Bayes' Rule With Python

A Tutorial Introduction to Bayesian Analysis

James V Stone

Title:
Bayes' Rule With Python
A Tutorial Introduction to Bayesian Analysis
Author: James V Stone
Published by Sebtel Press

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the author.

First Edition, 2015.
Typeset in LaTeX 2_ε.
Cover Design by Stefan Brazzo.
Copyright ©2016 by James V Stone
First printing.

ISBN 9780993367939

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
Pierre Simon Laplace, 1812.

Contents

Preface

This introductory text is intended to provide a straightforward explanation of Bayes' rule, using plausible and accessible examples. It is written specifically for readers who have little mathematical experience, but who are nevertheless willing to acquire the required mathematics on a 'need to know' basis.

Lecturers (and authors) like to teach using a top-down approach, so they usually begin with abstract general principles, and then move on to more concrete examples. In contrast, students usually like to learn using a bottom-up approach, so they like to begin with examples, from which abstract general principles can then be derived. As this book is not designed to teach lecturers or authors, it has been written using a bottom-up approach. Accordingly, the first chapter contains several accessible examples of how Bayes' rule can be useful in everyday situations, and these examples are examined in more detail in later chapters. The reason for including many examples in this book is that, whereas one reader may grasp the essentials of Bayes' rule from a medical example, another reader may feel more comfortable with the idea of flipping a coin to find out if it is 'fair'. One side-effect of so many examples is that the book may appear somewhat repetitive. For this, I make no apology. As each example is largely self-contained, it can be read in isolation. The obvious advantages of this approach inevitably lead to a degree of repetition, but this is a small price to pay for the clarity that an introductory text should possess.

Computer Code in MatLab, Python and R

MatLab, Python and R code snippets can be downloaded from here:
jim-stone.staff.shef.ac.uk/BookBayes2012/BayesRuleMatlabCode.html
jim-stone.staff.shef.ac.uk/BookBayes2012/BayesRulePythonCode.html
jim-stone.staff.shef.ac.uk/BookBayes2012/BayesRuleRCode.html

Code Snippets Included in Text

This book contains exactly the same text as the book *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*, but also includes additional code snippets printed close to relevant equations and figures. For readers with some proficiency in programming, these snippets should aid understanding of the relevant equations.

The code snippets included within the text work on their own, but the corresponding online files contain additional lines which clean up the figures drawn by the code (e.g. by making plotted lines bold). These code snippets can be downloaded from the web site given above.

Corrections

Please email any corrections to j.v.stone@sheffield.ac.uk. A list of corrections is at http://jim-stone.staff.shef.ac.uk/BayesBook/Corrections.

Acknowledgments

Thanks to friends and colleagues for reading draft chapters, including David Buckley, Nikki Hunkin, Danielle Matthews, Steve Snow, Tom Stafford, Stuart Wilson, Paul Warren, Charles Fox, and to John de Pledge for suggesting the particular medical example in Section 2.6. Special thanks to Royston Sellman for providing most of the Python computer code, and to Patricia Revest for the R computer code. Thanks to those readers who have emailed me to point out errors. Finally, thanks to my wife, Nikki Hunkin, for sound advice on the writing of this book, during which she tolerated Bayesian reasoning being applied to almost every aspect of our lives.

Jim Stone,
Sheffield, England.

Chapter 1

An Introduction to Bayes' Rule

> "... we balance probabilities and choose the most likely. It is the scientific use of the imagination ... "
> Sherlock Holmes, The Hound of the Baskervilles.
> AC Doyle, 1901.

Introduction

Bayes' rule is a rigorous method for interpreting evidence in the context of previous experience or knowledge. It was discovered by Thomas Bayes (c. 1701-1761), and independently discovered by Pierre-Simon Laplace (1749-1827). After more than two centuries of controversy, during which Bayesian methods have been both praised and pilloried, Bayes' rule has recently emerged as a powerful tool with a wide range

(a) Bayes

(b) Laplace

Figure 1.1: The fathers of Bayes' rule. a) Thomas Bayes (c. 1701-1761). b) Pierre-Simon Laplace (1749-1827).

of applications, which include: genetics[2], linguistics[12], image processing[15], brain imaging[33], cosmology[17], machine learning[5], epidemiology[26], psychology[31;44], forensic science[43], human object recognition[22], evolution[13], visual perception[23;41], ecology[32] and even the work of the fictional detective Sherlock Holmes[21]. Historically, Bayesian methods were applied by Alan Turing to the problem of decoding the German enigma code in the Second World War, but this remained secret until recently[16;29;37].

In order to appreciate the inner workings of any of the above applications, we need to understand why Bayes' rule is useful, and how it constitutes a mathematical foundation for reasoning. We will do this using a few accessible examples, but first, we will establish a few ground rules, and provide a reassuring guarantee.

Ground Rules

In the examples in this chapter, we will not delve into the precise meaning of probability, but will instead assume a fairly informal notion based on the frequency with which particular events occur. For example, if a bag contains 40 white balls and 60 black balls then the probability of reaching into the bag and choosing a black ball is the same as the proportion of black balls in the bag (ie 60/100=0.6). From this, it follows that the probability of an event (eg choosing a black ball) can adopt any value between zero and one, with zero meaning it definitely will not occur, and one meaning it definitely will occur. Finally, given a set of mutually exclusive events, such as the outcome of choosing a ball, which has to be either black or white, the probabilities of those events have to add up to one (eg 0.4+0.6=1). We explore the subtleties of the meaning of probability in Section 7.1.

A Guarantee

Before embarking on these examples, we should reassure ourselves with a fundamental fact regarding Bayes' rule, or *Bayes' theorem*, as it is also called: Bayes' theorem is not a matter of conjecture. By definition, a theorem is a mathematical statement that has been proved to be true. This is reassuring because, if we had to establish the rules for

calculating with probabilities, we would insist that the result of such calculations must tally with our everyday experience of the physical world, just as surely as we would insist that $1 + 1 = 2$. Indeed, if we insist that probabilities must be combined with each other in accordance with certain common sense principles then Cox(1946)[7] showed that this leads to a unique set of rules, a set which includes Bayes' rule, which also appears as part of Kolmogorov's(1933)[24] (arguably, more rigorous) theory of probability.

1.1. Example 1: Poxy Diseases

The Patient's Perspective

Suppose that you wake up one day with spots all over your face, as in Figure 1.2. The doctor tells you that 90% of people who have smallpox have the same symptoms as you have. In other words, the probability of having these symptoms given that you have smallpox is 0.9 (ie 90%). As smallpox is often fatal, you are naturally terrified.

However, after a few moments of contemplation, you decide that you do not want to know the probability that you have these symptoms (after all, you already know you have them). Instead, what you really want to know is the probability that you have smallpox.

So you say to your doctor, "Yes, but what is the probability that I have smallpox given that I have these symptoms?". "Ah", says your doctor, "a very good question." After scribbling some equations, your doctor looks up. "The probability that you have smallpox given that you have these symptoms is 1.1%, or equivalently, 0.011." Of course,

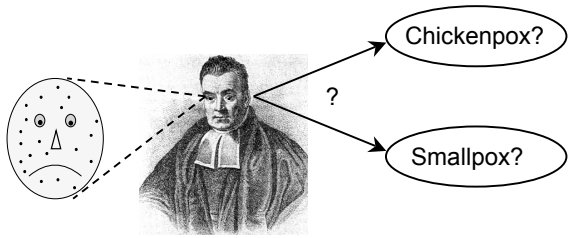

Figure 1.2: Thomas Bayes diagnosing a patient.

this is not good news, but it sounds better than 90%, and (more importantly) it is at least useful information. This demonstrates the stark contrast between the probability of the symptoms given a disease (which you do not want to know) and the probability of the disease given the symptoms (which you do want to know).

Bayes' rule transforms probabilities that look useful (but are often not) into probabilities that are useful. In the above example, the doctor used Bayes' rule to transform the uninformative probability of your symptoms given that you have smallpox into the informative probability that you have smallpox given your symptoms.

The Doctor's Perspective

Now, suppose you are a doctor, confronted with a patient who is covered in spots. The patient's symptoms are consistent with chickenpox, but they are also consistent with another, more dangerous, disease, smallpox. So you have a dilemma. You know that 80% of people with chickenpox have spots, but also that 90% of people with smallpox have spots. So the probability (0.8) of the symptoms given that the patient has chickenpox is similar to the probability (0.9) of the symptoms given that the patient has smallpox (see Figure 1.2).

If you were a doctor with limited experience then you might well think that both chickenpox and smallpox are equally probable. But, as you are a knowledgeable doctor, you know that chickenpox is common, whereas smallpox is rare. This knowledge, or *prior information*, can be used to decide which disease the patient probably has. If you had to guess (and you do have to guess because you are the doctor) then you would combine the possible diagnoses implied by the symptoms with your prior knowledge to arrive at a conclusion (ie that the patient probably has chickenpox). In order to make this example more tangible, let's run through it again, this time with numbers.

The Doctor's Perspective (With Numbers)

We can work out probabilities associated with a disease by making use of public health statistics. Suppose doctors are asked to report the number of cases of smallpox and chickenpox, and the symptoms

observed. Using the results of such surveys, it is a simple matter to find the proportion of patients diagnosed with smallpox and chickenpox, and each patient's symptoms (eg spots). Using these data, we might find that the probability that a patient has spots given that the patient has smallpox is 90% or 0.9. We can write this in an increasingly succinct manner using a special notation

$$p(\text{symptoms are spots} \mid \text{disease is smallpox}) = 0.9, \tag{1.1}$$

where the letter p stands for probability, and the vertical bar $\mid$ stands for "given that". So, this short-hand statement should be read as

> "the probability that the patient's symptoms are spots given that he has smallpox is 90% or 0.9".

The vertical bar indicates that the probability that the patient has spots depends on the presence of smallpox. Thus, the probability of spots is said to be *conditional* on the disease under consideration. For this reason, such probabilities are known as *conditional probabilities*. We can write this even more succinctly as

$$p(\text{spots}|\text{smallpox}) = 0.9. \tag{1.2}$$

Similarly, we might find that spots are observed in 80% of patients who have chickenpox, which is written as

$$p(\text{spots}|\text{chickenpox}) = 0.8. \tag{1.3}$$

Equations 1.2 and 1.3 formalise why we should not use the symptoms alone to decide which disease the patient has. These equations take no account of our previous experience of the relative prevalence of smallpox and chickenpox, and are based only on the observed symptoms. As we shall see later, this is equivalent to making a decision based on the (in this case, false) assumption that both diseases are equally prevalent in the population, and that they are therefore *a priori* equally probable.

Note that the conditional probability $p(\text{spots}|\text{smallpox})$ is the probability of spots given that the patient has smallpox, but it is called the *likelihood* of smallpox (which is confusing, but standard,

nomenclature). In this example, the disease smallpox has a larger likelihood than chickenpox. Indeed, as there are only two diseases under consideration, this means that, of the two possible alternatives, smallpox has the maximum likelihood. The disease with the maximum value of likelihood is known as the *maximum likelihood estimate* (MLE) of the disease that the patient has. Thus, in this case, the MLE of the disease is smallpox.

As discussed above, it would be hard to argue that we should disregard our knowledge or previous experience when deciding which disease the patient has. But exactly how should this previous experience be combined with current evidence (eg symptoms)? From a purely intuitive perspective, it would seem sensible to weight the likelihood of each disease according to previous experience of that disease, as in Figure 1.3. Since smallpox is rare, and is therefore intrinsically improbable, it might be sensible to weight the likelihood of smallpox by a small number. This would yield a small 'weighted likelihood', which would be a more realistic estimate of the probability that the patient has smallpox. For example, public health statistics may inform us that the prevalence of smallpox in the general population is 0.001, meaning that there is a one in a thousand chance that a

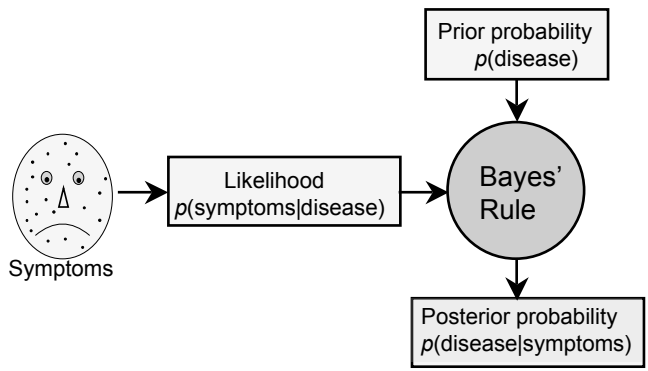

Figure 1.3: Schematic representation of Bayes' rule. Data, in the form of symptoms, are used find a likelihood, which is the probability of those symptoms given that the patient has a specific disease. Bayes' rule combines this likelihood with prior knowledge, and yields the posterior probability that the patient has the disease given that he has the symptoms observed.

randomly chosen individual has smallpox. Thus, the probability that a randomly chosen individual has smallpox is

$$p(\text{smallpox}) = 0.001. \tag{1.4}$$

This represents our prior knowledge about the disease in the population before we have observed our patient, and is known as the *prior probability* that any given individual has smallpox. As our patient (before we have observed his symptoms) is as likely as any other individual to have smallpox, we know that the prior probability that he has smallpox is 0.001.

If we follow our commonsense prescription, and simply weight (ie multiply) each likelihood by its prior probability then we obtain 'weighted likelihood' quantities which take account of the current evidence and of our prior knowledge of each disease. In short, this commonsense prescription leads to Bayes' rule. Even so, the equation for Bayes' rule given below is not obvious, and should be taken on trust for now. In the case of smallpox, Bayes' rule is

$$p(\text{smallpox}|\text{spots}) = \frac{p(\text{spots}|\text{smallpox}) \times p(\text{smallpox})}{p(\text{spots})}. \tag{1.5}$$

The term $p(\text{spots})$ in the denominator of Equation 1.5 is the proportion of people in the general population that have spots, and therefore represents the probability that a randomly chosen individual has spots. As will be explained on p16, this term is often disregarded, but we use a value that makes our sums come out neatly, and assume that $p(\text{spots}) = 0.081$ (ie 81 in every 1,000 individuals has spots). If we now substitute numbers into this equation then we obtain

$$\begin{aligned} p(\text{smallpox}|\text{spots}) &= 0.9 \times 0.001/0.081 & (1.6)\\ &= 0.011, & (1.7) \end{aligned}$$

which is the conditional probability that the patient has smallpox given that his symptoms are spots.

Crucially, the 'weighted likelihood' p(smallpox|spots) is also a conditional probability, but it is the probability of the disease smallpox given the symptoms observed, as shown in Figure 1.4. So, by making use of prior experience, we have transformed the conditional probability of the observed symptoms given a specific disease (the likelihood, which is based only on the available evidence) into a more useful conditional probability: the probability that the patient has a particular disease (smallpox) given that he has particular symptoms (spots).

In fact, we have just made use of Bayes' rule to convert one conditional probability, the likelihood p(spots|smallpox) into a more useful conditional probability, which we have been calling a 'weighted likelihood', but is formally known as the *posterior probability* p(smallpox|spots).

As noted above, both p(smallpox|spots) and p(spots|smallpox) are conditional probabilities, which have the same status from a mathematical viewpoint. However, for Bayes' rule, we treat them very differently.

Code Example 1.1: Smallpox
File: Ch1Eq06.py

```
# likelihood = prob of spots given smallpox
pSpotsGSmallpox = 0.9
# prior = prob of smallpox
pSmallpox = 0.001
# marginal likelihood = prob of spots
pSpots = 0.081
# find posterior = prob of smallpox given spots
pSmallpoxGSpots = pSpotsGSmallpox * pSmallpox / pSpots

print('Posterior, pSmallpoxGSpots = %.3f.' % pSmallpoxGSpots)
# Output: Posterior, pSmallpoxGSpots = 0.011.
```

The conditional probability p(spots|smallpox) is based only on the observed data (symptoms), and is therefore easier to obtain than the conditional probability we really want, namely p(smallpox|spots), which is also based on the observed data, but also on prior knowledge. For historical reasons, these two conditional probabilities have special names. As we have already seen, the conditional probability p(spots|smallpox) is the probability that a patient has spots given

that he has smallpox, and is known as the likelihood of smallpox. The complementary conditional probability $p(\text{smallpox}|\text{spots})$ is the posterior probability that a patient has smallpox given that he has spots. In essence, Bayes' rule is used to combine prior experience (in the form of a prior probability) with observed data (spots) (in the form of a likelihood) to interpret these data (in the form of a posterior probability). This process is known as *Bayesian inference*.

Code Example 1.2: Chickenpox
File: Ch1Eq09.py

```
#  likelihood = prob of spots given chickenpox
pSpotsGChickenpox = 0.8
#  prior = prob of chickenpox
pChickenpox = 0.1
#  marginal likelihood = prob of spots
pSpots = 0.081
#  find posterior = prob of chickenpox given spots
pChickenpoxGSpots = pSpotsGChickenpox * pChickenpox / pSpots
print('Posterior, pChickenpoxGSpots = %.3f.' % pChickenpoxGSpots)

# Output:   Posterior, pChickenpoxGSpots = 0.988.
```

The Perfect Inference Engine

Bayesian inference is not guaranteed to provide the correct answer. Instead, it provides the probability that each of a number of alternative answers is true, and these can then be used to find the answer that is most probably true. In other words, it provides an informed guess. While this may not sound like much, it is far from random guessing. Indeed, it can be shown that no other procedure can provide a better guess, so that Bayesian inference can be justifiably interpreted as the output of a perfect guessing machine, a perfect inference engine (see Section 4.9, p100). This perfect inference engine is fallible, but it is provably less fallible than any other.

Making a Diagnosis

In order to make a diagnosis, we need to know the posterior probability of both of the diseases under consideration. Once we have both

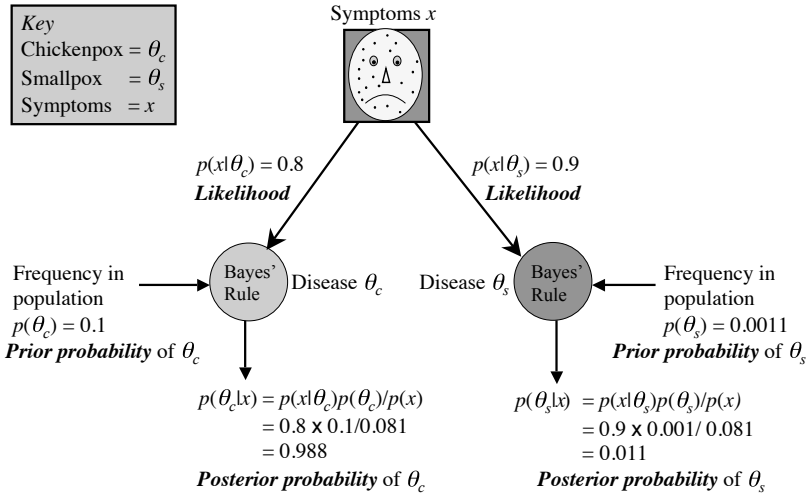

Figure 1.4: Comparing the probability of chickenpox and smallpox using Bayesian inference. The observed symptoms x seem to be more consistent with smallpox θ_s than chickenpox θ_c, as indicated by their likelihood values. However, the background rate of chickenpox in the population is higher than that of smallpox, which, in this case, makes it more probable that the patient has chickenpox, as indicated by its higher posterior probability.

posterior probabilities, we can compare them in order to choose the disease that is most probable given the observed symptoms.

Suppose that the prevalence of chickenpox in the general population is 10% or 0.1. This represents our prior knowledge about chickenpox before we have observed any symptoms, and is written as

$$p(\text{chickenpox}) = 0.1, \tag{1.8}$$

which is the prior probability of chickenpox. As was done in Equation 1.6 for smallpox, we can weight the likelihood of chickenpox with its

prior probability to obtain the posterior probability of chickenpox

$$\begin{aligned} p(\text{chickenpox}|\text{spots}) &= p(\text{spots}|\text{chickenpox}) \times p(\text{chickenpox})/p(\text{spots}) \\ &= 0.8 \times 0.1/0.081 \\ &= 0.988. \end{aligned} \tag{1.9}$$

The two posterior probabilities, summarised in Figure 1.4, are therefore

$$p(\text{smallpox}|\text{spots}) = 0.011 \tag{1.10}$$

$$p(\text{chickenpox}|\text{spots}) = 0.988. \tag{1.11}$$

Thus, the posterior probability that the patient has smallpox is 0.011, and the posterior probability that the patient has chickenpox is 0.988. Aside from a rounding error, these sum to one.

Notice that we cannot be certain that the patient has chickenpox, but we can be certain that there is a 98.8% probability that he does. This is not only our best guess, but it is provably the best guess that can be obtained; it is effectively the output of a perfect inference engine.

In summary, if we ignore all previous knowledge regarding the prevalence of each disease then we have to use the likelihoods to decide which disease is present. The likelihoods shown in Equations 1.2 and 1.3 would lead us to diagnose the patient as probably having smallpox. However, a more informed decision can be obtained by taking account of prior information regarding the diseases under consideration. When we do take account of prior knowledge, Equations 1.10 and 1.11 indicate that the patient probably has chickenpox. In fact, these equations imply that the patient is about 89 (=0.988/0.011) times more likely to have chickenpox than smallpox. As we shall see later, this ratio of posterior probabilities plays a key role in Bayesian statistical analysis (Section 1.1, p14).

Taking account of previous experience yields the diagnosis that is most probable, given the evidence (spots). As this is the decision associated with the maximum value of the posterior probability, it is known as the *maximum a posteriori* or MAP estimate of the disease.

The equation used to perform Bayesian inference is called Bayes' rule, and in the context of diagnosis is

$$p(\text{disease}|\text{symptoms}) = \frac{p(\text{symptoms}|\text{disease})p(\text{disease})}{p(\text{symptoms})}, \tag{1.12}$$

which is easier to remember as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \tag{1.13}$$

The *marginal likelihood* is also known as *evidence*, and we shall have more to say about it shortly.

Bayes' Rule: Hypothesis and Data

If we consider a putative disease to represent a specific hypothesis, and the symptoms to be some observed data then Bayes' rule becomes

$$p(\text{hypothesis}|\text{data}) \quad = \quad \frac{p(\text{data}|\text{hypothesis}) \times p(\text{hypothesis})}{p(\text{data})},$$

where the word "hypothesis" should be interpreted as, "hypothesis is true". Written in this form, the contrast between the likelihood and the posterior probability is more apparent. Specifically, the probability that the proposed hypothesis is true given some data that were actually observed is the posterior probability

$$p(\text{hypothesis}|\text{data}), \tag{1.14}$$

whereas the probability of observing the data given that the hypothesis is true is the likelihood

$$p(\text{data}|\text{hypothesis}). \tag{1.15}$$

A More Succinct Notation

We now introduce a succinct, and reasonably conventional, notation for the terms defined above. There is nothing new in the mathematics of this section, just a re-writing of equations used above. If we represent

the observed symptoms by x, and the disease by the Greek letter *theta* θ_s (where the subscript s stands for smallpox) then we can write the conditional probability (ie the likelihood of smallpox) in Equation 1.2

$$p(x|\theta_s) = p(\text{spots}|\text{smallpox}) = 0.9. \tag{1.16}$$

Similarly, the background rate of smallpox θ_s in the population can be represented as the prior probability

$$p(\theta_s) = p(\text{smallpox}) = 0.001, \tag{1.17}$$

and the probability of the symptoms (the marginal likelihood) is

$$p(x) = p(\text{spots}) = 0.081. \tag{1.18}$$

Substituting this notation into Equation 1.5 (repeated here)

$$p(\text{smallpox}|\text{spots}) = \frac{p(\text{spots}|\text{smallpox}) \times p(\text{smallpox})}{p(\text{spots})}, \tag{1.19}$$

yields

$$p(\theta_s|x) = \frac{p(x|\theta_s) \times p(\theta_s)}{p(x)}. \tag{1.20}$$

Similarly, if we define

$$\begin{aligned} p(x|\theta_c) &= p(\text{spots}|\text{chickenpox}) \\ p(\theta_c|x) &= p(\text{chickenpox}|\text{spots}) \\ p(\theta_c) &= p(\text{chickenpox}), \end{aligned} \tag{1.21}$$

then we can re-write Equation 1.9 to obtain the posterior probability of chickenpox as

$$p(\theta_c|x) = \frac{p(x|\theta_c) \times p(\theta_c)}{p(x)}. \tag{1.22}$$

If we use θ without a subscript to represent any disease (or hypothesis), and x to represent any observed symptoms (or data) then Bayes' rule can be written as (we now drop the use of the $\times$ symbol)

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}. \tag{1.23}$$

Finally, we should note that smallpox made history by being the first disease to be eradicated from the Earth in 1979, which makes the prior probability of catching it somewhat less than the value $p(\theta_s) = 0.001$ assumed in the above example.

Parameters and Variables: Notice that there is nothing special about which symbol stands for disease and which for symptoms, and that we could equally well have used θ to represent symptoms, and x to represent diseases. However, it is common to use a Greek letter like θ to represent the thing we wish to estimate, and x to represent the evidence (eg symptoms) on which our estimated value of θ will be based. Similarly, using an equally arbitrary but standard convention, the symbol that represents the thing we wish to estimate is usually called a *parameter* (θ), whereas the evidence used to estimate that thing is usually called a *variable* (x).

Model Selection, Posterior Ratios and Bayes Factors

As noted above, when we take account of prior knowledge, it turns out that the patient is about 90 times more likely (ie 0.988 vs 0.011) to have chickenpox than smallpox. Indeed, it is often the case that we wish to compare the relative probabilities of two hypotheses (eg diseases). As each hypothesis acts as a (simple) model for the data, and we wish to select the most probable model, this is known as *model selection*, which involves a comparison using a ratio of posterior probabilities.

The posterior ratio, which is also known as the *posterior odds* between the hypotheses θ_c and θ_s, is

$$R_{post} \quad = \quad \frac{p(\theta_c|x)}{p(\theta_s|x)}. \tag{1.24}$$

If we apply Bayes' rule to the numerator and denominator then

$$R_{post} = \frac{p(x|\theta_c)p(\theta_c)/p(x)}{p(x|\theta_s)p(\theta_s)/p(x)}, \tag{1.25}$$

where the marginal likelihood $p(x)$ cancels, so that

$$R_{post} = \frac{p(x|\theta_c)}{p(x|\theta_s)} \times \frac{p(\theta_c)}{p(\theta_s)}. \tag{1.26}$$

This is a product of two ratios, the ratio of likelihoods, or *Bayes factor*

$$B = \frac{p(x|\theta_c)}{p(x|\theta_s)}, \tag{1.27}$$

and the ratio of priors, or *prior odds* between θ_c and θ_s, which is

$$R_{prior} = p(\theta_c)/p(\theta_s). \tag{1.28}$$

Thus, the posterior odds can be written as

$$R_{post} = B \times R_{prior}, \tag{1.29}$$

which, in words, is: *posterior odds* = *Bayes factor* × *prior odds*. In this example, we have

$$R_{post} = \frac{0.80}{0.90} \times \frac{0.1}{0.001} = 88.9.$$

Code Example 1.3: Posterior Odds
File: Ch1Eq26.py

```
pSpotsGSmallpox = 0.9
pSmallpox = 0.001
pSpotsGChickenpox = 0.8
pChickenpox = 0.1
Rpost = pSpotsGChickenpox*pChickenpox / (pSpotsGSmallpox*pSmallpox)
print('Rpost = %.3f.' % Rpost)

# Output:   Rpost = 88.9
```

Note that the likelihood ratio (Bayes factor) is less than one (and so favours θ_s), whereas the prior odds is much greater than one (and favours θ_c), with the result that the posterior odds come out massively in favour of θ_c. If the posterior odds is greater than 3 or less than 1/3 (in both cases one hypothesis is more than 3 times more probable than the other) then this is considered to represent a substantial difference between the probabilities of the two hypotheses[19], so a posterior odds of 88.9 is definitely substantial.

Ignoring the Marginal Likelihood

As promised, we consider the marginal likelihood $p(\text{symptoms})$ or $p(x)$ briefly here (and in Chapter 2 and Section 4.5). The marginal likelihood refers to the probability that a randomly chosen individual has the symptoms that were actually observed, which we can interpret as the prevalence of spots in the general population.

Crucially, the decision as to which disease the patient has depends only on the relative sizes of different posterior probabilities (eg Equations 1.10, 1.11, and in Equations 1.20,1.22). Note that each of these posterior probabilities is proportional to $1/p(\text{symptoms})$ in Equations 1.10, 1.11, also expressed as $1/p(x)$ in Equations 1.20,1.22. This means that a different value of the marginal probability $p(\text{symptoms})$ would change all of the posterior probabilities by the same proportion, and therefore has no effect on their *relative* magnitudes. For example, if we arbitrarily decided to double the value of the marginal likelihood from 0.081 to 0.162 then both posterior probabilities would be halved (from 0.011 and 0.988 to about 0.005 and 0.494), but the posterior probability of chickenpox would still be 88.9 times larger than the posterior probability of smallpox. Indeed, the previous section on Bayes factors relies on the fact that the ratio of two posterior probabilities is independent of the value of the marginal probability.

In summary, the value of the marginal probability has no effect on which disease yields the largest posterior probability (eg Equations 1.10 and 1.11), and therefore has no effect on the decision regarding which disease the patient probably has.

1.2. Example 2: Forkandles

The example above is based on medical diagnosis, but Bayes' rule can be applied to any situation where there is uncertainty regarding the value of a measured quantity, such as the acoustic signal that reaches the ear when some words are spoken. The following example follows a similar line of argument as the previous one, and aside from the change in context, provides no new information for the reader to absorb.

If you walked into a hardware store and asked, *Have you got fork handles?*, then you would be surprised to be presented with four candles. Even though the phrases *fork handles* and *four candles* are acoustically almost identical, the shop assistant knows that he sells many more candles than fork handles (Figure 1.5). This in turn, means that he probably does not even hear the words *fork handles*, but instead hears *four candles*. What has this got to do with Bayes' rule?

The acoustic data that correspond to the sounds spoken by the customer are equally consistent with two interpretations, but the assistant assigns a higher weighting to one interpretation. This weighting is based on his prior experience, so he knows that customers are more likely to request four candles than fork handles. The experience of the assistant allows him to hear what was probably said by the customer, even though the acoustic data was pretty ambiguous. Without knowing it, he has probably used something like Bayes' rule to hear what the customer probably said.

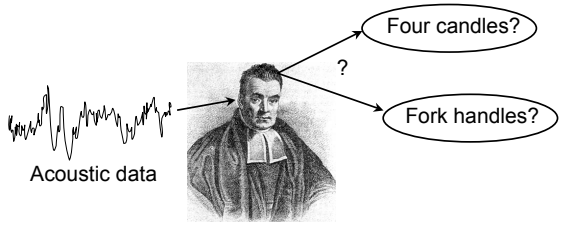

Figure 1.5: Thomas Bayes trying to make sense of a London accent, which removes the *h* sound from the word *handle*, so the phrase *fork handles* is pronounced *fork 'andles*, and therefore sounds like *four candles* (see Fork Handles YouTube clip by The Two Ronnies).

Likelihood: Answering the Wrong Question

Given that the two possible phrases are *four candles* and *fork handles*, we can formalise this scenario by considering the probability of the acoustic data given each of the two possible phrases. In both cases, the probability of the acoustic data depends on the words spoken, and this dependence is made explicit as two probabilities:

1) the probability of the acoustic data given *four candles* was spoken,
2) the probability of the acoustic data given *fork handles* was spoken.

A short-hand way of writing these is

$$\begin{aligned} &p(\text{acoustic data}|\text{four candles}) \\ &p(\text{acoustic data}|\text{fork handles}), \end{aligned} \tag{1.30}$$

where the expression $p(\text{acoustic data}|\text{four candles})$, for example, is interpreted as the likelihood that the phrase spoken was *four candles*. As both phrases are consistent with the acoustic data, the probability of the data is almost the same in both cases. That is, the probability of the data given that *four candles* was spoken is almost the same as the probability of the data given that *fork handles* was spoken. For simplicity, we will assume that these probabilities are

$$\begin{aligned} p(\text{data}|\text{four candles}) &= 0.6 \\ p(\text{data}|\text{fork handles}) &= 0.7. \end{aligned} \tag{1.31}$$

Knowing these two likelihoods does allow us to find an answer, but it is an answer to the wrong question. Each likelihood above provides an answer to the (wrong) question: *what is the probability of the observed acoustic data given that each of two possible phrases was spoken?*

Posterior Probability: Answering the Right Question

The right question, the question to which we would like an answer is: what is the probability that each of the two possible phrases was spoken given the acoustic data? The answer to this, the right question, is

implicit in two new conditional probabilities, the posterior probabilities

$$\begin{aligned} &p(\text{four candles}|\text{data}) \\ &p(\text{fork handles}|\text{data}), \end{aligned} \tag{1.32}$$

as shown in Figures 1.6 and 1.7. Notice the subtle difference between the pair of Equations 1.31 and the pair 1.32. Equations 1.31 tells us the likelihoods, the probability of the data given two possible phrases, which turn out to be almost identical for both phrases in this example. In contrast, Equations 1.32 tells us the posterior probabilities, the probability of each phrase given the acoustic data.

Crucially, each likelihood tells us the probability of the data given a particular phrase, but takes no account of how often that phrase has been given (ie has been encountered) in the past. In contrast, each posterior probability depends, not only on the data (in the form of the likelihood), but also on how frequently each phrase has been encountered in the past; that is, on prior experience.

So, we want the posterior probability, but we have the likelihood. Fortunately, Bayes' rule provides a means of getting from the likelihood to the posterior, by making use of extra knowledge in the form of prior experience, as shown in Figure 1.6.

Prior Probability

Let's suppose that the assistant has been asked for four candles a total of 90 times in the past, whereas he has been asked for fork handles only 10 times. To keep matters simple, let's also assume that the next customer will ask either for four candles or fork handles (we will revisit this simplification later). Thus, before the customer has uttered a single word, the assistant estimates that the probability that he will say each of the two phrases is

$$\begin{aligned} p(\text{four candles}) &= 90/100 = 0.9 \\ p(\text{fork handles}) &= 10/100 = 0.1. \end{aligned} \tag{1.33}$$

These two prior probabilities represent the prior knowledge of the assistant, based on his previous experience of what customers say.

When confronted with an acoustic signal that has one of two possible interpretations, the assistant naturally interprets this as *four candles*, because, according to his past experience, this is what such ambiguous acoustic data usually means in practice. So, he takes the two almost equal likelihood values, and assigns a weighting to each one, a weighting that depends on past experience, as in Figure 1.7. In other words, he uses the acoustic data, and combines it with his previous experience to make an inference about which phrase was spoken.

Inference

One way to implement this weighting (ie to do this inference) is to simply multiply the likelihood of each phrase by how often that phrase has occurred in the past. In other words, we multiply the likelihood of each putative phrase by its corresponding prior probability. The result

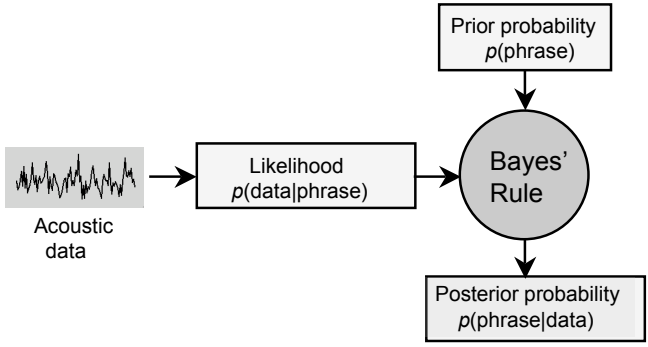

Figure 1.6: A schematic representation of Bayes' rule. Data alone, in the form of acoustic data, can be used to find a likelihood value, which is the conditional probability of the acoustic data given some putative spoken phrase. When Bayes' rule is used to combine this likelihood with prior knowledge then the result is a posterior probability, which is the probability of the phrase given the observed acoustic data.

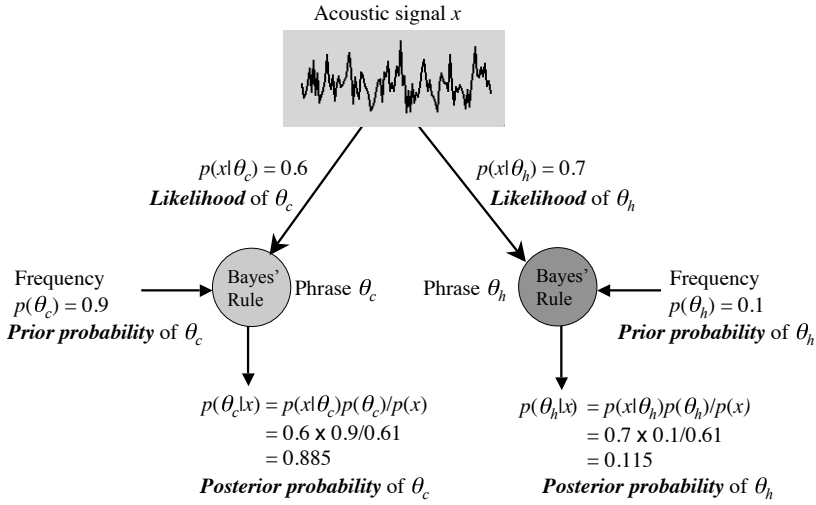

Figure 1.7: Bayesian inference applied to speech data.

yields a posterior probability for each possible phrase

$$p(\text{four candles}|\text{data}) = \frac{p(\text{data}|\text{four candles})p(\text{four candles})}{p(\text{data})}$$

$$p(\text{fork handles}|\text{data}) = \frac{p(\text{data}|\text{fork handles})p(\text{fork handles})}{p(\text{data})}, \quad (1.34)$$

where $p(\text{data})$ is the marginal likelihood, which is the probability of the observed data.

In order to ensure that the posterior probabilities sum to one, the value of $p(\text{data})$ is 0.61 in this example, but as we already know from Section 1.1 (p16), its value is not important for our purposes. If we substitute the likelihood and prior probability values defined in Equations 1.31 and 1.33 in 1.34 then we obtain their posterior

probabilities as

$$\begin{aligned} p(\text{four candles}|\text{data}) &= p(\text{data}|\text{four candles})p(\text{four candles})/p(\text{data}) \\ &= 0.6 \times 0.9/0.61 = 0.885, \end{aligned}$$

$$\begin{aligned} p(\text{fork handles}|\text{data}) &= p(\text{data}|\text{fork handles})p(\text{fork handles})/p(\text{data}) \\ &= 0.7 \times 0.1/0.61 = 0.115. \end{aligned}$$

As in the previous example, we can write this more succinctly by defining

$$\begin{aligned} x &= \text{acoustic data}, \\ \theta_c &= \text{four candles}, \\ \theta_h &= \text{fork handles}, \end{aligned}$$

so that

$$\begin{aligned} p(\theta_c|x) &= p(x|\theta_c)p(\theta_c)/p(x) = 0.885 \\ p(\theta_h|x) &= p(x|\theta_h)p(\theta_h)/p(x) = 0.115. \end{aligned} \tag{1.35}$$

Code Example 1.4: Four Candles
File: Ch1Eq34.py

```
pData = 0.61
pDataGFourCandles = 0.6
pFourCandles = 0.9
pFourCandlesGData = pDataGFourCandles * pFourCandles / pData
print 'pFourCandlesGData␣=␣%.3f' % pFourCandlesGData

pDataGForkHandles = 0.7
pForkHandles = 0.1
pForkHandlesGData = pDataGForkHandles * pForkHandles / pData
print 'pForkHandlesGData␣=␣%.3f\n' % pForkHandlesGData

# Output:   pFourCandlesGData = 0.885
#           pForkHandlesGData = 0.115
```

These two posterior probabilities represent the answer to the right question, so we can now see that the probability that the customer said *four candles* is 0.885 whereas the probability that the customer said *fork handles* was 0.115. As *four candles* is associated with the highest value of the posterior probability, it is the *maximum a posteriori* (MAP) estimate of the phrase that was spoken. The process that makes use of evidence (symptoms) to produce these posterior probabilities is called *Bayesian inference*.

1.3. Example 3: Flipping Coins

This example follows the same line of reasoning as those above, but also contains specific information on how to combine probabilities from independent events, such as coin flips. This will prove crucial in a variety of contexts, and in examples considered later in this book.

Here, our task is to decide how unfair a coin is, based on just two coin flips. Normally, we assume that coins are fair or unbiased, so that a large number of coin flips (eg 1000) yields an equal number of heads and tails. But suppose there was a fault in the machine that minted coins, so that each coin had more metal on one side or the other, with the result that each coin is biased to produce more heads than tails, or vice versa. Specifically, 25% of the coins produced by the machine have a bias of 0.4, and 75% have a bias of 0.6. By definition, a coin with a bias of 0.4 produces a head on 40% of flips, whereas a coin with a bias of 0.6 produces a head on 60% of flips (on average). Now, suppose we choose one coin at random, and attempt to decide which of the two bias

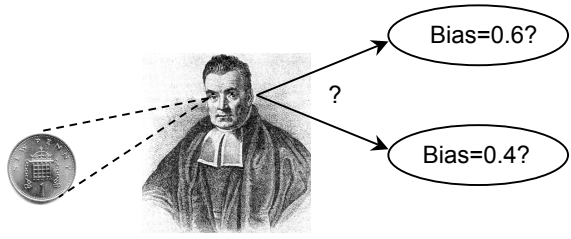

Figure 1.8: Thomas Bayes trying to decide the value of a coin's bias.

values it has. For brevity, we define the coin's bias with the parameter θ, so the true value of θ for each coin is either $\theta_{0.4} = 0.4$, or $\theta_{0.6} = 0.6$.

One Coin Flip: Here we use one coin flip to define a few terms that will prove useful below. For each coin flip, there are two possible outcomes, a head x_h, and a tail x_t. For example, if the coin's bias is $\theta_{0.6}$ then, by definition, the conditional probability of observing a head is $\theta_{0.6}$

$$p(x_h|\theta_{0.6}) = \theta_{0.6} = 0.6. \tag{1.36}$$

Similarly, the conditional probability of observing a tail is

$$p(x_t|\theta_{0.6}) \quad = \quad (1 - \theta_{0.6}) = 0.4, \tag{1.37}$$

where both of these conditional probabilities are likelihoods. Note that we follow the convention of the previous examples by using θ to represent the parameter whose value we wish to estimate, and x to represent the data used to estimate the true value of θ.

Two Coin Flips: Consider a coin with a bias θ (where θ could be 0.4 or 0.6, for example). Suppose we flip this coin twice, and obtain a head x_h followed by a tail x_t, which define the ordered list or *permutation*

$$\mathbf{x} \quad = \quad (x_h, x_t). \tag{1.38}$$

As the outcome of one flip is not affected by any other flip outcome, outcomes are said to be *independent* (see Section 2.2 or Appendix C). This independence means that the probability of observing any two

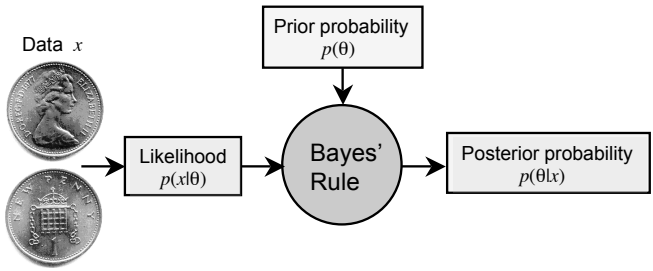

Figure 1.9: A schematic representation of Bayes' rule, applied to the problem of estimating the bias of a coin based on two coin flips.

outcomes can be obtained by multiplying their probabilities

$$p(\mathbf{x}|\theta) = p((x_h, x_t)|\theta) \tag{1.39}$$

$$= p(x_h|\theta) \times p(x_t|\theta). \tag{1.40}$$

More generally, for a coin with a bias θ, the probability of a head x_h is $p(x_h|\theta) = \theta$, and the probability of a tail x_t is therefore $p(x_t|\theta) = (1-\theta)$. It follows that Equation 1.40 can be written as

$$p(\mathbf{x}|\theta) = \theta \times (1 - \theta), \tag{1.41}$$

which will prove useful below.

The Likelihoods of Different Coin Biases: According to Equation 1.41, if the coin bias is $\theta_{0.6}$ then

$$p(\mathbf{x}|\theta_{0.6}) = \theta_{0.6} \times (1 - \theta_{0.6}) \tag{1.42}$$

$$= 0.6 \times 0.4 \tag{1.43}$$

$$= 0.24, \tag{1.44}$$

and if the coin bias is $\theta_{0.4}$ then (the result is the same)

$$p(\mathbf{x}|\theta_{0.4}) = \theta_{0.4} \times (1 - \theta_{0.4}) \tag{1.45}$$

$$= 0.4 \times 0.6 \tag{1.46}$$

$$= 0.24. \tag{1.47}$$

Note that the only difference between these two cases is the reversed ordering of terms in Equations 1.43 and 1.46, so that both values of θ have equal likelihood values. In other words, the observed data $\mathbf{x}$ are equally probable given the assumption that $\theta_{0.4} = 0.4$ or $\theta_{0.6} = 0.6$, so they do not help in deciding which bias our chosen coin has.

Prior Probabilities of Different Coin Biases: We know (from above) that 25% of all coins have a bias of $\theta_{0.4}$, and that 75% of all coins have a bias of $\theta_{0.6}$. Thus, even before we have chosen our coin, we know (for example) there is a 75% chance that it has a bias of 0.6. This information defines the prior probability that any coin has one of two bias values, either $p(\theta_{0.4}) = 0.25$, or $p(\theta_{0.6}) = 0.75$.

Posterior Probabilities of Different Coin Biases: As in previous examples, we adopt the naïve strategy of simply weighting each likelihood value by its corresponding prior (and dividing by $p(\mathbf{x})$) to obtain Bayes' rule

$$\begin{aligned} p(\theta_{0.4}|\mathbf{x}) &= p(\mathbf{x}|\theta_{0.4})p(\theta_{0.4})/p(\mathbf{x}) \\ &= 0.24 \times 0.25/0.24 \\ &= 0.25, \qquad (1.48) \\ p(\theta_{0.6}|\mathbf{x}) &= p(\mathbf{x}|\theta_{0.6})p(\theta_{0.6})/p(\mathbf{x}) \\ &= 0.24 \times 0.75/0.24 \\ &= 0.75. \qquad (1.49) \end{aligned}$$

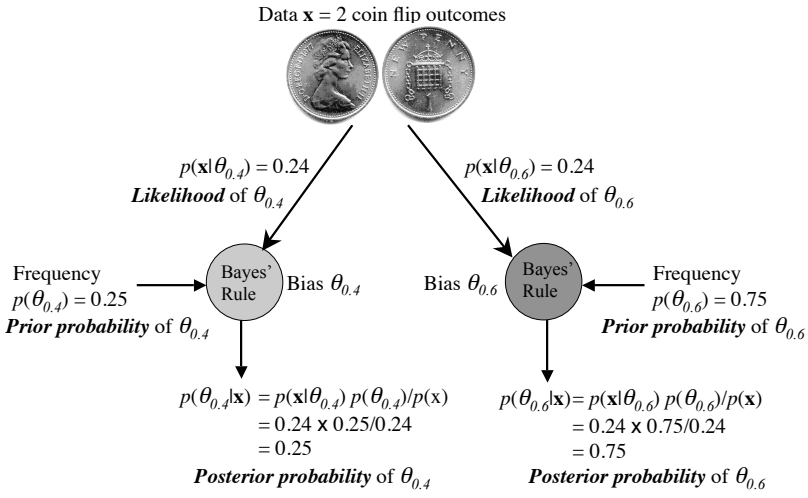

Figure 1.10: Bayesian inference applied to coin flip data.

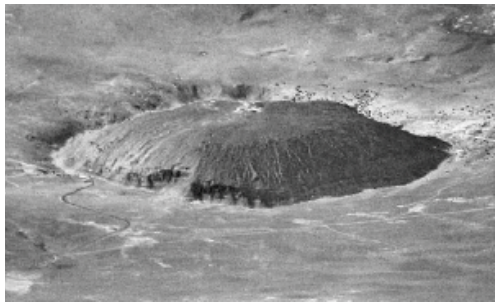

Figure 1.11: Is this a hill or a crater? Try turning the book upside-down. (Barringer crater, with permission, United States Geological Survey).

In order to ensure posterior probabilities sum to one, we have assumed a value for the marginal probability of $p(\mathbf{x}) = 0.24$ (but we know from p16 that its value makes no difference to our final decision about coin bias). As shown in Figures 1.9 and 1.10, the probabilities in Equations 1.48 and 1.49 take account of both the data and of prior experience, and are therefore posterior probabilities. In summary, whereas the equal likelihoods in this example (Equations 1.44 and 1.47) did not allow us to choose between the coin biases $\theta_{0.4}$ and $\theta_{0.6}$, the values of the posterior probabilities (Equations 1.48 and 1.49) imply that a bias of $\theta_{0.6}$ is 3 (=0.75/0.25) times more probable than a bias is $\theta_{0.4}$.

1.4. Example 4: Light Craters

When you look at Figure 1.11, do you see a hill or a crater? Now turn the page upside-down. When you invert the page, the content of the picture does not change, but what you see does change (from a hill to a crater). This illusion almost certainly depends on the fact that your visual system assumes that the scene is lit from above. This, in turn, forces you to interpret the Figure 1.11 as a hill, and the inverted version as a crater (which it is, in reality).

In terms of Bayes' rule, the image data are equally consistent with a hill and a crater, where each interpretation corresponds to a different maximum likelihood value. Therefore, in the absence of any prior assumptions on your part, you should see the image as depicting either a hill or a crater with equal probability. However, the assumption

that light comes from above corresponds to a prior, and this effectively forces you to interpret the image as a hill or a crater, depending on whether the image is inverted or not. Note that there is no uncertainty or noise; the image is perfectly clear, but also perfectly ambiguous without the addition of a prior regarding the light source. This example demonstrates that Bayesian inference is useful even when there is no noise in the observed data, and that even the apparently simple act of seeing requires the use of prior information[10;40;41;42]:

> Seeing is not a direct apprehension of reality, as we often like to pretend. Quite the contrary: *seeing is inference from incomplete information ...*
> ET Jaynes, 2003 (p133)[18].

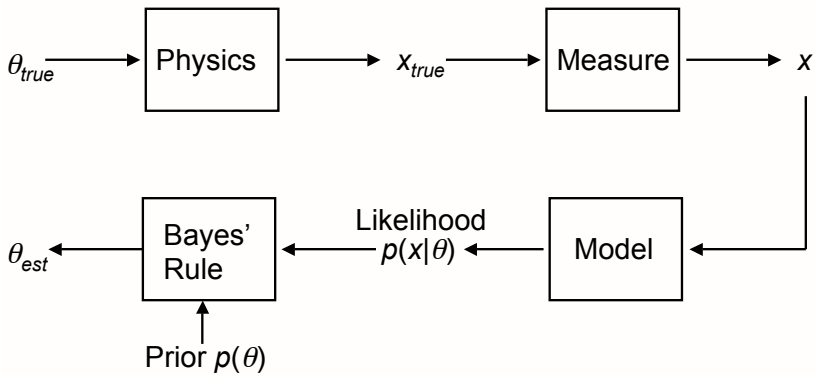

Figure 1.12: Forward and inverse probability.
Top: Forward probability. A parameter value θ_{true} (eg coin bias) which is implicit in a physical process (eg coin flipping) yields a quantity x_{true} (eg proportion of heads), which is measured as x, using an imperfect measurement process.
Bottom: Inverse probability. Given a mathematical model of the physics that generated x_{true}, the measured value x implies a range of possible values for the parameter θ. The probability of x given each possible value of θ defines a likelihood. When combined with a prior, each likelihood yields a posterior probability $p(\theta|x)$, which allows an estimate θ_{est} of θ_{true} to be obtained.

1.5. Forward and Inverse Probability

If we are given a coin with a known bias of say, $\theta = 0.6$, then the probability of a head for each coin flip is given by the likelihood $p(x_h|\theta) = 0.6$. This is an example of a *forward probability*, which involves calculating the probability of each of a number of different consequences (eg obtaining two heads) given some known cause or fact, see Figure 1.12. If this coin is flipped a 100 times then the number of heads could be 62, so the actual proportion of heads is $x_{true} = 0.62$. But, because no measurement is perfectly reliable, we may mis-count 62 as 64 heads, so the measured proportion is $x = 0.64$. Consequently, there is a difference, often called *noise*, between the true coin bias and the measured proportion of heads. The source of this noise may be due to the probabilistic nature of coin flips or to our inability to measure the number of heads accurately. Whatever the cause of the noise, the only information we have is the measured number of heads, and we must use this information as wisely as possible.

The converse of reasoning forwards from a given physical parameter or scenario involves a harder problem, also illustrated in Figure 1.12. Reasoning backwards from measurements (eg coin flips or images) amounts to finding the posterior or *inverse probability* of the value of an unobserved variable (eg coin bias, 3D shape), which is usually the cause of the observed measurement. By analogy, arriving at the scene of a crime, a detective must reason backwards from the clues, as eloquently expressed by Sherlock Holmes:

> Most people, if you describe a train of events to them, will tell you what the result would be. They can put those events together in their minds, and argue from them that something will come to pass. There are few people, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were that led to that result. This power is what I mean when I talk of reasoning backward, or analytically.
>
> Sherlock Holmes, from A Study in Scarlet. AC Doyle, 1887.

Indeed, finding inverse probabilities is precisely the problem Bayes' rule is designed to tackle.

Summary

All decisions should be based on evidence, but the best decisions should also be based on previous experience. The above examples demonstrate not only that prior experience is crucial for interpreting evidence, but also that Bayes' rule provides a rigorous method for doing so.

Chapter 2

Bayes' Rule in Pictures

> Probability theory is nothing but common sense reduced to calculation.
> Pierre-Simon Laplace (1749-1827).

Introduction

Some people understand mathematics through the use of symbols, but additional insights are often obtained if those symbols can be translated into geometric diagrams and pictures. In this chapter, once we have introduced random variables and the basic rules of probability, several different pictorial representations of probability will be used to encourage an intuitive understanding of the logic that underpins those rules. Having established a firm understanding of these rules, Bayes' rule follows using a few lines of algebra.

2.1. Random Variables

As in the previous chapter, we define the bias of a coin as its propensity to land heads up. But now we are familiar with the idea of quantities (such as coin flip outcomes) that are subject to variability, we can consider the concept of a *random variable*. The term random variable continues in use for historical reasons, although random variables are not the same as the variables used in algebra, like the x in $3x + 2 = 5$, where the variable x has a definite, but unknown, value that we can solve for.

Further Reading

Bernardo JM and Smith A (2000)[4]. Bayesian Theory. *A rigorous account of Bayesian methods, with many real-world examples.*

Bishop C (2006)[5]. Pattern Recognition and Machine Learning. *As the title suggests, this is mainly about machine learning, but it provides a lucid and comprehensive account of Bayesian methods.*

Cowan G (1998)[6]. Statistical Data Analysis. *An excellent non-Bayesian introduction to statistical analysis.*

Dienes Z (2008)[8]. Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. *Provides tutorial material on Bayes' rule and a lucid analysis of the distinction between Bayesian and frequentist statistics.*

Gelman A, Carlin J, Stern H and Rubin D (2003)[14]. Bayesian Data Analysis. *A rigorous and comprehensive account of Bayesian analysis, with many real-world examples.*

Jaynes E and Bretthorst G (2003)[18]. Probability Theory: The Logic of Science. *The modern classic of Bayesian analysis. It is comprehensive and wise. Its discursive style makes it long (600 pages) but never dull, and it is packed full of insights.*

Khan S (2012). Conditional probability with Bayes Theorem.
Salman Khan's online mathematics videos make a good introduction to various topics, including Bayes' rule.
`www.khanacademy.org`

Lee PM (2004)[27]. Bayesian Statistics: An Introduction. *A rigorous and comprehensive text with a strident Bayesian style.*

MacKay DJC (2003)[28]. Information theory, inference, and learning algorithms. *The modern classic on information theory. A very readable text that roams far and wide over many topics, almost all of which make use of Bayes' rule.*

Migon HS and Gamerman D (1999)[30]. Statistical Inference: An Integrated Approach. *A straightforward (and clearly laid out) account of inference, which compares Bayesian and non-Bayesian approaches. Despite being fairly advanced, the writing style is tutorial in nature.*

Pierce JR (1980)[34], 2nd Edition. An introduction to information theory: symbols, signals and noise. *Pierce writes with an informal, tutorial style of writing, but does not flinch from presenting the fundamental theorems of information theory.*

Reza FM (1961)[35]. An introduction to information theory. *A more comprehensive and mathematically rigorous book than the Pierce book above, and should ideally be read only after first reading Pierce's more informal text.*

Sivia DS and Skilling J (2006)[38]. Data Analysis: A Bayesian Tutorial. *This is an excellent tutorial style introduction to Bayesian methods.*

Spiegelhalter D and Rice K (2009)[36]. Bayesian statistics. Scholarpedia, 4(8):5230. `www.scholarpedia.org/article/Bayesian_statistics`
A reliable and comprehensive summary of the current status of Bayesian statistics.

Appendices

Appendix A

Glossary

Bayes' rule Given some observed data x, the posterior probability that the parameter Θ has the value θ is $p(\theta|x) = p(x|\theta)p(\theta)/p(x)$, where $p(x|\theta)$ is the likelihood, $p(\theta)$ is the prior probability of the value θ, and $p(x)$ is the marginal probability of the value x.

conditional probability The probability that the value of one random variable Θ has the value θ given that the value of another random variable X has the value x; written as $p(\Theta = \theta|X = x)$ or $p(\theta|x)$.

forward probability Reasoning forwards from the known value of a parameter to the probability of some event defines the forward probability of that event. For example, if a coin has a bias of θ then the forward probability $p(x_h|\theta)$ of observing a head x_h is θ.

independence If two variables X and Θ are independent then the value x of X provides no information regarding the value θ of the other variable Θ, and vice versa.

inverse probability
Reasoning backwards from an observed measurement x_h (eg coin flip) involves finding the posterior or inverse probability $p(\theta|x_h)$ of an unobserved parameter θ (eg coin bias).

joint probability The probability that two or more quantities simultaneously adopt specified values. For example, the probability that a coin flip yields a head x_h and that a (possibly different) coin has a bias θ is the joint probability $p(x_h, \theta)$.

likelihood The conditional probability $p(x|\theta)$ that the observed data X has the value x given a putative parameter value θ is the likelihood of θ, and is often written as $L(\theta|x)$. When considered over all values Θ of θ, $p(x|\Theta)$ defines a *likelihood function*.

marginal distribution A distribution resulting from marginalisation of a multivariate (eg 2D) distribution. For example, the 2D distribution $p(X, \Theta)$ shown in Figure 3.4 has two marginal distributions, which, in this case, are the prior distribution $p(\Theta)$ and the distribution of marginal likelihoods $p(X)$.

maximum a posteriori (MAP) Given some observed data x, the value θ_{MAP} of an unknown parameter Θ that makes the posterior probability $p(\Theta|x)$ as large as possible is the maximum a posteriori or MAP estimate of the true value θ_{true} of that parameter. The MAP takes into account both the current evidence in the form of x as well as previous knowledge in the form of prior probabilities $p(\Theta)$ regarding each value of θ.

maximum likelihood estimate (MLE) Given some observed data x, the value θ_{MLE} of an unknown parameter Θ that makes the likelihood function $p(x|\Theta)$ as large as possible is the maximum likelihood estimate (MLE) of the true value of that parameter.

noise Usually considered to be the random jitter that is part of a measured quantity.

non-informative prior See reference prior, and Section 4.8.

parameter A variable (often a random variable), which is part of an equation which, in turn, acts as a model for observed data.

posterior The posterior probability $p(\theta|x)$ is the probability that a parameter Θ has the value θ, based on current evidence (data, x) and prior knowledge. When considered over all values of θ, it refers to the posterior probability distribution $p(\Theta|x)$.

prior The prior probability $p(\theta)$ is the probability that the random variable Θ adopts the value θ. When considered over all values Θ, it is the prior probability distribution $p(\Theta)$.

probability There are many definitions of probability. The two main ones are (using coin bias as an example): 1) Bayesian probability: an observer's estimate of the probability that a coin will land heads up is based on all the information the observer has, including the proportion of times it was observed to land heads up in the past. 2) Frequentist probability: the probability that a coin will land heads up is given by the proportion of times it lands heads up, when measured over a large number of coin flips.

probability density function (pdf) The function $p(\Theta)$ of a continuous random variable Θ defines the probability density of each possible value of Θ. The probability that $\Theta = \theta$ can be considered as the probability density $p(\theta)$ (it is actually the product $p(\theta) \times d\theta$).

probability distribution The distribution of probabilities of different values of a variable. The probability distribution of a continuous variable is a *probability density function*, and the probability distribution of a discrete variable is a *probability function*. When we refer to a case which includes either continuous or discrete variables, we use the term *probability distribution* in this text.

probability function (pf) A function $p(\Theta)$ of a discrete random variable Θ defines the probability of each possible value of Θ. The probability that $\Theta = \theta$ is $p(\Theta = \theta)$ or more succinctly $p(\theta)$. This is called a *probability mass function* (pmf) in some texts.

product rule The joint probability $p(x, \theta)$ is given by the product of the conditional probability $p(x|\theta)$ and the probability $p(\theta)$; that is, $p(x, \theta) = p(x|\theta)p(\theta)$. See Appendix C.

random variable (RV) Each value of a random variable can be considered as one possible outcome of an experiment that has a number of different possible outcomes, such as the throw of a die. The set of possible outcomes is the sample space of a

random variable. A discrete random variable has a probability function (pf), which assigns a probability to each possible value. A continuous random variable has a probability density function (pdf), which assigns a probability density to each possible value. Upper case letters (eg X) refer to random variables, and (depending on context) to the set of all possible values of that variable. See Section 2.1, p31.

real number A number that can have any value corresponding to the length of a continuous line.

regression A technique used to fit a parametric curve (eg a straight line) to a set of data points.

reference prior A prior distribution that is 'fair'. See Section 4.8, p99 and Appendix H, p169.

standard deviation The standard deviation of a variable is a measure of how 'spread out' its values are. If we have a sample of n values of a variable x then the standard deviation of our sample is

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{A.1}$$

where $\bar{x}$ is the mean of our sample. The sample's variance is σ^2.

sum rule This states that the probability $p(x)$ that $X = x$ is the sum of joint probabilities $p(x, \Theta)$, where this sum is taken over all N possible values of Θ,

$$p(x) = \sum_{i=1}^{N} p(x, \theta_i).$$

Also known as the law of total probability. See Appendix C.

variable A variable is essentially a 'container', usually for one number. We use the lower case (eg x) to refer to a particular value of a variable.

References

[1] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil Trans Roy Soc London*, 53:370–418.

[2] Beaumont, M. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, pages 251–261.

[3] Bernardo, J. (1979). Reference posterior distributions for Bayesian inference. *J. Royal Statistical Society B*, 41:113–147.

[4] Bernardo, J. and Smith, A. (2000). *Bayesian Theory*. John Wiley and Sons Ltd.

[5] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

[6] Cowan, G. (1998). *Statistical Data Analysis*. OUP.

[7] Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, 14:113.

[8] Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan.

[9] Donnelly, P. (2005). Appealing statistics. *Significance*, 2(1):46–48.

[10] Doya, K., Ishii, S., Pouget, A., and Rao, R. (2007). *The Bayesian Brain*. MIT, MA.

[11] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

[12] Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

[13] Geisler, W. and Diehl, R. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society London (B) Biology*, 357:419–448.

[14] Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis, Second Edition*. Chapman and Hall, 2nd edition.

[15] Geman, S. and Geman, D. (1993). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics*, 20:25–62.

[16] Good, I. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, 66(2):393–396.

[17] Hobson, M., Jaffe, A., Liddle, A., and Mukherjee, P. (2009). *Bayesian Methods in Cosmology*. Cambridge University Press.

[18] Jaynes, E. and Bretthorst, G. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

[19] Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

[20] Jones, M. and Love, B. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34:192–193.

[21] Kadane, J. (2009). Bayesian thought in early modern detective stories: Monsieur Lecoq, C. Auguste Dupin and Sherlock Holmes. *Statistical Science*, 24(2):238–243.

[22] Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. *Ann Rev Psychology*, 55(1):271–304.

[23] Knill, D. and Richards, R. (1996). *Perception as Bayesian inference*. Cambridge University Press, New York, NY, USA.

[24] Kolmogorov, A. (1933). *Foundations of the Theory of Probability*. Chelsea Publishing Company, (English translation, 1956).

[25] Land, M. and Nilsson, D. (2002). *Animal eyes*. OUP.

[26] Lawson, A. (2008). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. Chapman and Hall.

[27] Lee, P. (2004). *Bayesian Statistics: An Introduction*. Wiley.

[28] MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

[29] McGrayne, S. (2011). *The Theory That Would Not Die*. YUP.

[30] Migon, H. and Gamerman, D. (1999). *Statistical Inference: An Integrated Approach*. Arnold.

[31] Oaksford, M. and Chater, N. (2007). *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford University Press.

[32] Parent, E. and Rivot, E. (2012). *Bayesian Modeling of Ecological Data*. Chapman and Hall.

[33] Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362.

[34] Pierce, J. (1961 reprinted by Dover 1980). *An introduction to information theory: symbols, signals and noise*. Dover (2nd Edition).

[35] Reza, F. (1961). *Information Theory*. New York, McGraw-Hill.

[36] Rice, D. and Spiegelhalter, K. (2009). Bayesian statistics. *Scholarpedia*, 4(3):5230.

[37] Simpson, E. (2010). Edward Simpson: Bayes at Bletchley park. *Significance*, 7(2).

[38] Sivia, D. and Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial*. OUP.

[39] Stigler, S. (1983). Who discovered Bayes's theorem? *The American Statistician*, 37(4):290–296.

[40] Stone, J. (2011). Footprints sticking out of the sand (Part II): Children's Bayesian priors for lighting direction and convexity. *Perception*, 40(2):175–190.

[41] Stone, J. (2012). *Vision and Brain: How we perceive the world*. MIT Press.

[42] Stone, J., Kerrigan, I., and Porrill, J. (2009). Where is the light? Bayesian perceptual priors for lighting direction. *Proceedings Royal Society London (B)*, 276:1797–1804.

[43] Taroni, F., Aitken, C., Garbolino, P., and Biedermann, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science*. Wiley.

[44] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Index

About the Author

Dr James Stone is a Reader in Vision and Computational Neuroscience at the University of Sheffield, England.

Information Theory: A Tutorial Introduction, JV Stone, Sebtel Press, 2015.

Vision and Brain: How We Perceive the World, JV Stone, MIT Press, 2012.

Seeing: The Computational Approach to Biological Vision, JP Frisby and JV Stone, MIT Press, 2010.

Independent Component Analysis: A Tutorial Introduction, JV Stone, MIT Press, 2004.