Bayesian book

October 4, 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

Likelihoods

3.1 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.

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

3.2 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.

3.3 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:

$$P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)} \tag{3.2}$$

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:

$$P(data|\theta) = Probability(data|\theta, Model\ Choice) \tag{3.3}$$

What (3.3) 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, (3.3) 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 3.1: 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:

$$P(HH|\theta = \frac{1}{2}, \textit{Simple Model}) = \frac{1}{4} \tag{3.4}$$

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 3.1 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 3.1 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 3.2: An example posterior distribution for the probability of a heads.

Figure 3.3: 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.

3.4 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 (3.3) 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 3.2. Notice that the x-axis in figure 3.2 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 (3.3). 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 3.3. On first glances in might appear like 3.3 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 3.3 are non-negative; which is what we require. However, if we look at the area underneath the curve in figure 3.3, 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:

$$\mathcal{L}(\theta|data) = P(data|\theta) \tag{3.5}$$

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:

$$P(X = 0|\theta) = P(TT|\theta) = P(T|\theta) \times P(T|\theta) = (1 - \theta)^2 \tag{3.6}$$

$$P(X = 1|\theta) = P(HT|\theta) + P(TH|\theta) = 2 \times P(T|\theta) \times P(H|\theta) = 2\theta(1 - \theta) \tag{3.7}$$

$$P(X = 2|\theta) = P(HH|\theta) = P(H|\theta) \times P(H|\theta) = \theta^2 \tag{3.8}$$

In (3.6), 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 (3.8) since there are two ways of getting one head: {HT,TH}.

We can represent the corresponding values of likelihood/probability as in table 3.1. 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 3.1: 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.

3.5 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 [1]. 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.

3.6 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, practically how do we go about specifying a likelihood for a given situation? To answer this we will start with going through a simple example.

3.6.1 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:

$$X = \begin{cases} O & , \textit{No disease} \\ 1 & , \textit{Positive diagnosis} \end{cases} \tag{3.9}$$

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:

$$P(X = 0|\theta) = (1 - \theta) \tag{3.10}$$
$$P(X = 1|\theta) = \theta \tag{3.11}$$

However, we would like to be able to write down a single rule which yields (3.10) or (3.11) respectively, dependent on whether $X = 0$ or $X = 1$ respectively. It transpires that we can do this via the following rule:

$$P(X = \alpha|\theta) = \theta^{\alpha}(1 - \theta)^{1-\alpha} \tag{3.12}$$

Note that in (3.6.1) that $\alpha \in \{0, 1\}$ refers to the numeric value taken by the variable X. 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 (3.10) and (3.11) if the individual doesn't or does have the disease respectively:

$$P(X = 0|\theta) = \theta^{0}(1 - \theta)^{1} = (1 - \theta) \tag{3.13}$$
$$P(X = 1|\theta) = \theta^{1}(1 - \theta)^{0} = \theta \tag{3.14}$$

Since (3.6.1) 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 3.4 shows that for

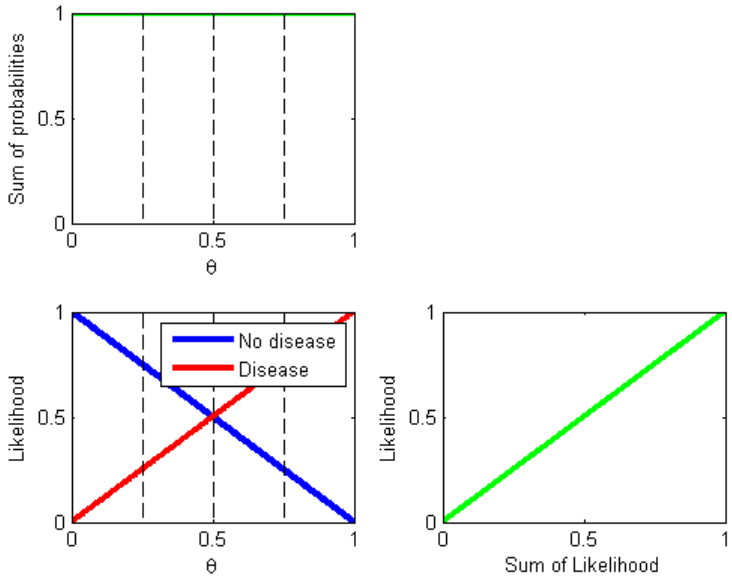

Figure 3.4: The likelihood function as theta varies for the case of the two possible data. For a fixed value of θ we find that the sum across the values of the probability is always 1. This is because when viewed in this way, the likelihood is really a discrete probability density across the two values 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.

3.6.2 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 3.6.1. So, for each individual we model the probability of their disease status X for a given value of *theta* as:

$$P(X = \alpha|\theta) = \theta^{\alpha}(1-\theta)^{1-\alpha} \tag{3.15}$$

[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 (3.6.2) 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.

$$\begin{aligned} P(X_1 = \alpha_1, X_2 = \alpha_2|\theta_1, \theta_2) &= P(X_1 = \alpha_1|\theta_1) \times P(X_2 = \alpha_2|\theta_2) \\ &= \theta_1^{\alpha_1}(1-\theta_1)^{1-\alpha_1} \times \theta_2^{\alpha_2}(1-\theta_2)^{1-\alpha_2} \end{aligned} \tag{3.16}$$

In (3.6.2) 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:

$$\begin{aligned} P(X_1 = \alpha_1, X_2 = \alpha_2|\theta) &= P(X_1 = \alpha_1|\theta) \times P(X_2 = \alpha_2|\theta) \\ &= \theta^{\alpha_1}(1-\theta)^{1-\alpha_1} \times \theta^{\alpha_2}(1-\theta)^{1-\alpha_2} \\ &= \theta^{\alpha_1+\alpha_2}(1-\theta)^{2-\alpha_1-\alpha_2} \end{aligned} \tag{3.17}$$

In (3.6.2) 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:

$$Z = X_1 + X_2 \tag{3.18}$$

We can then use (3.6.2) to generate the respective probabilities.

$$\begin{aligned} P(Z=0|\theta) &= P(X_1=0, X_2=0|\theta) = \theta^{0+0}(1-\theta)^{2-0-0} = (1-\theta)^2 \\ P(Z=1|\theta) &= P(X_1=1, X_2=0|\theta) + P(X_1=0, X_2=1|\theta) = 2\theta(1-\theta) \\ P(Z=2|\theta) &= P(X_1=1, X_2=1|\theta) = \theta^{1+1}(1-\theta)^{2-1-1} = \theta^2 \end{aligned} \tag{3.19}$$

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 (3.6.2) as:

$$\begin{aligned} P(Z=0|\theta) &= \theta^0(1-\theta)^2 \\ P(Z=1|\theta) &= 2\theta^1(1-\theta)^1 \\ P(Z=2|\theta) &= \theta^2(1-\theta)^0 \end{aligned} \tag{3.20}$$

In (3.6.2) 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:

$$P(Z=\beta|\theta) \sim \theta^\beta(1-\theta)^{2-\beta} \tag{3.21}$$

The only problem with matching (3.6.2) with the previously obtained result is the factor of 2 on the middle line of (3.6.2). However, we can get round this by taking an aside to note that when we expand a quadratic factor we get the following:

$$(a+b)^2 = a^2 + 2ab + b^2 = a^2 + 2a^1b + a^0b^2 \tag{3.22}$$

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:

$$\binom{n+1}{3} \tag{3.23}$$

3.6.3 A likelihood model of wage determinants

3.7 The subjectivity of model choice

3.8 Maximum likelihood - a short introduction

3.9 Chapter summary

Bibliography

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