About Chapter 3

If you are eager to get on to information theory, data compression, and noisy channels, you can skip to Chapter 4. Data compression and data modelling are intimately connected, however, so you'll probably want to come back to this chapter by the time you get to Chapter 6. Before reading Chapter 3, it might be good to look at the following exercises.

▷ Exercise 3.1.[2, p.59] A die is selected at random from two twenty-faced dice on which the symbols 1–10 are written with nonuniform frequency as follows.

Symbol	1	2	3	4	5	6	7	8	9	10
Number of faces of die A	6	4	3	2	1	1	1	1	1	0
Number of faces of die B	3	3	2	2	2	2	2	2	1	1

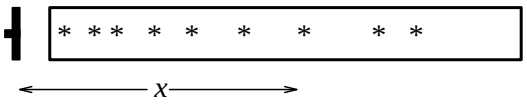
The randomly chosen die is rolled 7 times, with the following outcomes:

5, 3, 9, 3, 8, 4, 7.

What is the probability that the die is die A?

▷ Exercise 3.2.[2, p.59] Assume that there is a third twenty-faced die, die C, on which the symbols 1–20 are written once each. As above, one of the three dice is selected at random and rolled 7 times, giving the outcomes: 3, 5, 4, 8, 3, 9, 7.
What is the probability that the die is (a) die A, (b) die B, (c) die C?

Exercise 3.3.[3, p.48] Inferring a decay constant
Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ. Decay events can be observed only if they occur in a window extending from $x = 1\,\text{cm}$ to $x = 20\,\text{cm}$. N decays are observed at locations $\{x_1, \ldots, x_N\}$. What is λ?



▷ Exercise 3.4.[3, p.55] Forensic evidence
Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and of type 'AB' (a rare type, with frequency 1%). Do these data (type 'O' and 'AB' blood were found at scene) give evidence in favour of the proposition that Oliver was one of the two people present at the crime?

3

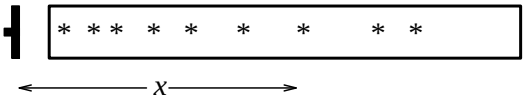
More about Inference

It is not a controversial statement that Bayes' theorem provides the correct language for describing the inference of a message communicated over a noisy channel, as we used it in Chapter 1 (p.6). But strangely, when it comes to other inference problems, the use of Bayes' theorem is not so widespread.

▶ 3.1 A first inference problem

When I was an undergraduate in Cambridge, I was privileged to receive supervisions from Steve Gull. Sitting at his desk in a dishevelled office in St. John's College, I asked him how one ought to answer an old Tripos question (exercise 3.3):

> Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ. Decay events can be observed only if they occur in a window extending from $x = 1\,\mathrm{cm}$ to $x = 20\,\mathrm{cm}$. N decays are observed at locations $\{x_1, \ldots, x_N\}$. What is λ?



I had scratched my head over this for some time. My education had provided me with a couple of approaches to solving such inference problems: constructing 'estimators' of the unknown parameters; or 'fitting' the model to the data, or to a processed version of the data.

Since the mean of an unconstrained exponential distribution is λ, it seemed reasonable to examine the sample mean $\bar{x} = \sum_n x_n/N$ and see if an estimator $\hat{\lambda}$ could be obtained from it. It was evident that the estimator $\hat{\lambda} = \bar{x} - 1$ would be appropriate for $\lambda \ll 20\,\mathrm{cm}$, but not for cases where the truncation of the distribution at the right-hand side is significant; with a little ingenuity and the introduction of ad hoc bins, promising estimators for $\lambda \gg 20$ cm could be constructed. But there was no obvious estimator that would work under all conditions.

Nor could I find a satisfactory approach based on fitting the density $P(x \mid \lambda)$ to a histogram derived from the data. I was stuck.

What is the general solution to this problem and others like it? Is it always necessary, when confronted by a new inference problem, to grope in the dark for appropriate 'estimators' and worry about finding the 'best' estimator (whatever that means)?

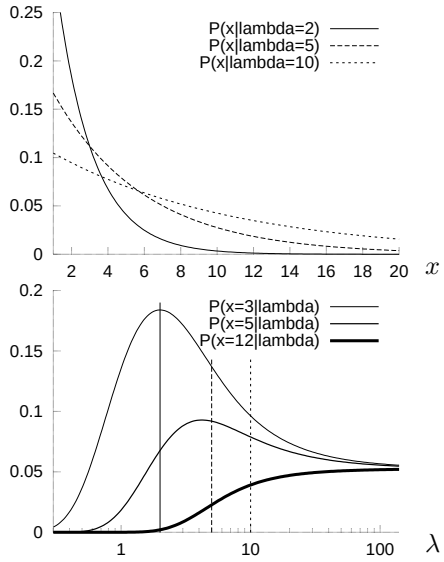


Figure 3.1. The probability density $P(x \mid \lambda)$ as a function of x.

Figure 3.2. The probability density $P(x \mid \lambda)$ as a function of λ, for three different values of x. When plotted this way round, the function is known as the *likelihood* of λ. The marks indicate the three values of λ, $\lambda = 2, 5, 10$, that were used in the preceding figure.

Steve wrote down the probability of one data point, given λ:

$$P(x \mid \lambda) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} / Z(\lambda) & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where

$$Z(\lambda) = \int_1^{20} \mathrm{d}x \, \tfrac{1}{\lambda} e^{-x/\lambda} = \left(e^{-1/\lambda} - e^{-20/\lambda} \right). \tag{3.2}$$

This seemed obvious enough. Then he wrote *Bayes' theorem*:

$$P(\lambda \mid \{x_1, \ldots, x_N\}) = \frac{P(\{x\} \mid \lambda) P(\lambda)}{P(\{x\})} \tag{3.3}$$

$$\propto \frac{1}{(\lambda Z(\lambda))^N} \exp\left(- \textstyle\sum_1^N x_n / \lambda \right) P(\lambda). \tag{3.4}$$

Suddenly, the straightforward distribution $P(\{x_1, \ldots, x_N\} \mid \lambda)$, defining the probability of the data given the hypothesis λ, was being turned on its head so as to define the probability of a hypothesis given the data. A simple figure showed the probability of a single data point $P(x \mid \lambda)$ as a familiar function of x, for different values of λ (figure 3.1). Each curve was an innocent exponential, normalized to have area 1. Plotting the same function as a function of λ for a fixed value of x, something remarkable happens: a peak emerges (figure 3.2). To help understand these two points of view of the one function, figure 3.3 shows a surface plot of $P(x \mid \lambda)$ as a function of x and λ.

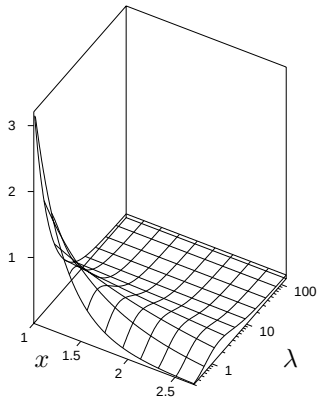


Figure 3.3. The probability density $P(x \mid \lambda)$ as a function of x and λ. Figures 3.1 and 3.2 are vertical sections through this surface.

For a dataset consisting of several points, e.g., the six points $\{x\}_{n=1}^N = \{1.5, 2, 3, 4, 5, 12\}$, the likelihood function $P(\{x\} \mid \lambda)$ is the product of the N functions of λ, $P(x_n \mid \lambda)$ (figure 3.4).

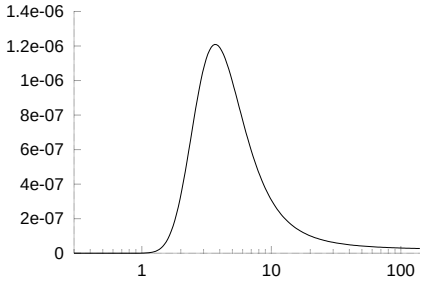


Figure 3.4. The likelihood function in the case of a six-point dataset, $P(\{x\} = \{1.5, 2, 3, 4, 5, 12\} \mid \lambda)$, as a function of λ.

Steve summarized Bayes' theorem as embodying the fact that

> what you know about λ after the data arrive is what you knew before $[P(\lambda)]$, and what the data told you $[P(\{x\} \mid \lambda)]$.

Probabilities are used here to quantify degrees of belief. To nip possible confusion in the bud, it must be emphasized that the hypothesis λ that correctly describes the situation is *not* a *stochastic* variable, and the fact that the Bayesian uses a probability distribution P does *not* mean that he thinks of the world as stochastically changing its nature between the states described by the different hypotheses. He uses the notation of probabilities to represent his *beliefs* about the mutually exclusive micro-hypotheses (here, values of λ), of which only one is actually true. That probabilities can denote degrees of belief, given assumptions, seemed reasonable to me.

The posterior probability distribution (3.4) represents the unique and complete solution to the problem. There is no need to invent 'estimators'; nor do we need to invent criteria for comparing alternative estimators with each other. Whereas orthodox statisticians offer twenty ways of solving a problem, and another twenty different criteria for deciding which of these solutions is the best, Bayesian statistics only offers one answer to a well-posed problem.

If you have any difficulty understanding this chapter I recommend ensuring you are happy with exercises 3.1 and 3.2 (p.47) then noting their similarity to exercise 3.3.

Assumptions in inference

Our inference is conditional on our assumptions [for example, the prior $P(\lambda)$]. Critics view such priors as a difficulty because they are 'subjective', but I don't see how it could be otherwise. How can one perform inference without making assumptions? I believe that it is of great value that Bayesian methods force one to make these tacit assumptions explicit.

First, once assumptions are made, the inferences are objective and unique, reproducible with complete agreement by anyone who has the same information and makes the same assumptions. For example, given the assumptions listed above, $\mathcal{H}$, and the data D, everyone will agree about the posterior probability of the decay length λ:

$$P(\lambda \mid D, \mathcal{H}) = \frac{P(D \mid \lambda, \mathcal{H}) P(\lambda \mid \mathcal{H})}{P(D \mid \mathcal{H})}. \tag{3.5}$$

Second, when the assumptions are explicit, they are easier to criticize, and easier to modify – indeed, we can quantify the sensitivity of our inferences to the details of the assumptions. For example, we can note from the likelihood curves in figure 3.2 that in the case of a single data point at $x = 5$, the likelihood function is less strongly peaked than in the case $x = 3$; the details of the prior $P(\lambda)$ become increasingly important as the sample mean $\bar{x}$ gets closer to the middle of the window, 10.5. In the case $x = 12$, the likelihood function doesn't have a peak at all – such data merely rule out small values of λ, and don't give any information about the relative probabilities of large values of λ. So in this case, the details of the prior at the small–λ end of things are not important, but at the large–λ end, the prior is important.

Third, when we are not sure which of various alternative assumptions is the most appropriate for a problem, we can treat this question as another inference task. Thus, given data D, we can compare alternative assumptions $\mathcal{H}$ using Bayes' theorem:

$$P(\mathcal{H} \mid D, I) = \frac{P(D \mid \mathcal{H}, I) P(\mathcal{H} \mid I)}{P(D \mid I)}, \tag{3.6}$$

where I denotes the highest assumptions, which we are not questioning.

Fourth, we can take into account our uncertainty regarding such assumptions when we make subsequent predictions. Rather than choosing one particular assumption $\mathcal{H}^*$, and working out our predictions about some quantity $\mathbf{t}$, $P(\mathbf{t} \mid D, \mathcal{H}^*, I)$, we obtain predictions that take into account our uncertainty about $\mathcal{H}$ by using the sum rule:

$$P(\mathbf{t} \mid D, I) = \sum_{\mathcal{H}} P(\mathbf{t} \mid D, \mathcal{H}, I) P(\mathcal{H} \mid D, I). \tag{3.7}$$

This is another contrast with orthodox statistics, in which it is conventional to 'test' a default model, and then, if the test 'accepts the model' at some 'significance level', to use exclusively that model to make predictions.

Steve thus persuaded me that

> probability theory reaches parts that ad hoc methods cannot reach.

Let's look at a few more examples of simple inference problems.

▶ 3.2 The bent coin

A bent coin is tossed F times; we observe a sequence $\mathbf{s}$ of heads and tails (which we'll denote by the symbols $\mathtt{a}$ and $\mathtt{b}$). We wish to know the bias of the coin, and predict the probability that the next toss will result in a head. We first encountered this task in example 2.7 (p.30), and we will encounter it again in Chapter 6, when we discuss adaptive data compression. It is also the original inference problem studied by Thomas Bayes in his essay published in 1763.

As in exercise 2.8 (p.30), we will assume a uniform prior distribution and obtain a posterior distribution by multiplying by the likelihood. A critic might object, 'where did this prior come from?' I will not claim that the uniform prior is in any way fundamental; indeed we'll give examples of nonuniform priors later. The prior is a subjective assumption. One of the themes of this book is:

> you can't do inference – or data compression – without making assumptions.

We give the name $\mathcal{H}_1$ to our assumptions. [We'll be introducing an alternative set of assumptions in a moment.] The probability, given $p_{\mathtt{a}}$, that F tosses result in a sequence $\mathbf{s}$ that contains $\{F_{\mathtt{a}}, F_{\mathtt{b}}\}$ counts of the two outcomes is

$$P(\mathbf{s} \mid p_{\mathtt{a}}, F, \mathcal{H}_1) = p_{\mathtt{a}}^{F_{\mathtt{a}}} (1 - p_{\mathtt{a}})^{F_{\mathtt{b}}}. \tag{3.8}$$

[For example, $P(\mathbf{s} = \mathtt{aaba} \mid p_{\mathtt{a}}, F = 4, \mathcal{H}_1) = p_{\mathtt{a}} p_{\mathtt{a}} (1 - p_{\mathtt{a}}) p_{\mathtt{a}}$.] Our first model assumes a uniform prior distribution for $p_{\mathtt{a}}$,

$$P(p_{\mathtt{a}} \mid \mathcal{H}_1) = 1, \quad p_{\mathtt{a}} \in [0, 1] \tag{3.9}$$

and $p_{\mathtt{b}} \equiv 1 - p_{\mathtt{a}}$.

Inferring unknown parameters

Given a string of length F of which $F_{\mathtt{a}}$ are $\mathtt{a}$s and $F_{\mathtt{b}}$ are $\mathtt{b}$s, we are interested in (a) inferring what $p_{\mathtt{a}}$ might be; (b) predicting whether the next character is

an a or a b. [Predictions are always expressed as probabilities. So 'predicting whether the next character is an a' is the same as computing the probability that the next character is an a.]

Assuming $\mathcal{H}_1$ to be true, the posterior probability of $p_\mathtt{a}$, given a string $\mathbf{s}$ of length F that has counts $\{F_\mathtt{a}, F_\mathtt{b}\}$, is, by Bayes' theorem,

$$P(p_\mathtt{a} \,|\, \mathbf{s}, F, \mathcal{H}_1) \;=\; \frac{P(\mathbf{s} \,|\, p_\mathtt{a}, F, \mathcal{H}_1) P(p_\mathtt{a} \,|\, \mathcal{H}_1)}{P(\mathbf{s} \,|\, F, \mathcal{H}_1)}. \tag{3.10}$$

The factor $P(\mathbf{s} \,|\, p_\mathtt{a}, F, \mathcal{H}_1)$, which, as a function of $p_\mathtt{a}$, is known as the likelihood function, was given in equation (3.8); the prior $P(p_\mathtt{a} \,|\, \mathcal{H}_1)$ was given in equation (3.9). Our inference of $p_\mathtt{a}$ is thus:

$$P(p_\mathtt{a} \,|\, \mathbf{s}, F, \mathcal{H}_1) \;=\; \frac{p_\mathtt{a}^{F_\mathtt{a}} (1 - p_\mathtt{a})^{F_\mathtt{b}}}{P(\mathbf{s} \,|\, F, \mathcal{H}_1)}. \tag{3.11}$$

The normalizing constant is given by the beta integral

$$P(\mathbf{s} \,|\, F, \mathcal{H}_1) = \int_0^1 \mathrm{d}p_\mathtt{a}\, p_\mathtt{a}^{F_\mathtt{a}} (1 - p_\mathtt{a})^{F_\mathtt{b}} = \frac{\Gamma(F_\mathtt{a} + 1)\Gamma(F_\mathtt{b} + 1)}{\Gamma(F_\mathtt{a} + F_\mathtt{b} + 2)} = \frac{F_\mathtt{a}! F_\mathtt{b}!}{(F_\mathtt{a} + F_\mathtt{b} + 1)!}. \tag{3.12}$$

Exercise 3.5.[2, p.59] Sketch the posterior probability $P(p_\mathtt{a} \,|\, \mathbf{s} = \mathtt{aba}, F = 3)$. What is the most probable value of $p_\mathtt{a}$ (i.e., the value that maximizes the posterior probability density)? What is the mean value of $p_\mathtt{a}$ under this distribution?

Answer the same questions for the posterior probability $P(p_\mathtt{a} \,|\, \mathbf{s} = \mathtt{bbb}, F = 3)$.

From inferences to predictions

Our prediction about the next toss, the probability that the next toss is an a, is obtained by integrating over $p_\mathtt{a}$. This has the effect of taking into account our uncertainty about $p_\mathtt{a}$ when making predictions. By the sum rule,

$$P(\mathtt{a} \,|\, \mathbf{s}, F) \;=\; \int \mathrm{d}p_\mathtt{a}\, P(\mathtt{a} \,|\, p_\mathtt{a}) P(p_\mathtt{a} \,|\, \mathbf{s}, F). \tag{3.13}$$

The probability of an a given $p_\mathtt{a}$ is simply $p_\mathtt{a}$, so

$$P(\mathtt{a} \,|\, \mathbf{s}, F) = \int \mathrm{d}p_\mathtt{a}\, p_\mathtt{a} \frac{p_\mathtt{a}^{F_\mathtt{a}} (1 - p_\mathtt{a})^{F_\mathtt{b}}}{P(\mathbf{s} \,|\, F)} \tag{3.14}$$

$$= \int \mathrm{d}p_\mathtt{a}\, \frac{p_\mathtt{a}^{F_\mathtt{a}+1} (1 - p_\mathtt{a})^{F_\mathtt{b}}}{P(\mathbf{s} \,|\, F)} \tag{3.15}$$

$$= \left[\frac{(F_\mathtt{a} + 1)!\, F_\mathtt{b}!}{(F_\mathtt{a} + F_\mathtt{b} + 2)!}\right] \Bigg/ \left[\frac{F_\mathtt{a}!\, F_\mathtt{b}!}{(F_\mathtt{a} + F_\mathtt{b} + 1)!}\right] \;=\; \frac{F_\mathtt{a} + 1}{F_\mathtt{a} + F_\mathtt{b} + 2}, \tag{3.16}$$

which is known as *Laplace's rule*.

▶ 3.3 The bent coin and model comparison

Imagine that a scientist introduces another theory for our data. He asserts that the source is not really a bent coin but is really a perfectly formed die with one face painted heads ('a') and the other five painted tails ('b'). Thus the parameter $p_\mathtt{a}$, which in the original model, $\mathcal{H}_1$, could take any value between 0 and 1, is according to the new hypothesis, $\mathcal{H}_0$, not a free parameter at all; rather, it is equal to 1/6. [This hypothesis is termed $\mathcal{H}_0$ so that the suffix of each model indicates its number of free parameters.]

How can we compare these two models in the light of data? We wish to infer how probable $\mathcal{H}_1$ is relative to $\mathcal{H}_0$.

Model comparison as inference

In order to perform model comparison, we write down Bayes' theorem again, but this time with a different argument on the left-hand side. We wish to know how probable $\mathcal{H}_1$ is given the data. By Bayes' theorem,

$$P(\mathcal{H}_1 \,|\, \mathbf{s}, F) = \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1)}{P(\mathbf{s} \,|\, F)}. \tag{3.17}$$

Similarly, the posterior probability of $\mathcal{H}_0$ is

$$P(\mathcal{H}_0 \,|\, \mathbf{s}, F) = \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0)}{P(\mathbf{s} \,|\, F)}. \tag{3.18}$$

The normalizing constant in both cases is $P(\mathbf{s} \,|\, F)$, which is the total probability of getting the observed data. If $\mathcal{H}_1$ and $\mathcal{H}_0$ are the only models under consideration, this probability is given by the sum rule:

$$P(\mathbf{s} \,|\, F) = P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1) + P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0). \tag{3.19}$$

To evaluate the posterior probabilities of the hypotheses we need to assign values to the prior probabilities $P(\mathcal{H}_1)$ and $P(\mathcal{H}_0)$; in this case, we might set these to 1/2 each. And we need to evaluate the data-dependent terms $P(\mathbf{s} \,|\, F, \mathcal{H}_1)$ and $P(\mathbf{s} \,|\, F, \mathcal{H}_0)$. We can give names to these quantities. The quantity $P(\mathbf{s} \,|\, F, \mathcal{H}_1)$ is a measure of how much the data favour $\mathcal{H}_1$, and we call it the *evidence* for model $\mathcal{H}_1$. We already encountered this quantity in equation (3.10) where it appeared as the normalizing constant of the first inference we made – the inference of $p_{\mathtt{a}}$ given the data.

> **How model comparison works:** The evidence for a model is usually the normalizing constant of an earlier Bayesian inference.

We evaluated the normalizing constant for model $\mathcal{H}_1$ in (3.12). The evidence for model $\mathcal{H}_0$ is very simple because this model has no parameters to infer. Defining p_0 to be 1/6, we have

$$P(\mathbf{s} \,|\, F, \mathcal{H}_0) = p_0^{F_{\mathtt{a}}} (1 - p_0)^{F_{\mathtt{b}}}. \tag{3.20}$$

Thus the posterior probability ratio of model $\mathcal{H}_1$ to model $\mathcal{H}_0$ is

$$\frac{P(\mathcal{H}_1 \,|\, \mathbf{s}, F)}{P(\mathcal{H}_0 \,|\, \mathbf{s}, F)} = \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1)}{P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0)} \tag{3.21}$$

$$= \frac{F_{\mathtt{a}}! F_{\mathtt{b}}!}{(F_{\mathtt{a}} + F_{\mathtt{b}} + 1)!} \Big/ p_0^{F_{\mathtt{a}}} (1 - p_0)^{F_{\mathtt{b}}}. \tag{3.22}$$

Some values of this posterior probability ratio are illustrated in table 3.5. The first five lines illustrate that some outcomes favour one model, and some favour the other. No outcome is completely incompatible with either model. With small amounts of data (six tosses, say) it is typically not the case that one of the two models is overwhelmingly more probable than the other. But with more data, the evidence against $\mathcal{H}_0$ given by any data set with the ratio $F_{\mathtt{a}}{:}\,F_{\mathtt{b}}$ differing from 1:5 mounts up. You can't predict in advance how much data are needed to be pretty sure which theory is true. It depends what $p_{\mathtt{a}}$ is.

The simpler model, $\mathcal{H}_0$, since it has no adjustable parameters, is able to lose out by the biggest margin. The odds may be hundreds to one against it. The more complex model can never lose out by a large margin; there's no data set that is actually *unlikely* given model $\mathcal{H}_1$.

F	Data $(F_\mathtt{a}, F_\mathtt{b})$	$\frac{P(\mathcal{H}_1 \mid \mathbf{s}, F)}{P(\mathcal{H}_0 \mid \mathbf{s}, F)}$	
6	$(5, 1)$	222.2	
6	$(3, 3)$	2.67	
6	$(2, 4)$	0.71	$= 1/1.4$
6	$(1, 5)$	0.356	$= 1/2.8$
6	$(0, 6)$	0.427	$= 1/2.3$
20	$(10, 10)$	96.5	
20	$(3, 17)$	0.2	$= 1/5$
20	$(0, 20)$	1.83	

Table 3.5. Outcome of model comparison between models $\mathcal{H}_1$ and $\mathcal{H}_0$ for the 'bent coin'. Model $\mathcal{H}_0$ states that $p_\mathtt{a} = 1/6$, $p_\mathtt{b} = 5/6$.

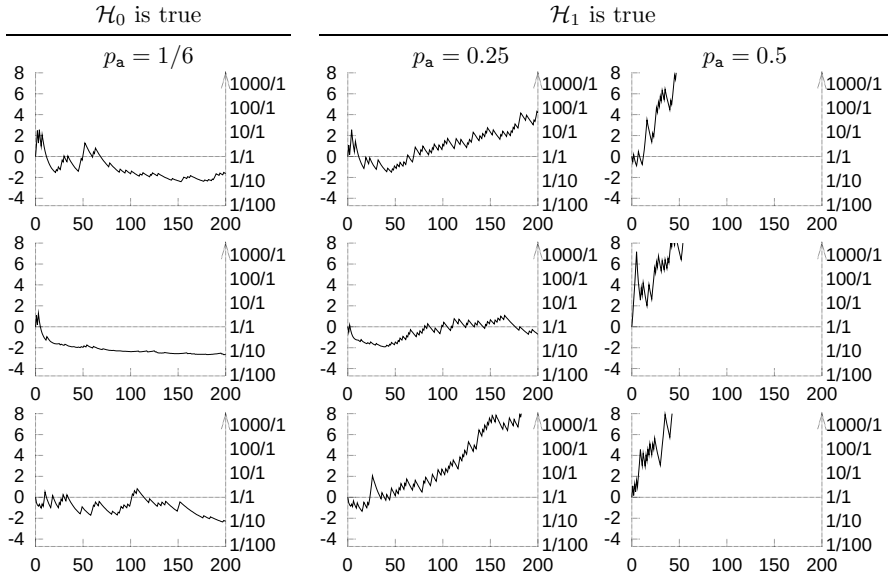


Figure 3.6. Typical behaviour of the evidence in favour of $\mathcal{H}_1$ as bent coin tosses accumulate under three different conditions (columns 1, 2, 3). Horizontal axis is the number of tosses, F. The vertical axis on the left is $\ln \frac{P(\mathbf{s} \mid F, \mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_0)}$; the right-hand vertical axis shows the values of $\frac{P(\mathbf{s} \mid F, \mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_0)}$. The three rows show independent simulated experiments. (See also figure 3.8, p.60.)

▷ Exercise 3.6.[2] Show that after F tosses have taken place, the biggest value that the log evidence ratio

$$\log \frac{P(\mathbf{s} \mid F, \mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_0)} \tag{3.23}$$

can have scales *linearly* with F if $\mathcal{H}_1$ is more probable, but the log evidence in favour of $\mathcal{H}_0$ can grow at most as $\log F$.

▷ Exercise 3.7.[3, p.60] Putting your sampling theory hat on, assuming $F_\mathtt{a}$ has not yet been measured, compute a plausible range that the log evidence ratio might lie in, as a function of F and the true value of $p_\mathtt{a}$, and sketch it as a function of F for $p_\mathtt{a} = p_0 = 1/6$, $p_\mathtt{a} = 0.25$, and $p_\mathtt{a} = 1/2$. [Hint: sketch the log evidence as a function of the random variable $F_\mathtt{a}$ and work out the mean and standard deviation of $F_\mathtt{a}$.]

Typical behaviour of the evidence

Figure 3.6 shows the log evidence ratio as a function of the number of tosses, F, in a number of simulated experiments. In the left-hand experiments, $\mathcal{H}_0$ was true. In the right-hand ones, $\mathcal{H}_1$ was true, and the value of $p_\mathtt{a}$ was either 0.25 or 0.5.

We will discuss model comparison more in a later chapter.

▶ 3.4 An example of legal evidence

The following example illustrates that there is more to Bayesian inference than the priors.

> Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and of type 'AB' (a rare type, with frequency 1%). Do these data (type 'O' and 'AB' blood were found at scene) give evidence in favour of the proposition that Oliver was one of the two people present at the crime?

A careless lawyer might claim that the fact that the suspect's blood type was found at the scene is positive evidence for the theory that he was present. But this is not so.

Denote the proposition 'the suspect and one unknown person were present' by S. The alternative, $\bar{S}$, states 'two unknown people from the population were present'. The prior in this problem is the prior probability ratio between the propositions S and $\bar{S}$. This quantity is important to the final verdict and would be based on all other available information in the case. Our task here is just to evaluate the contribution made by the data D, that is, the likelihood ratio, $P(D \mid S, \mathcal{H})/P(D \mid \bar{S}, \mathcal{H})$. In my view, a jury's task should generally be to multiply together carefully evaluated likelihood ratios from each independent piece of admissible evidence with an equally carefully reasoned prior probability. [This view is shared by many statisticians but learned British appeal judges recently disagreed and actually overturned the verdict of a trial because the jurors *had* been taught to use Bayes' theorem to handle complicated DNA evidence.]

The probability of the data given S is the probability that one unknown person drawn from the population has blood type AB:

$$P(D \mid S, \mathcal{H}) = p_{\mathrm{AB}} \tag{3.24}$$

(since given S, we already know that one trace will be of type O). The probability of the data given $\bar{S}$ is the probability that two unknown people drawn from the population have types O and AB:

$$P(D \mid \bar{S}, \mathcal{H}) = 2\, p_{\mathrm{O}}\, p_{\mathrm{AB}}. \tag{3.25}$$

In these equations $\mathcal{H}$ denotes the assumptions that two people were present and left blood there, and that the probability distribution of the blood groups of unknown people in an explanation is the same as the population frequencies.

Dividing, we obtain the likelihood ratio:

$$\frac{P(D \mid S, \mathcal{H})}{P(D \mid \bar{S}, \mathcal{H})} = \frac{1}{2p_{\mathrm{O}}} = \frac{1}{2 \times 0.6} = 0.83. \tag{3.26}$$

Thus the data in fact provide weak evidence *against* the supposition that Oliver was present.

This result may be found surprising, so let us examine it from various points of view. First consider the case of another suspect, Alberto, who has type AB. Intuitively, the data do provide evidence in favour of the theory S'

that this suspect was present, relative to the null hypothesis $\bar{S}$. And indeed the likelihood ratio in this case is:

$$\frac{P(D \mid S', \mathcal{H})}{P(D \mid \bar{S}, \mathcal{H})} = \frac{1}{2\, p_{\mathrm{AB}}} = 50. \tag{3.27}$$

Now let us change the situation slightly; imagine that 99% of people are of blood type O, and the rest are of type AB. Only these two blood types exist in the population. The data at the scene are the same as before. Consider again how these data influence our beliefs about Oliver, a suspect of type O, and Alberto, a suspect of type AB. Intuitively, we still believe that the presence of the rare AB blood provides positive evidence that Alberto was there. But does the fact that type O blood was detected at the scene favour the hypothesis that Oliver was present? If this were the case, that would mean that regardless of who the suspect is, the data make it more probable they were present; everyone in the population would be under greater suspicion, which would be absurd. The data may be *compatible* with any suspect of either blood type being present, but if they provide evidence *for* some theories, they must also provide evidence *against* other theories.

Here is another way of thinking about this: imagine that instead of two people's blood stains there are ten, and that in the entire local population of one hundred, there are ninety type O suspects and ten type AB suspects. Consider a particular type O suspect, Oliver: without any other information, and before the blood test results come in, there is a one in 10 chance that he was at the scene, since we know that 10 out of the 100 suspects were present. We now get the results of blood tests, and find that *nine* of the ten stains are of type AB, and *one* of the stains is of type O. Does this make it more likely that Oliver was there? No, there is now only a one in ninety chance that he was there, since we know that only one person present was of type O.

Maybe the intuition is aided finally by writing down the formulae for the general case where n_{O} blood stains of individuals of type O are found, and n_{AB} of type AB, a total of N individuals in all, and unknown people come from a large population with fractions $p_{\mathrm{O}}, p_{\mathrm{AB}}$. (There may be other blood types too.) The task is to evaluate the likelihood ratio for the two hypotheses: S, 'the type O suspect (Oliver) and $N-1$ unknown others left N stains'; and $\bar{S}$, 'N unknowns left N stains'. The probability of the data under hypothesis $\bar{S}$ is just the probability of getting $n_{\mathrm{O}}, n_{\mathrm{AB}}$ individuals of the two types when N individuals are drawn at random from the population:

$$P(n_{\mathrm{O}}, n_{\mathrm{AB}} \mid \bar{S}) = \frac{N!}{n_{\mathrm{O}}!\, n_{\mathrm{AB}}!} p_{\mathrm{O}}^{n_{\mathrm{O}}} p_{\mathrm{AB}}^{n_{\mathrm{AB}}}. \tag{3.28}$$

In the case of hypothesis S, we need the distribution of the $N-1$ other individuals:

$$P(n_{\mathrm{O}}, n_{\mathrm{AB}} \mid S) = \frac{(N-1)!}{(n_{\mathrm{O}}-1)!\, n_{\mathrm{AB}}!} p_{\mathrm{O}}^{n_{\mathrm{O}}-1} p_{\mathrm{AB}}^{n_{\mathrm{AB}}}. \tag{3.29}$$

The likelihood ratio is:

$$\frac{P(n_{\mathrm{O}}, n_{\mathrm{AB}} \mid S)}{P(n_{\mathrm{O}}, n_{\mathrm{AB}} \mid \bar{S})} = \frac{n_{\mathrm{O}}/N}{p_{\mathrm{O}}}. \tag{3.30}$$

This is an instructive result. The likelihood ratio, i.e. the contribution of these data to the question of whether Oliver was present, depends simply on a comparison of the frequency of his blood type in the observed data with the background frequency in the population. There is no dependence on the counts of the other types found at the scene, or their frequencies in the population.

If there are more type O stains than the average number expected under hypothesis $\bar{S}$, then the data give evidence in favour of the presence of Oliver. Conversely, if there are fewer type O stains than the expected number under $\bar{S}$, then the data reduce the probability of the hypothesis that he was there. In the special case $n_{\mathrm{O}}/N = p_{\mathrm{O}}$, the data contribute no evidence either way, regardless of the fact that the data are compatible with the hypothesis S.

▶ 3.5 Exercises

Exercise 3.8.[2, p.60] The three doors, normal rules.

On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will *not* be opened. Instead, the gameshow host will open one of the other two doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference?

Exercise 3.9.[2, p.61] The three doors, earthquake scenario.

Imagine that the game happens again and just as the gameshow host is about to open one of the doors a violent earthquake rattles the building and one of the three doors flies open. It happens to be door 3, and it happens not to have the prize behind it. The contestant had initially chosen door 1.

Repositioning his toupée, the host suggests, 'OK, since you chose door 1 initially, door 3 is a valid door for me to open, according to the rules of the game; I'll let door 3 stay open. Let's carry on as if nothing happened.'

Should the contestant stick with door 1, or switch to door 2, or does it make no difference? Assume that the prize was placed randomly, that the gameshow host does not know where it is, and that the door flew open because its latch was broken by the earthquake.

[A similar alternative scenario is a gameshow whose *confused host* forgets the rules, and where the prize is, and opens one of the unchosen doors at random. He opens door 3, and the prize is not revealed. Should the contestant choose what's behind door 1 or door 2? Does the optimal decision for the contestant depend on the contestant's beliefs about whether the gameshow host is confused or not?]

▷ Exercise 3.10.[2] Another example in which the emphasis is not on priors. You visit a family whose three children are all at the local school. You don't

know anything about the sexes of the children. While walking clumsily round the home, you stumble through one of the three unlabelled bedroom doors that you know belong, one each, to the three children, and find that the bedroom contains girlie stuff in sufficient quantities to convince you that the child who lives in that bedroom is a girl. Later, you sneak a look at a letter addressed to the parents, which reads 'From the Headmaster: we are sending this letter to all parents who have male children at the school to inform them about the following boyish matters...'.

These two sources of evidence establish that at least one of the three children is a girl, and that at least one of the children is a boy. What are the probabilities that there are (a) two girls and one boy; (b) two boys and one girl?

▷ Exercise 3.11.[2, p.61] Mrs S is found stabbed in her family garden. Mr S behaves strangely after her death and is considered as a suspect. On investigation of police and social records it is found that Mr S had beaten up his wife on at least nine previous occasions. The prosecution advances this data as evidence in favour of the hypothesis that Mr S is guilty of the murder. 'Ah no,' says Mr S's highly paid lawyer, '*statistically*, only one in a thousand wife-beaters actually goes on to murder his wife.[1] So the wife-beating is not strong evidence at all. In fact, given the wife-beating evidence alone, it's extremely *unlikely* that he would be the murderer of his wife – only a 1/1000 chance. You should therefore find him innocent.'

Is the lawyer right to imply that the history of wife-beating does not point to Mr S's being the murderer? Or is the lawyer a slimy trickster? If the latter, what is wrong with his argument?

[Having received an indignant letter from a lawyer about the preceding paragraph, I'd like to add an extra inference exercise at this point: *Does my suggestion that Mr. S.'s lawyer may have been a slimy trickster imply that I believe* all *lawyers are slimy tricksters?* (Answer: No.)]

▷ Exercise 3.12.[2] A bag contains one counter, known to be either white or black. A white counter is put in, the bag is shaken, and a counter is drawn out, which proves to be white. What is now the chance of drawing a white counter? [Notice that the state of the bag, after the operations, is exactly identical to its state before.]

▷ Exercise 3.13.[2, p.62] You move into a new house; the phone is connected, and you're pretty sure that the phone number is `740511`, but not as sure as you would like to be. As an experiment, you pick up the phone and dial `740511`; you obtain a 'busy' signal. Are you now more sure of your phone number? If so, how much?

▷ Exercise 3.14.[1] In a game, two coins are tossed. If either of the coins comes up heads, you have won a prize. To claim the prize, you must point to one of your coins that is a head and say 'look, that coin's a head, I've won'. You watch Fred play the game. He tosses the two coins, and he

[1] In the U.S.A., it is estimated that 2 million women are abused each year by their partners. In 1994, 4739 women were victims of homicide; of those, 1326 women (28%) were slain by husbands and boyfriends.
(Sources: `http://www.umn.edu/mincava/papers/factoid.htm`, `http://www.gunfree.inter.net/vpc/womenfs.htm`)

points to a coin and says 'look, that coin's a head, I've won'. What is the probability that the *other* coin is a head?

▷ Exercise 3.15.[2, p.63] A statistical statement appeared in *The Guardian* on Friday January 4, 2002:

> When spun on edge 250 times, a Belgian one-euro coin came up heads 140 times and tails 110. 'It looks very suspicious to me', said Barry Blight, a statistics lecturer at the London School of Economics. 'If the coin were unbiased the chance of getting a result as extreme as that would be less than 7%'.

But *do* these data give evidence that the coin is biased rather than fair? [Hint: see equation (3.22).]

▶ 3.6 Solutions

Solution to exercise 3.1 (p.47). Let the data be D. Assuming equal prior probabilities,

$$\frac{P(A\,|\,D)}{P(B\,|\,D)} = \frac{1}{2}\frac{3}{2}\frac{1}{1}\frac{3}{2}\frac{1}{2}\frac{2}{2}\frac{1}{2} = \frac{9}{32} \tag{3.31}$$

and $P(A\,|\,D) = 9/41$.

Solution to exercise 3.2 (p.47). The probability of the data given each hypothesis is:

$$P(D\,|\,A) = \frac{3}{20}\frac{1}{20}\frac{2}{20}\frac{1}{20}\frac{3}{20}\frac{1}{20}\frac{1}{20} = \frac{18}{20^7}; \tag{3.32}$$

$$P(D\,|\,B) = \frac{2}{20}\frac{2}{20}\frac{2}{20}\frac{2}{20}\frac{2}{20}\frac{1}{20}\frac{2}{20} = \frac{64}{20^7}; \tag{3.33}$$

$$P(D\,|\,C) = \frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{20} = \frac{1}{20^7}. \tag{3.34}$$

So

$$P(A\,|\,D) = \frac{18}{18+64+1} = \frac{18}{83}; \qquad P(B\,|\,D) = \frac{64}{83}; \qquad P(C\,|\,D) = \frac{1}{83}. \tag{3.35}$$

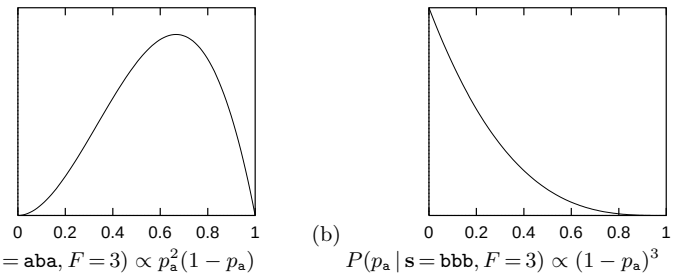


Figure 3.7. Posterior probability for the bias p_a of a bent coin given two different data sets.

Solution to exercise 3.5 (p.52).

(a) $P(p_a\,|\,s=aba, F=3) \propto p_a^2(1-p_a)$. The most probable value of p_a (i.e., the value that maximizes the posterior probability density) is 2/3. The mean value of p_a is 3/5.

See figure 3.7a.