# 5 Parameter estimation: single parameter

In chapter 3 I introduced the basic concepts of inference. In section 3.2 we used Bayes' theorem to interpret medical test results, and in section 3.5 we looked at the simple problem of estimating one parameter from one measurement. Here I take these ideas further to estimate the posterior probability density over a model parameter given a set of data, focusing on a single parameter problem and the use of so-called conjugate priors. We will see specifically how the prior and the likelihood combine to make the posterior, and how this depends on the amount of data available. I will also discuss assigning priors and summarizing distributions.

## 5.1 Bayesian analysis of coin tossing

We are given a coin and we toss it $n$ times. It lands heads in $r$ of them. Is the coin fair?

This question has no definitive answer; we can only answer it probabilistically. Let $p$ be the unknown probability that the coin lands heads in any one toss. We can interpret the question to mean "what is the posterior PDF over $p$?", i.e. "what is $P(p\,|\,n, r, M)$?", where $M$ describes our assumptions. Bayes' theorem (equation 3.12) tells us that the posterior is the product of the likelihood and prior, divided by a normalization constant.

While coin tossing is not a very scientific enterprise, this example stands in for any repeated process that has just two alternative outcomes, such as detecting or not detecting a type of source in a survey, or a disease in a test.

Let's first identify the likelihood, which is the probability of the data given the parameter. We set up a model $M$ for the phenomenon and the measurement. This specifies that the coin lands heads in a single toss with probability $p$, that all tosses are independent, and that $p$ is constant (does not change with the number of tosses already done). $p$ is the only parameter of the model. The appropriate likelihood is the binomial distribution (section 1.4.1)

$$P(r\,|\,p, n, M) \;=\; \binom{n}{r}\, p^r (1-p)^{n-r} \tag{5.1}$$

where $r \leq n$. My notation here explicitly shows the conditioning on $p$, $n$, and $M$. Although $n$ is part of the data we are given, the relevant likelihood describes the variation of $r$ for given $n$, which is why $n$ is on the right of the conditioning bar.[1] Thus in terms of Bayes'

[1] We could imagine a different problem in which $r$ is given (fixed) and we are interested in the variation of $n$, e.g. how many times do we toss the coin until we achieve $r$ heads. In that case we need the likelihood

theorem in equation 3.12, $\theta = p$, $D = r$, and $n$ is an additional quantity which all terms are conditioned on. In this case Bayes' theorem is

$$P(p|r,n,M) = \frac{P(r|p,n,M)P(p|M)}{P(r|n,M)}. \tag{5.2}$$

I have removed the $n$ in the conditioning of the prior $P(p|M)$ because the prior is independent of the number of coin tosses. In parameter estimation problems everything is conditioned on the model $M$, so to make the expressions less cumbersome authors sometimes remove this explicit conditioning. Sometimes $n$ is removed too. It is important when manipulating probabilistic equations to know exactly what you are conditioning on. In this section I will retain all conditioning.

I will now adopt two different forms of the prior, derive and plot the posterior, and also work out the expectation value of $p$ from this.

### 5.1.1 Uniform prior

Let us first adopt a uniform prior $\mathcal{U}(0,1)$ over $p$. In that case the posterior PDF is just proportional to the likelihood

$$P(p|r,n,M) = \frac{1}{Z} p^r (1-p)^{n-r} \tag{5.3}$$

where $Z$ is a normalization constant which does not depend on $p$. I have absorbed the binomial coefficient into this normalization constant, since it depends only on $n$ and $r$. Imagine we had $n = 20$, $r = 7$. We can plot the posterior PDF using the following code.

```
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
n <- 20
r <- 7
p <- seq(from=0,to=1,length.out=201)
plot(p, dbinom(x=r, size=n, prob=p), ylim=c(0, 0.2), xaxs="i", yaxs="i",
  xlab="p", ylab=expression(paste(P^symbol("*"), "(p | r,n,M)")), type="l",
  lwd=1.5)
```

The result is in figure 5.1. We see that the PDF peaks well away from a fair coin with $p = 1/2$, although a fair coin is not strongly disfavoured. The mode is at $r/n = 0.35$, which is probably most people's intuitive estimate for the value of $p$. Remember that this curve is *not* binomial in $p$ (it is binomial in $r$). Furthermore, the posterior as plotted – as a function of $p$ – is *unnormalized*; the integral over $p$ is not unity. As normalization just rescales the whole curve by a scalar multiple, it does not change the peak (the mode) or relative probabilities of solutions. But normalization is necessary if we want to calculate expectation values, e.g. the mean or variance.[2] If we denote the unnormalized posterior PDF as $P^*(p|r,n,M)$ – which in this case is $p^r(1-p)^{n-r}$ – then the expectation value

$P(n|p,r,M)$, which is not the binomial distribution, but rather the negative binomial distribution (see equation 11.50).

[2] Normalization for this purpose will no longer be necessary if we were to *draw samples from* the distribution, as we shall see in chapter 8.

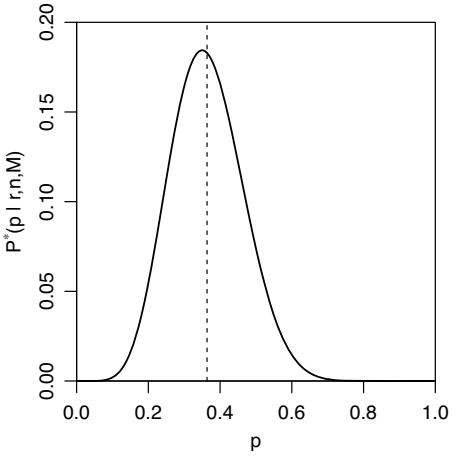


**Fig. 5.1** The unnormalized posterior PDF for the probability $p$ of a coin giving heads on a single toss, when $r = 7$ of $n = 20$ observed tosses are heads. A uniform prior over $p$ is used. The mode is at $r/n = 0.35$. The dashed line is the mean, which is at $0.364$.

of $p$ is

$$E[p] \;=\; \int_0^1 p\, P(p\,|\,r,n,M)\, dp \;=\; \frac{1}{Z}\int_0^1 p\, P^*(p\,|\,r,n,M)\, dp \tag{5.4}$$

where $Z = \int_0^1 P^*(p\,|\,r,n,M)dp$. We can approximate these integrals with a sum. Here I use the rectangle rule (section 3.3.4), so for example

$$Z \;=\; \int_0^1 P^*(p\,|\,r,n,M)\, dp \;\simeq\; \sum_i P^*(p_i\,|\,r,n,M)\, \delta p \tag{5.5}$$

for some fixed (small) step size $\delta p$. Thus the normalized PDF is

$$P(p\,|\,r,n,M) \;\simeq\; \frac{1}{\sum_i P^*(p_i\,|\,r,n,M)\, \delta p}\, P^*(p\,|\,r,n,M). \tag{5.6}$$

When using such sums the step size appears in both the numerator and denominator of equation 5.4, so it cancels to leave

$$E[p] \;\simeq\; \frac{1}{\sum_i P^*(p_i\,|\,r,n,M)} \sum_i p_i\, P^*(p_i\,|\,r,n,M). \tag{5.7}$$

Although not used in the code below, we could estimate the variance (equation 1.30) in a similar way as

$$\mathrm{Var}(p) \;\simeq\; \frac{1}{\sum_i P^*(p_i\,|\,r,n,M)} \sum_i (p_i - E[p])^2\, P^*(p_i\,|\,r,n,M). \tag{5.8}$$

The following R code implements the mean as just described and overplots it in figure 5.1 with the dashed line.

```
pdense <- dbinom(x=r, size=n, prob=p) # unnormalized in p
p.mean <- sum(p*pdense)/sum(pdense)
abline(v=p.mean, lty=2)
```

It is instructive to repeat this example for $n = 20$ for a range of values of $r$. This is done by the following R code, which produces figure 5.2. The more heads we toss, the more our inference of $p$ shifts towards larger values. But since we only have 20 tosses, a degree of uncertainty remains, as reflected by the finite width of the posterior PDF.

R file: `coin1.R`

```
##### Compute the posterior PDF for coin problem with a uniform prior for a
##### range of r

n <- 20
Nsamp <- 200 # no. of points to sample at
pdf("coin1.pdf", 9, 7)
par(mfrow=c(3,4), mgp=c(2,0.8,0), mar=c(3.5,3.5,1.5,1), oma=0.5*c(1,1,1,1))
deltap <- 1/Nsamp # width of rectangles used for numerical integration
p <- seq(from=1/(2*Nsamp), by=1/Nsamp, length.out=Nsamp) # rectangle centres
for(r in seq(from=0, to=20, by=2)) {
  pdense <- dbinom(x=r, size=n, prob=p)
  pdense <- pdense/(deltap*sum(pdense)) # normalize posterior
  plot(p, pdense, type="l", lwd=1.5, xlim=c(0,1), ylim=c(0,1.1*max(pdense)),
       xaxs="i", yaxs="i", xlab="p", ylab="P(p | r,n,M)")
  title(main=paste("r =",r), line=0.3, cex.main=1.2)
  p.mean <- deltap*sum(p*pdense)
  abline(v=p.mean, lty=2)
}
dev.off()
```

In the above code I divide the range 0–1 into `Nsamp` equal-sized intervals and compute the (at first unnormalized) posterior density `pdense` at the centre of each of these. Thus the first point in this grid is at `1/(2*Nsamp)`, the next is offset by $\delta p =$ `1/Nsamp`, etc. This grid is constructed by the following statement.

```
p <- seq(from=1/(2*Nsamp), by=1/Nsamp, length.out=Nsamp)
```

I then normalize the posterior – make the area under the curve equal to one – by dividing `pdense` by `deltap*sum(pdense)` (equation 5.6). It is then a probability *density* function, so `sum(pdense)` is not equal to one. As $\sum_i P(p_i|r,n,M)\delta p = 1$, it follows that `sum(pdense)` is just equal to `1/deltap`, which is `Nsamp`.

## 5.1.2 Beta prior

In practice you are unlikely to have a personal prior for $p$ which is uniform. Given a random coin, I suspect you are far more likely to believe that the coin is fair, or close to fair.[3] An appropriate density function for parameters constrained to lie in the range 0–1 is the beta

[3] If you were attending a magicians' convention you might well adopt a broader prior. Background information is important when analysing data. We do not interpret data independent of context.

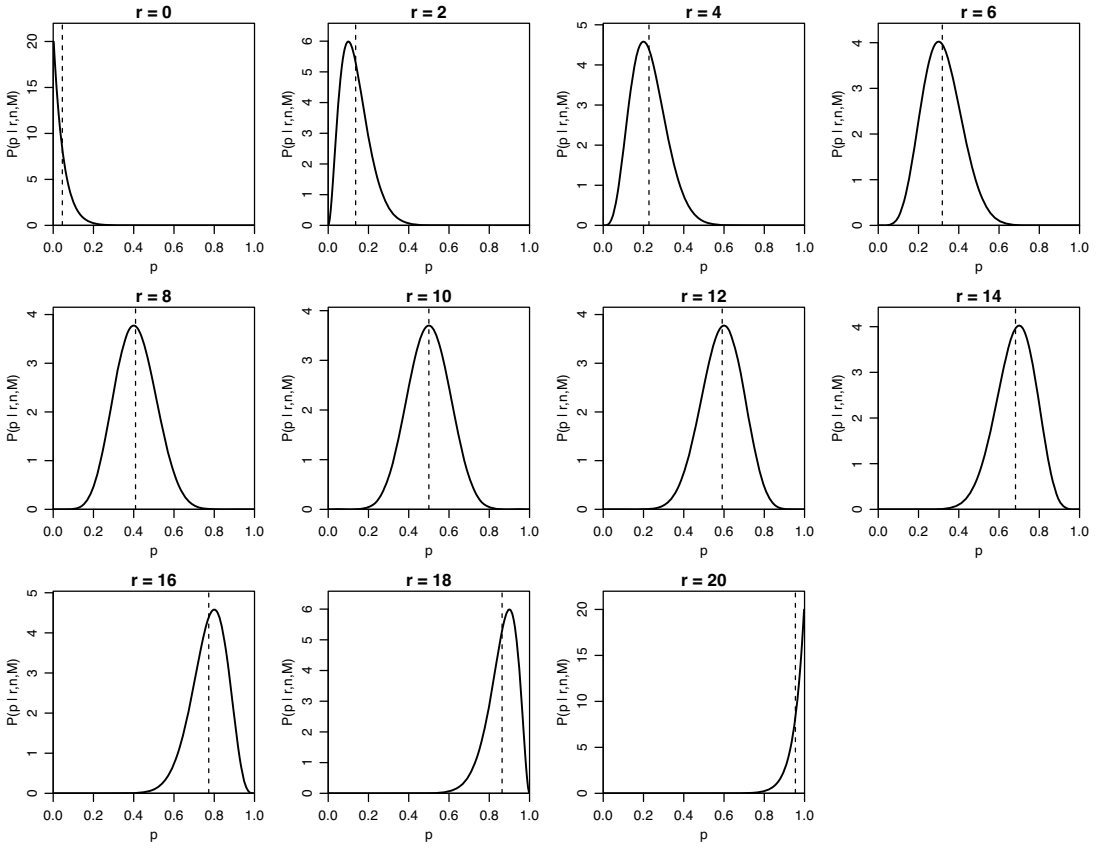


**Fig. 5.2** The posterior PDF for the probability $p$ of a coin giving heads on a single toss, when $r$ of $n = 20$ observed tosses are heads. A uniform prior over $p$ is used. The dashed line is the mean. Note the different vertical scales on the individual panels (the posterior is normalized).

distribution (section 1.4.3). As a reminder, this is described by two parameters, $\alpha$ and $\beta$. Its PDF is

$$P(p) \;=\; \frac{1}{B(\alpha,\beta)}\, p^{\alpha-1}(1-p)^{\beta-1} \quad \text{where} \quad \alpha > 0,\ \beta > 0,\ 0 \le p \le 1. \tag{5.9}$$

If $\alpha = \beta$ the function is symmetric, the mean and mode are 0.5, and the larger the value of $\alpha$ (when $\alpha \ge 1$) the narrower the distribution. An example is shown in figure 5.3. Multiplying this prior by the likelihood, and absorbing terms independent of $p$ into the normalization constant $Z$, gives the posterior

$$\begin{aligned} P(p\,|\,r,n,M) \;&=\; \frac{1}{Z}\, p^{r}(1-p)^{n-r}\, p^{\alpha-1}(1-p)^{\beta-1} \\ &=\; \frac{1}{Z}\, p^{r+\alpha-1}(1-p)^{n-r+\beta-1}. \end{aligned} \tag{5.10}$$

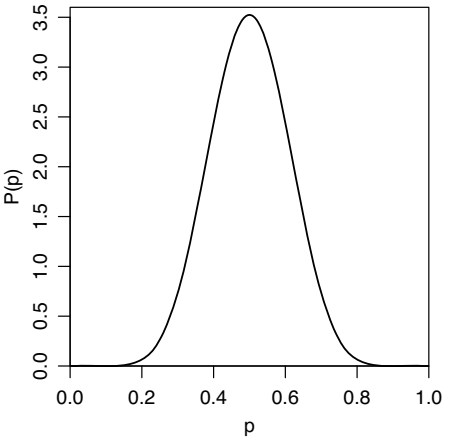


**Fig. 5.3** The beta distribution prior with $\alpha = \beta = 10$.

The posterior has the same form as the prior (for this likelihood); it is also a beta distribution.[4] When this happens, we say that the prior and posterior are *conjugate distributions*. The prior is the *conjugate prior* for this likelihood function. This is not only convenient, but also instructive, because it helps us to appreciate how the data update the prior to produce the posterior. If we start off with a prior with parameters $\alpha_p$ and $\beta_p$, and then measure $r$ heads from $n$ tosses, the posterior is a beta distribution with parameters

$$\begin{aligned} \alpha &= \alpha_p + r \\ \beta &= \beta_p + n - r. \end{aligned} \tag{5.11}$$

As larger values of these parameters correspond to narrower distributions,[5] we see that the data improve our knowledge of $p$ by narrowing the posterior. The mean and mode of the posterior (from equation 1.4.3) are

$$\text{mean} = \frac{\alpha_p + r}{\alpha_p + \beta_p + n} \quad \text{and} \tag{5.12}$$

$$\text{mode} = \frac{\alpha_p + r - 1}{\alpha_p + \beta_p + n - 2} \tag{5.13}$$

respectively. We could use either of these as a single "best" estimate of $p$.

Equation 5.9 tells us that the uniform prior is a beta distribution with $\alpha = \beta = 1$. Thus with a uniform prior, the posterior (equation 5.3) is also a beta distribution, the mean and

[4] It follows from the definition of the beta distribution that the normalization constant in equation 5.10 is $Z = B(r + \alpha, n - r + \beta)$.

[5] If we set $\alpha = \beta$ into the expression for the variance of the beta distribution (equation 1.59) then we see that for large $\alpha$, $\text{Var}(p) \sim 1/\alpha$.

mode of which are

$$\text{mean} = \frac{1+r}{2+n} \tag{5.14}$$

$$\text{mode} = \frac{r}{n}. \tag{5.15}$$

Given $r$ heads from $n$ tosses, what would *your* estimate of $p$ be? The intuitive answer, $r/n$, is equal to the mode of the posterior when adopting a uniform prior. (This is also the maximum likelihood solution, which you can verify by differentiation; see section 4.4.) But if you tossed the coin three times and you got zero heads, would you really estimate $p$ to be zero? Aren't you still more likely to think $p$ is nearer to $1/2$ than to zero, given that the vast majority of coins you encounter are very close to fair?[6] Instead of the mode you could use the mean of the posterior as your estimator, which for zero heads from three tosses gives an estimate for $p$ of 0.2 (and the standard deviation is 0.16, indicating there is considerable uncertainty). The nice thing about the mean in this case is that it gives sensible answers in the limit of little data. Indeed, even if $r = n = 0$ (no data, so the posterior is just the prior) the mean is $1/2$, whereas the mode is undefined (because the posterior is uniform). The standard deviation (of the uniform distribution) is $1/\sqrt{12} = 0.29$.

Equation 5.14 is sometimes called *Laplace's rule of succession*. The larger the number of coin tosses, the more peaked the posterior becomes, and the mean and mode both converge to $r/n$.

Let us now adopt a beta prior with $\alpha = \beta = 10$. This is plotted in figure 5.3. I think this is still a very conservative (broad) prior for a randomly encountered coin, but it is useful for illustrating the influence of the prior.[7] The following code performs the same experiment as before of varying $r$ for $n = 20$, and produces figure 5.4.

R file: `coin2.R`

```
##### Compute posterior PDF for coin problem with a beta prior for a
##### range of r

n <- 20
alpha.prior <- 10
beta.prior  <- 10
Nsamp <- 200 # no. of points to sample at
pdf("coin2.pdf", 9, 7)
par(mfrow=c(3,4), mgp=c(2,0.8,0), mar=c(3.5,3.5,1.5,1), oma=0.5*c(1,1,1,1))
deltap <- 1/Nsamp # width of rectangles used for numerical integration
p <- seq(from=1/(2*Nsamp), by=1/Nsamp, length.out=Nsamp) # rectangle centres
for(r in seq(from=0, to=20, by=2)) {
  pdense <- dbeta(x=p, shape1=alpha.prior+r, shape2=beta.prior+n-r)
  plot(p, pdense, type="l", lwd=1.5, xlim=c(0,1), ylim=c(0, 6.5),
```

[6] If you still think inference should not depend on prior information, do this experiment with someone with a real coin, and bet real money on the outcome of the next toss of the coin. I'm sure that after three tails in three tosses you won't put a large amount of money on the next toss being tails.

[7] If $\alpha > 1$ and $\beta > 1$ the prior is zero at $p = 0$ and $p = 1$, meaning that the posterior will always be zero at these values no matter how much (finite) data we collect. It is generally a bad idea to assign a prior density of zero to a value of a parameter which is not actually impossible, even if it is highly implausible. On the other hand, with enough data, the posterior can be non-zero arbitrarily close to $p = 0$ and $p = 1$, so this is of little practical consequence (in this case).

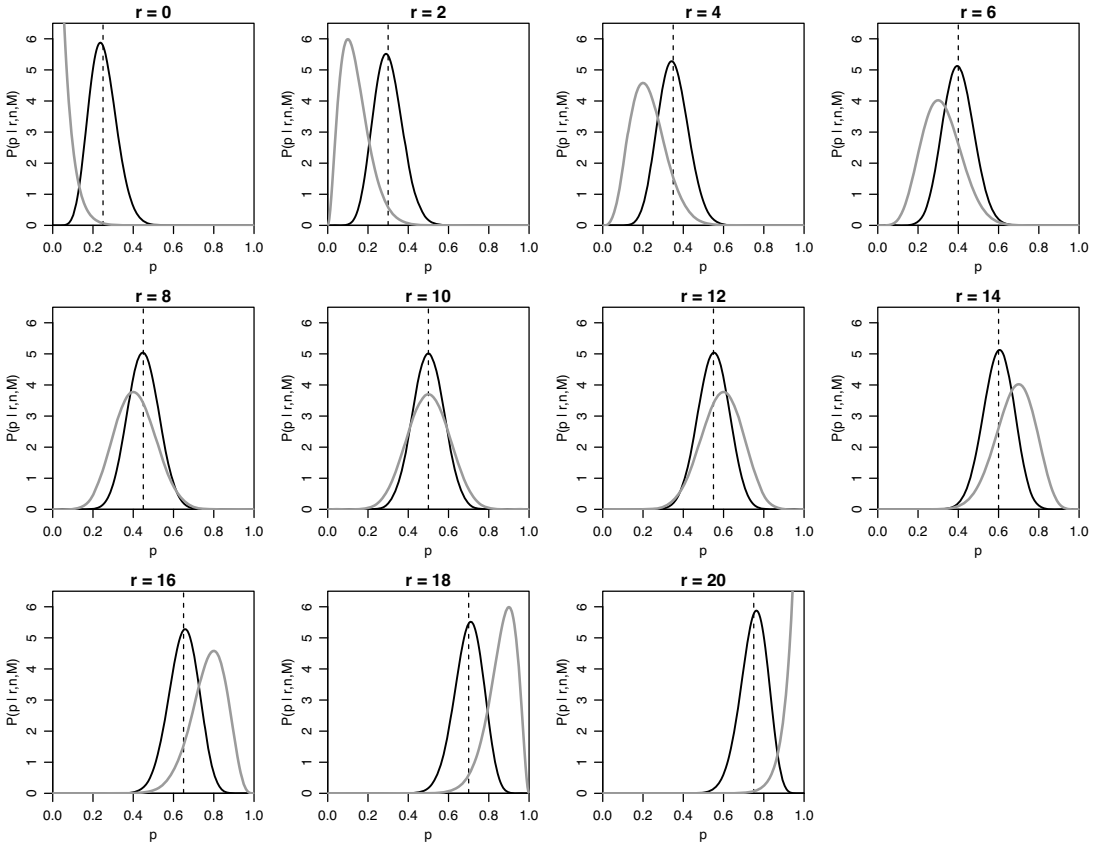


**Fig. 5.4** As figure 5.2, but now using a beta prior on $p$ with $\alpha = \beta = 10$. The black line is the posterior. For comparison, the grey line shows the posterior when using the uniform prior. Note that this extends beyond the top of the plots in the top left and bottom right panels.

```
      xaxs="i", yaxs="i", xlab="p", ylab="P(p | r,n,M)")
  title(main=paste("r =",r), line=0.3, cex.main=1.2)
  p.mean <- deltap*sum(p*pdense)
  abline(v=p.mean, lty=2)
  # overplot posterior obtained from a uniform prior
  pdense.uniform <- dbinom(x=r, size=n, prob=p)
  lines(p, pdense.uniform/(deltap*sum(pdense.uniform)), lwd=2,
        col="grey60")
  # Can verify that pdense can also be found by direct calculation
  #pdense2 <- dbinom(x=r, size=n, prob=p) *
  #           dbeta(x=p, shape1=alpha.prior, shape2=beta.prior)
  #pdense2 <- pdense2/(deltap*sum(pdense2)) # normalize posterior
  #lines(p, pdense2, col="red", lty=2)
}
dev.off()
```

In the above code we no longer need to normalize `pdense`, because it is a beta distribution in $p$ calculated by the R function `dbeta`, which is a normalized density function. The direct calculation of the posterior in the code (commented out) produces `pdense2` which is not normalized, so we must normalize this manually.

It is instructive to plot the likelihood, prior, and posterior together, which is done by the following code. The result is shown in figure 5.5.

R file: `coin3.R`

```
##### Plot prior, likelihood, and posterior PDF for coin problem with a
##### beta prior for a range of r with n fixed

n <- 20
alpha.prior <- 10
beta.prior  <- 10
Nsamp <- 200 # no. of points to sample at
pdf("coin3.pdf", 9, 7)
par(mfrow=c(3,4), mgp=c(2,0.8,0), mar=c(3.5,3.5,1.5,1), oma=0.5*c(1,1,1,1))
deltap <- 1/Nsamp # width of rectangles used for numerical integration
p <- seq(from=1/(2*Nsamp), by=1/Nsamp, length.out=Nsamp) # rectangle centres
prior <- dbeta(x=p, shape1=alpha.prior, shape2=beta.prior)
for(r in seq(from=0, to=20, by=2)) {
  like  <- dbinom(x=r, size=n, prob=p)
  like  <- like/(deltap*sum(like)) # for plotting convenience only
  post  <- dbeta(x=p, shape1=alpha.prior+r, shape2=beta.prior+n-r)
  plot(p, prior, type="l", lwd=1.5, lty=2, xlim=c(0,1), ylim=c(0, 6.5),
       xaxs="i", yaxs="i", xlab="p", ylab="density")
  lines(p, like, lwd=1.5, lty=3)
  lines(p, post, lwd=1.5)
  title(main=paste("r =",r), line=0.3, cex.main=1.2)
}
dev.off()
```

We see from this plot how the (fixed) prior combines with the likelihood to form the posterior: the posterior (solid line) is just the product of the likelihood (dotted line) and the prior (dashed line), which is then renormalized. While the prior and posterior are automatically normalized by the R function `dbeta`, I explicitly normalize the likelihood before plotting. This doesn't really make sense, because it is not a PDF over $p$, but rather a PDF over $r$ given $p$. I normalize it just to ease visual comparison (so the area under all curves is unity).

You might think that when the prior density is small, the posterior density must also be small. But the absolute value of the product of prior and likelihood is not important, because this product is renormalized to make the posterior. This is crucial. Even if the likelihood is small everywhere over the range of the prior, the posterior must still integrate to one.

As we get more and more (useful) data, the prior stays the same, but the likelihood becomes more peaked, so the posterior will be influenced more by the likelihood than by the prior. I demonstrate this using the same prior as above, a beta distribution with $\alpha = \beta = 10$. Let's start with $(r, n) = (2, 3)$, i.e. two heads and one tail. I will increase the amount of data in steps of factors of two, but keeping the proportion of heads to tails the

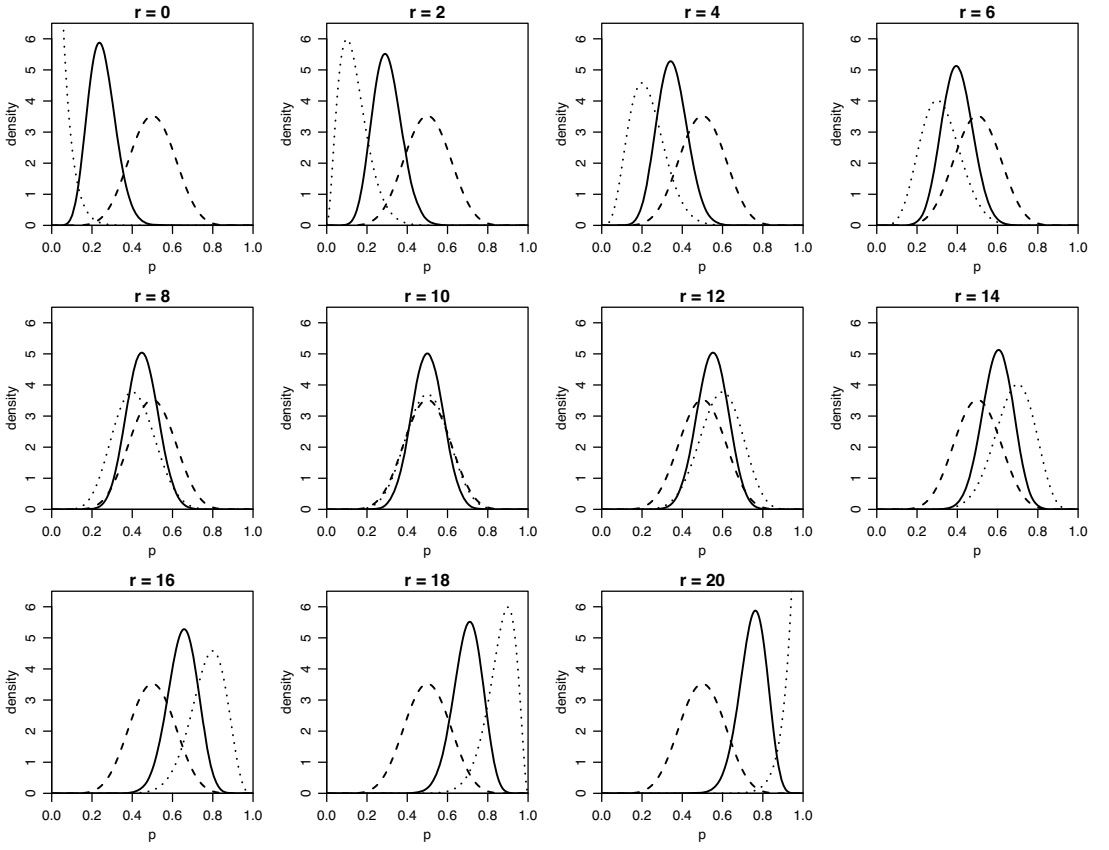


**Fig. 5.5** As figure 5.4, but now showing in addition to the posterior (solid line), the likelihood (dotted line), and the prior (dashed line). The likelihood has been normalized to have unit area under the curve.

same, so $(r, n) = (2, 3), (4, 6), (8, 12), \ldots$. That is, I increase the signal-to-noise ratio in the data. The following code implements this; the results are shown in figure 5.6.

R file: `coin4.R`

```
##### Plot prior, likelihood, and posterior PDF for coin problem with a
##### beta prior for a range of r with ratio r/n fixed

alpha.prior <- 10
beta.prior  <- 10
Nsamp <- 200 # no. of points to sample at
pdf("coin4.pdf", 7, 7)
par(mfrow=c(3,3), mgp=c(2,0.8,0), mar=c(3.5,3.5,1.5,1), oma=0.5*c(1,1,1,1))
deltap <- 1/Nsamp # width of rectangles used for numerical integration
p <- seq(from=1/(2*Nsamp), by=1/Nsamp, length.out=Nsamp) # rectangle centres
prior <- dbeta(x=p, shape1=alpha.prior, shape2=beta.prior)
for(r in 2^(1:9)) {
```

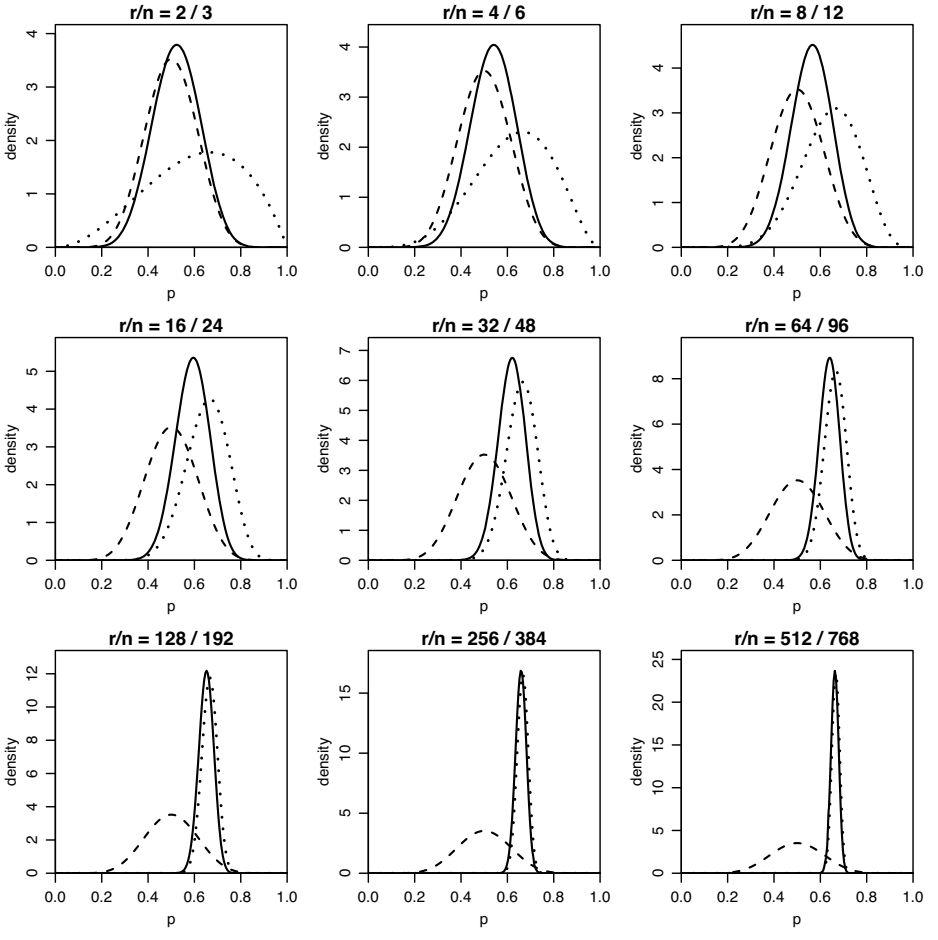


**Fig. 5.6** The evolution of the likelihood (dotted line) and posterior (solid line) as the amount of data is increased, with $r/n$ kept constant at $2/3$. The beta prior (dashed line) is kept fixed (with $\alpha = \beta = 10$). Note the different vertical scales on the individual panels.

```
  n <- (3/2)*r
  like  <- dbinom(x=r, size=n, prob=p)
  like  <- like/(deltap*sum(like)) # for plotting convenience only
  post  <- dbeta(x=p, shape1=alpha.prior+r, shape2=beta.prior+n-r)
  plot(p, prior, type="l", lwd=1.5, lty=2, xlim=c(0,1),
       ylim=c(0,1.1*max(prior,post,like)), xaxs="i", yaxs="i",
       xlab="p", ylab="density")
  lines(p, like, lwd=2, lty=3)
  lines(p, post, lwd=1.5)
  title(main=paste("r/n =",r,"/",n), line=0.3, cex.main=1.2)
}
dev.off()
```

The prior (dashed line) is the same in all panels (just the vertical scale changes). Initially we have few data, so the likelihood (dotted line) is broad, and the posterior (solid line) is dominated by the prior. As the amount of data increases, the posterior is increasingly dominated by the likelihood: the posterior shifts from the prior towards the likelihood. By the time we have a lot of data, the prior is basically irrelevant. This is all a consequence of the posterior being the product of the prior and likelihood (then renormalized).

From equation 5.11 we see that the effect of the prior is equivalent to adding $\alpha_{\rm prior}$ successes to the actual observed number of successes, and adding $\beta_{\rm prior}$ failures to the actual observed number of failures.

In none of this have we considered the order of heads and tails. If we wanted to drop the assumption of the independence of the coin tosses, then we would need to take into account this ordering. This corresponds to a more complex (but legitimate) model for the coin, in which the binomial distribution is no longer the appropriate likelihood function.

## 5.2 Likelihoods can be arbitrarily small and their absolute values are irrelevant

We have seen in the previous section how the posterior is the product of the likelihood and prior, which is then renormalized. This renormalization is critical because the posterior must integrate to unity. Even if the prior and likelihood have little overlap, the posterior can still be strongly concentrated and have a large peak. We see this when comparing the panels labelled $r = 0$ and $r = 10$ in figure 5.5: the degree of overlap is very different, yet the posteriors have similar heights.

This highlights the fact that the absolute values of the likelihood are not relevant. What counts is their relative values. If we toss a coin twice, there is a probability of $1/4$ of getting a particular sequence of heads and tails. If we toss it $n$ times, the probability is $2^{-n}$. The larger the data set, the lower the likelihood of any particular outcome. Even if we have the true model with the true parameters, the probability (likelihood) of any *particular* data set can be very small, and for all but the smallest data sets, the likelihood is invariably very small for *any* data set. The same is true for continuous data when the likelihood is a probability density rather than an actual probability.

We can demonstrate this using the coin example. If we toss a fair coin $n$ times, then the expected number of heads (the mean of the distribution) is $E[r] = n/2$. The probability that we observe $r$ heads – the likelihood – is given by the binomial distribution. The left panel of figure 5.7 shows the variation of the logarithm of this likelihood as a function of $r/E[r]$ for four different fixed values of $E[r]$. The curve labelled $E[r] = 10$ means we toss the coin 20 times – for which we expect to get 10 heads – and shows the probability of getting 7, 8, ..., 13 heads. (The distribution is discrete, so this "curve" can only be plotted as a set of points.) I plot against $r/E[r]$ rather than $r$ so that the curves are all centred on 1.0. Looking at the different curves we see how quickly the likelihood drops for larger sample sizes. For example, the probability of getting seven heads in 20 tosses (top curve

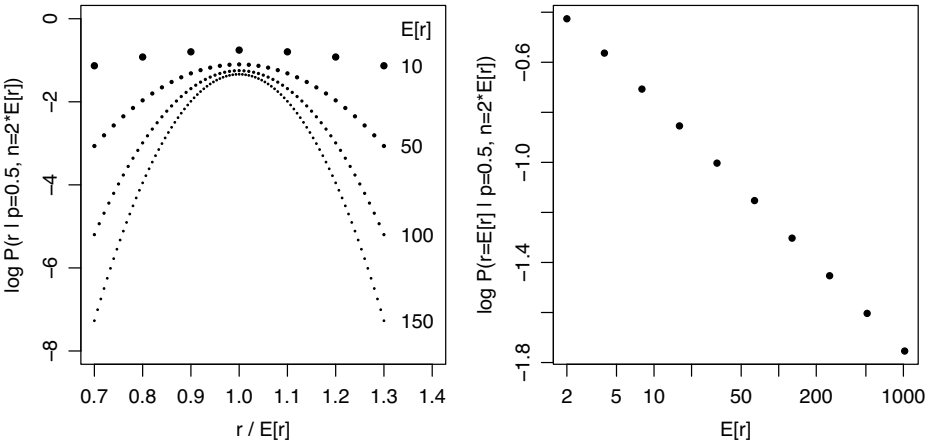


**Fig. 5.7** The likelihood is smaller for larger data sets. This is shown here using the binomial likelihood with $p = 0.5$. Left: the variation of the log (base 10) likelihood as a function of $r/E[r]$, where $E[r]$ is the expected number of successes, which is half the sample size, $E[r] = n/2$. This is shown for four sample sizes, each plotted with a different point size and labelled with the value of $E[r]$. Right: the log (base 10) likelihood of the expected number of successes as a function of that number. This is the sequence of values at $r/E[r] = 1$ in the left panel, but now extended to larger sample sizes (the horizontal axis in the right panel is logarithmic).

for $r/E[r] = 7/10$) is 0.07 (log probability is $-1.2$), whereas the probability of getting 70 heads in 200 tosses (third curve down) is $6 \times 10^{-6}$ (log probability is $-5.2$).

Yet the likelihood does not drop just for "unlikely" outcomes of the data. The right panel of figure 5.7 shows how the likelihood of the expected number of tosses varies with that expected number. We see that even at the expected value of the data (which is also the single most likely outcome: half the tosses are heads), the likelihood is small, and gets smaller the larger the data set.

The take-home message is that absolute values of the likelihood on their own tell us nothing about how good the model is.

This fact is central to the problem of frequentist hypothesis testing using p values, which we will cover in chapter 10. The appropriate p value in our example is the probability that the number of heads obtained in $n$ tosses would be equal to or less than some particular threshold value, $r$. This is given by the cumulative probability of the binomial likelihood up to $r$

$$\text{p value} = \sum_{r'=0}^{r} \binom{n}{r'} p^{r'} (1-p)^{n-r'}. \tag{5.16}$$

In the frequentist paradigm, a hypothesis – in this case a fair coin, $p = 1/2$ – is rejected when the p value is small, because this indicates that the probability is small of getting as few heads as we did ($r$) or even fewer (i.e. even more "extreme" data). The problem with

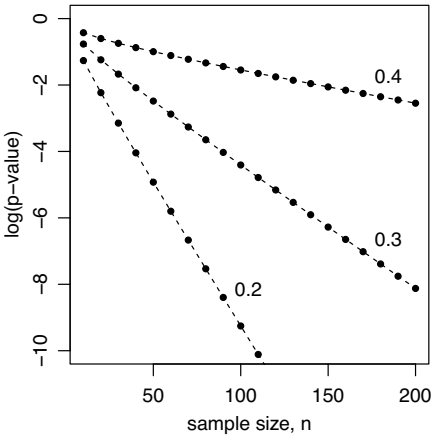


**Fig. 5.8** In the coin tossing example the p value is the probability of getting $r$ or fewer heads in $n$ tosses, which is given by the cumulative binomial distribution (equation 5.16). The logarithm (base 10) of the p value is shown here as a function of $n$ (in steps of 10) as the black points, for three different values of $r/n$: 0.4, 0.3, 0.2, for a fair coin (parameter $p = 1/2$; this "$p$" has nothing to do with the "p" in p value). These values and the step size were chosen to ensure that $r$ is always an integer (the binomial is a discrete distribution). The dashed lines just connect the points for common $r/n$.

this idea, however, is that the larger the data set, the smaller the p value. We can see this by plotting the p value as a function of $n$ for a fixed fraction of heads, $r/n$. That is, we ask ourselves what is the probability of getting, say, 30% or fewer heads in $n$ tosses. This is shown in figure 5.8 for three values of $r/n$ with $p = 1/2$, i.e. the true hypothesis. For example, if $n = 50$, then the probability of getting 15 or fewer heads (30%, the line marked 0.3) is $3.3 \times 10^{-3}$ (log probability is $-2.5$), but for $n = 100$, the probability of getting 30 or fewer heads (also 30%) is $3.9 \times 10^{-5}$ (log probability is $-4.4$). This is what we expect. But the usual approach in the frequentist paradigm is to reject a hypothesis once the p value drops below some fixed threshold, say $10^{-3}$, without explicitly taking into account the size of the data set. Yet we see from figure 5.8 that whether or not the true hypothesis is rejected depends on the size of the data set. The larger the data set, the more likely we are to incorrectly reject the true hypothesis. This tells us that the p value is insufficient for deciding whether or not a hypothesis can be rejected. We must also take into account the sample size. Yet in real problems involving more complicated likelihood functions, it can be difficult to decide what threshold to use.

This problem with p values arises because it is using the absolute values of the likelihood, or sums (or integrals) thereof. We will see in chapter 11 how we can avoid this problem by instead considering the *ratios* of probabilities for different models.

The following two R scripts can be used to reproduce the experiments and plots in this section.

R file: `likelihoods_small.R`

```
##### Illustration with the binomial distribution that likelihoods are small

pdf("likelihoods_small.pdf", 8, 4)
par(mfrow=c(1,2), mgp=c(2.2,0.8,0), mar=c(3.5,3.5,0.5,1),oma=0.5*c(1,1,1,0))
expVal <- c(10, 50, 100, 150)
plot(0:1, 0:1, type="n", xlim=c(0.7, 1.4), ylim=c(-8, 0), xlab="r / E[r]",
     ylab=expression(paste(log," P(r | p=0.5, n=2*E[r])")))
for(i in 1:length(expVal)) {
  rVec <- seq(from=0.7*expVal[i], to=1.3*expVal[i], by=1)
  loglike <- (1/log(10))*dbinom(rVec, size=2*expVal[i], prob=0.5, log=TRUE)
  points(rVec/expVal[i], loglike, pch=20, cex=1/i)
}
text(1.31, -0.3, "E[r]", pos=4)
text(1.31, c(-1.18, -3.13, -5.25, -7.32), expVal, pos=4)
expVal  <- 2^(1:10)
loglike <- (1/log(10))*dbinom(expVal, size=2*expVal, prob=0.5, log=TRUE)
plot(expVal, loglike, log="x", pch=20, xlab="E[r]",
     ylab=expression(paste(log," P(r=E[r] | p=0.5, n=2*E[r])")))
dev.off()
```

R file: `pvalues_binomial.R`

```
##### Illustration with the binomial distribution that p-values depend on
##### sample size

pdf("pvalues_binomial.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
sampSize <- seq(from=10, to=200, by=10)
fracLim  <- c(0.4, 0.3, 0.2)
plot(range(sampSize), c(-10,0), type="n", xlab="sample size, n",
     ylab="log(p-value)")
for(f in fracLim) {
  pVal <- pbinom(q=f*sampSize, size=sampSize, prob=0.5)
  points(sampSize, log10(pVal), pch=20)
  lines(sampSize,  log10(pVal), lty=2)
}
text(c(170, 170, 90), c(-1.8, -6.7, -8.2), fracLim, pos=4)
dev.off()
```

## 5.3 Assigning priors

Probabilistic inference provides answers to well-posed problems. But it does not define our models, or our priors, or tell us which data to collect or how to collect it. These are all important questions which affect the inference and thus the conclusions. Here we look at how to assign priors.

We saw in section 3.5 with the parallax problem that a prior can be essential to making

any sensible inference; there was no general answer without one. Both this example and the coin example in section 5.1 show how the posterior PDF over a parameter depends on both the prior and the likelihood. As the data become more informative (e.g. the fractional parallax error is smaller or we make more coin tosses), the posterior becomes increasingly dominated by the likelihood, and the prior becomes less and less relevant. This is because as the data get better, the likelihood – when viewed as a function of the parameters – becomes narrower. So no matter what the shape of the prior, it will be approximately constant across this narrow stretch of likelihood (unless the prior has finite support and partially excludes the likelihood). In such cases we need not be too concerned with the exact specification of the prior. When the data are poor, on the other hand, the prior plays a more dominant role. As indeed it should: it is what we know about the parameter in the limit of no data. So when the data are not highly informative we do need to be concerned about the prior. How, then, do we assign a prior?

The prior should incorporate any relevant information we have about the problem. One invariably has some kind of prior information on the model parameters, just from the very fact of having chosen the model. Sometimes one knows the parameter cannot occupy a certain range by definition, e.g. the distance must be positive, or the probability of a coin landing heads lies between 0 and 1. These may seem trivial, but we already saw in section 3.5 how the positivity of distance enabled us to sensibly interpret arbitrarily small and negative parallaxes. Limits of measurement or observability may also provide bounds.

We often have information beyond such hard constraints. For the coin example in section 5.1, a flat prior on $p$ for a coin chosen at random is very conservative. We saw this from the inferred posterior: with 20 tosses and seven heads, adopting a uniform prior led to a very broad posterior (figure 5.1). I suspect most people, given these data, would implicitly conclude on a much narrower posterior than this. This is because we actually have a strong prior that the coin is very close to fair. No coin is perfectly symmetric, but deviations from this are small. You would need a lot of data to convince you otherwise. The prior quantifies this.

We implicitly use priors all the time in every day life. If a brown animal dashes past you in the forest one day, but you didn't get a good look at it, are you more likely to think it was a deer or a dinosaur? Yet while few people seriously doubt the existence or importance of priors, it can be difficult to specify them exactly in practice. For the coin example we would presumably adopt a prior symmetric about $p = 0.5$, but how narrow should it be? There is no rule here, because priors depend on what you know, believe, and understand about the problem. Turning this notion into a probability distribution is hard and rarely unique.

Often we adopt standard distributions for priors, like the beta distribution, and we'll see more examples of using standard distributions below and in the next chapter. If the parameter has finite support – or if we want our prior to enforce this – then we could even define a discrete prior ourselves using a histogram.

There are some principles which can help us to adopt appropriate priors. One is the *principle of insufficient reason*, also called the *principle of indifference*. This essentially says that if we have a set of mutually exclusive outcomes, and we don't expect any one of them to be more likely to occur than the other, then we should assign them equal probabilities.

This is almost a tautology, but it can nonetheless help (it leads us to derive the binomial distribution, for example – see section 1.4.1).

However, if we are unable to decide on the set of outcomes, we will run into problems. Consider the proposition – which I'll call $R$ – "Norah is wearing a red dress today". If we have no idea whether the dress is red or not, then we might conclude that $P(R) = 1/2$, and $P(R') = 1/2$, where $R'$ is the complementary proposition "Norah is not wearing a red dress today". But if we then consider an additional proposition $W$ – "Norah is wearing a white dress today" – do we now change our probabilities to $P(R) = 1/3$, $P(W) = 1/3$, and $P(\text{neither } R \text{ nor } W) = 1/3$? Our probabilities appear to change based on how we choose to enumerate the possibilities. How many colours are there? Are they equally probable a priori? Is she even wearing a dress?

In general we will want to adopt a prior which is as conservative (uninformative) as possible, consistent with the information we have. Below we will look at some simple but useful cases of this. There are principled approaches for determining priors given certain information, perhaps the best known being the principle of *maximum entropy*. The idea is to find the least informative (most entropic) distribution given certain information.[8]

Another approach is *empirical Bayes*, in which we estimate the priors from some general properties of the data (or perhaps a subset thereof). Note that we should not iterate the Bayesian procedure, using the posterior as the prior for a reanalysis of the same data, and then repeating this $N$ more times in the hope of removing the effect of the original prior. This will result in a posterior which is proportional to $P(\theta)P(D\,|\,\theta)^{N+1}$. This is indeed dominated by the data and will make the prior irrelevant for large $N$, but only because we have have reused the same data $N + 1$ times by erroneously multiplying non-independent (in fact identical) likelihoods.

In contrast, we *can* take the posterior from one analysis to be the prior of the next analysis if these two analyses involve independent data. The final posterior will be identical to having combined the two data sets together with the original prior. This follows because the likelihood factorizes. Let $D_1$ and $D_2$ be the two independent data sets. Then

$$\begin{aligned} P(\theta\,|\,D_1, D_2) &\propto P(D_1, D_2\,|\,\theta)P(\theta) \\ &\propto P(D_2\,|\,\theta)P(D_1\,|\,\theta)P(\theta) \\ &\propto \underbrace{P(D_2\,|\,\theta)}_{\text{likelihood for } D_2} \times \underbrace{P(\theta\,|\,D_1)}_{\text{posterior from } D_1} . \end{aligned} \tag{5.17}$$

This highlights again that the prior and posterior are both PDFs on the parameters, but based on different information.

Another way of assigning priors is *hierarchical Bayes*. This involves selecting a form for the prior parameter distribution, but then putting a prior distribution over its parameters. We then use this to marginalize over the parameter prior PDF. That is, we introduce a higher level prior with fixed parameters to average over the lower level prior. This topic is covered by several text books, such as Kruschke (2015) and McElreath (2016).

A pragmatic approach to dealing with an inability to specify a unique prior is to test

[8] If only the mean and the variance of a probability distribution are known, then the principle of maximum entropy shows that the Gaussian is the least informative distribution.

how sensitive our results are to the choice of prior. We may vary the bounds selected, or the shape of the function. Given two or more priors, all of which seem equally reasonable, we can derive the posteriors and compare them. If they give results which we consider to be indistinguishable from the point of view of their interpretation, then we can consider the choice of prior as unimportant. If not, then this is because the data are insufficiently conclusive (by definition of what we have just considered to be reasonable priors). We should also not forget that the posterior PDF, just like the prior, is a probability *distribution*: it does not give certainty. In science, as in life, there is none.

### 5.3.1 Location and scale parameters

Suppose we have a model that specifies the location (in an abstract sense) $x_0$ of some quantity, and we have no prior knowledge of its value other than that it lies between some limits. We adopt a prior $P(x_0)$. Our result (posterior PDF) should be independent of the origin of the coordinate system we adopt for this location. This requires that the prior be invariant with respect to a linear translation of $x_0$. We can write this translation as $x_1 = x_0 + c$ for some constant $c$. This change of variables requires

$$P(x_o)\,dx_o \;=\; P(x_1)\,dx_1 \;=\; P(x_0 + c)\,dx_0. \tag{5.18}$$

This can only be satisfied if $P(x_0)$ is constant over the limits, and zero outside. This is the uniform prior, which is therefore the invariant distribution with respect to additive changes in a quantity.

Suppose now we wish to infer the size or scale of something. Complete ignorance here means that we know nothing about this scale, other than that it must be positive. Our prior PDF should therefore be invariant with respect to being stretched. Let this scale be a length, $w_0$. If we specify a prior over $w_0$ in units of metres, the prior should be the same if we then decide to express it in centimetres instead (as the units should not matter if we are ignorant of the scale). This transformation is $w_1 = aw_0$ for a positive constant $a$. This demands

$$P(w_0)\,dw_0 \;=\; P(w_1)\,dw_1 \;=\; P(aw_0)\,a\,dw_0. \tag{5.19}$$

This can only be satisfied if $P(w_0) \propto 1/w_0$, as then $P(aw_0) \propto 1/(aw_0)$. Thus a prior proportional to the inverse of a parameter is invariant with respect to rescalings of that parameter. Through a change of variables we see that this is equivalent to $P(\log w_0) = \text{constant}$ (see the example in section 1.9.1). This tells us that equal multiples of $w_0$ have the same probability (e.g. the probability between 1 and 10 is the same as the probability between 10 and 100), which is just what ignorance of scale means. This type of prior is often called a *Jeffreys prior* (but see the next section).

Location and scale parameters are common. If we want to infer the mean of a Gaussian, the above principle tells us to use a uniform distribution for the prior. When inferring its variance, we should use a Jeffreys prior. Note that

$$P(\log \sigma^2) = \text{constant} \quad \Rightarrow \quad P(\log \sigma) = \text{constant} \tag{5.20}$$

so it does not matter whether we talk about a Jeffreys prior in the variance or in the standard deviation.

We may have one of these priors on a transformed version of the parameter. For example, if we are trying to infer a length $x$ from an angular measurement $\phi$, and we assign the angle a uniform prior, then a transformation of variables shows us that the prior over $x \propto \tan\phi$ has a Cauchy distribution (section 1.4.7).

One disadvantage of these particular uniform priors is that they are improper (cannot be integrated). We must be careful when we use such priors: as the posterior is the product of the prior and likelihood, it is possible (depending on the data) that the posterior will also be improper, in which case it is meaningless. We saw with the parallax problem in section 3.5 that improper priors can lead to wild inferences if the data are poor. But often the likelihood – when seen as a function of the parameters – will drop rapidly enough beyond a finite-sized region so that the resulting posterior is proper. The likelihood may even be quite peaked, in which case the exact form of the prior will have little influence on the resulting posterior. However, improper priors are unusable when we do model comparison with the evidence, because then we must integrate the unnormalized posterior over the parameter (this topic will be covered in chapter 11). A pragmatic solution is to apply upper and lower limits to make it proper, and then to test the sensitivity of our results to these limits. Doing this for the Jeffreys prior between limits $\theta_{\min}$ and $\theta_{\max}$ and normalizing gives

$$P(\theta) = \frac{1}{\theta \ln(\theta_{\max}/\theta_{\min})}. \tag{5.21}$$

Setting priors for location and scale parameters are examples of setting priors that are invariant with respect to irrelevant reparametrizations of the problem. We will not always want such invariances, however. A more fundamental consideration is the nature of the problem. In the astronomical distance inference problem in section 3.5, for example, we know that the Galaxy has a characteristic scale length, so our prior should respect this. In general, priors should be independent of irrelevant information. A nice illustration of this is the *Bertrand paradox*, in which different – but apparently equally valid – assumptions about what constitutes "random" leads to different answers to a probabilistic problem. Jaynes (1973) describes this problem and shows that a proper consideration of the invariances in the problem leads to a unique solution.

### 5.3.2 Jeffreys prior

What I referred to in the previous section as the Jeffreys prior can be seen as a case of a more general approach to setting priors introduced by Harold Jeffreys (Jeffreys 1961; see also Robert *et al.* 2009).[9] His goal was to produce posteriors that are invariant under reparametrizations. As the likelihood is motivated by the problem context – the generative model and the measurement model – we can imagine that achieving this invariance will put conditions on the prior that depend (only) on the likelihood. Let $\boldsymbol{\theta}$ be a $J$-dimensional vector of parameters and $P(x\,|\,\boldsymbol{\theta})$ be the likelihood. The Jeffreys prior is defined to be

[9] In fact, Harold Jeffreys introduced various rules for specifying priors. The general rule, described in this section, does not always give the location and scale priors of the previous section, and there were situations where Jeffreys did not recommend the general rule. For a discussion see Kass & Wasserman (1996).

proportional to the square root of the determinant of the *Fisher information matrix* $\mathcal{I}(\boldsymbol{\theta})$,

$$P(\boldsymbol{\theta}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\theta})|}. \tag{5.22}$$

The Fisher information matrix is the $J \times J$ matrix with elements $\mathcal{I}_{i,j}$ equal to the expectation of the second derivatives of the log likelihood[10]

$$\mathcal{I}_{i,j} = -E\left[\frac{\partial^2 \ln P(x|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] \tag{5.24}$$

where the expectation is taken with respect to the data $x$, i.e.

$$E\left[\frac{\partial^2 \ln P(x|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] = \int \frac{\partial^2 \ln P(x|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} P(x|\boldsymbol{\theta})\, dx. \tag{5.25}$$

The second derivative is a measure of the curvature of a function. Thus if the log likelihood has a sharply peaked maximum (as a function of $\theta$), the second derivative at this point will be large and negative, so the information will be large and positive.[11]

The origin of the Jeffreys prior for a one parameter problem can be understood as follows. If we make a smooth, monotonic change of variables $\theta = \theta(\psi)$, then the information for $\psi$ is

$$\begin{aligned}
\mathcal{I}(\psi) &= -E\left[\frac{\partial^2 \ln P(x|\psi)}{\partial\psi^2}\right] \\
&= -E\left[\frac{\partial^2 \ln P(x|\theta)}{\partial\theta^2}\right]\left|\frac{d\theta}{d\psi}\right|^2 \\
&= \mathcal{I}(\theta)\left|\frac{d\theta}{d\psi}\right|^2
\end{aligned} \tag{5.26}$$

where $\theta(\psi)$ is independent of $x$ so could be taken out of the expectation. This equation has the same form as the expression for the prior under the change of variables (section 1.9.1)

$$P(\psi) = P(\theta)\left|\frac{d\theta}{d\psi}\right|. \tag{5.27}$$

Thus if we set $P(\theta) \propto \sqrt{|\mathcal{I}(\theta)|}$ in this equation, we see that the transformation of this prior is still proportional to the square root of the information (because the Jacobian cancels). As the information depends only on the likelihood, this shows that the information contained in the prior is invariant under the transformation.

[10] By carrying out the differentiation and taking expectations you can show that this may also be written as

$$\mathcal{I}_{i,j} = E\left[\frac{\partial \ln P(x|\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \ln P(x|\boldsymbol{\theta})}{\partial\theta_j}\right]. \tag{5.23}$$

[11] If you return to section 4.5 you will recall that the log likelihood for multi-dimensional linear regression was $-\frac{1}{2}\mathrm{SS}_{\mathrm{res}}$ to within an additive constant. The negative second derivative of this with respect to the parameters $\boldsymbol{\beta}$ is $X^{\mathsf{T}}X$, the (Fisher) information matrix.

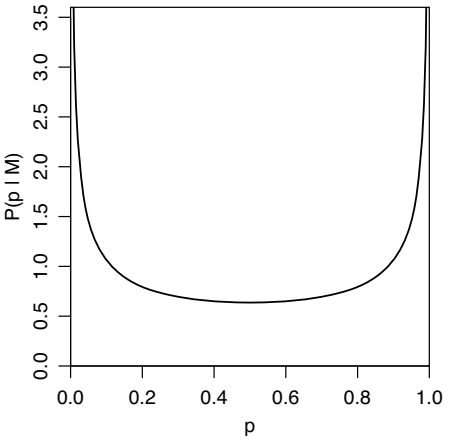


**Fig. 5.9** The beta distribution prior with $\alpha = \beta = 1/2$ (also called the arcsine distribution).

## Example: Jeffreys prior for a binomial likelihood

As an example we compute the Jeffreys prior for the binomial likelihood, $P(r\,|\,p, n)$ in equation 1.43, which has a single parameter $p$. Taking the logarithm of this likelihood and differentiating twice with respect to $p$ gives

$$\frac{\partial^2 \ln P(r\,|\,p,n)}{\partial p^2} = -\frac{r}{p^2} - \frac{(n-r)}{(1-p)^2}. \tag{5.28}$$

Using the definition of expectation for discrete distributions (equation 1.22), the information is

$$\mathcal{I}(p) = -\sum_{r=0}^{n} \frac{\partial^2 \ln P(r\,|\,p,n)}{\partial p^2} P(r\,|\,p,n). \tag{5.29}$$

As the Fisher information is linear in $r$, and $E[r] = \sum_r r P(r\,|\,p,n) = np$, this sum is easy to evaluate

$$\begin{aligned} \mathcal{I}(p) &= \frac{np}{p^2} + \frac{(n-np)}{(1-p)^2} \\ &= \frac{n}{p(1-p)} \\ P(p) &\propto p^{-1/2}(1-p)^{-1/2} \end{aligned} \tag{5.30}$$

where in the last line I have used equation 5.22 and have absorbed $n$ into the proportionality constant. This prior is a beta distribution with parameters $(\alpha, \beta) = (1/2, 1/2)$ (this is also called an arcsine distribution) and is shown in figure 5.9. It is symmetric about $p = 1/2$, of course, and goes to infinity at $p = 0$ and $p = 1$. Note that in this case the Jeffreys prior is also a conjugate prior.

Taking the same approach we can show that the Jeffreys prior for the standard deviation

$\sigma$ of a Gaussian distribution is $1/\sigma$ (when the mean is known).[12] This is the same result which we derived in equation 5.19 to achieve the invariance of a parameter with respect to multiplicative scalings. This makes sense, as the standard deviation of a Gaussian acts like a scale parameter. You can likewise show that the Jeffreys prior for the Poisson rate parameter ($\lambda$ in equation 1.51) is $1/\sqrt{\lambda}$.

### 5.3.3 Are priors subjective?

Some people are uncomfortable with using probabilistic inference (Bayesian statistics) because it seems to depend on the existence of subjective priors, and not just on the data. But if we think a bit more carefully, we will realise that the data are subjective in some sense too. What data did we choose to collect? What did we discard? How did we process them? We also make a decision when we choose the likelihood function, a decision that is based on how we understand the data to have been generated and what we assume their noise properties to be. This choice is independent of the measured values of the data, and different people could plausibly make different choices. Is the choice of the likelihood function any more "objective" than the choice of the prior? The data analysis and interpretation depend on more than just the data themselves. They also depend on what we know and what we assume, as examples in this and previous chapters have shown.

Probabilities reflect our own state of knowledge. Suppose I write "X" on one card, "Y" on another, and put each into a different sealed envelope. I then shuffle them and give one envelope to person A and the other to person B. Before opening anything, both A and B think there is a probability of 0.5 that A has card X. If B now opens her card and sees she has Y (but says nothing), then she knows (probability 1) that A has card X. But A's state of knowledge has not changed. Different people can correctly assign different probabilities to the same proposition because they have different knowledge. This may seem like an irrelevant example, but we have precisely this situation in science: the data we obtain give us incomplete knowledge of the true situation.

So in the sense that priors do not come from the data, they are indeed subjective. Yet so is the process of analysing and interpreting data, because people with different knowledge about a problem are justified to analyse and interpret the data differently. If I have information which tells me that some data can be rejected as outliers, then surely I should use that information. But if you do not have this information, then you should not start rejecting data in an ad hoc manner.

The important point about assigning priors is that two people with the same knowledge should assign the same prior. Given the same data and choice of the likelihood, they will then agree on the posterior.

Priors are unavoidable when doing inference. Many statistical approaches that pretend

[12] When the standard deviation is known, this general approach gives a uniform prior for the mean, which is the same result derived in equation 5.18. When both mean and standard deviation are unknown, however, the general approach gives the prior $P(\mu, \sigma) \propto \sigma^{-2}$. Jeffreys (1961) argues that this is inappropriate, and that we should instead adopt independent priors even when both parameters are unknown. This is what we will do in section 6.2.3.

not to employ them actually do so implicitly. The probabilistic approach allows one to express priors openly, to quantify them, and to investigate the sensitivity of results to them.

So the answer to the question posed by this section title is "yes", to which we should add "just like the rest of data analysis".

## 5.4 Some other conjugate priors

We saw in section 5.1.2 that the conjugate prior for the binomial likelihood is a beta distribution. Many likelihood distributions have conjugate priors. The key to finding the conjugate prior is to identify a distribution which, when multiplied by the likelihood, will have the same form of dependence on the model parameters. If we know what the conjugate prior is, then finding out the parameters of the posterior just involves some algebra. We will look at a couple more examples here.

### Poisson likelihood

Comparing the Poisson distribution, equation 1.51, with the gamma distribution, equation 1.61, we see that both have the same dependence on $\lambda$ if we set $x = \lambda$ in the latter. If we measure $N$ data points $\{r_i\}$ independently from a process described by a Poisson distribution with rate parameter $\lambda$, then the likelihood is the product of $N$ Poisson distributions. Dropping terms independent of $\lambda$ we can write this as

$$P(\{r_i\}\,|\,\lambda) \;\propto\; \lambda^{N\overline{r}}\, e^{-N\lambda} \tag{5.31}$$

where $\overline{r}$ is the mean of the data. Likewise the gamma distribution prior can be written as

$$P(\lambda\,|\,k,\theta) \;\propto\; \lambda^{k-1}\, e^{-\lambda/\theta}. \tag{5.32}$$

Taking their product to get the (unnormalized) posterior, we can easily see that the result is a gamma distribution with parameters $k'$ and $\theta'$ in which

$$\begin{aligned} k' - 1 &= k - 1 + N\overline{r} \quad \text{so} \\ k' &= k + N\overline{r} \quad \text{and} \qquad (5.33) \\ \frac{1}{\theta'} &= \frac{1}{\theta} + N \quad \text{so} \\ \theta' &= \frac{\theta}{1 + N\theta}. \qquad (5.34) \end{aligned}$$

In the limit of a large amount of data, the posterior has parameters $k' \simeq N\overline{r}$ and $\theta' \simeq 1/N$, which are independent of the prior, as we would expect.

### Gaussian likelihood

Consider now a Gaussian likelihood in the variable $x$, $\mathcal{N}(\mu, \sigma)$, for which $\sigma$ is known and we want to infer $\mu$. If we adopt a Gaussian prior $\mathcal{N}(m, s)$ on $\mu$, then because the Gaussian

is invariant under an exchange of its variable with its mean, the posterior is proportional to the product of two Gaussians in $\mu$, i.e.

$$P(\mu|x,\sigma,s,m) \propto \frac{1}{\sigma s}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\exp\left[-\frac{(m-\mu)^2}{2s^2}\right]. \tag{5.35}$$

I have written the posterior conditioned on the known parameters of the prior and likelihood to emphasise that $\mu$ is the only unknown quantity. You can show in a few lines of algebra that this product is in fact another Gaussian in $\mu$, with

$$\text{mean} = \left(\frac{x}{\sigma^2}+\frac{m}{s^2}\right)\left(\frac{1}{\sigma^2}+\frac{1}{s^2}\right)^{-1} \tag{5.36}$$

$$\text{standard deviation} = \left(\frac{1}{\sigma^2}+\frac{1}{s^2}\right)^{-1/2}. \tag{5.37}$$

This looks plausible: the mean is the inverse variance weighted average of the terms in the prior and likelihood (which are treated equally). If we had a set of $N$ independent measurements $\{x\}$ drawn from this likelihood, then the expressions for the mean and standard deviation of the posterior are similar except with $x$ replaced by the mean of the data, $\overline{x}$, and $\sigma$ replaced by $\sigma/\sqrt{N}$ (see section 6.2.1). In the limit $s \to \infty$, the Gaussian prior becomes an (improper) uniform prior, and the posterior is equal to the likelihood, but now seen as a function of $\mu$ rather than $x$.

If, for the Gaussian likelihood, we instead know $\mu$ and want to infer $\sigma^2$, then the conjugate prior on $\sigma^2$ (not on $\sigma$) is the inverse gamma distribution. We shall come back to this distribution in section 6.2.3.

Conjugate priors are convenient, but they are only useful if our prior knowledge really can be represented by the corresponding distribution.

## 5.5 Summarizing distributions

The posterior PDF over a parameter is *the* answer to an inference problem. One should never forget that we always have a distribution. Nonetheless, it is convenient – and often necessary – to summarize it with a few numbers. For some problems the posterior will be a standard distribution, such as a Poisson or Gaussian, which is defined entirely by its parameters (the variance for the Poisson, the mean and variance for the Gaussian). This will occur when we use a conjugate prior for the likelihood, as we just saw. Such cases are rare in practice, unfortunately, so we must think carefully about appropriate distribution summaries.

We already looked at some summary metrics in chapters 1 and 2. The mean and variance are common choices. However, these both require that the posterior is normalizable, which may not be the case if we used an improper prior (for example). We have also seen examples where the mean is an inefficient estimator – for the uniform distribution (see section 2.5) – and an inconsistent estimator – for the Cauchy distribution (see section 2.3.2). The standard deviation also does not make sense if the distribution does not have infinite

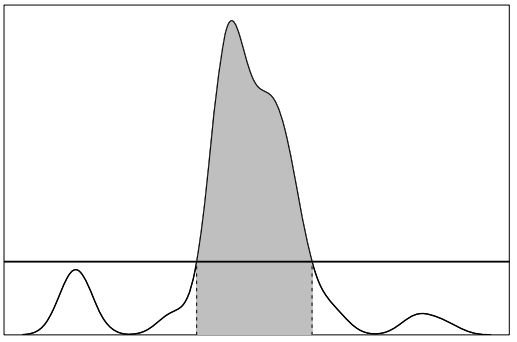

**Fig. 5.10** The highest density interval (HDI) for a specified probability $p$ is the span of the variable that encloses the region of highest probability density, the integral of which is $p$. It can be found by lowering a horizontal line (or plane in higher dimensions) which is parallel to the parameter axis until the grey region shown contains probability $p$.

support, such as when it is constrained to be positive, as was the case when estimating distances from parallaxes (see section 3.5): it is meaningless to summarize a quantity as $2 \pm 5$ if we know it cannot be negative.[13]

Considering first one-dimensional distributions, a more robust choice for the location of a distribution may be the median or the mode. The mode of a theoretical distribution can be found by differentiation, but in many cases this will not have an analytic solution, in which case we will have to use a numerical method (see section 12.6). The distribution may have more than one substantial mode, in which case we will want to report all of them (quoting a mean or median in this case would make little sense). If we have a sample of discrete data described as real variables, then each value occurs exactly once (finite numerical precision aside). To calculate the mode from samples we first need to make a density estimate, a procedure I will explain in section 7.2.

A robust choice for the scale (width) of a distribution may be the full-width at half-maximum (FWHM), although this alone does not tell us how much probability is enclosed within this width. Note that the mode and FWHM may be defined even if the probability distribution is improper. When the distribution is skew, or a quantity has a natural bound (e.g. must be positive), then we might be better off reporting quantiles of the distribution (see section 1.5). If we want to summarize such a distribution with two numbers, then the 5% and 95% quantiles (for example) may be more meaningful than the location and scale parameters. The difference between these define a 90% confidence interval. If this distribution is a posterior PDF, this tells us there is a 90% probability that the parameter lies within this interval.[14]

A $p \times 100\%$ confidence interval is not uniquely defined: there are an infinite number of

[13] Even $2 \pm 1$ makes no sense if – as is usually the case – we interpret these numbers as the mean and standard deviation of a Gaussian, because negative deviations beyond $2\sigma$ would be impossible.

[14] I am referring here to Bayesian confidence intervals, which are sometimes called credible intervals to distinguish them from frequentist confidence intervals (which we will encounter in section 10.2.3).

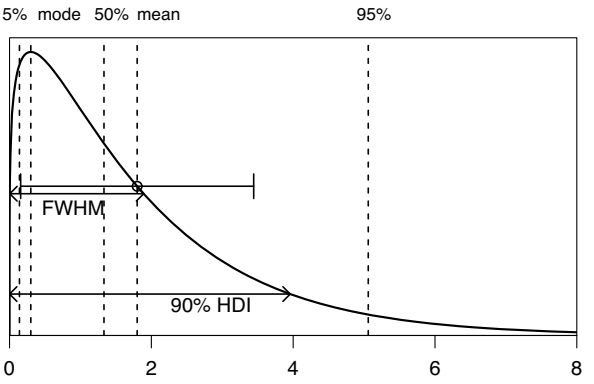


**Fig. 5.11** Various metrics for summarizing a distribution. The vertical dashed lines are labelled by the metric names, the numbers in percentages being the respective quantiles. The 5% and 95% quantiles together form the 90% equal-tailed confidence interval. The error bar shows the range of the mean $\pm$ one standard deviation. The upper double-headed arrow shows the full-width at half-maximum (FWHM), the lower one the 90% highest density interval (HDI).

choices of bounds that enclose a fraction $p$ of the probability ($0 \leq p \leq 1$). The one just mentioned is the equal-tailed interval. While intuitive and easy to compute, it may not be suitable for a skew distribution; it could even exclude the mean or mode, for example. Another common choice is the *highest density interval* (HDI), which – as the name suggests – encloses the regions of highest probability density, including the mode. If the distribution is unimodal, then it can be found by lowering a horizontal line parallel to the parameter axis until the region it defines by interception with the distribution includes the fraction $p$ of the probability (see figure 5.10). If the distribution is multimodal then we can do the same thing, but we must consider what to do if we include a second (local) maximum. For a unimodal distribution, the HDI is also the shortest confidence interval corresponding to the specified value of $p$. This follows because it includes the regions of highest density. The HDI and several other summaries of a distribution are illustrated in figure 5.11.

Summaries of multivariate distributions necessarily involve more numbers. The mean and mode are now vectors and the variance is replaced by the covariance matrix. Quantiles are no longer uniquely defined, and it is not trivial to identify the equivalent of equal-tailed intervals. The multivariate generalization of the HDI is the highest density region. It is still uniquely defined for a unimodal distribution, and it is common practice to use it to define volumes that include a fraction $p$ of the probability. For a two-dimensional distribution this corresponds to finding contours that include a fraction $p$ of the probability. Such regions could be complicated and must normally be found numerically in practice.

Which summaries of a distribution are most appropriate in any situation depends on the shape of the distribution and on our objectives. Visual inspection is always recommended, but this will be impractical if we have a lot of data sets, and will provide limited insight if the distribution is more than two-dimensional. Whenever you read summaries of a pos-

terior, always ask yourself what the full posterior PDF might look like, and thus what the limitations of the summaries might be.

How easy it is to calculate the above summaries depends on how we calculate the posterior. If we have samples *drawn from* the posterior – as opposed to densities *calculated at* particular values of it – then, as we will see in section 8.5.3, it is straightforward to calculate the mean, variance and higher moments, as well as quantiles for one-dimensional PDFs.