Bayesian book

Ben Lambert

October 25, 2014

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

Chapter 3

The posterior - the goal of Bayesian inference

Chapter 4

Likelihoods

4.1 Chapter Mission statement

At the end of this chapter a reader will know how to choose an appropriate likelihood model for most situations. 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 for Bayesians. Further the choice over which likelihood to use for a given situation is often difficult to those unfamiliar with statistics. This chapter will provide practical guidance to likelihood choice, which should allow the student to be confident in their choice of model. 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

Figure 4.1: Insert bar chart here of the number of heads along the x axis - 0,1,2 - and the associated probability of each of these outcomes as being the bar height - (1/4,1/2,1/4).

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. In this idealised model[1] of the situation the probability of the coin coming up as heads twice is simply $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Written mathematically, this is simply the likelihood:

$$P(HH|\theta = \frac{1}{2}, \textit{Simple Model}) = \frac{1}{4} \tag{4.4}$$

Hence, the likelihood simply summarises the possibility of obtaining a given set of data given a choice of model *and* choice of the model's parameter(s). If we continue to assume that the probability of a head, θ, is given by $\frac{1}{2}$, 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. The important thing to note however about figure 4.1 isn't the individual probabilities, it is that it represents as a whole a *proper* probability distribution. What do we mean by this? Well, the individual event probabilities are all greater than 0 and less than 1, and when we sum the individual probabilities together we get 1 overall. So in the case where we assume a particular value for θ, and keep it fixed there, the likelihood really is simply just a probability distribution. So, why do we bother changing its names from a 'probability' to a 'likelihood'? That is to be explained in the next section...

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

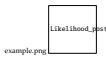

Figure 4.2: An example posterior distribution for the probability of a heads.

Figure 4.3: The x-axis here is theta, ranging between 0 and 1, assuming that one head is obtained this graphs the likelihood, which does not sum to 1.

4.4 Why is a likelihood not a probability for Bayesians?

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

The reason is that in Bayesian inference, we *don't* keep the parameters of our model fixed! In Bayesian analysis, it is the *data* that is fixed, and we vary the parameters. Why do we do this? It is because a posterior probability distribution is a probability of a parameter in a model taking on a particular value, across a range of different parameter values. 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. Notice that the x-axis in figure 4.2 is the value of θ - the probability of a heads being obtained. In order to get this posterior probability, $P(\theta|data)$, for each value of theta, we use Bayes' rule in (4.3). This means that for each *different* value of θ, we calculate the first part of the numerator which is $P(data|\theta)$; meaning that we calculate this across a range of θ. If we assume that we have obtained two heads, and vary θ between 0 and 1 we can obtain the likelihood, which is shown in figure 4.3. On first glances in might appear like 4.3 is 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,

when we vary θ we find that, $P(data|\theta)$ is not a valid probability distribution! We thus introduce the term 'likelihood' to represent value of $P(data|\theta)$ 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: $\theta \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. In this circumstance, we can calculate the probability of the number of heads, X, occurring as:

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

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

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

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 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). The important thing to see here is 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 proper probability distribution. By contrast, when we hold the number of heads fixed, and vary the parameter θ, moving down each column,

[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.

summing the entries, we find that the values do not sum to 1. Hence, when we vary θ, we are not dealing with a proper probability distribution, thus 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}. In each of the rows, the value of *theta* is held constant, meaning that $P(data|\theta)$ is a proper probability distribution and thus the probabilities sum to 1. However, in the columns, the data - the number of heads thrown - is held constant, and thus the probabilities do not sum to 1, and we thus we are better off viewing these data as likelihoods, since they do not satisfy the properties of a proper probability distribution.

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 [2]. 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 a 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 goes on to suggest 16 reasons, other than prediction, to build a model, of which I list a selected few below:

- Explain
- 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 I believe that this list encapsulates the majority of them. 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 model for likelihood?

Bayesians are acutely aware that their models are wrong. At best, they represent an abstraction from reality, and at worst, they can provide very misleading descriptions of a real phenomenon. Before we use a model for prediction, we should always require that a model is capable of *explanation* of the past and present. With this in mind, I propose the following course of action to specifying a statistical model of which likelihood forms a core part.

1. Write down the real life behaviour/data patterns that your model should be capable of explaining.
2. Write down the assumptions that we believe are reasonable in order to arrive at such a model.
3. Search for an appropriate/similar model in the literature.
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 simple statistical model which is used 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 assume that we start off by assuming that we have a sample of only one person, of which 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 a likelihood model which we specify is to be able to explain probabilistically the relative likelihood that an individual has a disease, as well as make predictions about disease incidence in new samples of individuals. 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 specify 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 be able to write down a single rule which yields (4.10) or (4.11) respectively, dependent on whether $X = 0$ or $X = 1$ respectively. It transpires that we can do this via the following rule:

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

Note that in (4.12) that $\alpha \in \{0, 1\}$ refers to the numeric value taken by the variable X. The function given in (4.12) is known as a Bernoulli probability density.

Although this rule for calculating a probability of obtaining a disease status of value α at first looks relatively complex, we see that it reduces to (4.10) and (4.11) if the individual doesn't or does have the disease respectively:

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

Since (4.12) yields the probability of obtaining any possible value of data, for a given θ, we conclude that this expression is the likelihood. However, we need to be careful, strictly this expression is only a likelihood if we hold the data fixed and vary the parameter θ. Figure 4.4 shows that for

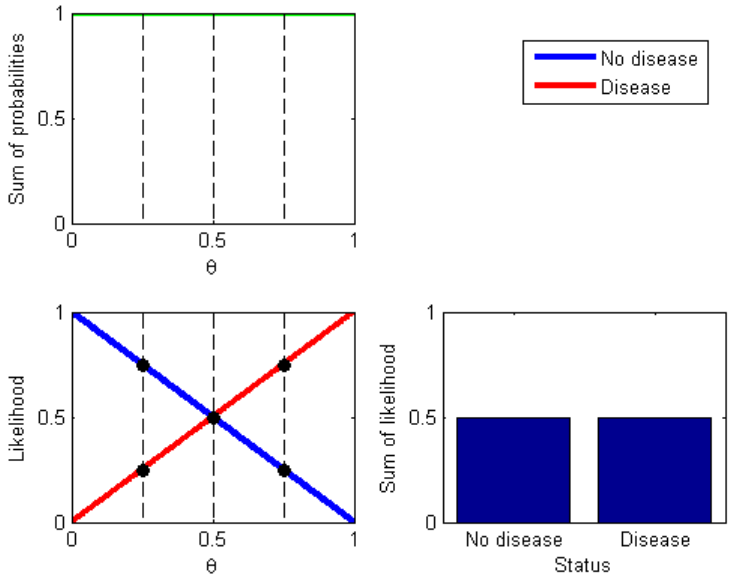

Figure 4.4: The likelihood function as theta varies for the case of the two possible data. For a fixed value of θ we find that the sum across the values of the probability is always 1. This is because when viewed in this way, the likelihood is really a discrete probability density across the two values which x can take on. However, when we hold the data fixed (choose either the red or blue line) and sum the likelihood horizontally across the values of *theta* we do not find that the sum is generally equal to 1.

a fixed value of *theta* the sum (here we mean the vertical sum) of the two probability densities is always equal to 1. However, when we hold the data fixed, (therefore choosing either the red or green line), we find that the sum of likelihoods across values of *theta* does not in general equal 1; again demonstrating that when varying parameters that likelihood is not a valid probability density.

4.6.2 A likelihood model for disease prevalence of a group

Now we imagine that instead of this solitary individual, we have a group of N individuals. What we would like to do is to calculate the develop a model which will tell us the probability of obtaining Z disease cases within

our sample. We would also like to be able to use our model to predict the most likely number of individuals who have the disease in a sample, for a given value of the parameters[3].

In order to write down an idealised model we first of all need to make some assumptions to simplify the situation. 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 to make 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 pre-existing ailment. For example, if the advert that attracted participants reads 'Psychological experiment on sleep disorders: participants wanted'. In this case we might suspect that there would be an over-presence of insomniacs than is found in the population as a whole. In other words one individual's status would convey extra information about the probability that another in the sample has the disease. 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. In statistical language, we assume that individuals in our sample are 'identically distributed'.

With our two assumptions in hand, that the individuals in our sample are independent and identically distributed, we can begin to formulate a model for the probability of obtaining Z disease-positive individuals out of a total of N individuals. Since we have assumed that the individuals are independent of one another[5], we can treat each person individually and re-use our model that we found in section 4.6.1. So, for each individual we model the probability of their disease status X for a given value of *theta* as:

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

[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 status

[5]Apart from their individual dependence on the population disease prevalence θ.

Note that in (4.15) the $\alpha \in \{0, 1\}$ refers to a particular numeric value taken by the variable X. If we treat the individuals as independent of one another this means that we can get the overall probability by multiplying together the individual probabilities (include reference back to discussion of probabilities). In words, we need the probability that the first person has disease status $X1$ *and* that the second person has status $X2$. When we use the word *and* in probability, this is normally translated into *multiply* in mathematical language.

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

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.

Now we use our second assumption - that of identically distributed individuals - to simplify the above expression by assuming that all individuals have the same pre-experimental disposition to the disease θ:

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

In (4.17) we have obtained the third line merely by using the simple exponent rule: $a^b \times a^c = a^{b+c}$, for the two 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 $X1$ and $X2$ to Z by:

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

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

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

In order to complete our probability model we need to try 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, we can get round this by taking an aside to note that when we expand a quadratic factor we get the following:

$$(a+b)^2 = a^2 + 2ab + b^2 = a^2 + 2a^1b + a^0b^2 \tag{4.22}$$

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

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

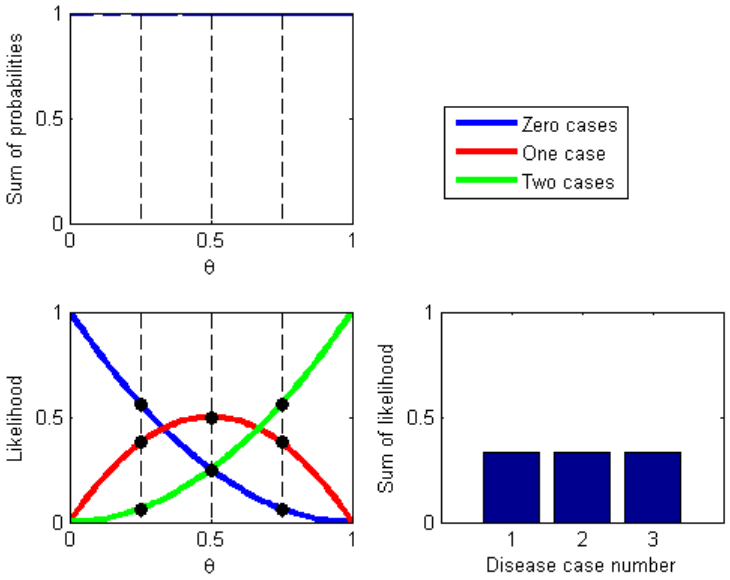

Figure 4.5: The likelihood function as theta varies for the case of the two possible data. For a fixed value of θ we find that the sum across the values of the probability is always 1. This is because when viewed in this way, the likelihood is really a discrete probability density across the two values which x can take on. However, when we hold the data fixed (choose either the red or blue line) and sum area under the likelihood curve across the values of *theta* we do not find that the sum is generally equal to 1.

This likelihood function is illustrated for the three possible numbers of disease cases found in our sample of 2 individuals, in figure 4.5.

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

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

If we continue to assume independence then we can simplify the probabilities to:

$$
\begin{aligned}
P(Z=0|\theta) &= \ P(X_1=0|\theta)P(X_2=0|\theta)P(X_3=0|\theta) \\
P(Z=1|\theta) &= 3P(X_1=1|\theta)P(X_2=0|\theta)P(X_3=0|\theta) \\
P(Z=2|\theta) &= 3P(X_1=1|\theta)P(X_2=1|\theta)P(X_3=0|\theta) \\
P(Z=3|\theta) &= \ P(X_1=1|\theta)P(X_2=1|\theta)P(X_3=1|\theta)
\end{aligned}
\tag{4.26}
$$

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 $(1+x)^3$. Hence, we can again rewrite the likelihood using the binomial expansion notation:

$$
P(Z=\beta|\theta) = \binom{3}{\beta}\theta^{\beta}(1-\theta)^{3-\beta} \tag{4.27}
$$

We recognise a pattern in the likelihoods of (4.24) and (4.27) which allows us to deduce that, for a sample size of N, that the model likelihood is given by:

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

The likelihood function given in (4.28) is known as the binomial probability distribution.

If we had data, then we could test whether the assumptions which we have made are appropriate by calculating the model-implied-probability of this outcome. For example, if we had a sample of 100 people of which 10 had the

disease, 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 the likelihood in 4.28:

$$P(Z \geq 10|\theta = 0.01) = \sum_{Z=10}^{100} \binom{100}{Z} 0.01^Z (1 - 0.01)^{100-Z} = 7.63 \times 10^{-8} \quad (4.29)$$

We have summed over all the disease cases from 10 to 100 here, because we wanted to work out 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 a classical hypothesis test, which we will dispense with later on, but for now it seems a reasonable way of testing our model.

The probability which we found in this case was 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 that given by 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

Here we are tasked with formulating a model for an intelligence test score for a group of individuals for which we have data. We are told that the

Figure 4.6: A figure with LHS showing a normal distribution with the area to the right of 140 indicated by shading. The RHS shows a normal CDF with the value of 140 indicated.

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 $\sigma^2 = 10$ 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[6] which average out to result in an intelligence score, then the normal distribution provides a reasonable approximation to the distribution of test scores. In which case, we might assume that a normal distribution might be reasonable to assume for our likelihood function for an individual's test score, X:

$$P(X = \alpha|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha-\mu)^2}{2\sigma^2}} \tag{4.30}$$

If we obtain an individual within our sample who has achieved a test score of 140, we can ask what is the probability of achieving a result as extreme as this, using our idealised model, by simply integrating the probability density (this is the continuous analogue to the discrete summing that we did in (4.6.2)):

[6]Technically a infinite number of factors which additively result in the overall score.

$$\begin{aligned} P(X \geq 140|\mu = 70, \sigma^2 = 10) &= \int_{140}^{\infty} \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(\alpha-70)^2}{2\times 10}} \mathrm{d}\alpha \\ &= 1 - \Phi\left(\frac{140-70}{\sqrt{10}}\right) \approx 0 \end{aligned} \tag{4.31}$$

In (4.31), Φ stands for the value of the *standard* normal cumulative distribution function[7] at the value of 140 (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[8], then we might assume *independence* for our data. We might also assume that all individuals come from the same population. These two assumptions - independence and identical distribution - are equivalent to saying that the individually were randomly sampled from the population. We can use these two assumptions to calculate the joint probability density for a sample of N individuals by simply multiplying together the individual densities:

$$P(X_1 = \alpha_1, X_2 = \alpha_2, ..., X_N = \alpha_N|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha_i-\mu)^2}{2\sigma^2}} \tag{4.32}$$

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 (4.31) since we would have to integrate across all individual's variables. For example if we obtained a sample of $\{75, 71\}$ then we could obtain the probability of obtaining a sample as extreme as this under our model:

[7] 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.

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

$$
\begin{aligned}
P(X_1 \geq 75, X_2 \geq 71 | \mu = 70, \sigma^2 = 10) &= \int_{75}^{\infty} \int_{71}^{\infty} \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(\alpha_1 - 70)^2}{2\times 10}} \times e^{-\frac{(\alpha_2 - 70)^2}{2\times 10}} \mathrm{d}\alpha_1 \mathrm{d}\alpha_2 \\
&= \int_{75}^{\infty} \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(\alpha_1 - 70)^2}{2\times 10}} \mathrm{d}\alpha_1 \int_{71}^{\infty} \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(\alpha_2 - 70)^2}{2\times 10}} \mathrm{d}\alpha_2 \\
&= (1 - \Phi\left(\frac{75 - 70}{\sqrt{(10)}}\right)) \times (1 - \Phi\left(\frac{71 - 70}{\sqrt{(10)}}\right)) \approx 0.02
\end{aligned}
\tag{4.33}
$$

In (4.6.3), the integrals are easily separated because we have assumed that the individuals' test scores are independent. If we had not assumed this, then we would have to have done a double integral of a multivariate normal distribution. Again, we have found that it is unlikely that we would have obtained two observations as extreme as this, if our model's assumptions are in fact true. This should prompt us to go back and re-examine these steps.

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 making assumptions which allow us to simplify the situation, then look for a likelihood which is a good fit to the simplified model, which is then used to test the validity of the assumptions with the sample of data obtained. If the model struggles to explain the data, then we should go back and iteratively modify, then test our model, until it adequately can be thought to represent the real life data obtained.

However, it should be re-emphasised that by its nature, a model is always a simplification of reality. As such, no one model is *correct*. There are often many models that could be used to explain the data which we have to hand. We should always take care to test each of these against its ability to explain the aspect of the data 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 can not, 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, 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 look at how this estimation strategy can yield estimates of parameters, and we will also go on to look at how these estimates can be used to make inferences about what is going on in the population as a whole.

The principal used in 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. The resultant joint probability distribution gives the highest probability of achieving those observations given this choice of parameters. 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 the population from which our sampled individuals came. As mentioned previously, it is uncommon that such a thing be known before carrying out an analysis. In fact, the reason for undertaking a statistical analysis of the situation is often to try to estimate the underlying prevalence of disease within the population.

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.

If we find that under a sample of 100 individuals, that 10 tested positively (and assuming for simplicity that there are no false-positives), and we make the same assumptions as in section 4.6.2 - that of independence and identically-distributed - then we can write down the overall likelihood function using (4.28) as:

$$L(\theta|data) = \binom{100}{10}\theta^{10}(1-\theta)^{100-10} \tag{4.34}$$

Remember, that since we are varying θ and holding the data constant here, that (4.34) is a *likelihood*, not a probability. We then need to simply choose θ so that we can maximise the likelihood. We could simply differentiate (4.34) 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):

$$l(\theta|data) = log\binom{100}{10} + 10log(\theta) + 90log(1-\theta) \tag{4.35}$$

Where to get the result (4.35), we have used the log rules:

$$\begin{aligned} log(ab) &= log(a) + log(b) \\ log(a^b) &= blog(a) \end{aligned} \tag{4.36}$$

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

Figure 4.8: A figure displaying log-likelihood as a function of theta, maximised at 1/10.

$$\frac{\partial l}{\partial \theta} = \frac{10}{\hat{\theta}} - \frac{90}{1-\hat{\theta}} = 0 \tag{4.37}$$

If we set the derivative to 0 we then obtain the maximum likelihood *estimator*, $\hat{\theta} = \frac{1}{10}$ (see figure 4.8). We have placed hats on θ to emphasise that this is an estimator for the variable. An estimator is a mathematical function which outputs an estimate of a parameter in our model.

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 of the disease prevalence exactly equal to that in our sample:

$$\hat{\theta} = \frac{\beta}{N} \tag{4.38}$$

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:

$$L(\mu, \sigma^2 | X_1 = 75, X_2 = 71) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(75-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(71-\mu)^2}{2\sigma^2}} \tag{4.39}$$

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

$$l(\mu, \sigma^2 | X_1 = 75, X_2 = 71) = 2log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(75-\mu)^2}{2\sigma^2} - \frac{(71-\mu)^2}{2\sigma^2} \quad (4.40)$$

Where we have again used the log rules in (4.36) to achieve (4.39). We can now proceed to differentiate (4.40) with respect to both variables, holding the other constant, setting each to 0, to achieve the following:

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{(75-\hat{\mu})}{\hat{\sigma^2}} + \frac{(71-\hat{\mu})}{\hat{\sigma^2}} = 0 \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{1}{\hat{\sigma^2}} + \frac{(75-\hat{\mu})^2 + (71-\hat{\mu})^2}{2\hat{\sigma^4}} = 0 \end{aligned} \quad (4.41)$$

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

$$\hat{\sigma^2} = \frac{1}{2}\left[(75-73)^2 + (71-73)^2)\right] = 4 \quad (4.42)$$

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:

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X} \quad (4.43)$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N} (X_i - \bar{X})^2 = s^2 \quad (4.44)$$

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

[9] Albeit a biased estimator of the population variance. The unbiased estimator would divide by 1.

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 **??**). 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 likelihood near the ML value. 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 the more certain we can be about making conclusions on the population. 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 complete description of parameters, and thus a less opaque manor in which to draw inferences.

4.10 Chapter summary

We should now understand what is meant by a likelihood, and how to build probabilistic models of real life processes. Of course, 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 from 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.

[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 construct and understand 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. The likelihood tells us what the data implies are most likely values of the parameters, but without a prior weighting of the probability of each of these parameter values, it is not that useful. Bayesian analysis requires that we specify this pre-data weighting ascribed to different parameter values in a full 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 *sub-*

jectivity. 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 less so. 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?

Two representations of a prior. The data converts our prior knowledge in a logical way to a posterior distribution. Alternatively, the prior tells us how to convert the likelihood into a proper probability distribution.

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 [4]. 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.

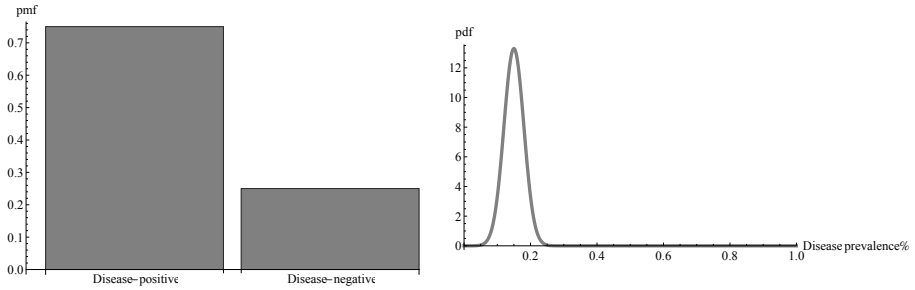

Figure 5.1: Left - a prior for a doctor's pre-testing diagnostic probability of an individual having a disease. Right - a prior which represents pre-sample uncertainty in disease prevalence.

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 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 *population* interpretation where the current value of a parameter is the result of a draw from a true population distribution; alternatively, in the more subjective *state of knowledge* interpretation, we specify our knowledge and uncertainty in a parameter as if regarding it as a draw from a probability distribution [3]. 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 likelihood before the results of a blood test become available, then probability 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 the 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 illicit our prior expectations as to the value of a parameter. For example, we could use the probability distribution for the proportion of individuals having a particular disorder in figure 5.1 to estimate what we ascribe to be the probability of its prevalence in the population; finding it to have a mean of approximately 15%.

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 agnostic with regards to final face up side 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 be 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:

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

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

This would surely negate the need for specification of a prior, and thwart all attempts to denounce Bayesian statistics as *subjective*. So won't don't we do just that? There is a pedantic, mathematical argument against this, which is $P(\theta)$ 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 $\int_{-\infty}^{+\infty} P(\theta)\mathrm{d}\theta \rightarrow \infty$, 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 that the majority of coins are relatively unbiased - that 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 would by looking at the likelihood alone, be able to make a more reasonable conclusion on the fairness of the coin. 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.

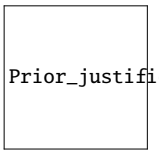

Figure 5.2: A figure showing in the left hand panel the Venn diagrams when we assume that a) the coin is fair with a quarter of the area shaded, and the other half not. b) Biased, and the majority of the figure is shaded - corresponding to a high overlap between data and parameter. The right hand panel 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.

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* nature of the 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 the chapter changes 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 [5] and [1], 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

[2]Taken from Bolstad's great introduction to Bayesian statistics [1].

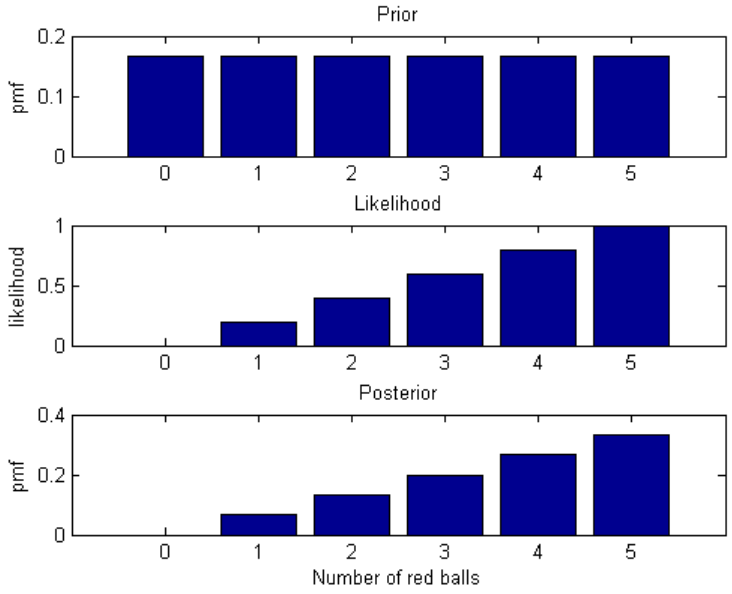

Figure 5.3: The prior, likelihood and posterior for the urn of balls example described in 5.6.1. The prior in the upper panel gives uniform weighting to all possible numbers of red balls. This is then multiplied by the likelihood (in the middle panel) at each number of balls, and normalised to make the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

on summation gives us $P(data) = \frac{1}{2}$, which we use to create a proper probability distribution for the posterior in the last column. For a mathematical description of this process see section 5.10.1.

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).

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**

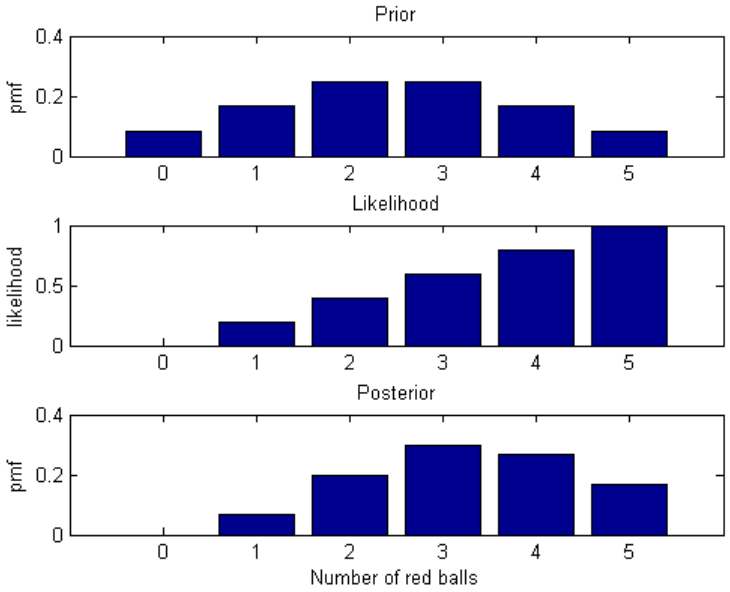

Figure 5.4: The prior, likelihood and posterior for the urn of balls example described in 5.6.1. The prior in the upper panel gives more weighting to more equal numbers of red and white balls. This is then multiplied by the likelihood (in the middle panel) at each number of balls, and normalised to make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

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 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 we deem to be 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.

5.6.3 The numerator of Bayes' rule determines the shape

We notice that 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:

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

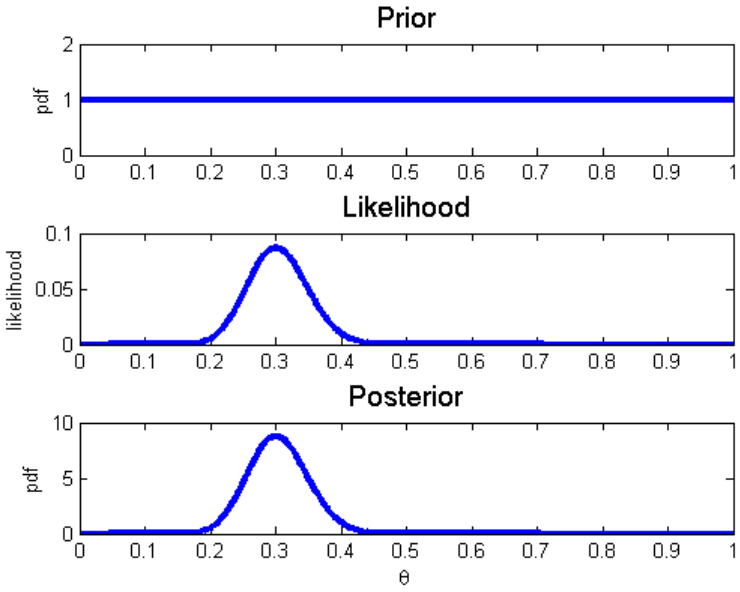

Figure 5.5: The prior, likelihood and posterior for the disease proportion example described in section 5.6.2. Each point in θ along the continuous prior curve (top panel) is multiplied by the corresponding value of likelihood (middle panel), to form the numerator of Bayes' rule. The numerator is then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

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

The shape of the posterior is given 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 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 can be used to assemble pre-experimental knowledge in a logical manner.

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:

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

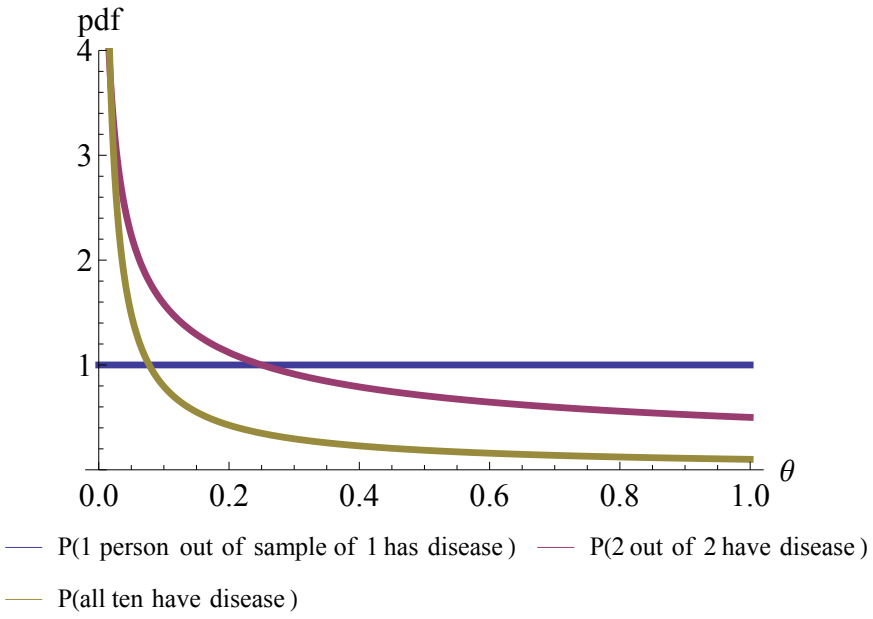

Figure 5.6: The probability density for obtaining all diseased individuals within sample sizes of 1, 2 and 10 respectively. Starting out with a flat prior for the probability that one individual has a disease has resulted in non-flat priors for the other 2 probabilities.

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 the type of prior shown in figure 5.6 for the probability of these two individuals having the disease. Furthermore, when we consider the probability that within a sample of ten individuals, all are diseased, we see that a flat prior for θ implies an even more accentuated prior for 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. In general, 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. However, it is important to key this property of flat priors in mind when conducting analyses. All priors contain some information, so we prefer the use of the terms "vague" or "diffuse" to represent situations where a premium is placed on drawing conclusions from only the data at hand.

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

Whilst uniform priors are relatively straightforward to specify when we aim to infer about a parameter which is bounded - such as in the previous example where $\theta \in \{0, 1\}$, or in the case of discrete parameters - we run into issues for parameters which have no predefined range. An example of this would be if we were aiming to determine the mean, μ, time of onset of lung cancer for individuals who develop the disease, after they begin to smoke. If we remove all background cases (assumed not to be caused by smoking), then μ has a lower bound of 0. However, there is no obvious point at which to draw an upper bound. A naive solution to this would be to use a prior for $\mu \sim Unif(0, \infty)$. This solution, although at first appears to be reasonable, is not viable for two reasons; one statistical, another which is practical. The statistical reason is that $\mu \sim Unif(0, \infty)$ is not a valid probability density, because any non-zero constant value for the pdf will mean that the area under the curve is ∞ because the μ axis stretches out forever. The common sense argument is that we would never ascribe the same likelihood to an individual having the onset of lung cancer after 10 years as we would to 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 be a proper 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

[4] Although not assured.

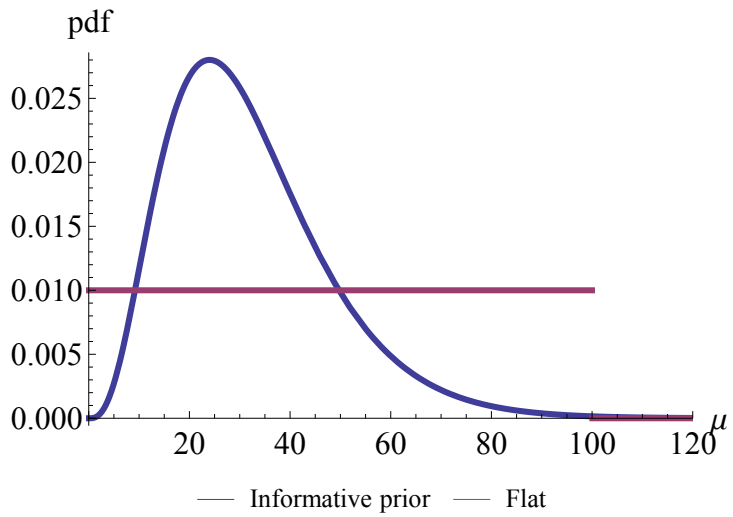

Figure 5.7: Two viable prior distributions for the average time taken before the onset of lung cancer after patients begin smoking.

range of μ, and allow the pdf to be zero elsewhere (see the right hand panel of 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 on conclusions drawn due to existing prejudices. There are however occasions when the choice of prior acknowledges that the analysis is based on more than the latest data. This choice of prior can be used to incorporate previous data, conclusions from older studies, or to include expert opinion.

In cases where data is available from previous studies, the construction of a prior can proceed methodically via a method that is known as *moment-matching*. Suppose that we have the data shown in figure 5.8 for SAT scores of past participants of a particular class. We might think that to a reasonable approximation the data could be modelled as having come from a normal distribution[5]. We typically characterise normal distributions

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

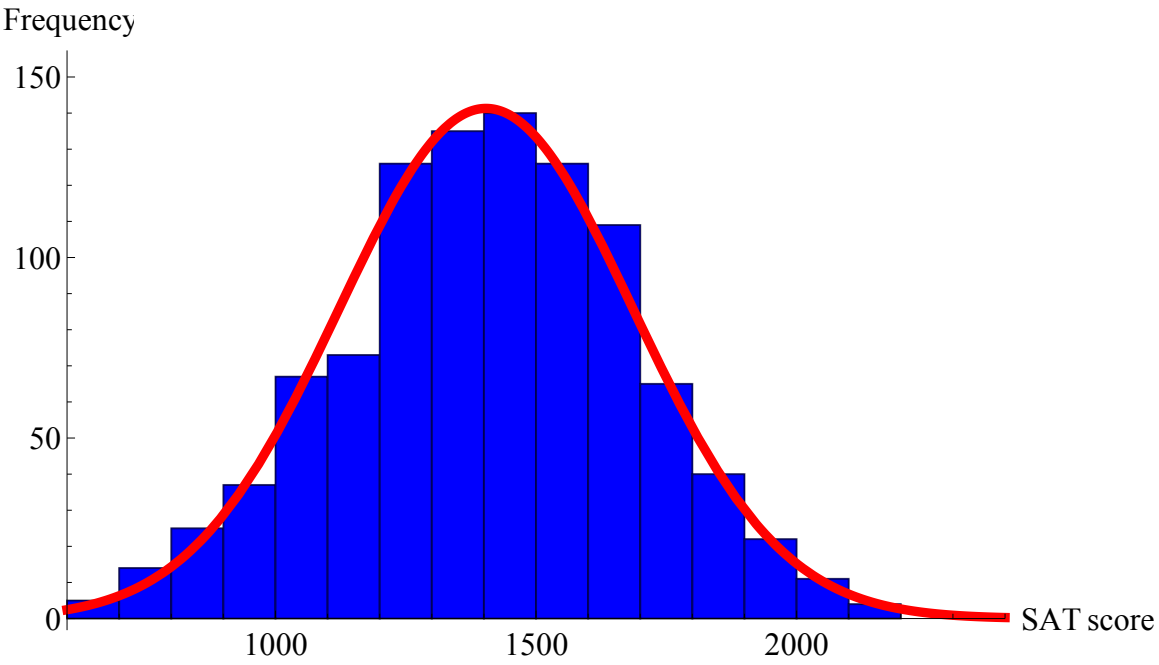

Figure 5.8: The SAT scores for past students of a class. The mean and variance of this hypothetical sample are 1404, and 79,716 respectively, which are used to fit a normal distribution to the data, and is shown in red.

via two parameters: its mean, μ, and variance, σ^2. In moment-matching a normal prior to this previous data, we choose the mean and variance to be equal to their sample equivalents, in this case $\mu = 1404$, and $\sigma^2 = 79,716$, respectively.

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

5.7.3 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 [4] 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} correspond to the values of a standardised normal variable which satisfy $\Phi(z_{25}) = 0.25$ and $\Phi(z_{75}) = 0.75$ (Φ being the standard normal cumulative density function). 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).

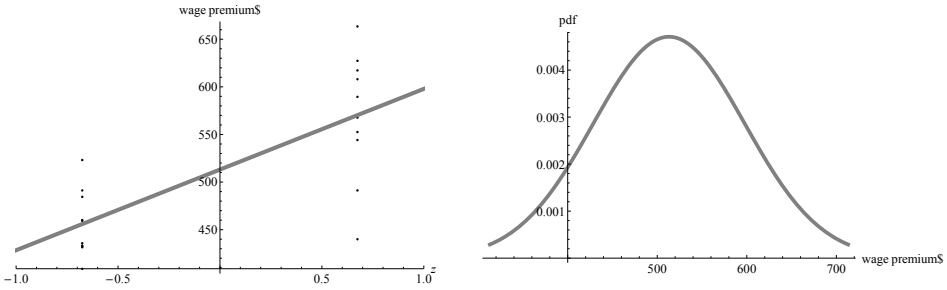

Figure 5.9: Hypothetical data for the 25th and 75th percentiles of the estimated wage premium from 10 experts. In the left hand panel we regress these percentiles on the corresponding percentiles from a standard normal distribution, yielding estimates of the mean and variance of a normal prior, which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

5.8 A strong model is not heavily influenced by priors

Returning to the example of section 5.6.2 of estimated the prevalence of a disease within a population, we now examine the effects of using an informative prior on the analysis. Suppose we choose a prior which represents our pre-data view that the prevalence of a disease within a particular population is high (see the topmost row of figure 5.10). If we only have a sample size of 10, and obtain 1 individual in our sample who tests positive for the disease we see that the posterior is located roughly equidistant between the peaks of the prior and likelihood functions respectively (see the left hand column of figure 5.10). Now if we increase the sample size to 100, keeping the same percentage of individuals who are disease-positive within our sample, we then find that the posterior is peaked much closer to the position of the likelihood (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

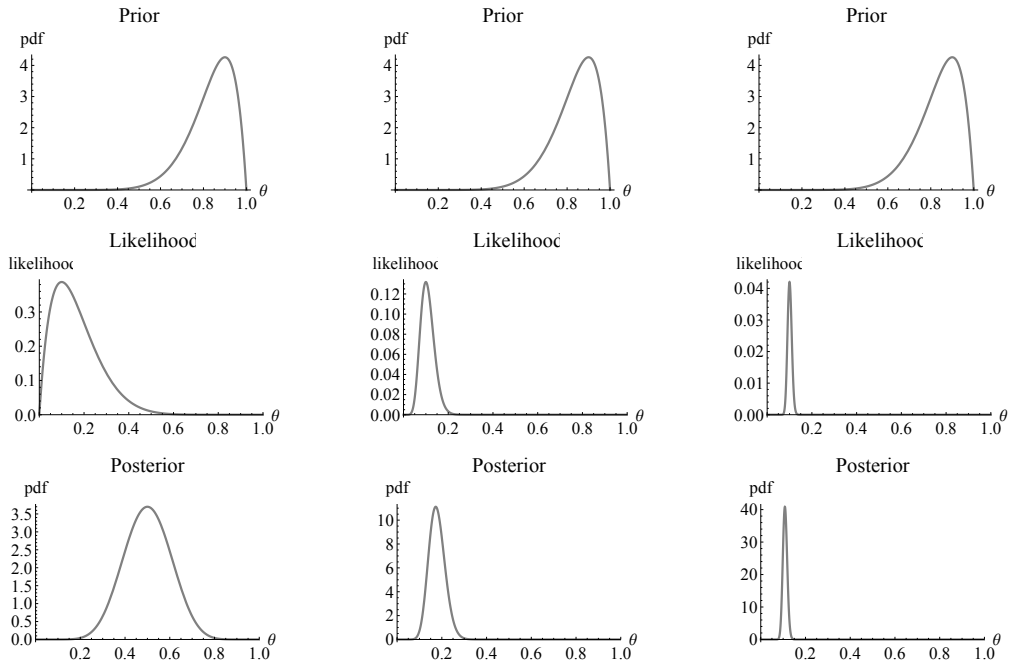

Figure 5.10: The effect of increasing sample size on the posterior density for the prevalence of a disease in a population. The leftmost column has N=10, the middle N=100, and the rightmost N=1,000. All three have the same proportion of disease cases in the sample.

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 priors.

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. In general, we also realise that as the amount of data increases, or a better model is chosen, then the posterior density is least 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 study 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 simply the inverse of $Y = \theta^2$. 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 difficulty is in the denominator

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

Chapter 10

Hierarchical models

Bibliography

[1] William M Bolstad. *Introduction to Bayesian statistics*. John Wiley & Sons, 2007.

[2] Joshua M Epstein. Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12, 2008.

[3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[4] Jeff Gill. *Bayesian methods: A social and behavioral sciences approach*. CRC press, 2007.

[5] Wayne Stewart and Sepideh Stewart. Teaching markov chain monte carlo: Revealing the basic ideas behind the algorithm. *PRIMUS*, 24(1):25–45, 2014.