Chapter 2

Introduction

Every day, throughout our lives, we are required to believe certain things and not to believe other things. This applies not only to the "big questions" of life, but also to trivial matters, and everything in between. For example, this morning I boarded the bus to university, sure that it would actually take me here and not to Wellington. How did I know the bus would not take me to Wellington? Well, for starters I have taken the same bus many times before and it has always taken me to the university. Another clue was that the bus said "Midtown" on it, and a bus to Wellington probably would have said Wellington, and would not have stopped at a minor bus stop in suburban Auckland. None of this evidence *proves* that the bus would take me to university, but it does makes it very *plausible*. Given all these pieces of information, I feel quite certain that the bus will take me to the city. I feel so certain about this that the possibility of an unplanned trip to Wellington never even entered my mind until I decided to write this paragraph.

Somehow, our brains are very often able to accurately predict the correct answer to many questions (e.g. the destination of a bus), even though we don't have all the available information that we would need to be 100% certain. We do this using our experience of the world and our intuition, usually without much conscious attention or problem solving. However, there are areas of study where we can't just use our intuition to make judgments like this. For example, most of science involves such situations. Does a new treatment work better than an old one? Is the expansion of the universe really accelerating? People tend to be interested in trying to answer questions that haven't been answered yet, so our attention is always on the questions where we're not sure of the answer. This is where statistics comes in as a tool to help us in this grey area, when we can't be 100% certain about things, but we still want to do the best we can with our incomplete information.

2.1 Certainty, Uncertainty and Probability

In the above example, I said things like "I couldn't be 100% certain". The idea of using a number to describe how certain you are is quite natural. For example, contestants on the TV show "Who Wants to be a Millionaire" often say things like "I'm 75% sure the answer

is A"[1].

There are some interesting things to notice about this statement. Firstly, it is a subjective statement. If someone else were in the seat trying to answer the question, she might say the probability that A is correct is 100%, because she knows the answer! A third person faced with the same question might say the probability is 25%, because he has no idea and only knows that one of the four answers must be correct.

In Bayesian statistics, the interpretation of what *probability* means is that it is a description of *how certain you are that some statement, or proposition, is true*. If the probability is 1, you are sure that the statement is true. So sure, in fact, that nothing could ever change your mind (we will demonstrate this in class). If the probability is 0, you are sure that the proposition is false. If the probability is 0.5, then you are as uncertain as you would be about a fair coin flip. If the probability is 0.95, then you're quite sure the statement is true, but it wouldn't be *too* surprising to you if you found out the statement was false. See Figure 2.1 for a graphical depiction of probabilities as degrees of certainty or plausibility.

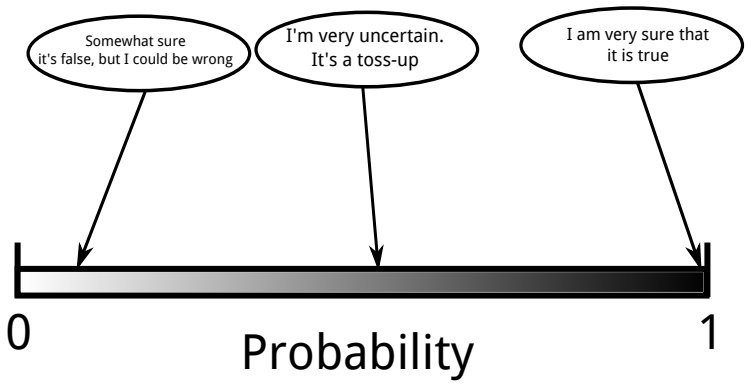

Figure 2.1: *Probability can be used to describe degrees of certainty, or how plausible some statement is. 0 and 1 are the two extremes of the scale and correspond to complete certainty. However, probabilities are not static quantities. When you get more information, your probabilities can change.*

In Bayesian statistics, probabilities are in the mind, not in the world.

It might sound like there is nothing more to Bayesian statistics than just thinking about a question and then blurting out a probability that feels appropriate. Fortunately for us, there's more to it than that! To see why, think about how you change your mind when new evidence (such as a data set) becomes available. For example, you may be on "Who Wants to be a Millionaire?" and not know the answer to a question, so you might think the probability that it is A is 25%. But if you call your friend using "phone a friend", and your friend says, "It's definitely A", then you would be much more confident that it is A! Your probability probably wouldn't go all the way to 100% though, because there is

[1]This reminds me of an amusing exchange from the TV show *Monk*. **Captain Stottlemeyer**: [about someone electrocuting her husband] Monk, are you sure? I mean, are you really sure? And don't give me any of that "95 percent" crap. **Monk**: Captain, I am 100% sure... that she *probably* killed him.

always the small possibility that your friend is mistaken.

When we get new information, we should *update* our probabilities to take the new information into account. Bayesian methods tell us exactly how to do this.

In this course, we will learn how to do *data analysis* from a Bayesian point of view. So while the discussion in this chapter might sound a bit like philosophy, we will see that using this kind of thinking can give us new and powerful ways of solving practical data analysis problems. The methods we will use will all have a common structure, so if you are faced with a completely new data analysis problem one day, you will be able to design your own analysis methods by using the Bayesian framework. Best of all, the methods make sense and perform extremely well in practice!

Chapter 3

First Examples

We will now look at a simple example to demonstrate the basics of how Bayesian statistics works. We start with some probabilities at the beginning of the problem (these are called *prior probabilities*), and how exactly these get updated when we get more information (these updated probabilities are called *posterior probabilities*). To help make things more clear, we will use a table that we will call a *Bayes' Box* to help us calculate the posterior probabilities easily.

Suppose there are two balls in a bag. We know in advance that at least one of them is black, but we're not sure whether they're both black, or whether one is black and one is white. These are the only two possibilities we will consider. To keep things concise, we can label our two competing hypotheses. We could call them whatever we want, but I will call them `BB` and `BW`. So, at the beginning of the problem, we know that *one and only one* of the following statements/hypotheses is true:

`BB`: Both balls are black `BW`: One ball is black and the other is white.

Suppose an experiment is performed to help us determine which of these two hypotheses is true. The experimenter reaches into the bag, pulls out one of the balls, and observes its colour. The result of this experiment is (drumroll please!):

D: The ball that was removed from the bag was black.

We will now do a Bayesian analysis of this result.

3.1 The Bayes' Box

A Bayesian analysis starts by choosing some values for the prior probabilities. We have our two competing hypotheses `BB` and `BW`, and we need to choose some probability values to describe how sure we are that each of these is true. Since we are talking about two hypotheses, there will be two prior probabilities, one for `BB` and one for `BW`. For simplicity,

we will assume that we don't have much of an idea which is true, and so we will use the following prior probabilities:

$$P(\texttt{BB}) = 0.5 \quad (3.1)$$
$$P(\texttt{BW}) = 0.5. \quad (3.2)$$

Pay attention to the notation. The upper case P stands for probability, and if we just write $P(\texttt{whatever})$, that means we are talking about the prior probability of `whatever`. We will see the notation for the posterior probability shortly. Note also that since the two hypotheses are mutually exclusive (they can't both be true) and exhaustive (one of these is true, it can't be some undefined third option). We will almost always consider mutually exclusive and exhaustive hypotheses in this course[1].

The choice of 0.5 for the two prior probabilities describes the fact that, before we did the experiment, we were very uncertain about which of the two hypotheses was true. I will now present a *Bayes' Box*, which lists all the hypotheses (in this case two) that might be true, and the prior probabilities. There are some extra columns which we haven't discussed yet, and will be needed in order to figure out the posterior probabilities in the final column. The first column of a Bayes' Box is just the list of hypotheses we are considering. In

Hypotheses	`prior`	`likelihood`	`prior` × `likelihood`	`posterior`
`BB`	0.5			
`BW`	0.5			
Totals:	1			

this case there are just two. If you need to construct a Bayes' box for a new problem, just think about what the possible answers to the problem are, and list them in the first column. The second column lists the prior probabilities for each of the hypotheses. Above, before we did the experiment, we decided to say that there was a 50% probability that `BB` is true and a 50% probability that `BW` is true, hence the 0.5 values in this column. The prior column should always sum to 1. Remember, the prior probabilities only describe our initial uncertainty, before taking the data into account. Hopefully the data will help by changing these probabilities to something a bit more decisive.

3.1.1 Likelihood

The third column is called *likelihood*, and this is a really important column where the action happens. The likelihood is a quantity that will be used for calculating the posterior probabilities. In colloquial language, likelihood is synonymous with probability. It means the same thing. However, in statistics, likelihood is a very specific kind of probability. To fill in the third column of the Bayes' Box, we need to calculate two likelihoods, so you can tell from this that the likelihood is something different for each hypothesis. But what is it exactly?

[1] If this does not appear to be true in a particular problem, it is usually possible to redefine the various hypotheses into a set that of hypotheses that *are* mutually exclusive and exhaustive.

The likelihood for a hypothesis is the probability that you would have observed the data, if that hypothesis were true. The values can be found by going through each hypothesis in turn, imagining it is true, and asking, "What is the probability of getting the data that I observed?".

Here is the Bayes' Box with the likelihood column filled in. I will explain how these numbers were calculated in a bit more detail in the next subsection. If you have taken

Hypotheses	`prior`	`likelihood`	`h = prior` × `likelihood`	`posterior`
`BB`	0.5	1		
`BW`	0.5	0.5		
Totals:	1			

STATS 210 and used the maximum likelihood method, where you find the value of a parameter that maximises the likelihood function, that is the same as the likelihood we use in this course! So you have a head start in understanding this concept.

3.1.2 Finding the Likelihood Values

We will first calculate the value of the likelihood for the `BB` hypothesis. Remember, the data we are analysing here is that we chose one of the balls in the bag "at random", and it was black. The likelihood for the `BB` hypothesis is therefore the probability that we would get a black ball if `BB` is true.

Imagine that `BB` is true. That means both balls are black. What is the probability that the experiment would result in a black ball? That's easy – it's 100%! So we put the number 1 in the Bayes Box as the likelihood for the `BB` hypothesis.

Now imagine instead that `BW` is true. That would mean one ball is black and the other is white. If this were the case and we did the experiment, what would be the probability of getting the black ball in the experiment? Since one of the two balls is black, the chance of choosing this one is 50%. Therefore, the likelihood for the `BW` hypothesis is 0.5, and that's why I put 0.5 in the Bayes' Box for the likelihood for `BW`.

In general, the likelihood is the *probability of the data that you actually got, assuming a particular hypothesis is true.* In this example it was fairly easy to get the likelihoods directly by asking "if this hypothesis is true, what is the probability of getting the black ball when we do the experiment?". Sometimes this is not so easy, and it can be helpful to think about ALL possible experimental outcomes/data you might have seen – even though ultimately, you just need to select the one that actually occurred. Table 3.1 shows an example of this process.

The fact that only the blue probabilities in Table 3.1 enter the Bayes' Box calculation is related to the *likelihood principle*, which we will discuss in lectures. Note also that in Table 3.1, the probabilities for the different possible data sets add to 1 within each hypothesis, but the sum of the blue "selected" likelihood values is not 1 (it is, in fact, meaningless).

Hypotheses	Possible Data	Probability
BB	Black Ball	1
	White Ball	0
BW	Black Ball	0.5
	White Ball	0.5

Table 3.1: *This table demonstrates a method for calculating the likelihood values, by considering not just the data that actually occurred, but all data that might have occurred. Ultimately, it is only the probability of the data which actually occurred that matters, so this is highlighted in blue.*

When we come to parameter estimation in later chapters, we will usually set up our problems in this way, by considering what data sets are possible, and assigning probabilities to them.

3.1.3 The Mechanical Part

The third column of the Bayes' Box is the product of the prior probabilities and the likelihoods, calculated by simple multiplication. The result will be called "prior times likelihood", but occasionally we will use the letter h for these quantities. This is the *unnormalised* posterior. It does not sum to 1 as the posterior probabilities should, but it is at least proportional to the actual posterior probabilities.

To find the posterior probabilities, we take the `prior` $\times$ `likelihood` column and divide it by its sum, producing numbers that do sum to 1. This gives us the final posterior probabilities, which were the goal all along. The completed Bayes' Box is shown below:

Hypotheses	`prior`	`likelihood`	`h = prior` $\times$ `likelihood`	`posterior`
BB	0.5	1	0.5	0.667
BW	0.5	0.5	0.25	0.333
Totals:	1		0.75	1

We can see that the posterior probabilities are not the same as the prior probabilities, because we have more information now! The experimental result made BB a little bit more plausible than it was before. Its probability has increased from 1/2 to 2/3.

3.1.4 Interpretation

The posterior probabilities of the hypotheses are proportional to the prior probabilies and the likelihoods. A high prior probability will help a hypothesis have a high posterior probability. A high likelihood value also helps. To understand what this means about reasoning, consider the meanings of the prior and the likelihood. There are two things that can contribute to a hypothesis being plausible:

- If the prior probability is high. That is, the hypothesis was *already* plausible, before we got the data.
- If the hypothesis *predicted the data* well. That is, the data was what we would have expected to occur if the hypothesis had been true.

I hope you agree that this is all very sensible.

In class we will also study variations on this problem, considering different assumptions about the prior probabilities and how they affect the results, and also considering what happens when we get more and/or different data.

3.2 Bayes' Rule

Bayes' rule is an equation from probability theory, shown in Figure 3.1. The various terms in Bayes' rule are all probabilities, but notice that there are conditional probabilities in there. For example, the left hand side of the equation is $P(A|B)$ and that means the probability of A **given** B. That is, it's the probability of A after taking into account the information B. In other words, $P(A|B)$ is a posterior probability, and Bayes' rule tells us how to calculate it from other probabilities. Bayes' rule is true for *any* statements A

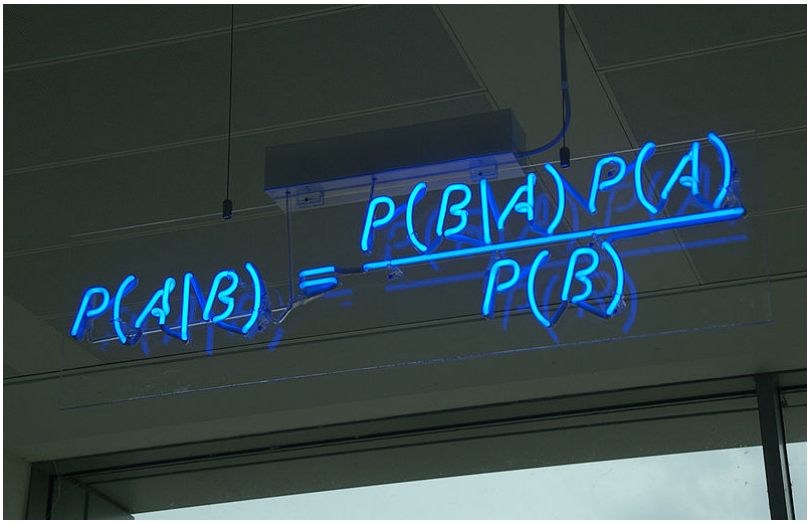

Figure 3.1: *A blue neon sign displaying Bayes' rule. You can use it to calculate the probability of A given B, if you know the values of some other probabilities on the right hand side. Image credit: Matt Buck. Obtained from Wikimedia Commons.*

and B. If you took the equation in Figure 3.1 and replaced A with "Kākāpō will survive beyond 2050" and B with "I had coffee this morning", the resulting equation would still be true[2].

It is helpful to relabel A and B in Bayes' rule to give a more clear interpretation of how the equation is to be used. In this version of Bayes' rule (which is one you should commit to memory), A has been replaced by H, and B has been replaced by D. The reason for these letters is that you should interpret H as *hypothesis* and D as *data*. Then you can interpret Bayes' rule as telling you the probability of a hypothesis given some data, in

[2]It would still be true, but it would not very interesting, because whether or not I had coffee doesn't tell you much about the survival prospects of endangered New Zealand parrots.

other words, a posterior probability.

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} \tag{3.3}$$

In Bayesian statistics, most of the terms in Bayes' rule have special names. Some of them even have more than one name, with different scientific communities preferring different terminology. Here is a list of the various terms and the names we will use for them:

- $P(H|D)$ is the **posterior probability**. It describes how certain or confident we are that hypothesis H is true, given that we have observed data D. Calculating posterior probabilities is the main goal of Bayesian statistics!
- $P(H)$ is the **prior probability**, which describes how sure we were that H was true, before we observed the data D.
- $P(D|H)$ is the **likelihood**. If you were to assume that H is true, this is the probability that you would have observed data D.
- $P(D)$ is the **marginal likelihood**. This is the probability that you would have observed data D, *whether H is true or not*.

Since you may encounter Bayesian methods outside of STATS 331, I have included an Appendix called "Rosetta Stone" that lists some common alternative terminology.

In the above example, we did some calculations to work out the numbers in the Bayes' Box, particularly the posterior probabilities, which are the ultimate goal of the calculation. *What we were actually doing in these calculations was applying Bayes' rule.* We actually applied Bayes' rule twice, once to compute $P(\texttt{BB}|D)$ and a second time to calculate $P(\texttt{BW}|D)$.

When you use a Bayes' Box to calculate posterior probabilities, you are really just applying Bayes' rule a lot of times: once for each hypothesis listed in the first column.

3.3 Phone Example

This example is based on Question 1 from the 2012 final exam. I got the idea for this question from an example in David MacKay's wonderful book "Information Theory, Inference and Learning Algorithms" (available online as a free PDF download. You're welcome to check it out, but it is a large book and only about 20% of the content is relevant to this course!).

You move into a new house which has a phone installed. You can't remember the phone number, but you suspect it might be `555-3226` (some of you may recognise this as being the phone number for Homer Simpson's "Mr Plow" business). To test this hypothesis, you carry out an experiment by picking up the phone and dialing `555-3226`.

If you are correct about the phone number, you will definitely hear a busy signal because you are calling yourself. If you are incorrect, the probability of hearing a busy signal is 1/100. However, all of that is only true if you assume the phone is working, and it might be broken! If the phone is broken, it will always give a busy signal.

When you do the experiment, the outcome (the data) is that you do actually get the busy signal. The question asked us to consider the following four hypotheses, and to calculate their posterior probabilities: Note that the four hypotheses are mutually exclusive and

Hypothesis	Description	Prior Probability
H_1	Phone is working and `555-3226` is correct	0.4
H_2	Phone is working and `555-3226` is incorrect	0.4
H_3	Phone is broken and `555-3226` is correct	0.1
H_4	Phone is broken and `555-3226` is incorrect	0.1

Table 3.2: *The four hypotheses about the state of the phone and the phone number. The prior probabilities are also given.*

exhaustive. If you were to come up with hypotheses yourself, "phone is working" and "`555-3226` is correct" might spring to mind. They wouldn't be mutually exclusive so you couldn't do a Bayes' Box with just those two, but it is possible to put these together (using "**and**") to make the four mutually exclusive options in the table.

3.3.1 Solution

We will go through the solution using a Bayes' Box. The four hypotheses listed in Table 3.2 and their prior probabilities are given, so we can fill out the first two columns of a Bayes' Box right away: The next thing we need is the likelihoods. The outcome of the experiment

Hypotheses	`prior`	`likelihood`	`prior` $\times$ `likelihood`	`posterior`
H_1	0.4			
H_2	0.4			
H_3	0.1			
H_4	0.1			
Totals:	1			

(the data) was the busy signal, so we need to work out $P(\text{busy signal}|H)$ for each H in the problem (there are four of them). Let's start (naturally!) with H_1.

If we assume H_1 is true, then the phone is working and `555-3226` is the correct phone number. In that case, we would definitely get a busy signal because we are calling ourselves. Therefore $P(\text{busy signal}|H_1) = 1$ is our first likelihood value.

Next, let's imagine that H_2 is true, so the phone is working, but `555-3226` is not the right phone number. In this case, it is given in the question that the probability of getting a busy signal is 1/100 or 0.01 (in reality, this would be based on some other data, or perhaps be a totally subjective judgement). Therefore $P(\text{busy signal}|H_2) = 0.01$, and that's our second likelihood value.

The likelihoods for H_3 and H_4 are quite straightforward because they both imply the phone is broken, and that means a busy signal is certain. Therefore $P(\text{busy signal}|H_3) = P(\text{busy signal}|H_4) = 1$. We have our four likelihoods, and can proceed to work out everything in the Bayes' Box, including the main goal – the posterior probabilities! Here it is:

Hypotheses	`prior`	`likelihood`	`prior` × `likelihood`	`posterior`
H_1	0.4	1	0.4	0.662
H_2	0.4	0.01	0.004	0.00662
H_3	0.1	1	0.1	0.166
H_4	0.1	1	0.1	0.166
Totals:	1		0.604	1

To conclude this phone problem, I should admit that I actually calculated the numbers in the Bayes' Box using R. My code is shown below. A lot of the code we write in labs will look like this. Obviously in the 2012 exam the students had to use their calculators instead.

```
prior = c(0.4, 0.4, 0.1, 0.1) # Vector of prior probs
lik = c(1, 0.01, 1, 1)        # Vector of likelihoods
h = prior*lik
Z = sum(h)                    # Sum of prior times likelihood
post = prior*lik/Z            # Normalise to get posterior
# Look at all the results
print(prior)
print(lik)
print(h)
print(Z)
print(post)
```

Now let's try to see if this makes sense. There are many things we could think about, but let's just consider the question of whether the phone is working or not. The first two hypotheses correspond to the phone being in a working state. If you want to calculate the probability of A **or** B, then you can just add the probabilities if they are mutually exclusive. The prior probability that the phone is working is therefore:

$$\begin{aligned} P(\text{phone working}) &= P(H_1 \vee H_2) && (3.4)\\ &= P(H_1) + P(H_2) && (3.5)\\ &= 0.4 + 0.4 && (3.6)\\ &= 0.8. && (3.7) \end{aligned}$$

Here, I have introduced the notation $\vee$, meaning "logical or": For any two propositions A, B, the proposition $(A \vee B)$ is true if either one of A or B is true (or both).

The posterior probability is worked out in a similar way, but using the posterior probabilities instead of the prior ones:

$$\begin{aligned} P(\text{phone working}|\text{busy signal}) &= P(H_1 \vee H_2|\text{busy signal}) && (3.8)\\ &= P(H_1|\text{busy signal}) + P(H_2|\text{busy signal}) && (3.9)\\ &= 0.662 + 0.00662 && (3.10)\\ &= 0.6689. && (3.11) \end{aligned}$$

Our probability that the phone is working has gone down a little bit as a result of this evidence! That makes sense to me. A busy signal is what you would expect to happen if the phone was broken. This data doesn't *prove* the phone is broken, but it does point in

that direction a little bit, and hence the probability that the phone is working has been reduced from 0.8 to 0.6689.

3.4 Important Equations

Posterior probabilities are calculated using Bayes' rule. For a single hypothesis H given data D, Bayes' rule is:

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} \tag{3.12}$$

This gives the posterior probability $P(H|D)$ in terms of the prior probability $P(H)$, the likelihood $P(D|H)$ and the marginal likelihood $P(D)$ in the denominator. To obtain $P(H)$, think about your prior beliefs (which may indicate a large amount of uncertainty, or may already be well informed based on previous data sets). To obtain $P(D|H)$, think about what the experiment is doing: If H is true, what data would you expect to see and with what probabilities?

The denominator is the probability of obtaining the data D but without assuming that H is either true or false. This is obtained using the sum rule. There are two ways that the data D could occur, either via the route of H being true (this has probability $P(H)P(D|H)$), or via the route of H being false (this has probability $P(\bar{H})P(D|\bar{H})$). These two ways are mutually exclusive, so we can add their probabilities:

$$P(D) = P(H)P(D|H) + P(\bar{H})P(D|\bar{H}). \tag{3.13}$$

Bayes' rule can be applied to a whole set of hypotheses (that are mutually exclusive and exhaustive) simultaneously. This is a more common way of using it, and it is the way we use it when we use a Bayes' Box. If we applied Equation 3.12 to N hypotheses $H_1, H_2, ..., H_N$, given data D, we would get the following for the posterior probability of each hypothesis H_i (for $i = 1, 2, ..., N$):

$$P(H_i|D) \;=\; \frac{P(H_i)P(D|H_i)}{P(D)} \tag{3.14}$$

The denominator $P(D)$ is a single number. It does not depend on the index i. It can again be obtained using the sum rule. There are N mutually exclusive ways that the data D could have occurred: via H_1 being true, or via H_2 being true, etc. Adding the probabilities of these gives:

$$P(D) \;=\; \sum_{i=1}^{N} P(H_i)P(D|H_i). \tag{3.15}$$

which just happens to be the sum of the prior times likelihood values. If you don't find equations particularly easy to read, just remember that following the steps for making a Bayes' Box is equivalent to applying Bayes' rule in this form! The $P(H_i)$ values are the prior probability column, the $P(D|H_i)$ values are the likelihood column, and the denominator is the sum of the prior times likelihood column. For example, the posterior

probability for H_1 (the top right entry in a Bayes' Box) is given by the prior probability for H_1 times the likelihood for H_1, divided by the sum of prior times likelihood values. That is, $P(H_1|D) = P(H_1)P(D|H_1)/P(D)$. The correspondence between the probabilities that go in a Bayes' Box (in general) and the terms in the Equations are given in Table 3.3.

Hypotheses	`prior`	`likelihood`	`prior` $\times$ `likelihood`	`posterior`
H_1	$P(H_1)$	$P(D\|H_1)$	$P(H_1) \times P(D\|H_1)$	$P(H_1\|D)$
H_2	$P(H_2)$	$P(D\|H_2)$	$P(H_2) \times P(D\|H_2)$	$P(H_2\|D)$
...	...	...	...	...
Totals:	1		$P(D)$	1

Table 3.3: *A general Bayes' Box. Using Bayes' rule or making a Bayes' Box are actually the same thing, and this table can be used to identify the terms.*