Chapter 3 - Likelihoods

October 3, 2014

Chapter 1

Likelihoods

1.1 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.

1.2 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.

1.3   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:

$$P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)} \tag{1.1}$$

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:

$$P(data|\theta) = Probability(data|\theta, Model\ Choice) \tag{1.2}$$

What (1.3) 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, (1.3) tells us the probability that we would have obtained our data.

But what does this mean in simple, everyday language? Imagine that we flip a *fair* coin. The most simple statistical model for coin flipping we can pick is to disregard the angle it was thrown at, as well as its height above the surface, along with any other details, and just pick the probability of the coin coming heads to be $\theta = \frac{1}{2}$. Furthermore, if a coin is thrown twice, we might choose to model the situation by assuming that the throwing technique is sufficiently similar between the two throws such that we can model each throw as independently having a probability of $\frac{1}{2}$. It's important to note that it is an assumption to forget about the throwing angle, as well as height of throw for each throw, and this forms part of our model of the situation.

Likelihood_fairCoin.png

Figure 1.1: 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).

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:

$$P(HH|\theta = \frac{1}{2}, \textit{Simple Model}) = \frac{1}{4} \tag{1.3}$$

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 1.1 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 1.1 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 1.2: An example posterior distribution for the probability of a heads.

Figure 1.3: The x-axis here is theta, ranging between 0 and 1, assuming that one he

1.4   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 (1.3) 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 1.2. Notice that the x-axis in figure 1.2 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 (1.3). 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 1.3. On first glances in might appear like 1.3 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 1.3 are non-negative; which is what we require. However, if we look at the area underneath the curve in figure 1.3, 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:

$$\mathcal{L}(\theta|data) = P(data|\theta) \qquad (1.4)$$

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:

$$P(X = 0|\theta) = (1 - \theta)^2 \qquad (1.5)$$
$$P(X = 1|\theta) = 2\theta(1 - \theta) \qquad (1.6)$$
$$P(X = 2|\theta) = \theta^2 \qquad (1.7)$$

In (1.5), 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 (1.7) since there are two ways of getting one head: {HT,TH}.

We can represent the corresponding values of likelihood/probability as in table **??**. 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 1.1: 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.

1.5 How to choose an appropriate likelihood

1.6 The subjectivity of model choice

1.7 Maximum likelihood - a short introduction

1.8 Chapter summary