Bayesian book

Ben Lambert

January 8, 2015

Contents

Chapter 1

How to best use this book

Part I

Understanding the Bayesian formula

Chapter 2

The subjective worlds of frequentist and Bayesian statistics

2.1 The purpose of statistics

2.2 The world according to classical statistics

2.3 The world according to Bayesian statistics

2.4 Similarities and differences in approaches between classical and Bayesian statistics

2.5 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 wait of us. In order to plan, as well as make sense of things, we frequently implicitly formulate an idea as to the relative likelihood of different out-

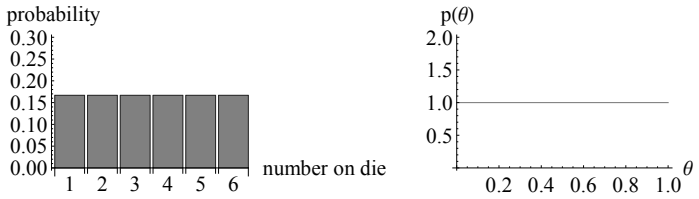

Figure 2.1: Left) A discrete uniform probability distribution for the number shown on a rolled fair die. Right) A continuous distribution to represent the probability of a coin coming up heads.**I want to remove the ticks from the discrete plot, simply placing a $\frac{1}{6}$ there instead.**

comes. 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*, in the form of 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 before we roll a die - which we believe to be fair - we assign an equal probability of $\frac{1}{6}$ to each of six the possible outcomes. 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 shown on the die (see figure 2.1).

2.5.1 What make a probability distribution *valid*? The most important thing today

The die example given in section 2.5 refers to discrete probability distribution, since the variable we were measuring - the number shown on the die after a throw - 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. Before we flip a coin we might be uncertain as to its innate bias, and might[2] imagine that any value for the probability of it coming up heads, θ, is equally likely. This results in the

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

[2] Perhaps foolishly since typically we expect that coins are designed to be relatively *fair* in this respect.

continuous analogue of the discrete die example (see figure 2.1).

The aforementioned examples are both examples of valid/proper probability distributions. So, what are its 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 die case, this is satisfied since $p(X) = \frac{1}{6} \geq 0$, and:

$$\sum_{i=1}^{6} \frac{1}{6} = 1 \tag{2.1}$$

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

$$\begin{aligned} \int_0^1 p(\theta) \mathrm{d}\theta &= \int_0^1 1 \mathrm{d}\theta \\ &= 1 \end{aligned} \tag{2.2}$$

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!

2.5.2 Interpreting discrete and continuous probability distributions

The discrete probability distribution for a fair die shown on the left hand side in figure 2.1, is straightforward to interpret. To calculate the probability that the die lands on a 1, we simply read off the probability from the graph corresponding to the height of the leftmost bar, and find that:

$$p(X = 1) = \frac{1}{6} \tag{2.3}$$

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 die lands on a number less than 4, we just add together the probabilities of it landing on 1, 2 and 3:

$$\begin{aligned} p(X < 4) &= p(X = 1) + p(X = 2) + p(X = 3) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{2} \end{aligned} \tag{2.4}$$

How can we use the continuous probability distribution such as the one shown on the right hand side of figure 2.1? If we want to calculate the probability that the bias of the coin is zero, in other words, the probability that $\theta = 0.5$, then we could simply draw a vertical line from this point on the θ axis up to the line of the distribution; concluding that $p(\theta = 0.5) = 1$! We know that this is not the intended result, since we supposed before that *all* values of θ were equally likely. This means that the probability that θ takes on any particular value between 0 and 1 is *actually* zero. Intuitively, this is because there are an infinity of particular values between 0 and 1 which θ can feasibly take on, meaning that picking one of them is infinitely unlikely. For example, the following values $\theta = \{0.5, 0.501, 0.50001, 0.5000001\}$ are all equally likely, and we could generate an infinity of these test values of θ, meaning that each one must have a zero likelihood of occurring. This means for a continuous random variable, we always have $p(\theta = \textit{number}) = 0$. 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 coin example, we can hence calculate the probability that $0.25 \leq \theta \leq 0.75$:

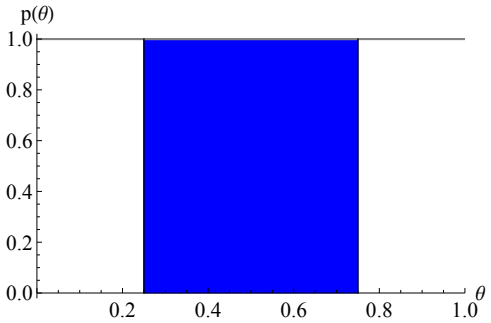

Figure 2.2: The probability that θ, representing the probability that a heads is thrown, lies between 0.25 and 0.75.

$$\begin{aligned} Pr(0.25 \leq \theta \leq 0.75) &= \int_{0.25}^{0.75} p(\theta) \mathrm{d}\theta \\ &= \int_{0.25}^{0.75} 1 \mathrm{d}\theta \\ &= [\theta]_{0.25}^{0.75} = (0.75 - 0.25) = 0.5 \end{aligned} \tag{2.5}$$

In (2.5), 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 (2.5), is equivalent to working out the area under the graph within those limits (see figure 2.2).

2.5.3 Generalising probability distributions to two dimensions

Life is often more complex than the examples of section 2.5. 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 comfort-

able 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 in the next few sections.

Biased coins: a 2-dimensional discrete probability example

Imagine that you are given two coins, and you are told that both are biased towards the same particular outcome - heads or tails - although, before flipping them you are unaware as to their inherent prejudice. We are also have an inherent feeling, due to our prior experience with the experimenter, that a bias towards heads is more likely. We use a binary variable to represent the outcome of each coin flip, which takes on the value of 0 if the coin falls tails-up, and 1 if it falls heads-up. We could then represent our beliefs beforehand in the form of a probability distribution shown in table 2.1.

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

Table 2.1: The probability distribution for the biased coins example described in section 2.5.3. $\{0, 1\}$ refers to the coin falling tails- or heads-up respectively.

How can we check whether this distribution satisfies the requirements for a valid probability distribution? We simply apply the rules described in section 2.5.1. 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:

$$\sum_{X_A=0}^{1} \sum_{X_B=0}^{1} p(X_A, X_B) = 0.3 + 0.1 + 0.1 + 0.5 = 1 \tag{2.6}$$

In (2.6), X_A and X_B are random variables[3] which refer to the outcome of Coin A and Coin 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, $p(X_A, X_B)$, 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 coins show tails (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 $p(X_A = 0, X_B = 0) = 0.3$. We ascribe a smaller likelihood to the coins coming up with different outcomes, $p(X_A = 0, X_B = 1) = 0.1$ or $p(X_A = 1, X_B = 0) = 0.1$, since we believe that the coins are biased in the same direction. We believe that the most likely outcome is that both coins show heads, since we believe that the experimenter is predisposed to this outcome, and hence ascribe the highest probability to this result, with $p(X_A, X_B) = 0.5$.

2.5.4 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 be 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 2.3).

We could verify that the distribution shown in figure 2.3 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[4]?

[3] 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.

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

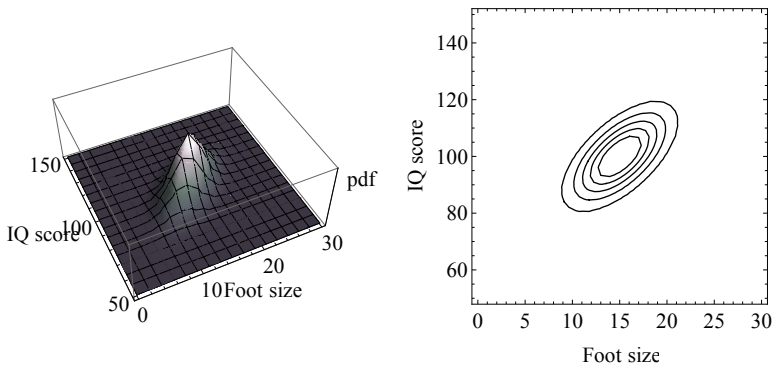

Figure 2.3: A probability distribution describing the foot size and IQ for an individual within our sample. Left) Represented as a 3-dimensional plot, and Right) Contour lines specify isolines of probability.

2.5.5 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 coins example, we might be interested in say, only the value obtained when we flip coin 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 coins example, if we are only interested in coin A, we can sum down the column values for coin B, obtaining the *marginal* distribution of coin A, shown at the bottom of table 2.2.

		Coin A		
		0	**1**	$p(X_B)$
Coin B	**0**	0.3	0.1	**0.4**
	1	0.1	0.5	**0.6**
	$p(X_A)$	**0.4**	**0.6**	

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

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

$$p(X_A = 1) = p(X_A = 1, X_B = 0) + p(X_A = 1, X_B = 1) \tag{2.7}$$

In (2.7), we see that A can come up 'heads' with B being 'tails', or alternatively A can land on 'heads' with B also on 'heads'.

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:

$$p(A = \alpha) = \sum_{\beta} p(A = \alpha, B = \beta) \tag{2.8}$$

In (2.8), α and β refer to the specific values taken on by the random variables A and B.

We can use (2.8) for the coin example to calculate the probability that coin B lands up on 'tails':

$$\begin{aligned} p(X_B = 0) &= \sum_{\alpha=0}^{1} p(X_B = 0, X_A = \alpha) \\ &= p(X_B = 0, X_A = 0) + p(X_B = 0, X_A = 1) \\ &= 0.3 + 0.1 = 0.4 \end{aligned} \tag{2.9}$$

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:

$$p_A(\alpha) = \int_{All\ \beta} p_{AB}(\alpha, \beta) \mathrm{d}\beta \tag{2.10}$$

In (2.10), $p_{AB}(\alpha, \beta)$ 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:

$$p(IQ) = \int_0^{30} p(IQ, FS) \mathrm{d}FS \tag{2.11}$$

The result of carrying out the step in (2.11) is that we are left with the distribution shown on the right of figure 2.4. We have rotated this graph to emphasise that it is the result of essentially summing[5] 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 height of the marginal density is given by the length of the path you have to walk along a particular line of constant IQ. 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.

An alternative way of thinking about marginal distributions is provided by the Venn diagram shown in figure 2.5.5. 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 arbitrary events A and B as sub-areas in the diagram, which overlap indicating a region of joint probability, $p(A, B)$. In this setup it is straightforward to calculate the marginal probability of events A or B; we find the area of the elliptic shapes A or B respectively. Considering event A, when we calculate the area of the entire ellipse, we are implicitly carrying out the sum of the form indicated in (2.8):

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

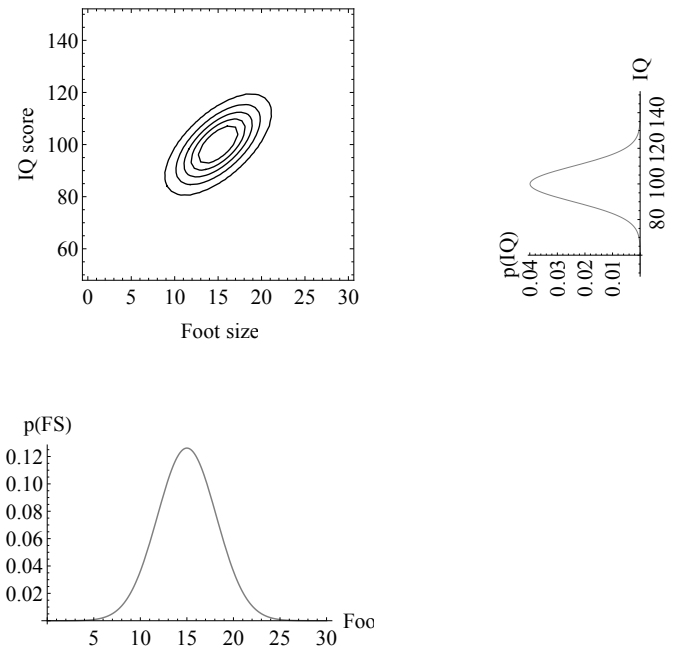

Figure 2.4: Top-left: the joint density of foot size and intelligence. Right: the marginal density of IQ. Bottom: the marginal density of foot size. **I want to add a line at a particular value of IQ, and at a particular value of FS, to illustrate the horizontal and vertical summing.**

$$p(A) = p(A, B) + p(A, not\ B) \tag{2.12}$$

In (2.12), the terms on the right hand side correspond to the overlap region, and the remaining part of A (where B does not occur) respectively.

2.5.6 Conditional distributions

We frequently receive partial information by observing only part of the system in which we are interested. In our coin example, we might flip of one of the coins finding lands heads-up, and on this basis update our probabilities of obtaining heads or tails for the other. 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 examples 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:

$$p(A|B) = \frac{p(A, B)}{p(B)} \tag{2.13}$$

In (2.13), $p(A|B)$ refers to the probability of A occurring, given that B has occurred. In the right hand side of (2.13), $p(B)$ is the *marginal* distribution of B occurring, and $p(A, B)$ is the joint probability of A and B occurring.

We can use (2.13) for the coins example to calculate the probability that *given* that coin A lands heads-up, what is the probability of coin B also landing heads-up when we flip it?

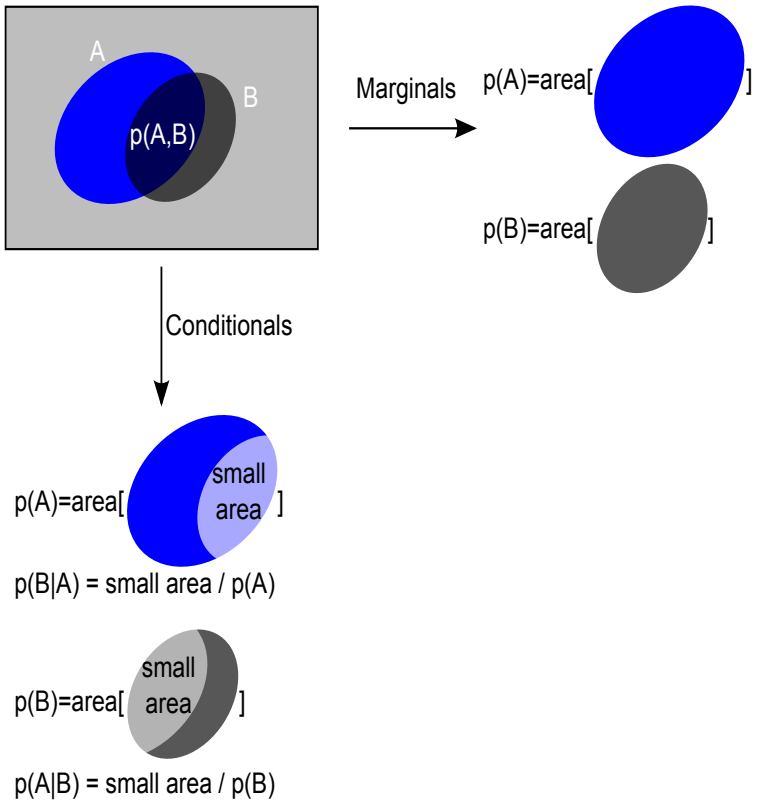

Figure 2.5: A Venn diagram showing one way of interpreting marginal and conditional distributions.

$$
\begin{aligned}
p(X_B = 1|X_A = 1) &= \frac{p(X_A = 1, X_B = 1)}{p(X_A = 1)} \\
&= \frac{p(X_A = 1, X_B = 1)}{p(X_A = 1, X_B = 0) + p(X_A = 1, X_B = 1)} \\
&= \frac{0.5}{0.1 + 0.5} \\
&= \frac{5}{6}
\end{aligned}
\tag{2.14}
$$

In (2.14), we have used the rule we discussed earlier for calculating marginal probabilities, shown in (2.8), to calculate the denominator, $p(X_A = 1)$.

Another way to see the workings of this calculation is shown in table 2.3. When we uncover that coin A is heads-up, 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.

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

Table 2.3: The highlighted region indicates the new solution space, since we know that coin A has landed heads-up.

The Venn diagram in figure 2.5.5 shows another way of interpreting conditional distributions. If we are told that a particular event B occurs, then our event space collapses to only the area specified by B. The conditional probability, $p(A|B)$ is then simply given by the ratio of the area of overlap between A and B to the total area of B. This makes intuitive sense, since this is the only way that event A can occur, given that B has already occurred.

We can also use (2.13) 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 (2.13).

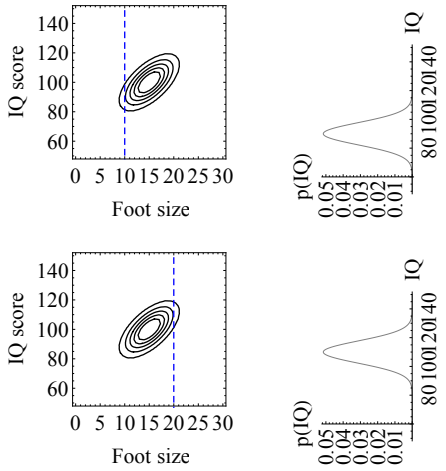

Figure 2.6: The dashed blue lines indicate the new event space in each case. The height walked following these lines is related to the magnitude of the conditional distributions shown on the right.

Figure 2.5.6 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 indicate the relative likelihood of different values of IQ.

2.6 Central Limit Theorems: the most important thing ever

Most distributions are not

2.7 The Bayesian formula

2.7.1 The intuition behind the formula

2.8 The Bayesian inference process from the Bayesian formula

2.9 Implicit vs Explicit subjectivity

2.10 What are the tangible benefits of Bayesian statistics?

2.11 Why don't more people use Bayesian statistics?

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 undertaking 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 section . After

this we will be able to appreciate how the posterior distribution of Bayes' formula can be put to its many uses.

3.3 Expressing uncertainty in a parameter 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.

3.3.1 Do parameters actually exist and have a point value? Bayesian vs classical standpoints

3.3.2 Failings of the classical confidence interval

3.3.3 The HDI as a better alternative

3.3.4 The central posterior interval

3.4 Prediction using a posterior distribution

3.4.1 Before experiment, using prior

3.4.2 After experiment, using posterior

3.5 Model comparison using the posterior

3.6 Model testing through the posterior

[1] Whether a true value for a parameter actually exists we leave until section 3.3.1.

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:

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

What (4.3) means is, what is 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:

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

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 *theta*, 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, 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 *theta*, 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.

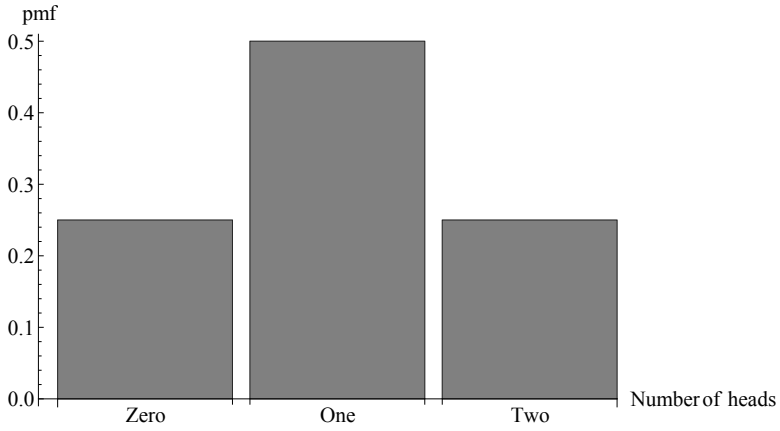

Figure 4.1: The probabilities of all possible numbers of heads for a fair coin.

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, 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 *theta*. If we assume we obtained two heads, and vary θ between 0 and 1, we can obtain the likelihood shown in figure 4.3. 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 look at the area underneath the curve in figure 4.3, we find that it does not integrate to 1. Thus we have a violation of the second condition for a valid probability distribution. Hence,

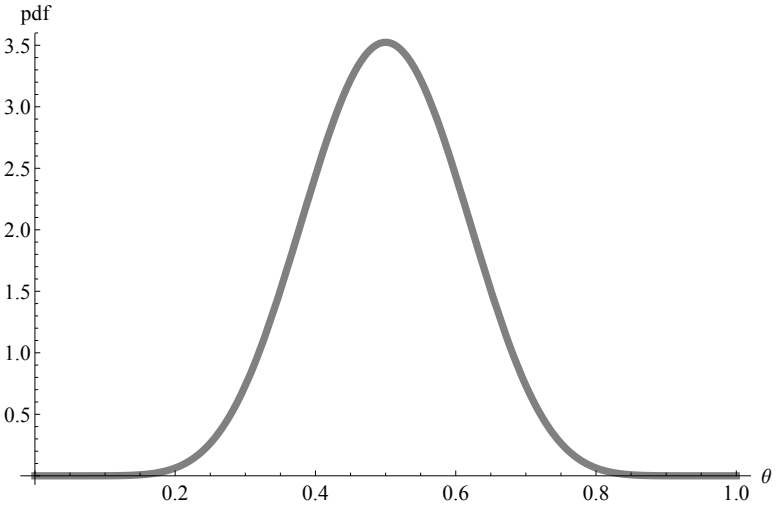

Figure 4.2: An example posterior distribution for the probability of obtaining a heads in a coin toss.

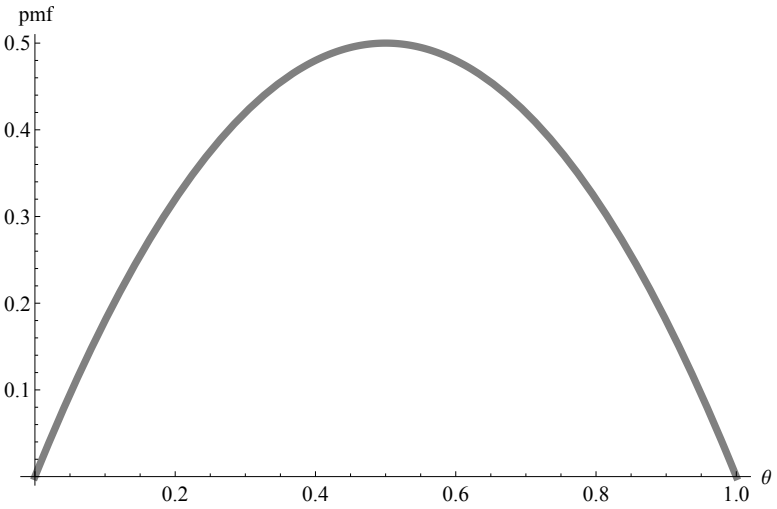

Figure 4.3: The likelihood function for obtaining a single head from two throws. The area under the curve is $\frac{1}{3}$.

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:

$$\mathcal{L}(\theta|data) = P(data|\theta) \tag{4.5}$$

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:

$$P(X = 0|\theta) = P(TT|\theta) = P(T|\theta) \times P(T|\theta) = (1 - \theta)^2 \tag{4.6}$$

$$P(X = 1|\theta) = P(HT|\theta) + P(TH|\theta) = 2 \times P(T|\theta) \times P(H|\theta) = 2\theta(1 - \theta) \tag{4.7}$$

$$P(X = 2|\theta) = P(HH|\theta) = P(H|\theta) \times P(H|\theta) = \theta^2 \tag{4.8}$$

In (4.6), 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[2], which is also $(1 - \theta)$. The factor of two arises in (4.8) 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

[2]Since we have assumed a model whereby the results of the first and second throws are independent, conditional on θ. In other words, all the similarity between the two throws is captured in the parameter θ.

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}.

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 of these models 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!

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 and calibration 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
- 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 NHS 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:

$$X = \begin{cases} O & , \textit{No disease} \\ 1 & , \textit{Positive diagnosis} \end{cases} \tag{4.9}$$

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:

$$P(X = 0|\theta) = (1 - \theta) \tag{4.10}$$
$$P(X = 1|\theta) = \theta \tag{4.11}$$

However, we would like to write down a single rule which yields (4.10) or (4.11) respectively, dependent on whether $X = 0$ or $X = 1$. This can be achieved with the following:

$$P(X = \alpha|\theta) = \theta^{\alpha}(1 - \theta)^{1-\alpha} \tag{4.12}$$

Note that in (4.12) that $\alpha \in \{0, 1\}$ refers to the numeric value taken by the variable X. The function (4.12) 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 (4.10) and (4.11) if the individual is disease -negative/-positive respectively:

$$P(X = 0|\theta) = \theta^{0}(1 - \theta)^{1} = (1 - \theta) \tag{4.13}$$
$$P(X = 1|\theta) = \theta^{1}(1 - \theta)^{0} = \theta \tag{4.14}$$

When we hold the datum X fixed, and vary θ (4.12) 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 (4.12) 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

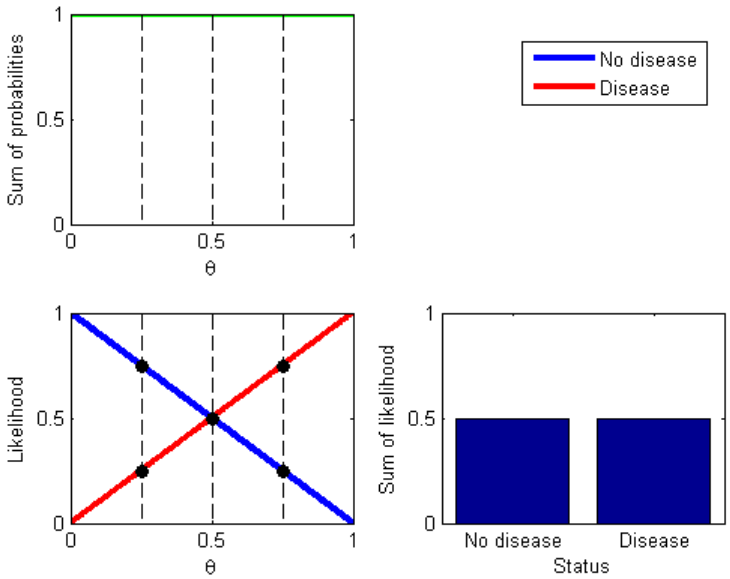

Figure 4.4: The likelihood function as theta varies for the case of the two possible data. The sum of likelihoods is found by the area under each line, whereas the sum of probabilities is a discrete sum.

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[3].

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[4]. 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 (4.12):

$$P(X = \alpha|\theta) = \theta^{\alpha}(1-\theta)^{1-\alpha} \tag{4.15}$$

Note that in (4.15) 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 (include reference back to discussion of probabilities). In words, we obtain the probability that the first person has disease status X_1 *and* the second

[3]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.

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

person has status X_2:

$$\begin{aligned} P(X_1 = \alpha_1, X_2 = \alpha_2|\theta_1, \theta_2) &= P(X_1 = \alpha_1|\theta_1) \times P(X_2 = \alpha_2|\theta_2) \\ &= \theta_1^{\alpha_1}(1-\theta_1)^{1-\alpha_1} \times \theta_2^{\alpha_2}(1-\theta_2)^{1-\alpha_2} \end{aligned} \tag{4.16}$$

In (4.16) 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 $\theta_1 = \theta_2$:

$$\begin{aligned} P(X_1 = \alpha_1, X_2 = \alpha_2|\theta) &= \theta^{\alpha_1}(1-\theta)^{1-\alpha_1} \times \theta^{\alpha_2}(1-\theta)^{1-\alpha_2} \\ &= \theta^{\alpha_1+\alpha_2}(1-\theta)^{2-\alpha_1-\alpha_2} \end{aligned} \tag{4.17}$$

In (4.17) 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:

$$Z = X_1 + X_2 \tag{4.18}$$

We can then use (4.17) to generate the respective probabilities.

$$\begin{aligned} P(Z = 0|\theta) &= P(X_1 = 0, X_2 = 0|\theta) = \theta^{0+0}(1-\theta)^{2-0-0} = (1-\theta)^2 \\ P(Z = 1|\theta) &= P(X_1 = 1, X_2 = 0|\theta) + P(X_1 = 0, X_2 = 1|\theta) = 2\theta(1-\theta) \\ P(Z = 2|\theta) &= P(X_1 = 1, X_2 = 1|\theta) = \theta^{1+1}(1-\theta)^{2-1-1} = \theta^2 \end{aligned} \tag{4.19}$$

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 (4.19) as:

$$\begin{aligned} P(Z = 0|\theta) &= \theta^0(1-\theta)^2 \\ P(Z = 1|\theta) &= 2\theta^1(1-\theta)^1 \\ P(Z = 2|\theta) &= \theta^2(1-\theta)^0 \end{aligned} \tag{4.20}$$

In (4.20) 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:

$$P(Z = \beta|\theta) \sim \theta^{\beta}(1-\theta)^{2-\beta} \tag{4.21}$$

The only problem with matching (4.21) with the previously obtained result is the factor of 2 on the middle line of (4.20). However, as a complete aside we note that when we expand a quadratic factor we get the following:

$$(x+1)^2 = x^2 + 2x + 1 \tag{4.22}$$

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 ${}^{n}C_{r}$. The expansion coefficients are normally written in compact form:

$$\binom{2}{\beta} = \frac{2!}{(2-\beta)!\beta!} \tag{4.23}$$

In (4.23) 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:

$$P(Z = \beta|\theta) = \binom{2}{\beta}\theta^{\beta}(1-\theta)^{2-\beta} \tag{4.24}$$

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

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:

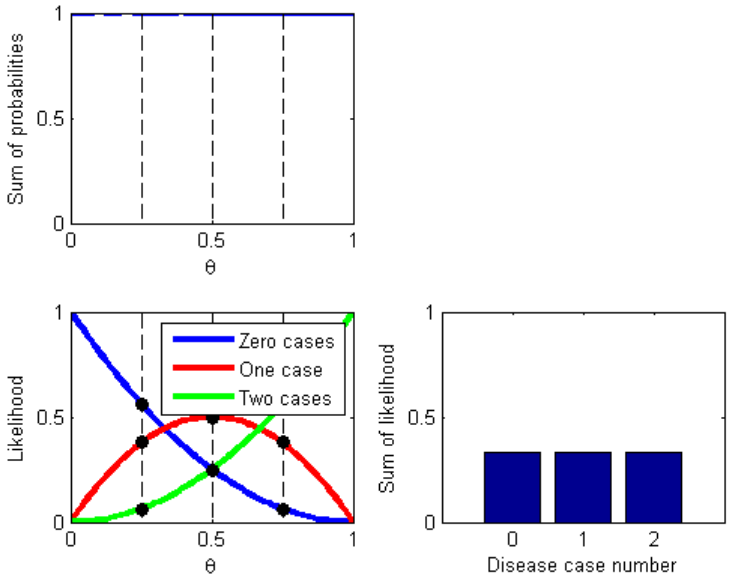

Figure 4.5: The likelihood function as theta varies for a sample of 2 individuals.

$$\begin{aligned} P(Z=0|\theta) &= P(X_1=0|\theta)P(X_2=0|\theta)P(X_3=0|\theta) \\ P(Z=1|\theta) &= 3P(X_1=1|\theta)P(X_2=0|\theta)P(X_3=0|\theta) \\ P(Z=2|\theta) &= 3P(X_1=1|\theta)P(X_2=1|\theta)P(X_3=0|\theta) \\ P(Z=3|\theta) &= P(X_1=1|\theta)P(X_2=1|\theta)P(X_3=1|\theta) \end{aligned} \tag{4.25}$$

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:

$$P(Z=\beta|\theta) = \binom{3}{\beta}\theta^{\beta}(1-\theta)^{3-\beta} \tag{4.26}$$

We recognise a pattern in the likelihoods of (4.24) and (4.26) which allows us to deduce that, for a sample size of N, the likelihood is given by:

$$P(Z=\beta|\theta) = \binom{N}{\beta}\theta^{\beta}(1-\theta)^{N-\beta} \tag{4.27}$$

(4.27) 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 (4.27):

$$P(Z \geq 10|\theta = 0.01) = \sum_{Z=10}^{100}\binom{100}{Z}0.01^{Z}(1-0.01)^{100-Z} = 7.63\times10^{-8} \tag{4.28}$$

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 *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 4.8). 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 test scores. Don't worry if you are not aware of this theorem, we will cover it in due course, but basically it says, if there are a large number of factors whose average results in an intelligence score, then the

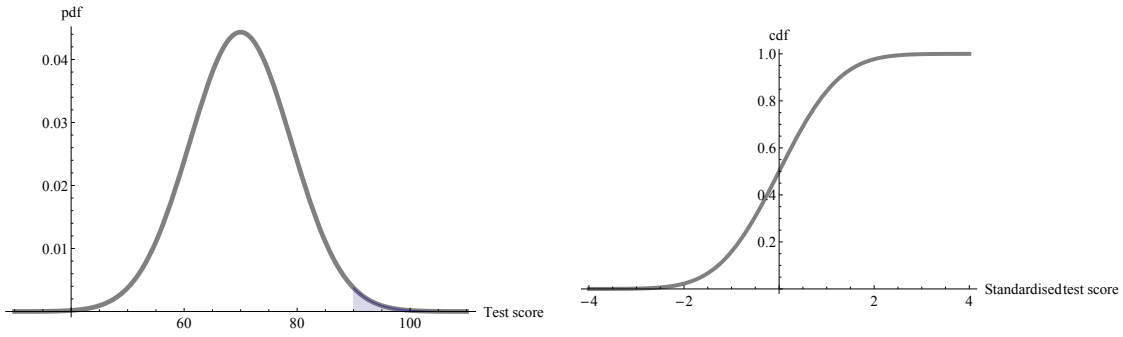

Figure 4.6: Left panel shows a normal with $\mu = 70$ and $\sigma^2 = 81$, with the area corresponding to a result as extreme as 90 indicated. This translates into a standard normal cdf shown in the right panel, which can be used to calculate this area from the first figure. This translation to the standard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

normal distribution provides a reasonable approximation to the distribution of test scores. In which case, we assume that a normal distribution for our likelihood function for an individual's test score, X:

$$P(X = \alpha|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha-\mu)^2}{2\sigma^2}} \tag{4.29}$$

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)):

$$\begin{aligned} P(X \geq 90|\mu = 70, \sigma^2 = 81) &= \int_{90}^{\infty} \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(\alpha-70)^2}{2\times 10}} \, \mathrm{d}\alpha \\ &= 1 - \Phi\left(\frac{90-70}{9}\right) \approx 0.0131 \end{aligned} \tag{4.30}$$

In (4.30), Φ stands for the value of the *standard* normal cumulative distri-

bution function[5] 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 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[6], 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. We calculate the joint probability density for a sample of N individuals by multiplying together the individual densities:

$$P(X_1 = \alpha_1, X_2 = \alpha_2, ..., X_N = \alpha_N | \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha_i - \mu)^2}{2\sigma^2}} \tag{4.31}$$

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 (4.30) since we would have to integrate across all individuals' variables.

4.7 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

[5] 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.

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

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 with 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.

4.8 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, as well as how these estimates can be used to make inferences about the population.

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.

4.8.1 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

Figure 4.7: A figure with four panes. The top-left is log-likelihood as a function of likelihood. The bottom-left is likelihood as a function of theta, and log-likelihood plotted on the same axis. Top-right is a weird function of likelihood as a function of likelihood. Below it a graph of likelihood as a function of theta, with a different maximum reached for the weird function.

an analysis. If in a sample of 100 individuals, 10 test positively[7], 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 (4.27) as:

$$L(\theta|data) = \binom{100}{10}\theta^{10}(1-\theta)^{100-10} \tag{4.32}$$

Remember, that since we are varying θ and holding the data constant here, that (4.32) is a *likelihood*, not a probability. We then need to simply choose θ so that we can maximise the likelihood. We could simply differentiate (4.32) 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 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 4.7):

$$l(\theta|data) = LogL(\theta|data) = log\binom{100}{10} + 10log(\theta) + 90log(1-\theta) \tag{4.33}$$

Where to get the result (4.33), we have used the log rules:

$$\begin{aligned} log(ab) &= log(a) + log(b) \\ log(a^b) &= blog(a) \end{aligned} \tag{4.34}$$

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

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

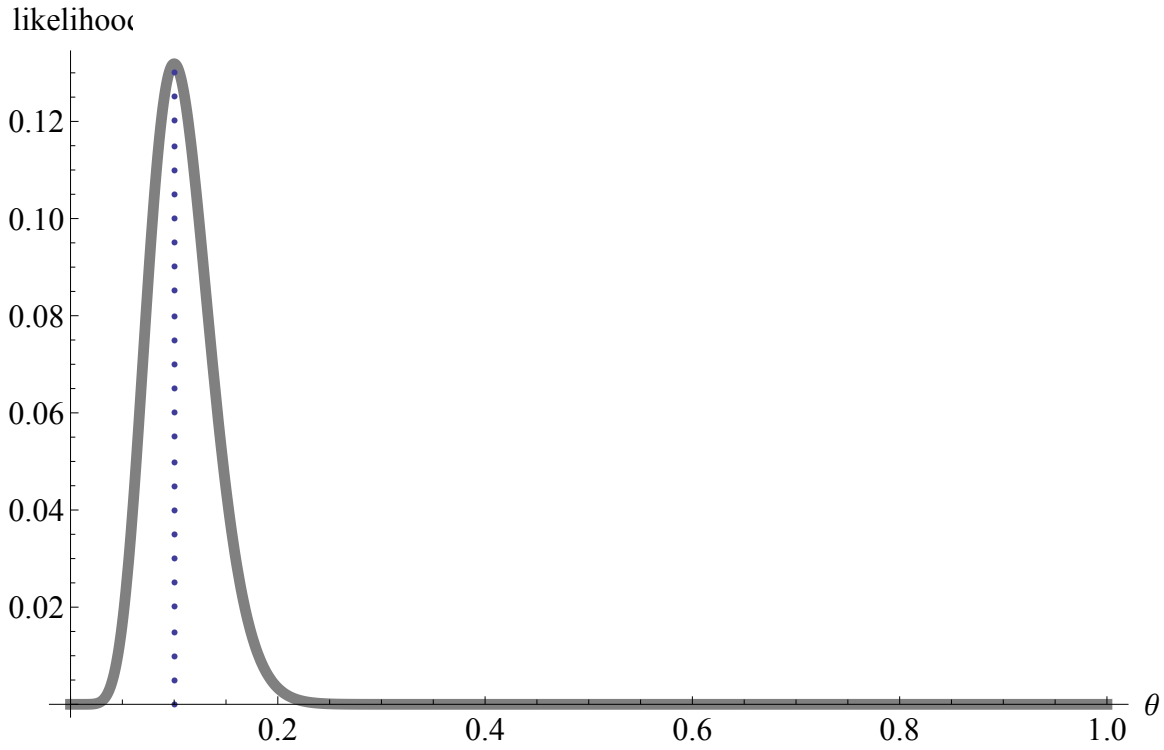

Figure 4.8: Log-likelihood of disease prevalence from section 4.8.1 as a function of theta, maximised at 1/10.

$$\frac{\partial l}{\partial \theta} = \frac{10}{\hat{\theta}} - \frac{90}{1-\hat{\theta}} = 0 \tag{4.35}$$

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[8] of the disease prevalence exactly equal to

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

that in our sample:

$$\hat{\theta} = \frac{\beta}{N} \tag{4.36}$$

4.8.2 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:

$$L(\mu, \sigma^2 | X_1 = 75, X_2 = 71) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(75-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(71-\mu)^2}{2\sigma^2}} \tag{4.37}$$

We can then proceed as we did in section 4.8.1 by taking the log of this expression before we differentiate it:

$$l(\mu, \sigma^2 | X_1 = 75, X_2 = 71) = 2log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(75-\mu)^2}{2\sigma^2} - \frac{(71-\mu)^2}{2\sigma^2} \tag{4.38}$$

Where we have again used the log rules in (4.34) to achieve (4.38). We can now proceed to differentiate (4.38) with respect to both variables, holding the other constant, setting each to 0:

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{(75-\hat{\mu})}{\hat{\sigma^2}} + \frac{(71-\hat{\mu})}{\hat{\sigma^2}} = 0 \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{1}{\hat{\sigma^2}} + \frac{(75-\hat{\mu})^2 + (71-\hat{\mu})^2}{2\hat{\sigma^4}} = 0 \end{aligned} \tag{4.39}$$

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

$$\hat{\sigma^2} = \frac{1}{2}\left[(75-73)^2 + (71-73)^2)\right] = 4 \tag{4.40}$$

Notice that the maximum likelihood estimators for the population mean

and variance are for this case the *sample mean* and *sample variance*[9]. In fact, this holds for the case of N individuals' data, then the maximum likelihood estimators for this case would be:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X} \tag{4.41}$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N} (X_i - \bar{X})^2 = s^2 \tag{4.42}$$

4.9 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[10] 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.

[9] Albeit a biased estimator of the population variance. The unbiased estimator would

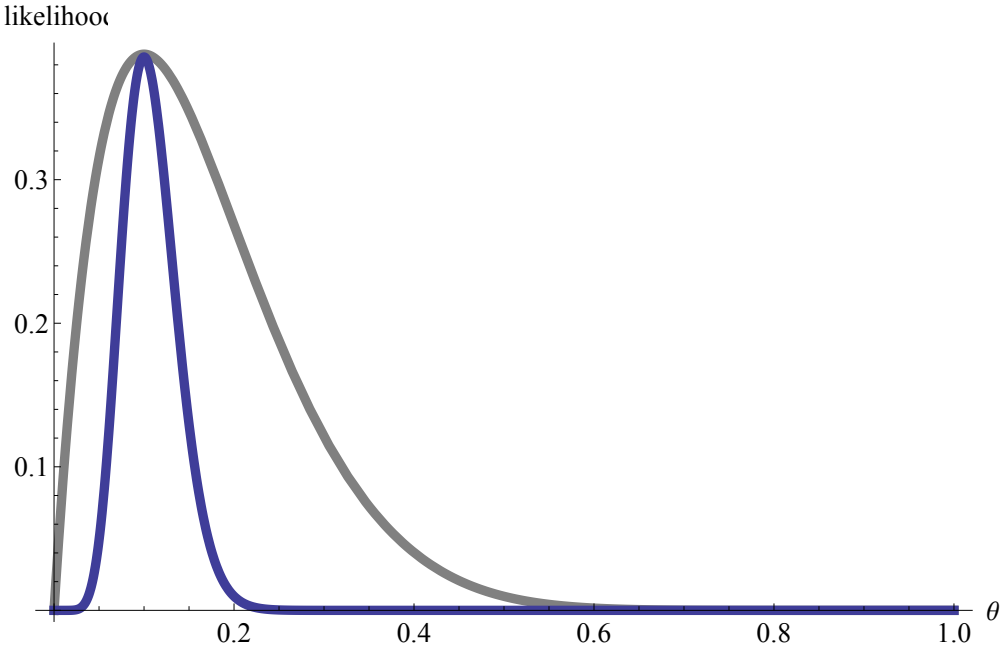

Figure 4.9: Two likelihoods which result in the same maximum likelihood estimates of parameters, at 0.1. The gray likelihood is less strongly-peaked, meaning we can be less confident about the estimates.

4.10 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.

divide by 1, rather than 2.

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

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

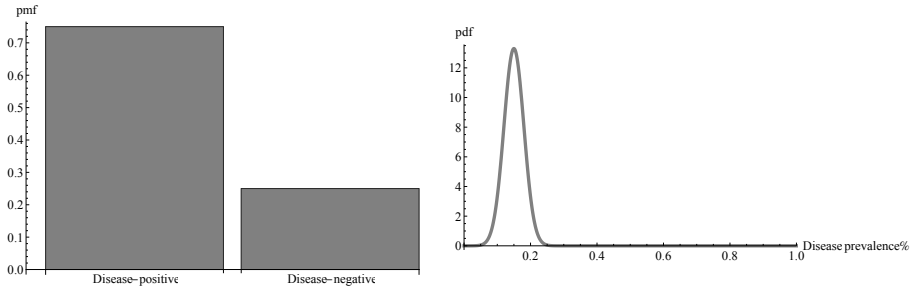

Figure 5.1: Left - a prior for a doctor's pre-testing diagnostic probability of an individual having a disease. Right - a prior which represents pre-sample uncertainty in disease prevalence.

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 a particular disorder. 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 described by Gelman, 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.

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

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:

$$P(\theta|data) = \frac{P(data|\theta)}{P(data)} \tag{5.2}$$

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?

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

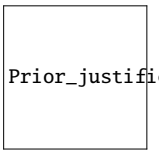

Figure 5.2: A figure showing in the left hand panel the Venn diagrams when we assume that a) the coin is fair with a quarter of the area shaded, and the other half not. b) Biased, and the majority of the figure is shaded - corresponding to a high overlap between data and parameter. The right hand panel then shows the joint probability of the data and the parameters θ, and we see that even though the ratio of the area is better for b), it is much less likely *a priori* that the coin is biased. The bottom panel shows the ML implied posterior distribution, with the bar for unfairness much higher than for fairness. The right shows a much more logical conclusion which takes into account their prior probabilities.

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, $P(\theta)$ 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 an improper 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.

5.5 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.

5.6 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.

5.6.1 An urn of balls[2]

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:

$$P(X = 1|Y = \alpha) = \frac{\alpha}{5} \tag{5.3}$$

In (5.3), $\alpha \in \{0, 1, 2, 3, 4, 5\}$ 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 (5.3). 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

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

description of this process see section 5.10.1.

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**

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).

5.6.2 Disease proportions revisited

Suppose that we substitute our urn from section 5.6.1 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

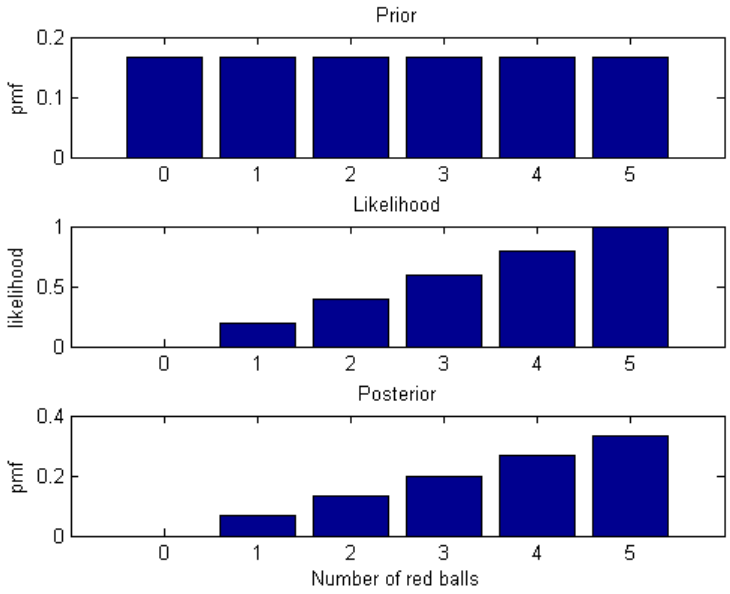

Figure 5.3: The prior, likelihood and posterior for the urn of balls example described in 5.6.1. The prior in the upper panel gives uniform weighting to all possible numbers of red balls. This is then multiplied by the likelihood (in the middle panel) at each number of balls, and normalised to make the posterior density shown in the bottom panel.

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**

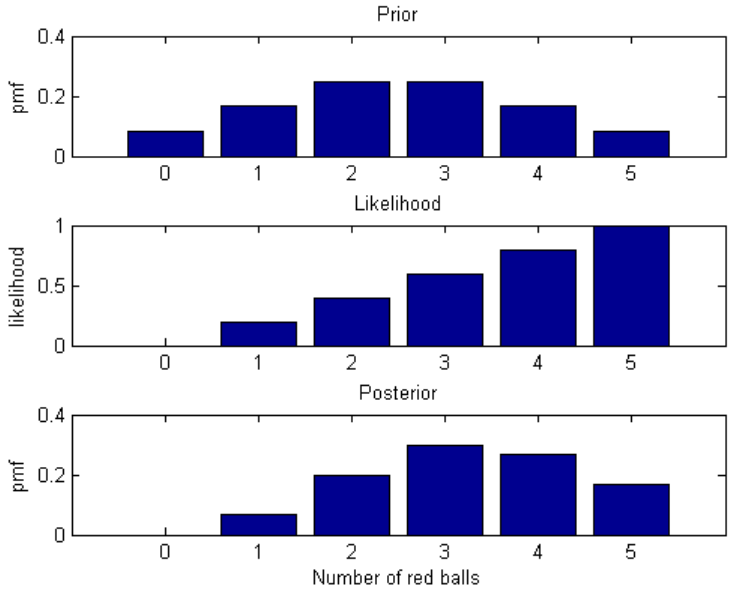

Figure 5.4: The prior, likelihood and posterior for the urn of balls example described in 5.6.1. The prior in the upper panel gives more weighting to more equal numbers of red and white balls. This is then multiplied by the likelihood (in the middle panel) at each number of balls, and normalised to make the posterior density shown in the bottom panel.

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[3]. 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:

$$P(Z = 3|\theta) = \binom{100}{3}\theta^{\beta}(1-\theta)^{100-3} \tag{5.4}$$

The reason for the $\binom{100}{3} = 161,700$ term at the beginning of (5.4) 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.

5.7 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.

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

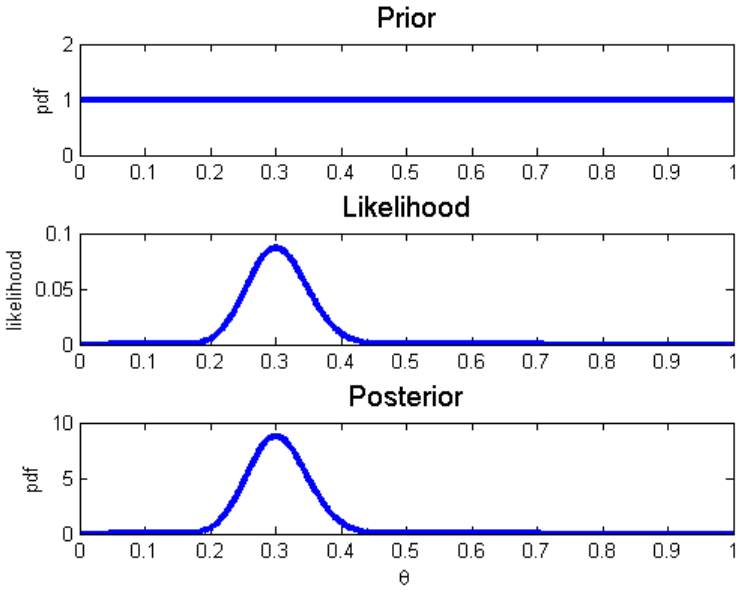

Figure 5.5: The prior, likelihood and posterior for the disease proportion example described in section 5.6.2. Each point in θ along the continuous prior curve (top panel) is multiplied by the corresponding value of likelihood (middle panel), to form the numerator of Bayes' rule. The numerator is then normalised to make the posterior probability density shown in the bottom panel.

5.7.1 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:

$$\begin{aligned} P(\theta|data) &= \frac{P(\theta) \times P(data|\theta)}{P(data)} \\ &\propto P(\theta) \times P(data|\theta) \\ &\propto P(data|\theta) \end{aligned} \tag{5.5}$$

In (5.7.1) 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 5.6.2, 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 mathematical details of these graphs see section 5.10.2.

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.

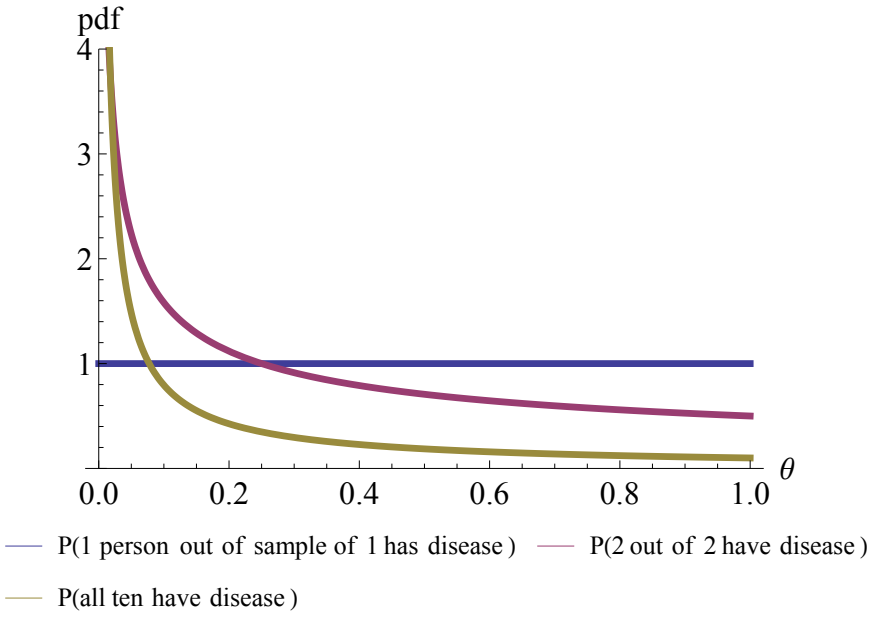

Figure 5.6: The probability density for obtaining all diseased individuals within sample sizes of 1, 2 and 10 respectively. Starting out with a flat prior for the probability that one individual has a disease has resulted in non-flat priors for the other 2 probabilities.

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[4], it would not actually be one (see section 5.4 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).

5.7.2 Informative priors

We have seen in section 5.7.1 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

[4] Although not assured.

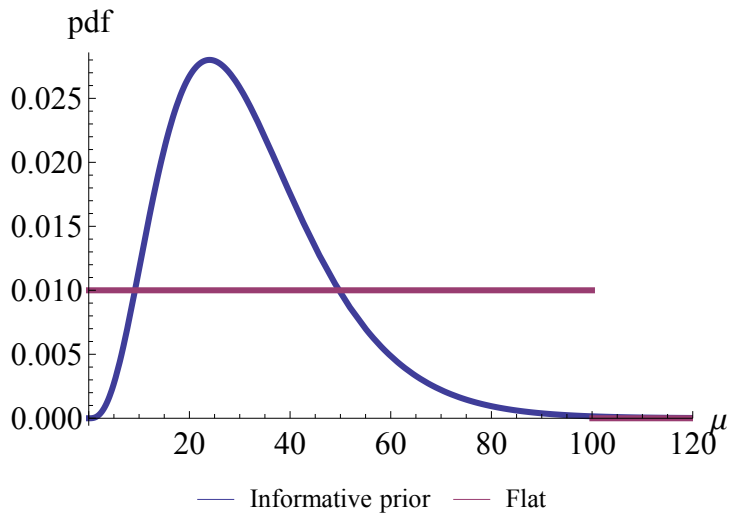

Figure 5.7: Two viable prior distributions for the average time taken before the onset of lung cancer after patients begin smoking.

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 a reasonable approximation the data could be modelled as having come from a normal distribution[5]. 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,

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

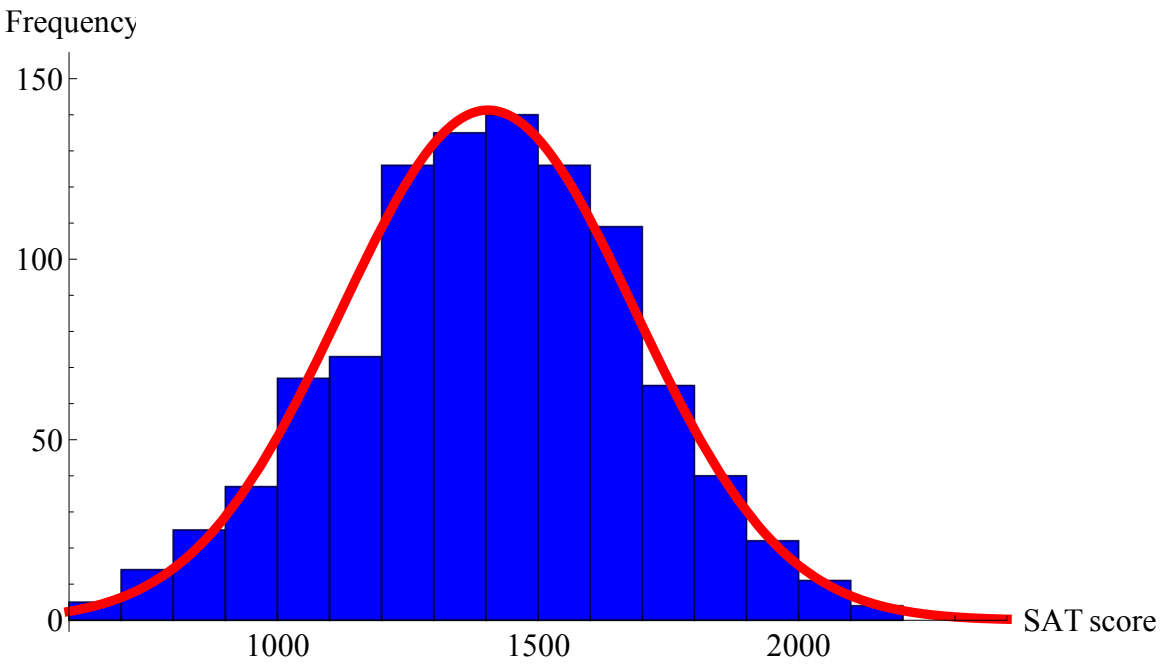

Figure 5.8: The SAT scores for past students of a class. The mean and variance of this hypothetical sample are 1404, and 79,716 respectively, which are used to fit a normal distribution to the data, and is shown in red.

we will learn a more pure Bayesian method which can be used to create prior densities.

5.7.3 The numerator of Bayes' rule determines the shape

We notice for both the examples described in sections 5.6.1 and 5.6.2 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:

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

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.

5.7.4 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:

$$\begin{aligned} z_{25} &= \frac{wage_{25} - \mu}{\sigma} \\ z_{75} &= \frac{wage_{75} - \mu}{\sigma} \end{aligned} \tag{5.7}$$

In (5.7), 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 (5.7) can be rearranged to the following:

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

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).

5.8 A strong model is not heavily influenced by priors

Returning to the example of section 5.6.2 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

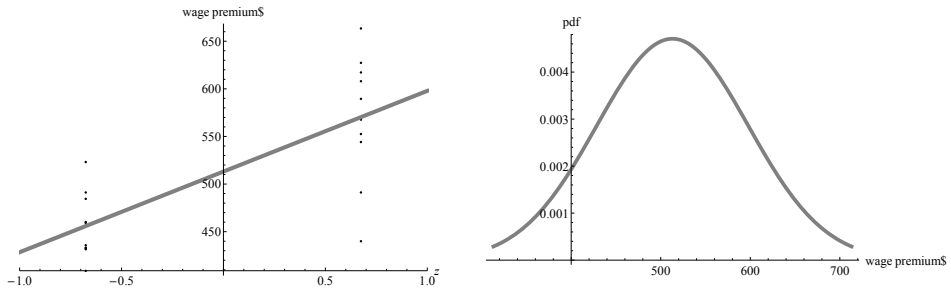

Figure 5.9: Hypothetical data for the 25th and 75th percentiles of the estimated wage premium from 10 experts. In the left hand panel we regress these percentiles on the corresponding percentiles from a standard normal distribution, yielding estimates of the mean and variance of a normal prior, which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

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.

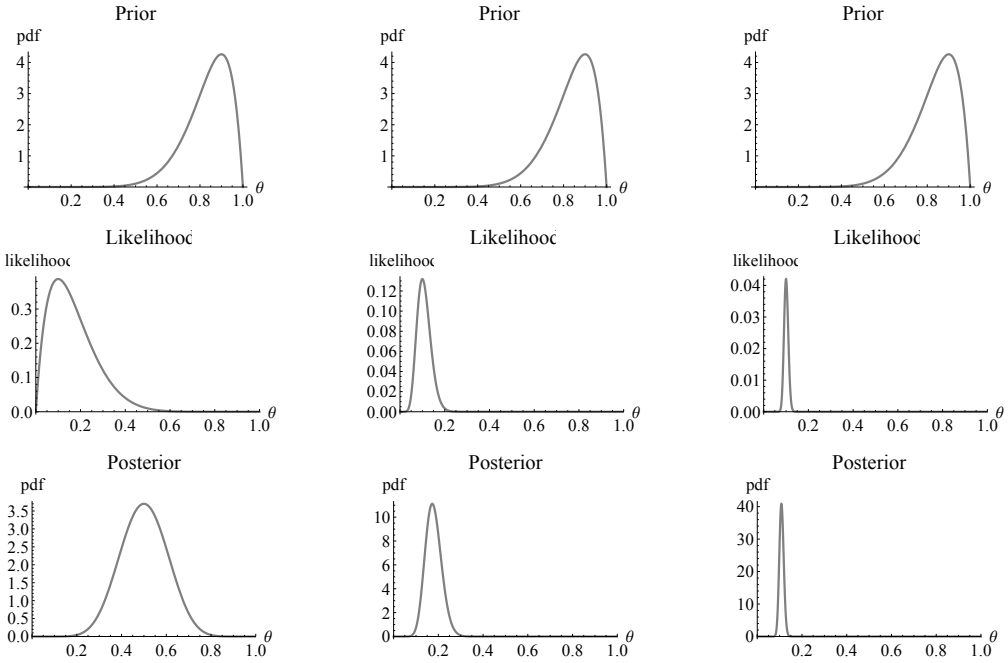

Figure 5.10: The effect of increasing sample size on the posterior density for the prevalence of a disease in a population. The leftmost column has N=10, the middle N=100, and the rightmost N=1,000. All three have the same proportion of disease cases in the sample.

5.9 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.

5.10 Appendix

5.10.1 Bayes' rule for the urn

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

$$\begin{aligned} P(Y=\alpha|X=1) &= \frac{P(X=1|Y=\alpha)\times P(Y=\alpha)}{P(X=1)} \\ &= \frac{P(X=1|Y=\alpha)\times P(Y=\alpha)}{\sum_{\alpha=0}^{5} P(X=1|Y=\alpha)\times P(Y=\alpha)} \\ &= \frac{\frac{\alpha}{5}\times\frac{1}{6}}{\sum_{\alpha=0}^{5}\frac{\alpha}{5}\times\frac{1}{6}} \end{aligned} \tag{5.9}$$

5.10.2 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:

$$P(Y) = P(\theta(Y)) \times |\theta'(Y)| \tag{5.10}$$

In (5.10), $\theta(Y) = Y^{-\frac{1}{2}}$ 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:

$$P(Y) = \frac{1}{2\sqrt{Y}} \tag{5.11}$$

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 5.7.3 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.

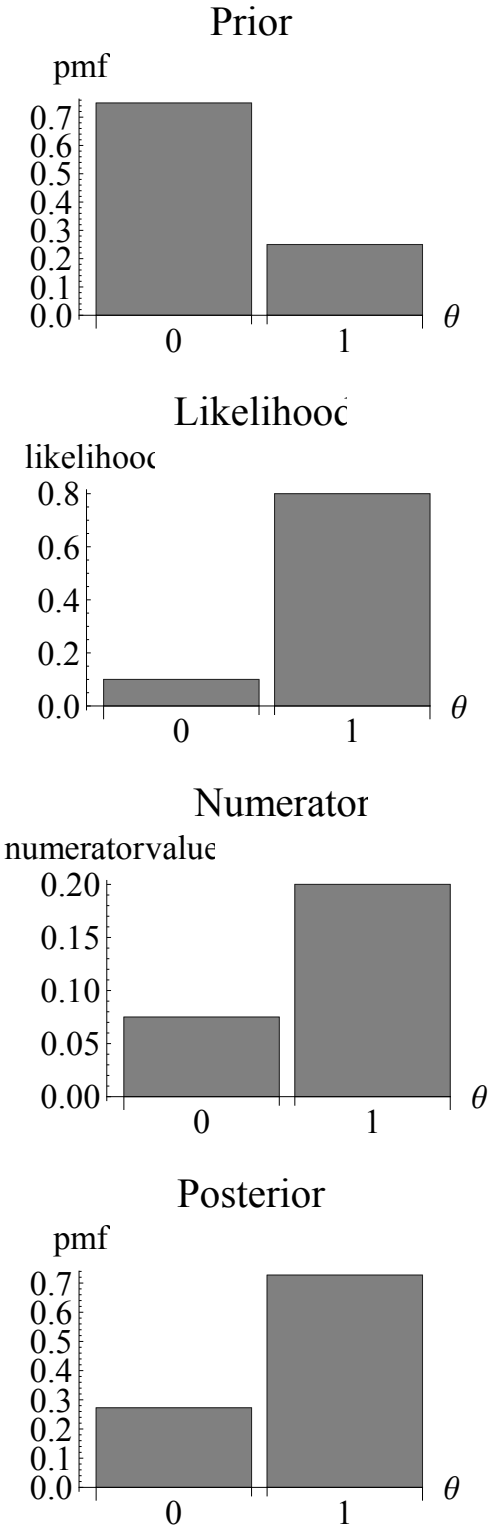

Figure 6.1: The prior is multiplied through by the likelihood, resulting in the numerator (the penultimate panel), which is then normalised by the sum over its values, to obtain the denominator.

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.

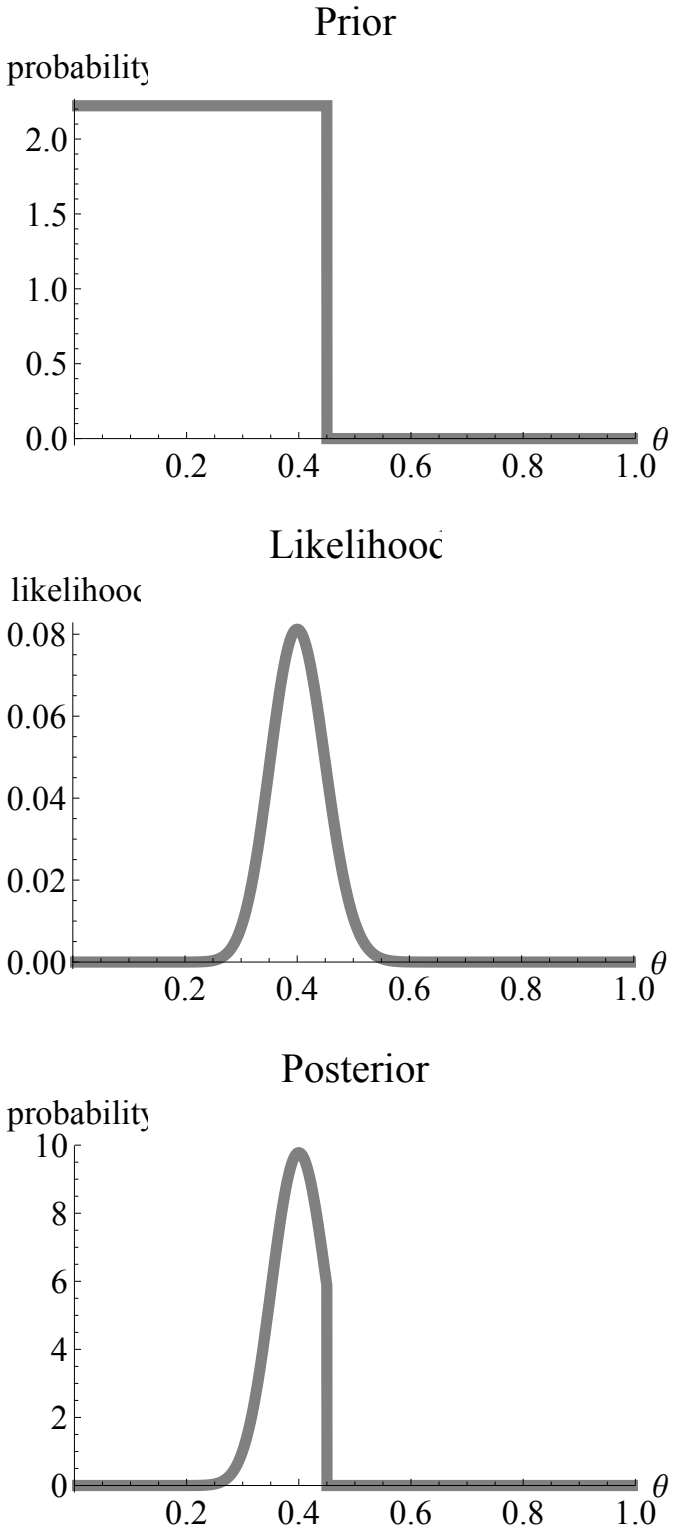

Figure 6.2: The prior, likelihood and posterior for the proportion of individuals voting for the conservative party in a general election, where we have found 40 people out of a sample of 100 voted conservative.

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 2.5:

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

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.

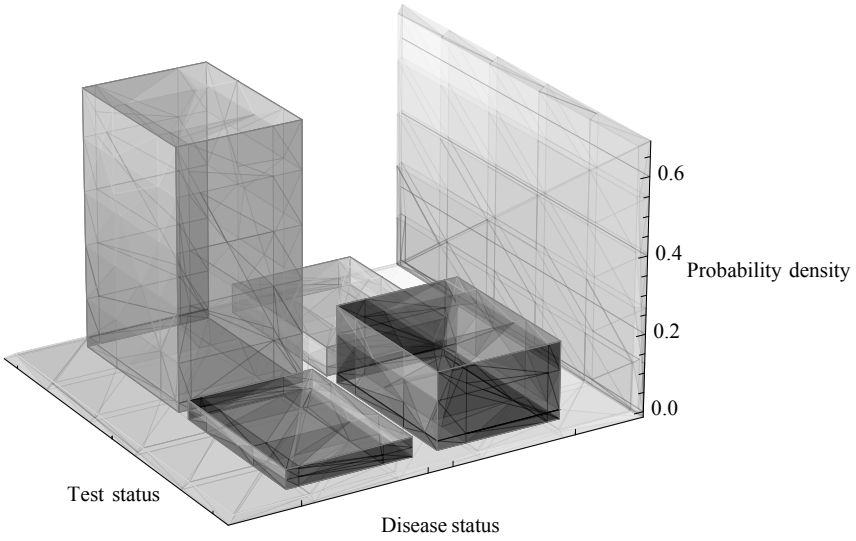

Figure 6.3: The joint density of the data and the parameter for the disease example described in section 6.3.2. When we uncover that the test result is positive, we are confined to look at the bars in dark grey; finding that the probability that an individual is diseased is significantly higher than 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.**

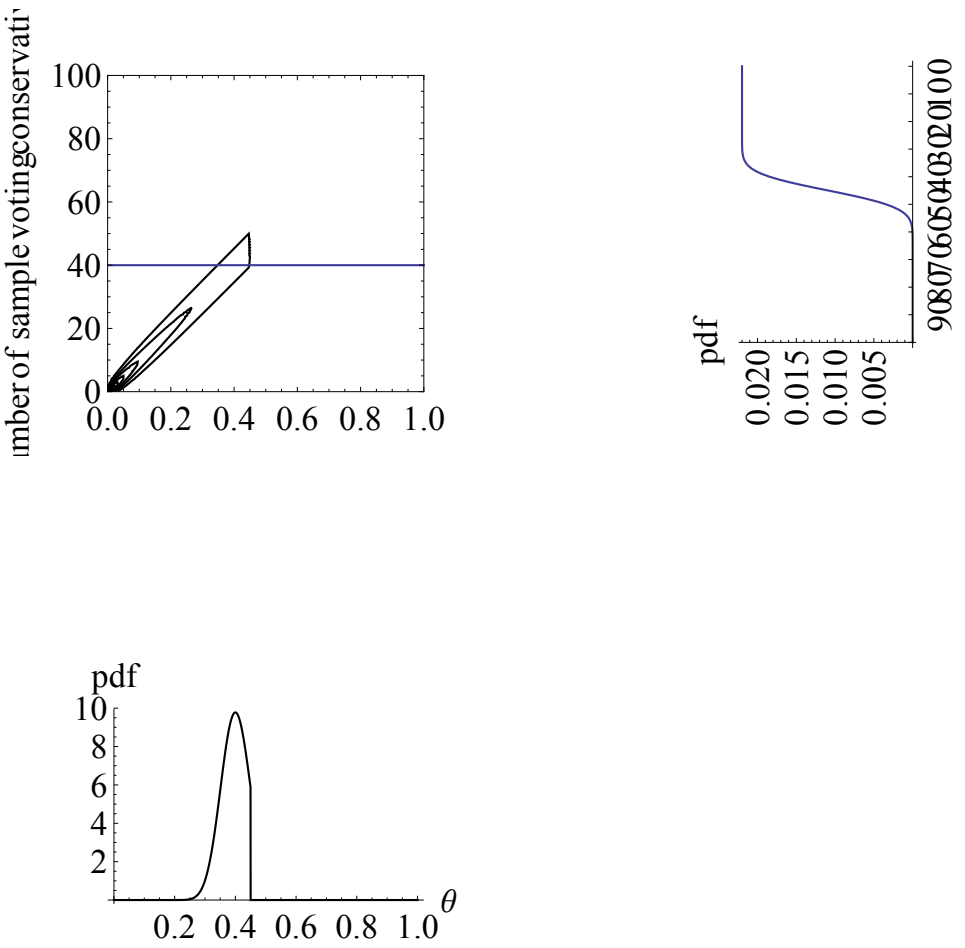

Figure 6.4: Top-left: a contour plot of the joint density of the voting example described in section 6.4. Top-right: the marginal density of $p(data)$ obtained by summing across all values of θ. Bottom-left: the posterior obtained by summing the joint density across the line shown at 40. Note that in 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 2.6 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 6.4.2. We importantly note that this prior is improper, since $\int_0^\infty \frac{1}{\sigma^2} d\sigma^2 \to \infty$, and hence must take case when interpreting the resultant 'posterior' distribution (see section 5.7.1).

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.

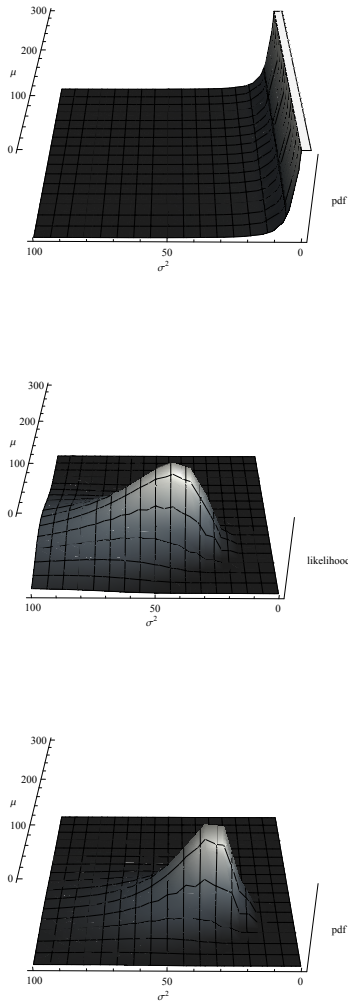

Figure 6.5: The prior, likelihood, and posterior distributions for the mean and variance of IQ example described in section 6.4.2.

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 6.4.2 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

6.7 Appendix

Part II

Analytic Bayesian methods

Chapter 7

An introduction to distributions for the mathematically-un-inclined

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

Chapter 10

Hierarchical models

Part IV

Regression analysis and 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.