(b) $P(p_{\mathtt{a}} \,|\, \mathbf{s}=\mathtt{bbb}, F=3) \propto (1-p_{\mathtt{a}})^3$. The most probable value of $p_{\mathtt{a}}$ (i.e., the value that maximizes the posterior probability density) is 0. The mean value of $p_{\mathtt{a}}$ is 1/5.

See figure 3.7b.

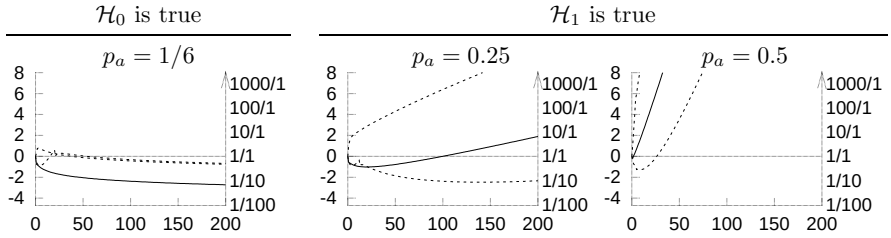

Figure 3.8. Range of plausible values of the log evidence in favour of $\mathcal{H}_1$ as a function of F. The vertical axis on the left is $\log \frac{P(\mathbf{s}\,|\,F,\mathcal{H}_1)}{P(\mathbf{s}\,|\,F,\mathcal{H}_0)}$; the right-hand vertical axis shows the values of $\frac{P(\mathbf{s}\,|\,F,\mathcal{H}_1)}{P(\mathbf{s}\,|\,F,\mathcal{H}_0)}$. The solid line shows the log evidence if the random variable F_a takes on its mean value, $F_a = p_a F$. The dotted lines show (approximately) the log evidence if F_a is at its 2.5th or 97.5th percentile. (See also figure 3.6, p.54.)

Solution to exercise 3.7 (p.54). The curves in figure 3.8 were found by finding the mean and standard deviation of F_a, then setting F_a to the mean $\pm$ two standard deviations to get a 95% plausible range for F_a, and computing the three corresponding values of the log evidence ratio.

Solution to exercise 3.8 (p.57). Let $\mathcal{H}_i$ denote the hypothesis that the prize is behind door i. We make the following assumptions: the three hypotheses $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ are equiprobable *a priori*, i.e.,

$$P(\mathcal{H}_1) = P(\mathcal{H}_2) = P(\mathcal{H}_3) = \frac{1}{3}. \tag{3.36}$$

The datum we receive, after choosing door 1, is one of $D=3$ and $D=2$ (meaning door 3 or 2 is opened, respectively). We assume that these two possible outcomes have the following probabilities. If the prize is behind door 1 then the host has a free choice; in this case we assume that the host selects at random between $D=2$ and $D=3$. Otherwise the choice of the host is forced and the probabilities are 0 and 1.

$$\left| \begin{array}{lll} P(D=2\,|\,\mathcal{H}_1)=1/2 & P(D=2\,|\,\mathcal{H}_2)=0 & P(D=2\,|\,\mathcal{H}_3)=1 \\ P(D=3\,|\,\mathcal{H}_1)=1/2 & P(D=3\,|\,\mathcal{H}_2)=1 & P(D=3\,|\,\mathcal{H}_3)=0 \end{array} \right| \tag{3.37}$$

Now, using Bayes' theorem, we evaluate the posterior probabilities of the hypotheses:

$$P(\mathcal{H}_i\,|\,D=3) = \frac{P(D=3\,|\,\mathcal{H}_i)P(\mathcal{H}_i)}{P(D=3)} \tag{3.38}$$

$$\left| P(\mathcal{H}_1\,|\,D=3) = \frac{(1/2)(1/3)}{P(D=3)} \right| P(\mathcal{H}_2\,|\,D=3) = \frac{(1)(1/3)}{P(D=3)} \left| P(\mathcal{H}_3\,|\,D=3) = \frac{(0)(1/3)}{P(D=3)} \right| \tag{3.39}$$

The denominator $P(D=3)$ is (1/2) because it is the normalizing constant for this posterior distribution. So

$$\left| P(\mathcal{H}_1\,|\,D=3) \;=\; 1/3 \right| P(\mathcal{H}_2\,|\,D=3) \;=\; 2/3 \left| P(\mathcal{H}_3\,|\,D=3) \;=\; 0. \right| \tag{3.40}$$

So the contestant should switch to door 2 in order to have the biggest chance of getting the prize.

Many people find this outcome surprising. There are two ways to make it more intuitive. One is to play the game thirty times with a friend and keep track of the frequency with which switching gets the prize. Alternatively, you can perform a thought experiment in which the game is played with a million doors. The rules are now that the contestant chooses one door, then the game

show host opens 999,998 doors in such a way as not to reveal the prize, leaving the *contestant's* selected door and *one other door* closed. The contestant may now stick or switch. Imagine the contestant confronted by a million doors, of which doors 1 and 234,598 have not been opened, door 1 having been the contestant's initial guess. Where do you think the prize is?

Solution to exercise 3.9 (p.57). If door 3 is opened by an earthquake, the inference comes out differently – even though visually the scene looks the same. The nature of the data, and the probability of the data, are both now different. The possible data outcomes are, firstly, that any number of the doors might have opened. We could label the eight possible outcomes $\mathbf{d} = (0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), \ldots, (1,1,1)$. Secondly, it might be that the prize is visible after the earthquake has opened one or more doors. So the data D consists of the value of $\mathbf{d}$, and a statement of whether the prize was revealed. It is hard to say what the probabilities of these outcomes are, since they depend on our beliefs about the reliability of the door latches and the properties of earthquakes, but it is possible to extract the desired posterior probability without naming the values of $P(\mathbf{d} \,|\, \mathcal{H}_i)$ for each $\mathbf{d}$. All that matters are the relative values of the quantities $P(D \,|\, \mathcal{H}_1)$, $P(D \,|\, \mathcal{H}_2)$, $P(D \,|\, \mathcal{H}_3)$, for the value of D that actually occurred. [This is the *likelihood principle*, which we met in section 2.3.] The value of D that actually occurred is '$\mathbf{d} = (0,0,1)$, and no prize visible'. First, it is clear that $P(D \,|\, \mathcal{H}_3) = 0$, since the datum that no prize is visible is incompatible with $\mathcal{H}_3$. Now, assuming that the contestant selected door 1, how does the probability $P(D \,|\, \mathcal{H}_1)$ compare with $P(D \,|\, \mathcal{H}_2)$? Assuming that earthquakes are not sensitive to decisions of game show contestants, these two quantities have to be equal, by symmetry. We don't know how likely it is that door 3 falls off its hinges, but however likely it is, it's just as likely to do so whether the prize is behind door 1 or door 2. So, if $P(D \,|\, \mathcal{H}_1)$ and $P(D \,|\, \mathcal{H}_2)$ are equal, we obtain:

$$\left| \begin{array}{c|c|c} P(\mathcal{H}_1|D) = \frac{P(D|\mathcal{H}_1)(1/3)}{P(D)} & P(\mathcal{H}_2|D) = \frac{P(D|\mathcal{H}_2)(1/3)}{P(D)} & P(\mathcal{H}_3|D) = \frac{P(D|\mathcal{H}_3)(1/3)}{P(D)} \\ = 1/2 & = 1/2 & = 0. \end{array} \right| \tag{3.41}$$

The two possible hypotheses are now equally likely.

If we assume that the host knows where the prize is and might be acting deceptively, then the answer might be further modified, because we have to view the host's words as part of the data.

Confused? It's well worth making sure you understand these two gameshow problems. Don't worry, I slipped up on the second problem, the first time I met it.

There is a general rule which helps immensely when you have a confusing probability problem:

> Always write down the probability of everything.
> *(Steve Gull)*

From this joint probability, any desired inference can be mechanically obtained (figure 3.9).

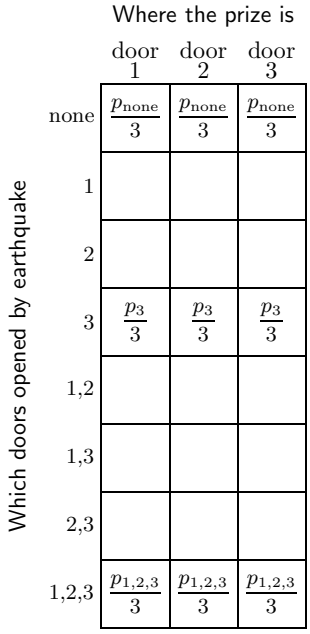

Figure 3.9. The probability of everything, for the second three-door problem, assuming an earthquake has just occurred. Here, p_3 is the probability that door 3 alone is opened by an earthquake.

Solution to exercise 3.11 (p.58). The statistic quoted by the lawyer indicates the probability that a randomly selected wife-beater will also murder his wife. The probability that the husband was the murderer, *given that the wife has been murdered*, is a completely different quantity.

To deduce the latter, we need to make further assumptions about the probability that the wife is murdered by someone else. If she lives in a neighbourhood with frequent random murders, then this probability is large and the posterior probability that the husband did it (in the absence of other evidence) may not be very large. But in more peaceful regions, it may well be that the most likely person to have murdered you, if you are found murdered, is one of your closest relatives.

Let's work out some illustrative numbers with the help of the statistics on page 58. Let $m=1$ denote the proposition that a woman has been murdered; $h=1$, the proposition that the husband did it; and $b=1$, the proposition that he beat her in the year preceding the murder. The statement 'someone else did it' is denoted by $h=0$. We need to define $P(h \mid m=1)$, $P(b \mid h=1, m=1)$, and $P(b=1 \mid h=0, m=1)$ in order to compute the posterior probability $P(h=1 \mid b=1, m=1)$. From the statistics, we can read out $P(h=1 \mid m=1) = 0.28$. And if two million women out of 100 million are beaten, then $P(b=1 \mid h=0, m=1) = 0.02$. Finally, we need a value for $P(b \mid h=1, m=1)$: if a man murders his wife, how likely is it that this is the first time he laid a finger on her? I expect it's pretty unlikely; so maybe $P(b=1 \mid h=1, m=1)$ is 0.9 or larger.

By Bayes' theorem, then,

$$P(h=1 \mid b=1, m=1) = \frac{.9 \times .28}{.9 \times .28 + .02 \times .72} \simeq 95\%. \tag{3.42}$$

One way to make obvious the sliminess of the lawyer on p.58 is to construct arguments, with the same logical structure as his, that are clearly wrong. For example, the lawyer could say 'Not only was Mrs. S murdered, she was murdered between 4.02pm and 4.03pm. *Statistically*, only one in a *million* wife-beaters actually goes on to murder his wife between 4.02pm and 4.03pm. 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 murder his wife in this way – only a 1/1,000,000 chance.'

Solution to exercise 3.13 (p.58). There are two hypotheses. $\mathcal{H}_0$: your number is 740511; $\mathcal{H}_1$: it is another number. The data, D, are 'when I dialed 740511, I got a busy signal'. What is the probability of D, given each hypothesis? If your number is 740511, then we expect a busy signal with certainty:

$$P(D \mid \mathcal{H}_0) = 1.$$

On the other hand, if $\mathcal{H}_1$ is true, then the probability that the number dialled returns a busy signal is smaller than 1, since various other outcomes were also possible (a ringing tone, or a number-unobtainable signal, for example). The value of this probability $P(D \mid \mathcal{H}_1)$ will depend on the probability α that a random phone number similar to your own phone number would be a valid phone number, and on the probability β that you get a busy signal when you dial a valid phone number.

I estimate from the size of my phone book that Cambridge has about 75 000 valid phone numbers, all of length six digits. The probability that a random six-digit number is valid is therefore about $75\,000/10^6 = 0.075$. If we exclude numbers beginning with 0, 1, and 9 from the random choice, the probability α is about $75\,000/700\,000 \simeq 0.1$. If we assume that telephone numbers are clustered then a misremembered number might be more likely to be valid than a randomly chosen number; so the probability, α, that our guessed number would be valid, assuming $\mathcal{H}_1$ is true, might be bigger than

0.1. Anyway, α must be somewhere between 0.1 and 1. We can carry forward this uncertainty in the probability and see how much it matters at the end.

The probability β that you get a busy signal when you dial a valid phone number is equal to the fraction of phones you think are in use or off-the-hook when you make your tentative call. This fraction varies from town to town and with the time of day. In Cambridge, during the day, I would guess that about 1% of phones are in use. At 4am, maybe 0.1%, or fewer.

The probability $P(D \mid \mathcal{H}_1)$ is the product of α and β, that is, about $0.1 \times 0.01 = 10^{-3}$. According to our estimates, there's about a one-in-a-thousand chance of getting a busy signal when you dial a random number; or one-in-a-hundred, if valid numbers are strongly clustered; or one-in-10^4, if you dial in the wee hours.

How do the data affect your beliefs about your phone number? The posterior probability ratio is the likelihood ratio times the prior probability ratio:

$$\frac{P(\mathcal{H}_0 \mid D)}{P(\mathcal{H}_1 \mid D)} = \frac{P(D \mid \mathcal{H}_0)}{P(D \mid \mathcal{H}_1)} \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}. \tag{3.43}$$

The likelihood ratio is about 100-to-1 or 1000-to-1, so the posterior probability ratio is swung by a factor of 100 or 1000 in favour of $\mathcal{H}_0$. If the prior probability of $\mathcal{H}_0$ was 0.5 then the posterior probability is

$$P(\mathcal{H}_0 \mid D) = \frac{1}{1 + \frac{P(\mathcal{H}_1 \mid D)}{P(\mathcal{H}_0 \mid D)}} \simeq 0.99 \text{ or } 0.999. \tag{3.44}$$

Solution to exercise 3.15 (p.59). We compare the models $\mathcal{H}_0$ – the coin is fair – and $\mathcal{H}_1$ – the coin is biased, with the prior on its bias set to the uniform distribution $P(p|\mathcal{H}_1) = 1$. [The use of a uniform prior seems reasonable to me, since I know that some coins, such as American pennies, have severe biases when spun on edge; so the situations $p = 0.01$ or $p = 0.1$ or $p = 0.95$ would not surprise me.]

> When I mention $\mathcal{H}_0$ – the coin is fair – a pedant would say, 'how absurd to even consider that the coin is fair – any coin is surely biased to some extent'. And of course I would agree. So will pedants kindly understand $\mathcal{H}_0$ as meaning 'the coin is fair to within one part in a thousand, i.e., $p \in 0.5 \pm 0.001$'.

The likelihood ratio is:

$$\frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_0)} = \frac{\frac{140!110!}{251!}}{1/2^{250}} = 0.48. \tag{3.45}$$

Thus the data give scarcely any evidence either way; in fact they give weak evidence (two to one) in favour of $\mathcal{H}_0$!

'No, no', objects the believer in bias, 'your silly uniform prior doesn't represent *my* prior beliefs about the bias of biased coins – I was *expecting* only a small bias'. To be as generous as possible to the $\mathcal{H}_1$, let's see how well it could fare if the prior were presciently set. Let us allow a prior of the form

$$P(p|\mathcal{H}_1, \alpha) = \frac{1}{Z(\alpha)} p^{\alpha-1}(1-p)^{\alpha-1}, \quad \text{where } Z(\alpha) = \Gamma(\alpha)^2/\Gamma(2\alpha) \tag{3.46}$$

(a Beta distribution, with the original uniform prior reproduced by setting $\alpha = 1$). By tweaking α, the likelihood ratio for $\mathcal{H}_1$ over $\mathcal{H}_0$,

$$\frac{P(D|\mathcal{H}_1, \alpha)}{P(D|\mathcal{H}_0)} = \frac{\Gamma(140+\alpha)\,\Gamma(110+\alpha)\,\Gamma(2\alpha)2^{250}}{\Gamma(250+2\alpha)\,\Gamma(\alpha)^2}, \tag{3.47}$$

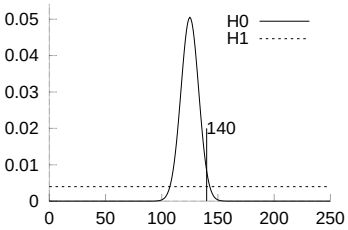

Figure 3.10. The probability distribution of the number of heads given the two hypotheses, that the coin is fair, and that it is biased, with the prior distribution of the bias being uniform. The outcome ($D = 140$ heads) gives weak evidence in favour of $\mathcal{H}_0$, the hypothesis that the coin is fair.

can be increased a little. It is shown for several values of α in figure 3.11. Even the most favourable choice of α ($\alpha \simeq 50$) can yield a likelihood ratio of only two to one in favour of $\mathcal{H}_1$.

In conclusion, the data are not 'very suspicious'. They can be construed as giving at most two-to-one evidence in favour of one or other of the two hypotheses.

> Are these wimpy likelihood ratios the fault of over-restrictive priors? Is there any way of producing a 'very suspicious' conclusion? The prior that is best-matched to the data, in terms of likelihood, is the prior that sets p to $f \equiv 140/250$ with probability one. Let's call this model $\mathcal{H}_*$. The likelihood ratio is $P(D|\mathcal{H}_*)/P(D|\mathcal{H}_0) = 2^{250} f^{140}(1-f)^{110} = 6.1$. So the strongest evidence that these data can possibly muster against the hypothesis that there is no bias is six-to-one.

α	$\frac{P(D\|\mathcal{H}_1,\alpha)}{P(D\|\mathcal{H}_0)}$
.37	.25
1.0	.48
2.7	.82
7.4	1.3
20	1.8
55	1.9
148	1.7
403	1.3
1096	1.1

Figure 3.11. Likelihood ratio for various choices of the prior distribution's hyperparameter α.

While we are noticing the absurdly misleading answers that 'sampling theory' statistics produces, such as the p-value of 7% in the exercise we just solved, let's stick the boot in. If we make a tiny change to the data set, increasing the number of heads in 250 tosses from 140 to 141, we find that the p-value goes below the mystical value of 0.05 (the p-value is 0.0497). The sampling theory statistician would happily squeak 'the probability of getting a result as extreme as 141 heads is smaller than 0.05 – we thus reject the null hypothesis at a significance level of 5%'. The correct answer is shown for several values of α in figure 3.12. The values worth highlighting from this table are, first, the likelihood ratio when $\mathcal{H}_1$ uses the standard uniform prior, which is 1:0.61 in favour of the *null hypothesis* $\mathcal{H}_0$. Second, the most favourable choice of α, from the point of view of $\mathcal{H}_1$, can only yield a likelihood ratio of about 2.3:1 in favour of $\mathcal{H}_1$.

α	$\frac{P(D'\|\mathcal{H}_1,\alpha)}{P(D'\|\mathcal{H}_0)}$
.37	.32
1.0	.61
2.7	1.0
7.4	1.6
20	2.2
55	2.3
148	1.9
403	1.4
1096	1.2

Figure 3.12. Likelihood ratio for various choices of the prior distribution's hyperparameter α, when the data are $D' = 141$ heads in 250 trials.

Be warned! A p-value of 0.05 is often interpreted as implying that the odds are stacked about twenty-to-one *against* the null hypothesis. But the truth in this case is that the evidence either slightly *favours* the null hypothesis, or disfavours it by at most 2.3 to one, depending on the choice of prior.

The p-values and 'significance levels' of classical statistics should be treated with *extreme caution*. Shun them! Here ends the sermon.