the log rules:

$$\begin{aligned} log(ab) &= log(a) + log(b) \\ log(a^b) &= blog(a) \end{aligned} \tag{4.38}$$

We can now simply differentiate the log-likelihood $l(\theta|data)$:

$$\frac{\partial l}{\partial \theta} = \frac{10}{\hat{\theta}} - \frac{90}{1-\hat{\theta}} = 0 \tag{4.39}$$

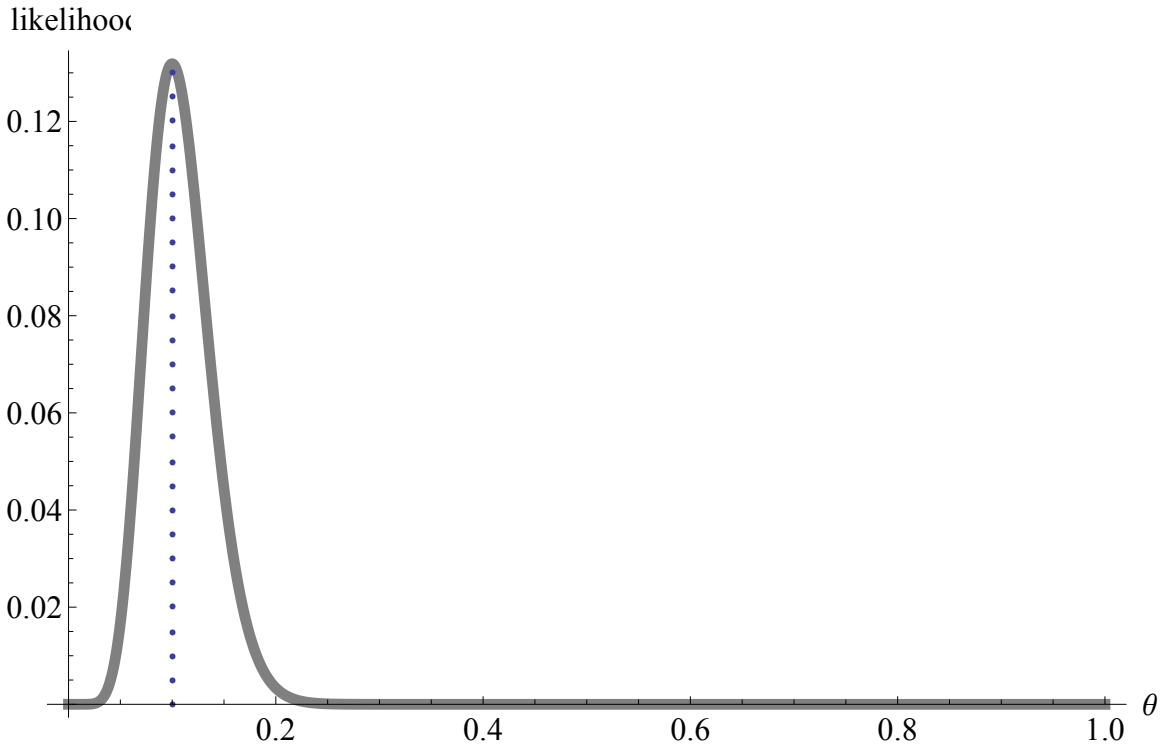

Figure 4.8: Log-likelihood of disease prevalence from section 4.9.1 as a function of the proportion of individuals which have the disease in a population, θ. The dotted line shows the maximum likelihood estimate $\hat{\theta} = 1/10$.

If we set the derivative to 0 we then obtain the maximum likelihood *estimate*, $\hat{\theta} = \frac{1}{10}$ (see figure 4.8).

This estimator makes sense intuitively. The value of the parameter which results in the highest likelihood of obtaining the data occurs when the population prevalence exactly matches that obtained in our sample. In general if we found a number β of individuals out of a sample of size N, who were disease-positive, then we would again find that the preceding analysis results in an estimator[15] of the disease prevalence exactly equal to that in our sample:

[15] An estimator is a mathematical function which outputs an estimate of a parameter in our model.

$$\hat{\theta} = \frac{\beta}{N} \tag{4.40}$$

4.9.2 Estimating the mean and variance in intelligence scores

We are given a sample of individuals with test scores $\{75, 71\}$, and we model the test scores using a normal likelihood as described in section 4.6.3:

$$L(\mu, \sigma^2 | X_1 = 75, X_2 = 71) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(75-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(71-\mu)^2}{2\sigma^2}} \tag{4.41}$$

We can then proceed as we did in section 4.9.1 by taking the log of this expression before we differentiate it:

$$l(\mu, \sigma^2 | X_1 = 75, X_2 = 71) = 2log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(75-\mu)^2}{2\sigma^2} - \frac{(71-\mu)^2}{2\sigma^2} \tag{4.42}$$

Where we have again used the log rules in (4.38) to achieve (4.42). We can now proceed to differentiate (4.42) with respect to both variables, holding the other constant, setting each to 0:

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{(75-\hat{\mu})}{\hat{\sigma^2}} + \frac{(71-\hat{\mu})}{\hat{\sigma^2}} = 0 \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{1}{\hat{\sigma^2}} + \frac{(75-\hat{\mu})^2 + (71-\hat{\mu})^2}{2\hat{\sigma^4}} = 0 \end{aligned} \tag{4.43}$$

The first of these expressions yields $\hat{\mu} = \frac{71+75}{2} = 73$, which when put into the second gives:

$$\hat{\sigma^2} = \frac{1}{2}\left[(75-73)^2 + (71-73)^2)\right] = 4 \tag{4.44}$$

Notice that the maximum likelihood estimators for the population mean and variance are for this case the *sample mean* and *sample variance*[16]. In fact,

[16] Albeit a biased estimator of the population variance. The unbiased estimator would divide by 1, rather than 2.

this holds for the case of N individuals' data, then the maximum likelihood estimators for this case would be:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X} \tag{4.45}$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N} (X_i - \bar{X})^2 = s^2 \tag{4.46}$$

4.10 Frequentist inference in Maximum Likelihood

We have now detailed how to derive point estimates of parameters using the method of maximum likelihood. However, at the moment we are unable to make any conclusions about the population. This is because we do not have any idea as to whether we obtained a particular estimate of a parameter due to picking a weird sample, or because it *actually* has a value in the population which is at this value. Frequentists get round this by examining a graph of log-likelihood near the maximum likelihood point estimate (see figure 4.9). If the log-likelihood is strongly peaked near the maximum likelihood estimate, then this suggests that only a small range of parameters would yield a similar valued likelihood. By contrast, if the log-likelihood is gently peaked near the ML estimate, then it is feasible that a large range of parameters would yield estimates close to this value. In the latter case, it seems logical that we should be less confident in the particular value of the parameter which is given by maximum likelihood. We can measure the 'peakedness' in the log-likelihood by looking at the magnitude of the second derivative[17] of the function at the ML point estimate value. The more curved the log-likelihood, the more confident we can be of our estimated parameter value, and any conclusions drawn from this. Note however, that the Frequentist inference is not based on proper probability distributions (since we infer based on a likelihood). This contrasts with the Bayesian method which, by its nature, allows for a more adequate description of parameters, using probability distributions.

[17]The first derivative gives the gradient, the second derivative gives the rate of change of the gradient - a measure of curvature.

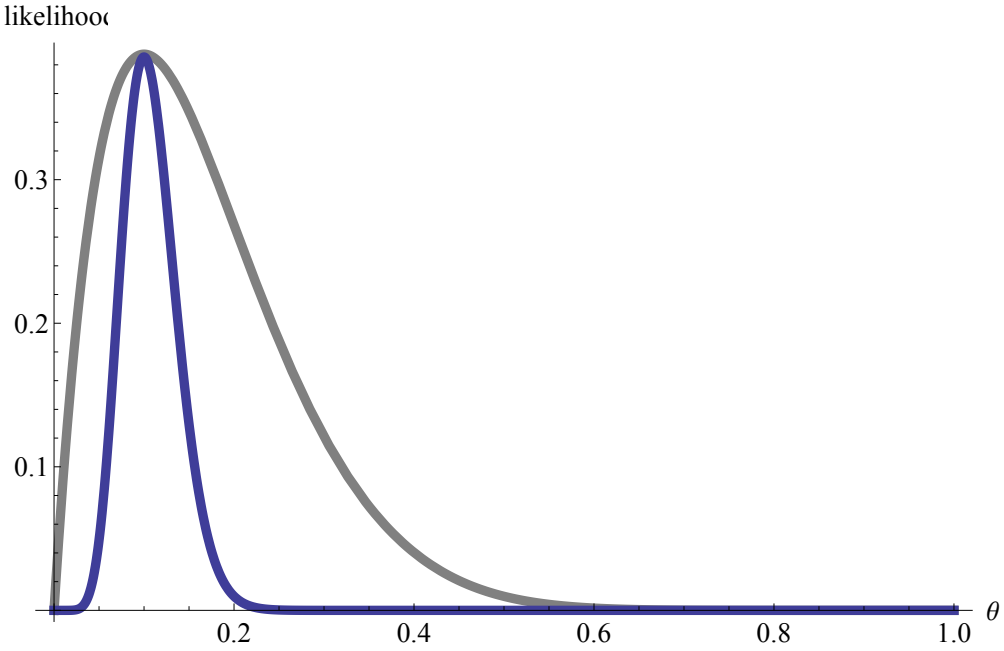

Figure 4.9: Two likelihoods which result in the same maximum likelihood estimates of parameters, at 0.1. The gray likelihood is less strongly-peaked, meaning we can be less confident about the estimates.

4.11 Chapter summary

We should now understand what is meant by a likelihood, and how to build probabilistic models of real life processes. However, the difficulty of modelling a process is governed by its degree of complexity and sensitivity to violations of assumptions. Further we should also understand how the Frequentist method of Maximum Likelihood can be used to yield point estimates of parameters. We are however, currently restricted in our ability to make inferences based on full probability distributions over parameters. Bayes' rule tells us how we can convert a likelihood - itself not a proper probability distribution - to a posterior (*correct*) probability distribution for parameters. In order to use to do this though, we need to understand what is meant by a *prior* distribution and how we can specify this distribution to suit the particular situation. This is what is covered in the next chapter.

Chapter 5

Priors

5.1 Chapter Mission statement

At the end of this chapter a reader will know what is meant by a prior, and the different philosophies that are used to understand and construct them.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)} \tag{5.1}$$

5.2 Chapter goals

Bayes' rule tells us how to convert a likelihood - itself not a proper probability distribution - into a posterior probability distribution for parameters, which can then be used for inference. We are required in the numerator to multiply the likelihood by a pre-experimental weighting of each set of parameter values described by a probability distribution, which is known as a *prior*. Priors are without doubt the most controversial aspect of Bayesian statistics, with its opponents criticising its inherent *subjectivity*. It is hoped that by the end of the chapter we will have convinced the reader that, not only is subjectivity inherent in *all* statistical models - both Frequentist

and Bayesian - but the explicit subjectivity of priors is more transparent, and hence open to interrogation, than the implicit subjectivity abound elsewhere.

This chapter will also explain the differing interpretations which are ascribed to priors. The reader will come to understand the types of method that can be used to construct prior distributions, and how they can be chosen to be minimally subjective, or otherwise to contain informative pre-experimental insights from data or opinion. Finally, the reader will understand that if significant data are available then the conclusions drawn should be insensitive to the initial choice of prior.

Inevitably, this chapter will be slightly more philosophical and abstract than other parts of this book, but it is hoped that the examples given will be sufficient to ensure its practical use.

5.3 What are priors, and what do they represent?

Chapter 4 introduced us to the concept of formulating a likelihood, and how this can be used to derive Frequentist estimates of parameters, using the method of maximum likelihood. This pre-supposes that the parameters in question are immutable, fixed quantities that actually exist, and can be estimated by methods that can be repeated, or imagined to be repeated many times [**?**]. As Gill (2007) indicates, this is unrealistic for the vast majority of social science research.

> It is simply not possible to rerun elections, repeat surveys under exactly the same conditions, replay the stock market with exactly matching market forces, or re-expose clinical subjects to identical stimuli.

Furthermore, parameters only exist because we have *invented* a model, hence we should innately be suspicious of any analysis which assumes an existence of a single certain value for any aspect of these abstractions.

For Bayesians, it is the data that are treated as fixed, and the parameters that vary. We know that the likelihood - however useful - is not a proper probability distribution. Bayes' rule tells us how to combine a likelihood with something called a *prior* to obtain a proper posterior distribution for

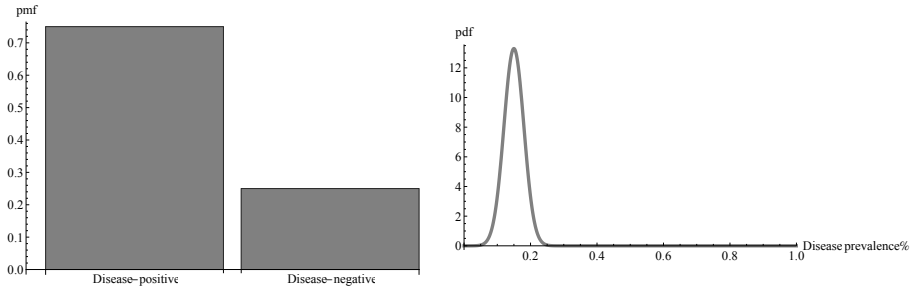

Figure 5.1: Left - a prior for a doctor's pre-testing diagnostic probability of an individual having a disease. Right - a prior which represents pre-sample uncertainty in disease prevalence.

the parameter in question, which can then be used for inference. But what does it actually mean for a parameter to have a prior distribution?

Gelman et al. (2013) suggests that there are two different interpretations of priors: the *state of knowledge* interpretation, where we specify our knowledge and uncertainty in a parameter as if regarding it as a draw from a probability distribution; alternatively in the more objective *population* interpretation where the current value of a parameter is the result of a draw from a true population distribution [**?**]. In both viewpoints the model parameters are not viewed as static, unwavering constants as they are taken to be in Frequentist theory.

If we adopt the *subjective* viewpoint above, then we can think of the prior as representing our pre-experimental/data certainty in the parameter in question. For example, imagine that a Doctor is asked to evaluate the probability before the results of a blood test become available, that a given individual has a particular disease. Using their knowledge of the patient's history and their expertise on the particular condition, they assign a prior disease probability of 75% (see figure 5.1).

Alternatively, imagine we are tasked with estimating the proportion of the UK population that has this disease. We may have some idea of its prevalence, as well as the variance in the mean prevalence of a disease across a range of previous samples of individuals which have been tested. In this case, the prior is continuous and represents our uncertainty in our estimate of the prevalence (see figure 5.1). In all cases a prior is a proper probability distribution, and hence can be used to elicit our prior expectations as to

the value of a parameter. For example, we could use the prior probability distribution for the proportion of individuals having a particular disorder in figure 5.1 to estimate a pre-experimental mean of approximately 15% prevalence.

Adopting the *population* perspective, we imagine the value of a parameter of current interest to be drawn from a population distribution. If we imagine the process of flipping a coin, we could if we knew the angle at which it is tossed, as well as the height from which it is thrown above the surface[1] predict deterministically the side on which the coin would fall face up. We could then hypothetically enumerate the (infinitely) many angles and heights of the coin, and for each set determine whether the coin would fall face up or down. Each time we throw the coin we are implicitly choosing an angle and height from the set of all possible combinations, which determines whether a heads or tails falls face up. Some ranges of the angle and the height will be more frequently chosen than others, albeit relatively agnostic with regards to final state of the coin. Hence we could think of this choice as the realisation from a distribution of all possible sets. Thus we could think about the choice of angle and height as being a realisation from this *population* distribution, and hence determines the fate of the coin toss.

Alternatively, going back to the disease prevalence example, we could imagine that each time we pick a sample, the data we obtain is partly determined by the exact characteristics of the sub-populations from which these individuals were drawn. The other part of variability is sampling variation within those sub-populations. Here we can view the particular sub-population characteristics can be viewed as draws from an overall population distribution of parameters, representing the entirety of the UK.

5.4 Why do we need priors at all?

A question we might ask is, why do we need priors at all? Can't we simply let the data speak for itself, without the need of these subjective beasts?

Frequentists without knowing it actually do use something equivalent to priors, by setting the size of statistical tests[2]. However, can't we as Bayesians side-step this subjective jump completely?

[1] Also assuming that we knew the physical properties of the coin and surface.

[2] See section 2.13 for a further discussion.

The answer to this question is provided by Bayes' rule. Its inclusion in Bayes' rule, which is the only correct way to update beliefs, means that if we are to be consistent with the laws of probability, we are required to provide this part for any Bayesian inferential procedure.

If you find this description somewhat unsatisfying, then another way of phrasing this same argument, is that Byaes' rule is really only a way to *update* our initial beliefs, to yield new beliefs which reflect the weight of the data obtained:

$$initial\ belief \xrightarrow{Bayes'\ rule} new\ beliefs \tag{5.2}$$

Viewed in this light, it is clear that we need to specify an initial belief, otherwise we have nothing to update! Bayes unfortunately doesn't tell us how to formulate this initial belief, but fear not, we will in this chapter delve into how we can go about setting priors in practice.

5.5 Why don't we just normalise likelihood by choosing a unity prior?

Another question that can be asked is, 'Why can't we simply let the prior be unity for all values of θ?'; in other words set $P(\theta) = 1$ in the numerator of Bayes' rule; resulting in a posterior that takes the form of a normalised likelihood:

$$P(\theta|data) = \frac{P(data|\theta)}{P(data)} \tag{5.3}$$

This would surely negate the need for specification of a prior, and thwart all attempts to denounce Bayesian statistics as *subjective*. So why don't we do just that?

There is a pedantic, mathematical argument against this, which is that $P(\theta)$ must be a proper probability distribution to ensure the same properness of the posterior. If we choose $P(\theta) = 1$ (or in fact any positive constant), then the integral $\int_{-\infty}^{+\infty} P(\theta)\mathrm{d}\theta \to \infty$, and we can no longer think of the distribution,

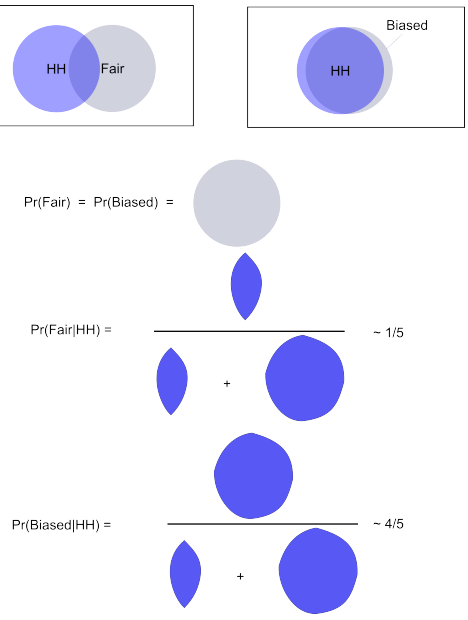

Figure 5.2: Ignoring common sense, by setting a uniform prior between a coin being fair and biased results in unrealistic posteriors.

$P(\theta)$ as representing a probability. It may still be possible that even if the prior is improper, that the resultant posterior also satisfies the required properties of a proper probability distribution, but care must be taken when using these distributions for inference, as technically they are *not* probability distributions, due to the abuse of Bayes' rule. In this case the posteriors can only be viewed at best as approximations to the result we would have obtained under some limiting prior distribution.

Another, perhaps more persuasive argument, is that by assuming all parameter sets have an equal likelihood of being chosen beforehand, then this can result in nonsensical resultant conclusions being drawn. Consider the following example:

We are given some data on a coin which has been flipped twice, with the result $\{H, H\}$. We are given the choice of deciding whether the coin is fair, with an equal chance of both heads and tails occurring, or biased with a very strong weighting towards heads. We denote fairness by a parameter $\theta = 1$, if the coin is fair, and $\theta = 0$ otherwise.

Figure 5.2 illustrates how assuming a uniform prior in this case results in

a very strong posterior weighting towards the coin being biased. This is because from a likelihood perspective - $P(data|\theta)$ - if we assume that the coin is biased, then the probability of obtaining two heads is high. Whereas if we assume that the coin is fair, then the probability of obtaining this data is only $\frac{1}{4}$. Thus, by ignoring common sense - that it is likely the majority of coins are relatively unbiased - we end up with a result that is nonsensical.

Of course, in this example we would hope that by collecting more data, in this case, throws of the coin, we could be confident in the conclusions drawn from the likelihood. However, Bayesian analysis allows us to achieve such a goal with a smaller sample size, should we be relatively confident about our pre-data knowledge.

5.6 The explicit subjectivity of priors

Opponents of Bayesian approaches to inference criticise the subjectivity inherent with choice of prior. However, all analysis involves a degree of subjectivity, particularly in regard to choice of statistical model. This choice is often formulated implicitly as being *objectively* correct, with little justification or discourse given to the underlying assumptions necessary to arrive there. The statement of a prior, necessary for any full description of a Bayesian analysis, is at least *explicit*; leaving this aspect of the modelling subject to the same interrogation and academic examination to which any analysis should be subjected. A word that is often used by protagonists of Bayesian methods, is that it is *honest* due to the *explicit* statement of assumptions. The statement of pre-experimental biases actually forces the analyst to self-examine, and perhaps also leads to a decline in the temptation to manipulate the analysis to one's own ends.

5.7 Interpreting priors through prior predictive distributions

5.8 Combining a prior and likelihood to form a posterior

This chapter thus far has given more attention to the philosophical and theoretical underpinnings of Bayesian analysis. Now we change tack to illustrate to the reader the mechanics behind Bayes' formula; specifically how the prior is combined with the likelihood to yield a posterior probability distribution. The following examples introduce an illustrative method, known as *Bayes' box* described in detail in [**?**] and [**?**], which illustrates the functioning of Bayes' rule, in which the parameter, prior, likelihood, and posterior are all displayed in a logical manner.

5.8.1 An urn of balls[3]

Imagine an urn of 5 balls, each of which is red or white, and suppose we are tasked with inferring the total number of red balls which are present in the urn, on the basis of a single ball which we pick out, and find to be red. Before we pull the ball out from the urn, we have no prejudice for a particular number of red balls, and so suppose that all possibilities - 0 to 5 - are equally likely, and hence have the probability of $\frac{1}{6}$ in our discrete prior. Our model for the likelihood is that a number Y of the balls are red, and that the result of an individual picking of a ball from the urn tells us nothing about future picks, apart from their joint dependence on Y. In this oversimplified example, this assumption of independence seems reasonable, particularly if the balls are picked out in a randomised manner and have no distinguishing features. Further suppose that the random variable $X \in \{0, 1\}$ indicates whether the ball is white or red respectively. The analogy with the disease status of an individual described in section 4.6.1 is evident, and hence we choose a likelihood of picking a red ball of the form:

$$P(X = 1|Y = \alpha) = \frac{\alpha}{5} \tag{5.4}$$

[3]Taken from Bolstad's great introduction to Bayesian statistics [**?**].

In (5.4), $\alpha \in \{0, 1, 2, 3, 4, 5\}$ represents the number of red balls in the urn.

We can then illustrate the functioning of Bayes' rule in the *Bayes' box* shown in table 5.1. We start by listing all the possible numbers of red balls that can exist in the Urn in the leftmost column. We then introduce our prior probabilities that we associate with each of the six potential numbers of red balls that can be in the urn. In the third column we then calculate the likelihoods for each of the outcomes using the simple rule given in (5.4). We then multiply the prior by the likelihood in the fourth column, which on summation gives us $P(data) = \frac{1}{2}$, which we use to create a proper probability distribution for the posterior in the last column. For a mathematical description of this process see section 5.12.1.

The Bayes' box illustrates the straightforward and mechanical working of Bayes' rule for the case of discrete data. We also note that when we sum the likelihood over all possible numbers of red balls in the urn - in this case the parameter which we are trying to infer - we find that this to be equal to 3; illustrating again that a likelihood is not a valid probability distribution. We also see that at a particular parameter value, if either the prior or the likelihood are found to be zero as is the case of 0 red balls being in the urn (impossible since we have at least one), then this ensures that the posterior distribution is zero at this point. This makes it important that we use a prior that gives a positive weight to *all* possible ranges of parameter values. The results are also displayed graphically in figure 5.3.

Table 5.1: A Bayes' box showing how to calculate the posterior for the case of drawing balls from an urn containing 5 red and white balls, one of which has been drawn and shown to be red. Here we assume that pre-experiment all possible numbers of red balls are equally likely, by adopting a uniform prior.

Number of red balls	**Prior**	**Likelihood**	**Prior x likelihood**	**Posterior**$= \frac{Prior \times Likelihood}{P(data)}$
0	1/6	0	0	0
1	1/6	1/5	1/30	1/15
2	1/6	2/5	1/15	2/15
3	1/6	3/5	1/10	3/15
4	1/6	4/5	2/15	4/15
5	1/6	1	1/6	5/15
Total	**1**	**3**	$P(data) = 1/2$	**1**

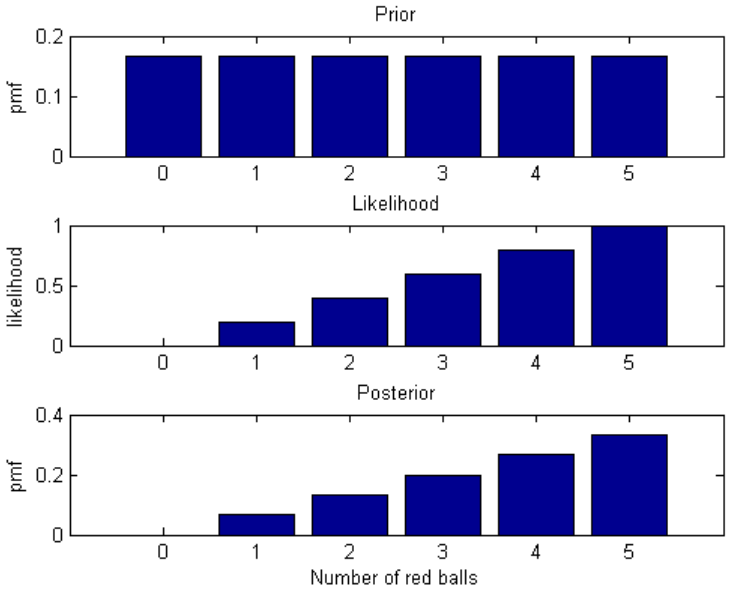

Figure 5.3: The prior, likelihood and posterior for the urn of balls example described in 5.8.1. The prior in the upper panel gives uniform weighting to all possible numbers of red balls. This is then multiplied by the likelihood (in the middle panel) at each number of balls, and normalised to make the posterior density shown in the bottom panel.

Now suppose that we had reason to believe that the urn-maker had a prejudice towards more equal numbers of both balls, and as a result we alter our prior to have a greater weight towards these numbers of red balls (see table 5.2 and figure 5.4).

Table 5.2: A Bayes' box showing how to calculate the posterior for the case of drawing balls from an urn containing 5 red and white balls, one of which has been drawn and shown to be red. Here a higher weighting is given to more equal numbers of red and white balls in the prior.

Number of red balls	**Prior**	**Likelihood**	**Prior x likelihood**	**Posterior=** $\frac{Prior \times Likelihood}{P(data)}$
0	1/12	0	0	0
1	1/6	1/5	1/30	1/15
2	1/4	2/5	1/10	1/5
3	1/4	3/5	3/20	3/10
4	1/6	4/5	2/15	4/15
5	1/12	1	1/12	1/6
Total	**1**	**3**	**1/2**	**1**

5.8.2 Disease proportions revisited

Suppose that we substitute our urn from section 5.8.1 for a sample of 100 individuals taken from the UK population. Suppose also that we continue to assert the independence of individuals within our sample, and make explicit the assumption that individuals are from the same population, and are hence identically-distributed. We are now interested in making conclusions about the overall proportion of individuals within the population who have the disease, θ. Since the parameter of interest is now continuous, we cannot use Bayes' box as there would be infinitely many rows (corresponding to the continuum of possible θ) over which to sum. Let's suppose that within our sample of 100 we find 3 of them who are disease-positive[4]. We could then use the assumptions of independence and identical-distribution to write down a likelihood of the form introduced in section 4.6.2:

$$P(Z = 3|\theta) = \binom{100}{3}\theta^{\beta}(1-\theta)^{100-3} \tag{5.5}$$

[4]We also suppose that there are no false-positives here.

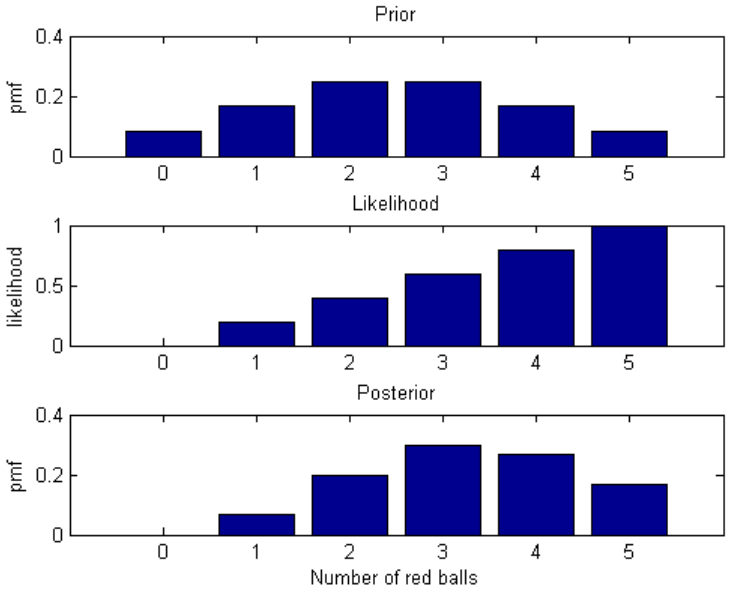

Figure 5.4: The prior, likelihood and posterior for the urn of balls example described in 5.8.1. The prior in the upper panel gives more weighting to more equal numbers of red and white balls. This is then multiplied by the likelihood (in the middle panel) at each number of balls, and normalised to make the posterior density shown in the bottom panel.

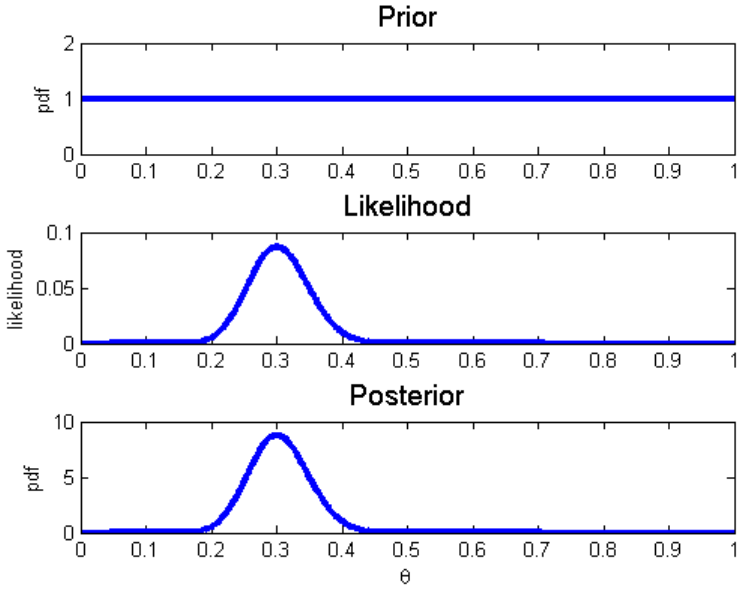

Figure 5.5: The prior, likelihood and posterior for the disease proportion example described in section 5.8.2. Each point in θ along the continuous prior curve (top panel) is multiplied by the corresponding value of likelihood (middle panel), to form the numerator of Bayes' rule. The numerator is then normalised to make the posterior probability density shown in the bottom panel.

The reason for the $\binom{100}{3} = 161,700$ term at the beginning of (5.5) is that we have to count the number of different permutations of getting 3 individuals who are disease-positive within a sample size of 100.

We suppose that at the beginning of the experiment all values of θ are equally likely. However, we would expect researchers to have a pre-experimental idea as to the most probable frequencies of the disease within the population, meaning that a flat prior which is given is likely understating a prejudice towards a certain range of θ values. Whilst, this is the case, it is often assumed in research papers - for the sake of objectivity - that priors are flat, in order to try to minimise the effect which assumptions here make on the outcome of an analysis.

5.9 Constructing priors

There are a number of different methodologies and philosophies when it comes to the construction of a prior density. In this section we consider briefly how priors can be engineered so as to be relatively uninformative - better-termed vague - or alternatively can be used to assemble pre-experimental knowledge in a logical manner.

5.9.1 Vague priors

When there is a premium placed on the objectivity of analysis, as is often the case in regulatory work - drug trials, public policy and the like - then use of a relatively 'uninformative' prior is often desired. If we were uncertain as to the proportion of individuals within a population who have a particular disease, then a uniform prior (see figure 5.6) is often employed to this end.

The use of a prior that has a constant value, $P(\theta) = constant$, is attractive because in this case:

$$\begin{aligned} P(\theta|data) &= \frac{P(\theta) \times P(data|\theta)}{P(data)} \\ &\propto P(\theta) \times P(data|\theta) \\ &\propto P(data|\theta) \end{aligned} \tag{5.6}$$

In (5.9.1) we thus see that the shape of the posterior distribution is solely determined by the likelihood function. This is seen as a merit of uniform priors since they 'let the data speak for itself' through the likelihood. This is used as the justification for using a flat prior in many analyses.

The flatness of the uniform prior distribution is often termed 'uninformative', but this is misleading. If we assume the same model as described in section 5.8.2, then the probability that one individual has the disease is θ, and the probability that two randomly sampled individuals both have the disease is θ^2. If we assume a flat prior for θ, then this implies a decreasing prior shown in figure 5.6 for θ^2. Furthermore, when we consider the probability that within a sample of ten individuals, all of whom are diseased, we see that a flat prior for θ implies an even more accentuated prior for this event; meaning that we beforehand give little weight to this event. For the

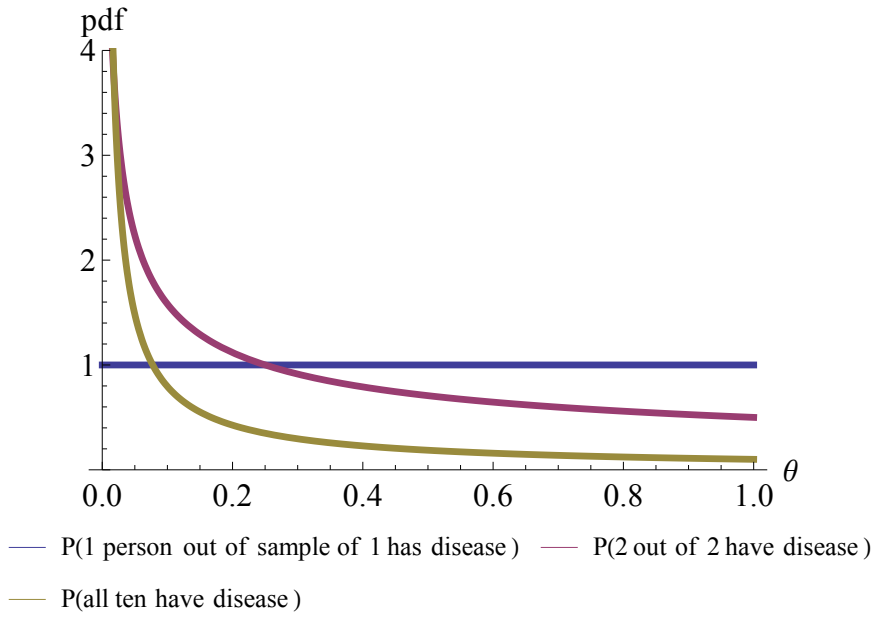

Figure 5.6: The probability density for obtaining all diseased individuals within sample sizes of 1, 2 and 10 respectively. Starting out with a flat prior for the probability that one individual has a disease has resulted in non-flat priors for the other 2 probabilities.

mathematical details of these graphs see section 5.12.2.

We can hence see that even though a uniform prior for an event looks, on first glances, to convey no information, we are actually making quite informative statements about other events. This aspect of choosing flat priors is swept under the carpet for most analyses, partly because often we care most about the particular parameter to which we create a prior. All priors contain some information, so we prefer the use of the terms "vague" or "diffuse" to represent situations where a premium is placed on drawing conclusions from only the data at hand.

There are methods for constructing priors that seek to limit the information contained within priors, so as to not colour the analysis with pre-experimental prejudices. However, we will leave a discussion of these methods until chapter 9 on *Objective Bayes*.

Whilst uniform priors are relatively straightforward to specify when we

aim to infer about a parameter which is bounded - such as in the previous example where $\theta \in \{0, 1\}$, or in the case of discrete parameters - we run into issues for parameters which have no predefined range. An example of this would be if we were aiming to determine the mean, μ, time of onset of lung cancer for individuals who develop the disease, after they begin to smoke. If we remove all background cases (assumed not to be caused by smoking), then μ has a lower bound of 0. However, there is no obvious point at which to draw an upper bound. A naive solution to this would be to use a prior for $\mu \sim Unif(0, \infty)$. This solution, although at first appears to be reasonable, is not viable for two reasons; one statistical, another which is practical. The statistical reason is that $\mu \sim Unif(0, \infty)$ is not a valid probability density, because any non-zero constant value for the pdf will mean that the area under the curve is ∞ because the μ axis stretches out forever. The common sense argument is that we would never ascribe the same likelihood to an individual having onset of lung cancer after 10 years as for it occurring after 250 years! The finiteness of human lifespan dictates that we select a more appropriate prior. If we were to ignore these two concerns although it is possible that the posterior could behave as a valid probability distribution[5], it would not actually be one (see section 5.5 for an explanation). A better choice of prior to use in this example would be one which ascribes zero probability to negative values of μ, and ever decreasing values of the pdf for high values of μ such as the one shown in figure 5.7. Alternatively, we could choose a uniform prior on a reasonable range of μ, and allow the pdf to be zero elsewhere (see figure 5.7).

5.9.2 Informative priors

We have seen in section 5.9.1 that priors are frequently chosen to give a strong voice to the recent data; minimising the impact of existing prejudices. There are however occasions when the choice of prior acknowledges that the analysis is based on more than the latest data. This choice of prior can be used to incorporate previous data, conclusions from older studies, or to include expert opinion.

In cases where data is available from previous studies, the construction of a prior can proceed methodically via a method that is known as *moment-matching*. Suppose that we have the data shown in figure 5.8 for SAT scores of past participants of a particular class. We might think that to

[5] Although not assured.

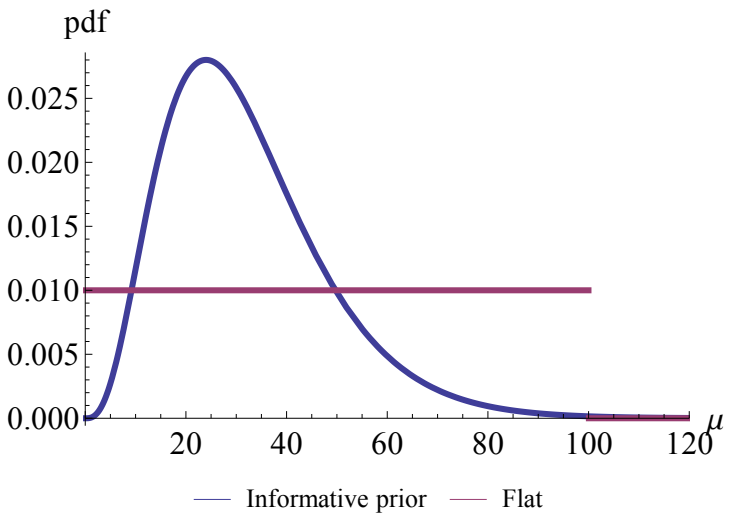

Figure 5.7: Two viable prior distributions for the average time taken before the onset of lung cancer after patients begin smoking.

a reasonable approximation the data could be modelled as having come from a normal distribution[6]. We typically characterise normal distributions via two parameters: its mean, μ, and variance, σ^2. In moment-matching a normal prior to this previous data, we choose the mean and variance to be equal to their sample equivalents, in this case $\mu = 1404$, and $\sigma^2 = 79,716$, respectively.

Whilst this simple methodology can result in priors that closely approximate pre-experimental datasets, note that it was a arbitrary choice to fit the first two moments of the sample. We could have used the skewness and kurtosis (measures related to the third and fourth centred moments respectively). Also, moment-matching is not Bayesian in nature, and can often be difficult to apply in practice. When we discuss hierarchical models in chapter 10, we will learn a more pure Bayesian method which can be used to create prior densities.

[6]A weakness of this model is that it allows for scores outside of the 600-2400 range of permissible SAT scores.

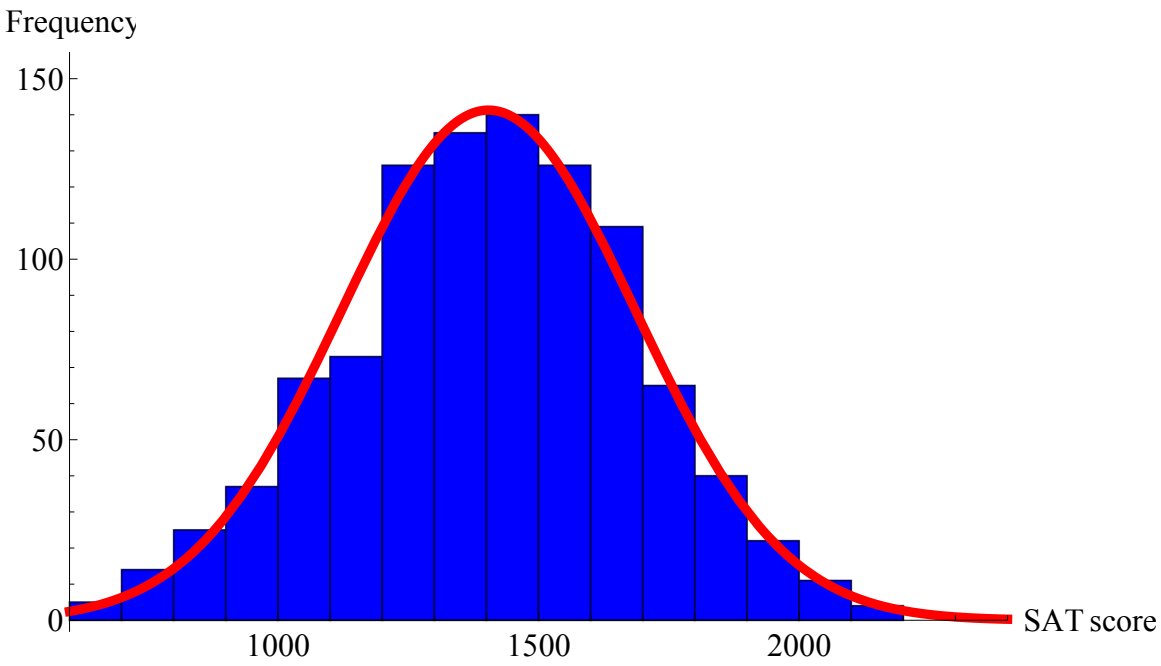

Figure 5.8: The SAT scores for past students of a class. The mean and variance of this hypothetical sample are 1404, and 79,716 respectively, which are used to fit a normal distribution to the data, and is shown in red.

5.9.3 The numerator of Bayes' rule determines the shape

We notice for both the examples described in sections 5.8.1 and 5.8.2 that the overall shape of the posterior distribution is determined by the prior, $P(\theta)$, multiplied by the likelihood, $P(data|\theta)$. This is the numerator of Bayes' rule:

$$P(\theta|data) = \frac{P(\theta) \times P(data|\theta)}{P(data)} \propto P(\theta) \times P(data|\theta) \tag{5.7}$$

The shape of the posterior is determined by how it varies with θ. Since the denominator is independent of θ, the numerator completely describes how the gradient and curvature of the posterior density varies with θ, which allows us to write the above $\propto P(\theta) \times P(data|\theta)$ statement. Viewed another way, the denominator is a nuisance normalisation factor which allows us to ensure that the posterior density when summed (discrete) or integrated (continuous) is equal to 1. We will return to a discussion of these concepts in depth in the chapter 6, but it doesn't hurt to see where we may be headed at present.

5.9.4 Eliciting priors

A different sort of informative prior is often required, which is not derived from prior data, but from expert opinions. In particular these priors are often required for evaluating clinical trials, and clinicians are interviewed before the trial is conducted. However, there is a raft of research in the social sciences which also make use of these methods for prior construction. Whilst there are a plethora of methods for creating priors from subjective views (see [**?**] for a detailed discussion), we go through a simplified example in order to explain a potential way in which these methods are used.

Suppose that we asked a range of economists to give their estimates of the 25th and 75th percentiles, $wage_{25}$ and $wage_{75}$, of the wage premium which one extra year of education spent at college commands on the job market on average. If we were to assume a normal prior for the data, then we can relate these two quantiles back to the corresponding values of a standardised normal distribution for each expert:

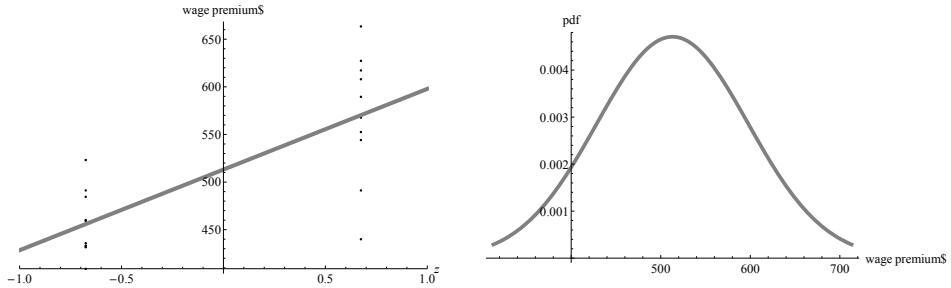

Figure 5.9: Hypothetical data for the 25th and 75th percentiles of the estimated wage premium from 10 experts. In the left hand panel we regress these percentiles on the corresponding percentiles from a standard normal distribution, yielding estimates of the mean and variance of a normal prior, which is shown on the right.

$$\begin{aligned} z_{25} &= \frac{wage_{25} - \mu}{\sigma} \\ z_{75} &= \frac{wage_{75} - \mu}{\sigma} \end{aligned} \tag{5.8}$$

In (5.8), z_{25} and z_{50} are the 25th and 75th percentiles of the standard normal distribution respectively. These two simultaneous equations can be solved for each expert, giving an estimate of the mean and variance of a normal variable. These could then be averaged to get estimates of the mean and variance across all the experts. However, a better method relies on linear regression. The expressions in (5.8) can be rearranged to the following:

$$\begin{aligned} wage_{25} &= \mu + \sigma z_{25} \\ wage_{75} &= \mu + \sigma z_{75} \end{aligned} \tag{5.9}$$

We now recognise that each equation is of the form of a straight line $y = mx + c$, where in this case $c = \mu$ and $m = \sigma$. If we then fit a linear regression line to the data from all the panel, we can then use the values of the y-intercept and gradient for μ and σ to estimate the mean and square root of the variance respectively (see figure 5.9).

5.10 A strong model is not heavily influenced by priors

Returning to the example of section 5.8.2 for estimated the prevalence of a disease within a population, we now examine the effects of using an informative prior on the analysis. Suppose we choose a prior which represents our pre-data view that the prevalence of a disease within a particular population is high (see the topmost row of figure 5.10). If we only have a sample size of 10, and obtain 1 individual in our sample who tests positive for the disease we see that the posterior is located roughly equidistant between the peaks of the prior and likelihood functions respectively (see the left hand column of figure 5.10). Now if we increase the sample size to 100, keeping the same percentage of individuals who are disease-positive within our sample, we then find that the posterior is peaked much closer to the position of the likelihood peak (see the middle column of figure 5.10). If we increase sample size further, maintaining the percentage of individuals with a disease in the sample, we see that the posterior peak's position appears indistinguishable from that of the likelihood (see the rightmost column of figure 5.10).

We can see from figure 5.10 that the effect of the prior on the posterior density decreases as we collect more data. Alternatively, we see that the likelihood - the effect of current data - increases as we have access to further data points. This makes intuitive sense, since when we collect more evidence that comes solely from the data we should lend this source more weight, and pay less attention to our pre-experimental prejudices.

In general, in Bayesian analysis, when we collect more data our conclusions become less influenced by priors. The use of a prior allows us to make inferences in small sample sizes by using pre-experimental knowledge of a situation, but in larger samples, and for more appropriate models, we should see the effect of choice of priors decline. We have an obligation to report when choice of priors heavily influences the conclusions that we draw from an analysis, and *sensitivity analysis* is a field which actually allows a range of priors to be specified, and combined into a single analysis. However, if we have sufficient data and a strong model, then we should see that the conclusions we draw are not heavily affected by choice of priors within a sensible range.

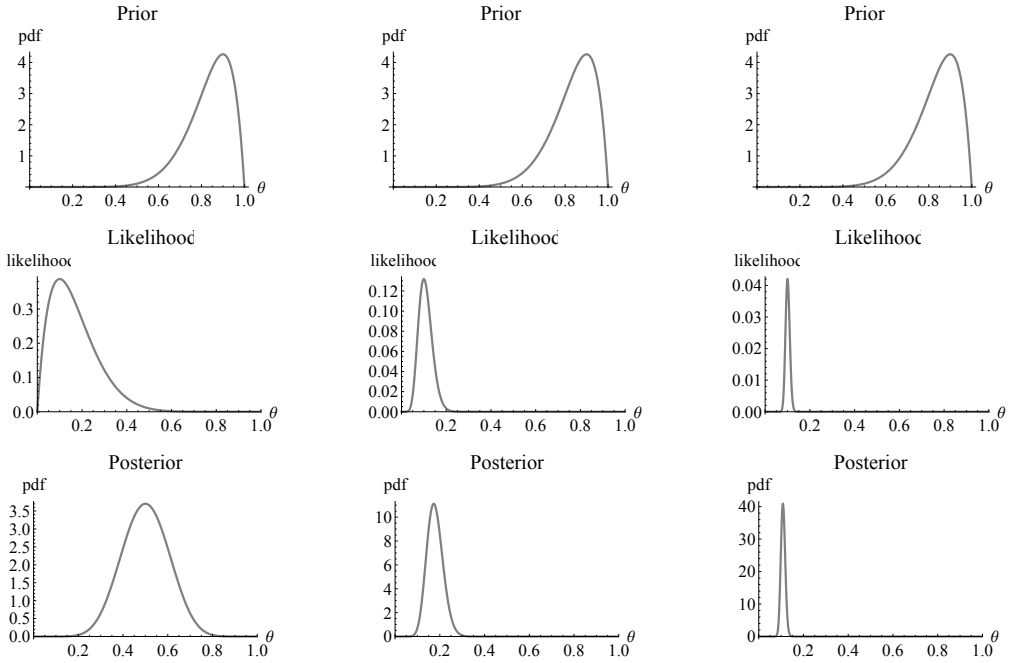

Figure 5.10: The effect of increasing sample size on the posterior density for the prevalence of a disease in a population. The leftmost column has N=10, the middle N=100, and the rightmost N=1,000. All three have the same proportion of disease cases in the sample.

5.11 Chapter summary

We now know that a *prior* is a probability distribution that represents our pre-experimental/-data knowledge about a particular situation. We also understand the importance of selecting a proper prior density, and the need to test and interpret a posterior carefully that results from using an improper prior. Further we understand that when an emphasis is placed on drawing conclusions solely from the data, that a vague prior may be most appropriate. This contrasts with situations in which we wish to use pre-experimental data or expert knowledge to help us to draw conclusions, in which case we may choose a more informative prior. In all cases however, we realise the need to be aware of the how sensitive our inferences are to choice of prior. We also realise that as the amount of data increases, or a better model is chosen, then the posterior density is less sensitive to choice of prior.

We are now nearly in a position to start doing Bayesian analysis, all that we have left to cover is the denominator of Bayes' rule. This aspect appears relatively benign on first glances, but is actually where the difficulty lies in Bayesian approaches to inference. Appropriately then we devote the next chapter to studying this final part of Bayes' rule.

5.12 Appendix

5.12.1 Bayes' rule for the urn

In this case the application of the discrete form Bayes' rule takes the following form:

$$\begin{aligned} P(Y=\alpha|X=1) &= \frac{P(X=1|Y=\alpha)\times P(Y=\alpha)}{P(X=1)} \\ &= \frac{P(X=1|Y=\alpha)\times P(Y=\alpha)}{\sum_{\alpha=0}^{5} P(X=1|Y=\alpha)\times P(Y=\alpha)} \\ &= \frac{\frac{\alpha}{5}\times\frac{1}{6}}{\sum_{\alpha=0}^{5}\frac{\alpha}{5}\times\frac{1}{6}} \end{aligned} \tag{5.10}$$

5.12.2 The probabilities of having a disease

We assume that the probability of an individual having a disease is θ, and we assume a uniform prior on this probability, $P(\theta) = 1$. We can calculate the probability that out of a sample of two, $P(Y) = P(\theta^2)$ by applying the change of variables rule:

$$P(Y) = P(\theta(Y)) \times |\theta'(Y)| \tag{5.11}$$

In (5.11), $\theta(Y) = Y^{-\frac{1}{2}}$ is the inverse of $Y = \theta^2$, and θ' means derivative wrt Y. Now substituting in this, we derive the probability density for two individuals having the disease:

$$P(Y) = \frac{1}{2\sqrt{Y}} \tag{5.12}$$

Chapter 6

The devil's in the denominator

6.1 Chapter mission

At the end of this chapter, the reader will understand what is represented by the denominator term, $P(data)$, in Bayes' rule. Furthermore, they will also have an appreciation of the inherent complexity of this term, as well as an idea of how modern computational methods can be used to bi-pass this.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)} \tag{6.1}$$

6.2 Chapter goals

Bayesian inference uses probability distributions, called *posteriors*, to make inferences about the world at large. However, to be able to use these powerful tools, we must ensure they are probability distributions. The denominator of Bayes' rule, $P(data)$, ensures that the posterior distribution is a *valid* probability distribution, by constraining the sum of its values to be 1.

$P(data)$ is a marginal probability density obtained by a sum across all parameter values of the numerator. The seeming simplicity of the previous statement belies the fact that for many circumstances its calculation can be complicated, and often practically intractable. In this chapter we will learn the circumstances when this difficultly arises, as well as a basic appreciation as to how modern computational methods sidestep this issue. We will leave the details of how these methods work in practice to part III, but this chapter will lay the foundations for this later study.

6.3 An introduction to the denominator

6.3.1 The denominator as a normalising factor

We know from chapter 4 that the likelihood is not a valid probability density, and hence we reason that the numerator of Bayes' rule - the likelihood multiplied by the prior - is similarly not constrained to be one either. The numerator will satisfy the first condition of a valid probability density: that its values are non-negative. However, the sum of the numerator across all parameter values will not generally be 1; meaning it fails the second test.

A natural way to normalise the numerator to ensure that the posterior is a valid probability density, is to divide by its sum; thus ensuring that its transformed variable's sum is always 1. The denominator of Bayes' rule, $P(data)$, is just this normalising factor. Notice that it does not contain the parameter, θ. This is because $P(data)$ is a *marginal* probability density (see section **??**), obtained by summing/integrating out all dependence on θ. This parameter-independence of the denominator ensures that the dependence of the posterior distribution $P(\theta|data)$ on θ is solely through the numerator (see sections 5.9.3 and 6.5).

There are two varieties of Bayes' rule which we will employ in this chapter, which use slightly different[1] formulations of the denominator. When θ is a discrete parameter we are required to *sum* over all possible parameter values, in order to obtain a factor which normalises the numerator:

$$P(data) = \sum_{All\ \theta} P(data|\theta) \times P(\theta) \tag{6.2}$$

[1] Although conceptually identical.

We will leave multiple-parameter inference largely to chapter 8, although will discuss how this leads to added complexity in section 6.4. However, the method proceeds in an analogous manner to (6.2), with the single sum replaced by a number of summations[2].

For continuous parameters we use the continuous analogue of the sum - an integral - resulting in a denominator of the form:

$$P(data) = \int_{All\ \theta} P(data|\theta) \times P(\theta) \mathrm{d}\theta \tag{6.3}$$

Similarly, for multiple-parameter systems the single integral is replaced by a multiple-integral. We will now demonstrate how to use (6.2) and (6.3) through two examples in sections 6.3.2 and 6.3.3 respectively.

6.3.2 Example: disease

Imagine that we are a medical practitioner tasked with evaluating the probability that a given patient has a particular disease. We use θ to represent the two possible outcomes:

$$\theta = \begin{cases} O & , Disease\ negative \\ 1 & , Disease\ positive \end{cases} \tag{6.4}$$

Using our experience and the patient's medical history we estimate that there is a probability of $\frac{1}{4}$ that this patient has the disorder; representing our prior. We then obtain test information, and are asked to re-evaluate the probability that the patient is disease-positive. In order to do this, we are required to state our likelihood. In this case we choose a likelihood of the form:

$$P(test\ positive|\theta) = \begin{cases} \frac{1}{10} & , \theta = 0 \\ \frac{4}{5} & , \theta = 1 \end{cases} \tag{6.5}$$

In (6.5), we implicitly assume that the probability of a negative test result is given by 1 minus the positive test probabilities. Also, by stating that

[2]The number of summations corresponds to the number of parameters in the model.

there is a non-zero probability for $P(positive|\theta = 0)$, we are assuming that false-positives do occur.

Suppose that the individual tests positive for the disease. We can now use (6.2) to calculate the denominator of Bayes' rule in this case:

$$\begin{aligned} P(test\ positive) &= \sum_{\theta=0}^{1} P(test\ positive|\theta) \times P(\theta) \\ &= P(test\ positive|\theta = 0) \times P(\theta = 0) + P(test\ positive|\theta = 1) \times P(\theta = 1) \\ &= \frac{1}{10} \times \frac{3}{4} + \frac{4}{5} \times \frac{1}{4} = \frac{11}{40} \end{aligned} \tag{6.6}$$

Furthermore, it turns out the denominator is also a valid probability density[3], meaning that we can calculate the counterfactual $P(test\ negative) = 1 - P(test\ positive) = \frac{29}{40}$. We need to be careful with interpreting this last result, since it didn't actually occur. It's best to think of $P(test\ negative)$ as the probability that we would assign to an individual testing negative before we carry out the test.

We can then use Bayes' rule to obtain the posterior probability that the individual has the disease, given that they tested positively:

$$\begin{aligned} P(\theta = 1|test\ positive) &= \frac{P(test\ positive|\theta = 1) \times P(\theta = 1)}{P(test\ positive)} \\ &= \frac{\frac{4}{5} \times \frac{1}{4}}{\frac{1}{10} \times \frac{3}{4} + \frac{4}{5} \times \frac{1}{4}} \\ &= \frac{8}{11} \end{aligned} \tag{6.7}$$

We see that in this case, even though we started off with a fairly optimistic prejudice - a probability that the individual has the disease of $\frac{1}{4}$ - the strength of the data has shone through, and we now are fairly confident of the alternative (see figure 6.1 for a graphical depiction of this change of heart). Bayesians are fickle by design!

[3]Due to the fact that we have removed the θ dependence that confounds attempts to view the numerator as one.

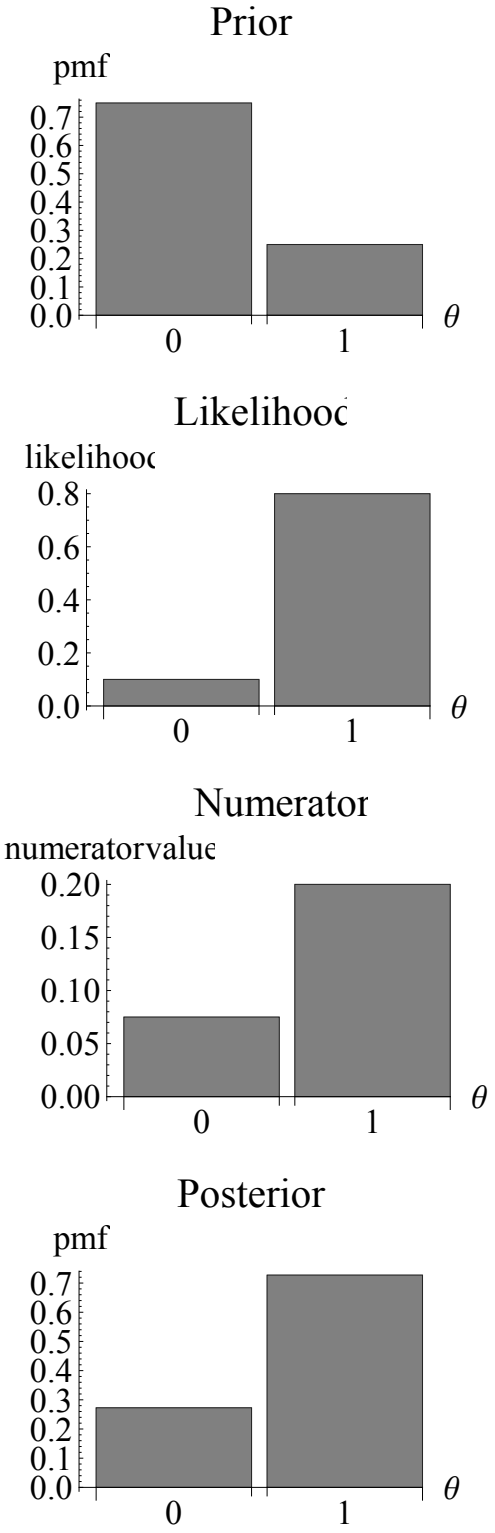

Figure 6.1: The prior is multiplied through by the likelihood, resulting in the numerator (the penultimate panel), which is then normalised by the sum over its values, to obtain the denominator.

6.3.3 Example: the proportion of people who vote for conservatively

We now are in the position of interpreting exit polls in a general election, and are tasked with inferring the proportion of voters, θ, that have voted for the conservative party. We suppose that conservatives are relatively unpopular at the time of the election, and hence we assume that, at most, 45% of the electorate will vote for them, meaning we choose a cut-off uniform prior of the form shown in figure 6.2[4]. For data we obtain voter preference data from 100 individuals leaving a particular polling station. To simplify the analysis, we will assume that there are only two political parties, and all voters must choose between either of these two options. We will assume that the polling station chosen is thought to be representative of the electorate as a whole, and voters' choices are independent of one another. In this situation we can use the results of section 4.6.2, and use a binomial likelihood function:

$$P(Z = \beta|\theta) = \binom{100}{\beta}\theta^{\beta}(1-\theta)^{100-\beta} \tag{6.8}$$

In (6.8), Z is a variable that represents the number of individuals who vote conservatively in the sample. $\beta \in [0, 100]$ is the value which corresponds to the number of conservative voters. We assume in this case that 40 people out of the sample of 100 voted conservatively resulting in the likelihood shown in figure 6.2, which is peaked at the Maximum Likelihood estimate of $\theta = 40\%$.

We then find the denominator by using (6.3), where $\theta \in \{0, 1\}$:

$$\begin{aligned} P(Z = 40) &= \int_0^1 P(Z = 40|\theta) \times P(\theta)\mathrm{d}\theta \\ &= \int_0^{0.45} \binom{100}{40}\theta^{40}(1-\theta)^{60} \times \frac{20}{9}\mathrm{d}\theta \\ &\approx 0.018 \end{aligned} \tag{6.9}$$

[4]This isn't really a reasonable prior in this case, since it is unrealistic to allow the probability density to jump from 0 at 46% to above 2 at 45%! However, we will stick with it to demonstrate its effect on inference.

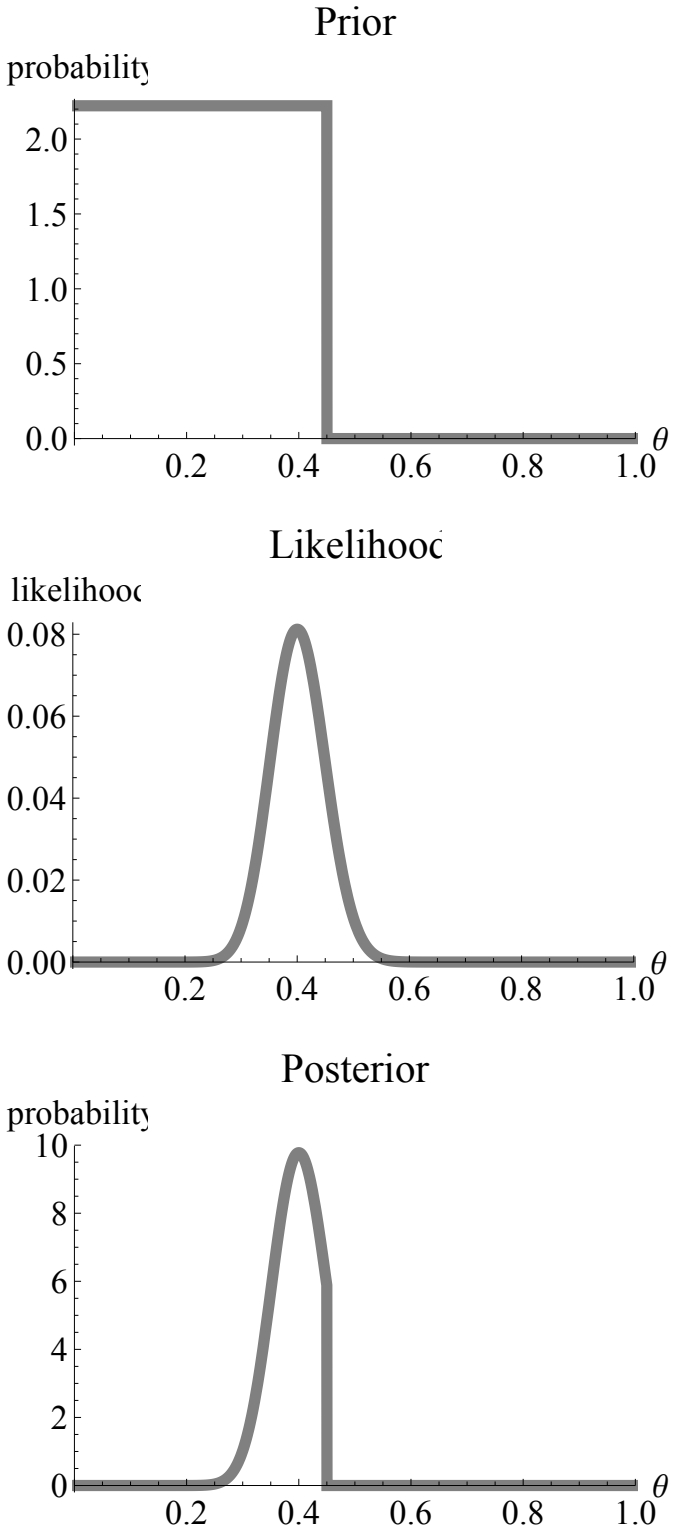

Figure 6.2: The prior, likelihood and posterior for the proportion of individuals voting for the conservative party in a general election, where we have found 40 people out of a sample of 100 voted conservative.

In (6.9), we have used the fact that since $P(\theta) = 0$ for $\theta > 0.45$, we can restrict the integral to only the region below that value. The value $\theta \approx 0.018$ has come by numerically integrating the second line.

Now that we have the value of the denominator, we can use it to normalise the product of the prior and the likelihood, resulting in the posterior distribution seen in figure 6.2. Notice the effect of truncating the uniform distribution at $\theta = 0.45$ is to truncate the posterior distribution at this value; resulting in a discontinuous jump in the posterior, which could be seen as an undesirable consequence of this prior.

6.3.4 The denominator as a probability

Another way to view the denominator is as the *probability of the data given choice of model*. Where *model* here encompasses both the likelihood and the prior. It is actually a *marginal* probability density that is obtained by summing/integrating out the dependence on the parameter(s) of the joint density $P(data, \theta)$:

$$\begin{aligned} p(data) &= \int_{All\ \theta} p(data|\theta) \times p(\theta) \mathrm{d}\theta \\ &= \int_{All\ \theta} p(data, \theta) \mathrm{d}\theta \end{aligned} \tag{6.10}$$

In (6.10) we have assumed that the parameter(s) is/are continuous. We have obtained the second line of (6.10) from the first by using the conditional probability formula introduced in section 2.7:

$$p(data|\theta) = \frac{p(data, \theta)}{p(\theta)} \tag{6.11}$$

We are thus able to characterise the joint density of the data and θ in Bayesian statistics. We can draw the joint density for each of the examples in sections 6.3.2 and 6.3.2 respectively, by taking the product of the likelihood and prior. In the disease example of section 6.3.2 this results in the discrete joint density shown in table 6.1, with a graphical depiction of the density shown

[4] The factor $\frac{20}{9}$ is from the uniform density for $\theta \leq 0.45$.

in figure 6.3. In the continuous case we obtain a joint probability density with a landscape of the form shown in figure 6.4.

		Disease status		
	Test Results	**Negative**	**Positive**	
Likelihood	0	0.90	0.20	
	1	0.10	0.80	
		×	×	
Prior		0.75	0.25	
		=	**=**	
	Test Results			$p(data)$
Joint density	0	0.675	0.05	**0.725**
	1	0.075	0.20	**0.275**

Table 6.1: Shows the derivation of the joint density for the disease example described in section 6.3.2. Each column of the likelihood - corresponding to a given disease status - is multiplied by the corresponding prior, resulting in the joint density. By summing the joint density across the different disease statuses of the patient, this results in $p(data)$. **Add pluses and equals to the calculation of $p(data)$. Also add in the posterior calculation.** See figure 6.3 for a graphical depiction of this joint density.

6.3.5 Using the denominator to choose between competing models

The denominator represents the accumulation of evidence for our particular model, with the result being a trade-off between our the data and our pre-experimental pre-conceptions. It represents the *average* fit of our model to the data across all parameter values. To see this note that the denominator is actually the expected value of the likelihood - the fit - given choice of prior[5]:

[5]This comes from the mathematical definition of the expected value of a quantity.

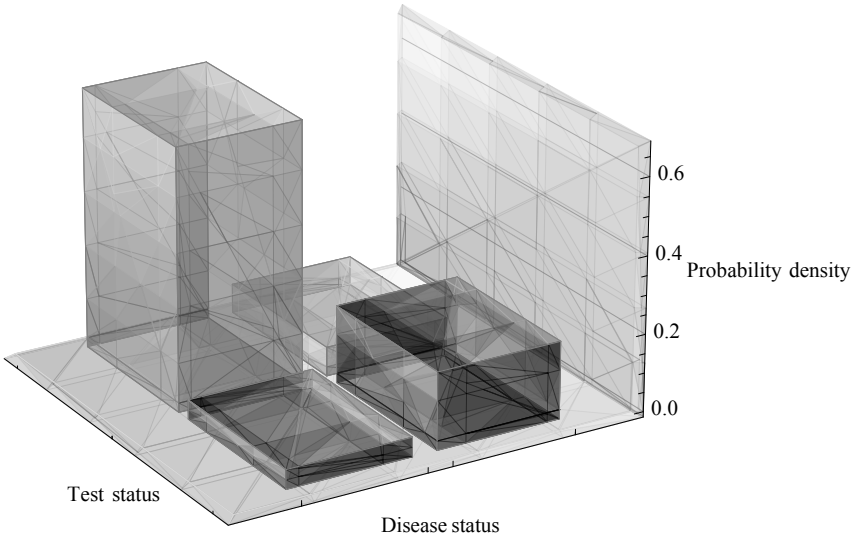

Figure 6.3: The joint density of the data and the parameter for the disease example described in section 6.3.2. When we uncover that the test result is positive, we are confined to look at the bars in dark grey; finding that the probability that an individual is diseased is significantly higher than the alternative (see the bottom panel of figure 6.1). **Perhaps redo this figure with a contour plot opposed to a 3D graph, and show how the posterior is obtained in another panel. Or just get rid of it, the table does pretty much cover it.**

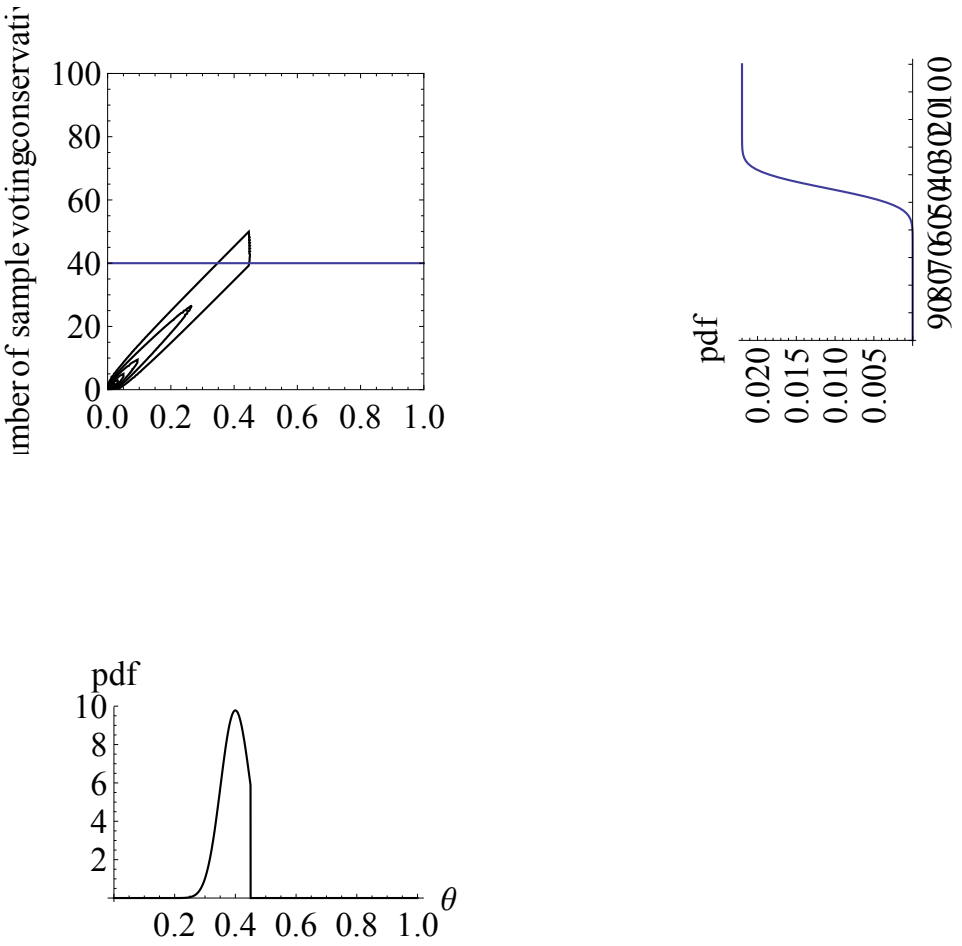

Figure 6.4: Top-left: a contour plot of the joint density of the voting example described in section 6.4. Top-right: the marginal density of $p(data)$ obtained by summing across all values of θ. Bottom-left: the posterior obtained by summing the joint density across the line shown at 40. Note that in reality the data variable is discrete, but I have drawn it here as continuous to make the plot simpler to interpret. **The line at 40 may be dashed in the final version. The axes all need to be aligned.**

$$p(data) = \mathbb{E}_\theta \left[p(data|\theta) \right] = \int_{All\ \theta} p(data|\theta) \times p(\theta) \mathrm{d}\theta \tag{6.12}$$

In (6.12) we have assumed that the parameter(s) are continuous, necessitating an integral rather than a discrete sum.

Since it really represents the evidence for our model, we can use it to compare two competing models. We could simply calculate the ratio of $\frac{p(data|model_1)}{p(data|model_2)}$, and use this as a guide to choose between models, but we would ideally like to do model selection in a more complete Bayesian manner. What we really care about for model choice is $p(model|data)$, rather than what we currently have $p(data|model)$. To obtain this we can use Bayes' rule, but now conditioning on choice of *model* rather than *parameter*:

$$p(model|data) = \frac{p(data|model) \times p(model)}{p(data)} \tag{6.13}$$

In (6.13), the denominator is *not* the same as that which we see in our previous applications of Bayes' rule, and represents the probability of obtaining the data across *all* models. Notice also that we also have introduced $p(model)$ which represents our prior faith in this particular model. We can now use (6.13) to choose between two models by calculating the ratio:

$$\frac{p(model_1|data)}{p(model_2|data)} = \frac{p(data|model_1)}{p(data|model_2)} \times \frac{p(model_1)}{p(model_2)} \tag{6.14}$$

If we have no prior leaning towards either of the two models then it seems reasonable to set $p(model_1) = p(model_2)$, and we are reduced to our previously proposed way of choosing between models. In fact the first ratio on the RHS of (6.14) is sufficiently used to merit its own name, the *Bayes' factor* is:

$$Bayes\ factor(model_1, model_2) = \frac{p(data|model_1)}{p(data|model_2)} \tag{6.15}$$

We shall come to discuss the usefulness of the Bayes factor in chapter 12 for choosing between models, as well as comparing hypotheses.

In both the examples discussed in sections 6.3.2 and 6.3.3, we found the denominator as a means to obtaining the posterior distribution through

Bayes' rule. However, as an ends in itself it is less useful, unless it is calculated across a number of models/hypotheses and then used to choose amongst them.

6.3.6 The denominator for improper priors

The difficulty calculating P(data) with an improper prior. Go through and correct P to p for probability.

6.4 The difficulty with the denominator

We have come to realise that the denominator of Bayes' rule is obtained by summing/integrating the joint density $p(data, \theta)$, where the latter is obtained by the product of the prior and the likelihood. The examples in section 6.3.4 indicate how this procedure works when there is a single parameter in the model. However, in most real-life applications of statistics, the likelihood is a function of a number of parameters. For the case of a two parameter discrete model, the denominator is given by a double sum:

$$p(data) = \sum_{All\ \theta_1} \sum_{All\ \theta_2} p(data, \theta_1, \theta_2) \tag{6.16}$$

And for a two-dimensional continuous parameter vector, we are now required to do a double integral:

$$p(data) = \int_{All\ \theta_1} \int_{All\ \theta_2} p(data, \theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2 \tag{6.17}$$

Whilst the two-parameter forms (6.16) and (6.17) may not look more intrinsically difficult than their single parameter counterparts, (6.2) and (6.3) respectively, this aesthetic similarity is misleading, particularly in the continuous case. Whilst in the discrete case, it is possible to enumerate all parameter values, and hence - by brute force - calculate the exact value of $p(data)$, for continuous parameters, the integral may be difficult to undertake. This difficulty is amplified the more parameters we include within the

model, rendering the analytic[6] calculation of the denominator practically impossible, for all but the simplest models.

6.4.1 Multi-parameter discrete model example: the comorbidity between depression and anxiety

In medicine comorbidity refers to the concurrence of two or more conditions. An example of this is the frequent coincidence of depression and anxiety in a patient. Let $D \in \{0, 1\}$ and $A \in \{0, 1\}$ be random variables representing the depression and anxiety statuses of a particular patient respectively. Now that we have two parameters, we must specify a joint prior distribution. An example prior is shown at of the top of table 6.2, in which we have also calculated the marginal prior distributions by summing over all values of the other variable. We suppose that *a priori* the clinician undertaking this case believes that the patient is unlikely to meet all the criteria necessary for them to be defined as having both disorders, which is reflected in a prior probability of $p(D = 1, A = 1) = 0.6$.

We can also use this joint distribution to calculate prior conditional probabilities. For example, we can calculate the probability that an individual has anxiety, *given* that they have depression:

$$\begin{aligned} p(A = 1|D = 1) &= \frac{p(A = 1, D = 1)}{p(D = 1)} \\ &= \frac{0.2}{0.35} \\ &= \frac{4}{7} \approx 0.57 \end{aligned} \tag{6.18}$$

This shows that it is considerably more likely that a patient has anxiety, if they are already depressed (compared with the unconditional $p(A = 1) = 0.25$), indicating our prior beliefs regarding the comorbidity of these two conditions.

We assume that the patient takes a personality diagnostic test which provides some extra information regarding whether the individual has either of these conditions. Let's assume for simplicity that the result of the test, $X \in \{0, 1\}$, has the likelihood shown in the second panel of table

[6] This just means to write down a relation for the denominator in closed form.

6.2. The maximum likelihood estimator would be that the individual has $(D = 1, A = 1)$, with the lowest likelihood going to the disorder-free case.

Prior			**A**		
			0	**1**	$p(D)$
	D	**0**	0.6	0.05	**0.65**
		1	0.15	0.2	**0.35**
		$p(A)$	**0.75**	**0.25**	
Likelihood (X=1)			**A**		
			0	**1**	
	D	**0**	0.05	0.4	
		1	0.4	0.8	
Numerator = Prior x Likelihood			**A**		
			0	**1**	
	D	**0**	0.03	0.02	
		1	0.06	0.16	
p(X=1)=0.03+0.03+0.06+0.16=0.27					
Posterior			**A**		
			0	**1**	$p(D\|X = 1)$
	D	**0**	0.11	0.07	**0.19**
		1	0.22	0.59	**0.81**
		$p(A\|X = 1)$	**0.33**	**0.67**	

Table 6.2

We would now like to calculate the joint posterior probability of the two conditions, given that an individual tests positive $(X = 1)$. We can write this using Bayes' rule, although now we must now make sure to condition the likelihood on both parameters. However, we can denote the parameter vector, $\boldsymbol{\theta} = (D, A)$, and apply Bayes' rule just as before:

$$\begin{aligned} p(\boldsymbol{\theta}|X = 1) &= \frac{p(X = 1|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(X = 1)} \\ &= \frac{p(X = 1|A, D) \times p(A, D)}{p(X = 1)} \end{aligned} \tag{6.19}$$

In (6.19), we have simply substituted the definition of, $\boldsymbol{\theta} = (D, A)$, into the top line to get the final expression. Therefore, just like before we multiply

the likelihood by the prior to obtain the numerator of Bayes' rule. We finally sum over all numerator values, and use this to obtain the posterior distribution (see table 6.2). In table 6.2 we have also calculated the marginal conditional posterior probabilities by using the law of conditional probability, and we find an 81 % probability that the individual has depression, and 67 % chance that they have anxiety. The probability that they have both disorders is 59%.

6.4.2 Continuous multi-parameter example: mean and variance of IQ

We now consider a situation where the parameters of interest are continuous. It is hoped that this section will provide evidence for the complexity of analytic multi-parameter inference in Bayesian statistics, and hence by its very nature, the material covered here may be difficult to fully grasp. However, we will cover it in more detail in part II.

We suppose we are interested in estimating the mean IQ of some population of interest, of which we only possess a sample of three persons' IQ data of $\boldsymbol{IQ} = \{100, 50, 150\}$. We suppose that since intelligence - as measured by IQ - is dependent on many additive factors, and hence as an approximation we assume a normal likelihood[7]:

$$p(IQ_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(IQ_i - \mu)^2}{2\sigma^2}\right) \tag{6.20}$$

For sake of simplicity, we will assume that IQ is measured on a fixed scale, $IQ \in [0, 300]$. We also assume that prior independence between μ and σ^2, which means that we can calculate the joint prior by multiplying together the individual probabilities:

$$p(\mu, \sigma^2) = p(\mu) \times p(\sigma^2) \tag{6.21}$$

Since $\sigma^2 \geq 0$, we might be tempted to specify a prior distribution for $\sigma^2 \sim Unif(0, \infty)$. However, this does not appear sensible because this would assign the same probability to an infinite variance, which is not possible on

[7] We have used the central limit theorem here - see section 2.10 for a full explanation.

finite-scaled data. A frequently-used alternative is to specify a prior as uniform in $log(\sigma^2)$ space. This serves two purposes, firstly, because the inverse of a log (the exponent) is always non-negative for real inputs, this ensures that this condition is satisfied by σ^2. Secondly, and most importantly, when we transform a uniform prior on $log(\sigma^2)$ back to σ^2 space, we find that the prior density is equivalent to[8]:

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{6.22}$$

This results in a joint prior for (μ, σ^2) shown in figure 6.5. We importantly note that this prior is improper, since $\int_0^\infty \frac{1}{\sigma^2} d\sigma^2 \rightarrow \infty$, and hence must take case when interpreting the resultant 'posterior' distribution (see section 5.9.1).

We imagine we only observe a sample of one individual, from which we would like to find the posterior distribution of the joint distribution of $(\mu, \sigma^2) = \boldsymbol{\theta}$. This is found by application of Bayes' rule:

$$\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{IQ}) &= \frac{p(\boldsymbol{IQ}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(\boldsymbol{IQ})} \\
&= \frac{p(\boldsymbol{IQ}|\mu, \sigma^2) \times p(\mu, \sigma^2)}{p(\boldsymbol{IQ})} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\sum_{i=1}^{3} (IQ_i - \mu)^2}{2\sigma^2} \right) \times \frac{1}{\sigma^2}}{\int_0^{300} \int_0^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\sum_{i=1}^{3} (IQ_i - \mu)^2}{2\sigma^2} \right) \times \frac{1}{\sigma^2} d\sigma^2 d\mu} \\
&\propto \sigma^{-3} exp\left(-\frac{\sum_{i=1}^{3} (IQ_i - \mu)^2}{2\sigma^2} \right)
\end{aligned} \tag{6.23}$$

In (6.23), the second line was obtained from the first by simply substituting

[8] See the chapter appendix for a full mathematical treatment of this result.

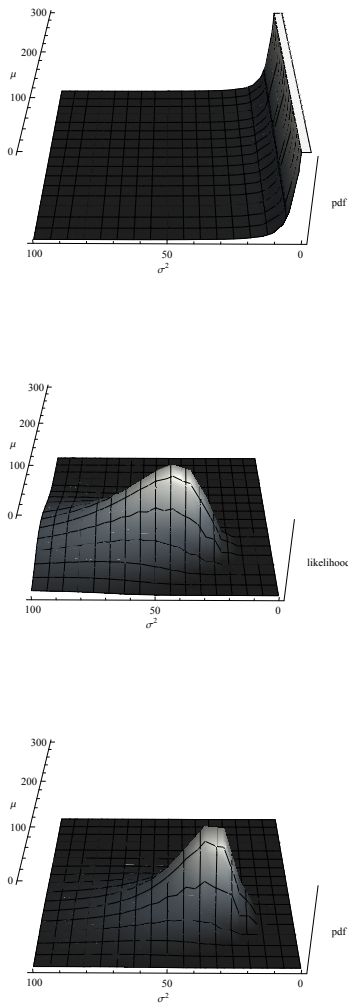

Figure 6.5: The prior, likelihood, and posterior distributions for the mean and variance of IQ example described in section 6.4.2.

in for $\boldsymbol{\theta} = (\mu, \sigma^2)$. We then substituted for the likelihood[9] and prior from (6.20) and (6.22) respectively.

We can choose, in this rather simplified example, to go through and actually evaluate the posterior exactly, by calculating the denominator by brute force. This results in a posterior density shown in figures 6.5 and **??**. We can then obtain the marginal densities by integrating out any dependence of the parameter not in interest (see figure **??**):

$$p(\mu|\boldsymbol{IQ}) = \int_0^\infty p(\mu, \sigma^2|\boldsymbol{IQ})\mathrm{d}\sigma^2 \tag{6.24}$$

$$p(\sigma^2|\boldsymbol{IQ}) = \int_0^{300} p(\mu, \sigma^2|\boldsymbol{IQ})\mathrm{d}\mu \tag{6.25}$$

$$\tag{6.26}$$

Although, here we could go through and analytically derive the posteriors[10], by evaluating the denominator, it is hoped that this example gives a little insight into the complexity of calculating the denominator in Bayesian models. The degree of difficulty of calculating the denominator increases rapidly in the number of unknown parameters within a model. In fact, at some point, the denominator becomes practically infeasible to calculate for models more complicated than only a few parameters.

However, all is not lost, as we discuss in section 6.5.

6.5 How to dispense with the difficulty: Bayesian computation

The Herculean task of calculating the denominator for continuous parameters would seem to put a real spanner in the works for Bayesian statistics,

[9] We have assumed independence for the data, meaning that to get the overall likelihood, we multiply together the three individual likelihoods.

[10] Although we have chosen to omit the exact closed-form results here for brevity. Postponing such a full derivation until part II.

such its reliance on the denominator of Bayes' rule. However, all is not lost. There are two solutions to the difficulty:

- Use priors conjugate to the likelihood (See chapter 8).
- Abandon analyticity, and opt to sample from the posterior instead.

The first of these workarounds still allows for exact derivation of an expression for the posterior distribution, by choosing a mathematically *nice* form for the prior distribution. This simplifies the analysis, since one can simply look up formulae for the posterior which have already been tabulated for us, avoiding to have to do any maths at all. However, frequently in real life applications of Bayesian statistics, we need to stray outside this realm of mathematical convenience. The price for a more varied choice of priors and likelihoods is that we have to give up our aspirations for closed-form calculation of the posterior density. However, it turns out in these circumstances we can still *exactly* sample from the posterior, and then use sample summary statistics to describe the posterior distribution in a very adequate way. We will leave a full description of these computational methods to part III, but to provide a clue as to where we may be heading, we note that the posterior density can be written:

$$\begin{aligned} p(\theta|data) &= \frac{p(data|\theta) \times p(\theta)}{p(data)} \\ &\propto p(data|\theta) \times p(\theta) \end{aligned} \tag{6.27}$$

In (6.27) we have arrived at the second line due to $p(data)$ being independent of θ; it is essentially a constant that we use to normalise the posterior. The numerator of Bayes' rule tells us everything that we need to know about the *shape*[11] of the posterior distribution, whereas the denominator merely tells us about its *height*. Computational methods use the shape of the posterior distribution to generate samples from it based on local comparison of relative probabilities.

This provides a little insight into the methodology of modern Bayesian methods, although we will cover this in more depth in Part III.

[11] It's dependence on θ.

6.6 Chapter summary

This chapter should have introduced you to the different interpretations of the denominator of Bayes' rule; firstly as a nuisance normalising factor; secondly as a probability of the data. We then discussed the difficulty that the denominator poses for Bayesians. Fortunately, this problem can be sidestepped via use of conjugate priors (see chapter 8), although this is often extremely limiting for all but the most simple of data processes. We then reasoned that since the curvature of the posterior is solely determined by the numerator of Bayes' rule - the prior multiplied by the likelihood - we can learn much of the posterior from this quantity. In particular modern computational methods use the numerator of Bayes' rule, to generate samples from the posterior distribution. These can then be used to summarise the posterior distribution, as well as for any of the uses of posterior distributions described in chapter 3. Before we introduce computational methods in full in part III, we firstly must understand the multitude of distributions which we have in our arsenal. We then discuss how conjugate priors can be used to analytically derive posteriors, which can be useful before moving to approximate computational methods.

6.7 Appendix

Part II

Analytic Bayesian methods

Chapter 7

An introduction to distributions for the mathematically-un-inclined

7.1 Chapter mission statement

7.2 Chapter goals

7.3 Sampling distributions for likelihoods

Bernoulli

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Discrete data.
2. A single trial.
3. *Only* two trial outcomes: *success* and failure[1]

[1]These do not need to literally represent successes and failures, but this shorthand is typically used, and we adopt it here.

Example uses: outcome of flipping a coin, a single clinical trial, or a presidential election!

Imagine you are interested in the outcome of a single horse race. For added simplicity, we suppose that we only care whether the horse wins, or loses; forgetting about its position if it doesn't win. In this framework, we formalise the outcomes of the race by creating a mathematical object - a *random variable* - which associates a numerical value with each of the outcomes, X:

$$X = \begin{cases} O & , \textit{horse wins} \\ 1 & , \textit{horse loses} \end{cases} \tag{7.1}$$

In this set-up it makes sense to model the outcome of a single race as being influenced by a background probability of success, $0 \leq \theta \leq 1$. We don't actually witness this probability, and after the discussion in chapter 2, we aren't sure it *really* exists. In the Frequentist paradigm tθ represents the proportion of races[2] that the horse has won historically[3]. Whereas to Bayesians, θ merely gauges our subjective confidence in the event of the horse winning[4].

Now that we have a model, we can work out the probability of the two distinct possible outcomes. The probability that the horse wins is straightforwardly $p(win) = \theta$; meaning the probability of a loss *must*[5] be given by $p(loss) = 1 - \theta$.

We now are in a position to work out the likelihood of a given outcome. Remember that a likelihood is *not* a valid probability density, and is found by holding the *outcome*, or in other words *data*, constant, whilst varying the parameters. Suppose that the horse has had a good meal of carrots this morning, and goes on to win by a country mile. We know that the likelihood of this event is given by $l(\theta|X = 1) = p(X = 1|\theta) = \theta$ (see the red line in figure 7.1).

[2]Really identical races, but who's counting?

[3]And will win in total if we were to continue re-racing forever.

[4]If any of this is news to you, then have a re-read of section 2.4.

[5]It *must* be given by $1 - \theta$, so that the Bernoulli distribution sums to 1, and hence is a valid probability density.

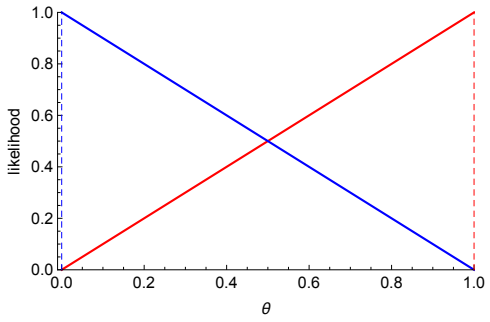

Figure 7.1: Bernoulli likelihoods for the event that the horse wins (red), and loses (blue). The maximum likelihood estimates are shown as dotted lines for each of the cases.

Alternatively, if our horse spent the night in the midst of night-mares[6], and went on to lose, then the likelihood is given by $l(\theta|X = 1) = 1 - \theta$ (see the red line in figure 7.1).

As you can see, in this situation, the maximum likelihood estimates in each case are given by $\hat{\theta} = 1$, and $\hat{\theta} = 0$, if the horse wins or loses respectively.

We can actually write down a single expression that yields the likelihood/probabilities[7] for both potentialities:

$$p(X|\theta) = \theta^X(1-\theta)^{1-X} \tag{7.2}$$

This distribution is known as the *Bernoulli*, after the Swiss mathematician Jacob Bernoulli who first probed this much-discussed distribution.

It is rare in life to come up against one-off events, and we typically have data for a number of races historically. However, this discussion has given us the necessary grounding to make the step to the next distribution in up the rung, the *binomial*, which is perfect for this more realistic setting.

Video : Need to make an intuitive video here. One exists but is overly mathematical.

[6]Get it? Sorry, that is terrible.

[7]Dependent on whether we vary the parameter or data respectively.

Binomial

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Discrete data.
2. Multiple trials.
3. Each trial has only two outcomes (see the Bernoulli): we for lack of better words, deem these *successes* and *failures*.
4. Individual outcome probabilities are not determined by any other factor that varies across individuals for which we have data.
5. Probability of success is same in each trial.
6. Overall data we measure is the aggregate number of successes.
7. Trials are independent[8].

Example uses: clinical drug trials, Democrat voters exiting a poll station, or the number of mosquito bites over the course of a day.

Now we jump immediately to a much more practically-relevant case where we have data on the outcome of a number of horse races, Republican votes, or in this case, we shall imagine we are examining the outcome of a clinical trial of a new flu drug. We have a sample of 10 willing students, who have chosen to take the pain of a sweaty week in pyjamas, for the betterment of humanity. At the week's start, the students are infected with the virus via an injection. At the end of the week, the consulting physician records the number of volunteers that still show flu symptoms, and those that do not. In order to build a model, we need some assumptions. Firstly, we assume that the students' data are exchangeable[9]. This might be violated if we knew that some of the volunteers have asthma, and are likely not to be so rapidly cured by the wonder drug. We also create a random variable

[8]Knowing the outcome of a single trial would not influence our opinion as to the likelihood of another trial's result, apart from any information it provides on θ. To be absolutely correct we should really say that the data are *conditionally*-independent.

[9]In this example, this really means that the students constitute a *random sample*; meaning the data are independent and identically-distributed.

X which represents the outcome of the trial for a single volunteer, which takes the value 1 if the trial is successful, in other words the volunteer is no longer symptomatic, and 0 if not.

However, we have data on the success of the trials for all 10 of our volunteers. Here, it makes sense to create another helper random variable, $0 \leq Z \leq 10$, which measures the aggregate outcome of the trial:

$$Z = \sum_{i=1}^{10} X_i \tag{7.3}$$

Carrying out the trial we find that 5 volunteers successfully recovered from the virus of the week. We reason that the following outcomes *could* have lead to this aggregate result: $\boldsymbol{X} = \{1, 1, 1, 1, 1, 0, 0, 0, 0, 0\}$; meaning the arbitrarily-chosen first 5 volunteers happened to react well to the treatment. Feeling satisfied we present our results to the pharma company executives who developed the drug in the first place. They look slightly perturbed, and say that $\boldsymbol{X} = \{1, 0, 1, 0, 1, 0, 1, 0, 1, 0\}$ would also have been possible. Realising our mistake, we counter with the argument that it is also possible that the latter 5 volunteers were the ones who recovered. This back-and-forth continues, until you realise that there is a unifying mathematical formula that can cover any situation - the binomial nCr formula[10]. You reason that there are possibly $\binom{10}{5} = 252$ possible overall combinations of results with the same aggregate outcome!

With this realisation you make the next step; writing down the likelihood. Since we have supposed that the observations are *independent*, we are able to calculate the overall probability by multiplying together the individual probabilities, taking into account the 252 possible combinations:

$$\begin{aligned} p(Z = 5|\theta) &= 252 \times p(X_1 = 1|\theta)...p(X_5 = 1|\theta)p(X_6 = 0|\theta)...p(X_10 = 0|\theta) \\ &= 252 \times \theta^5(1-\theta)^5 \end{aligned} \tag{7.4}$$

We can graph this likelihood as a function of θ (see the blue line of figure 7.2), and note that the maximum likelihood estimator of the parameter occurs at the intuitive result of $\theta = \frac{1}{2}$.

[10] See section 4.6.2 for a more complete discussion.

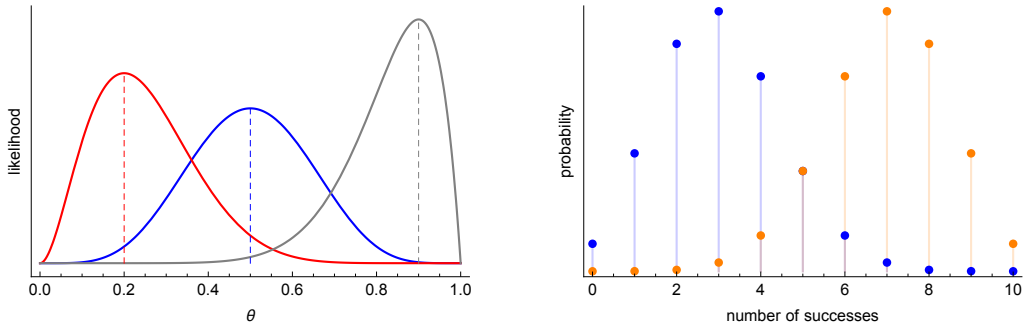

Figure 7.2: **Left**: Binomial likelihoods for the event that 5 (blue), 2 (red), and 9 (grey) volunteers recovered in the week period. The maximum likelihood estimates are shown as dotted lines for each of the cases. **Right:** the probability distribution of successful trials if $\theta = 0.3$ (blue) and $\theta = 0.7$ (orange).

If we were less lucky and only found 2 patients recovered, then we see that the likelihood is shifted leftwards; peaking now at $\theta = \frac{1}{5}$ (see the red line of figure 7.2). By contrast if our patients respond well, and 9 out of 10 recover in the week period, then the likelihood shifts right (see the grey line of figure 7.2).

As for the Bernoulli case, we would like a compact way of writing down the likelihood to cover any eventuality. We now suppose that the number of volunteers is mysteriously given by n, the probability of individual treatment success is θ, and k cases turn out to be successes:

$$p(Z = k|n, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{7.5}$$

This distribution is known as the Binomial distribution.

If we hold n constant, and increase the probability of success p, we see that the probability distribution of data shifts to the right (see the right-hand panel of figure 7.2).

Video : Need to make an intuitive video here. One exists but is overly mathematical.

Normal

Poisson

Negative binomial

Logistic

7.4 Prior distributions

Distributions for probabilities, proportions and percentages

Uniform

Beta

Dirichlet

7.5 Table of distributions, their uses, and reasonable priors

7.5.1 Distributions for means and medians

Normal

Student t

7.5.2 Distributions for variances, and shape parameters

Gamma

Half-Cauchy

Inverse Gamma

Inverse chi

7.5.3 Multinomial - or other regression

7.5.4 LBG prior - see Michael Betancourt video and Stan doc

Better alternative to Wishart.

7.5.5 Wishart

7.5.6 Distributions for categories

Categorical

7.6 Chapter summary

Chapter 8

Conjugate priors and their place in Bayesian analysis

Chapter 9

Objective Bayesian analysis

Part III

A practical guide to doing real life Bayesian analysis: Computational Bayes

Part IV

Regression analysis and hierarchical models

Chapter 10

Hierarchical models

Chapter 11

Hypothesis testing I: Classical Frequentist vs Bayesian approaches

Chapter 12

Evaluation of model fit

Formerly hypothesis testing II. Definitely use something similar to Kruscke's P67 example for choosing between models.