Chapter 2

Introduction to Bayesian statistics

This chapter provides some of the basic terminology for the rest of the module, and is written with the aim of conveying the main ideas upon which the following chapters will build. We shall introduce Bayes' rule (which you should have learned in your first or second year) in a probabilistic, non-statistical way, and thence move to Bayesian statistical inference. We conclude the chapter with a discussion of various forms of priors.

We shall look at three examples in this chapter: the use of mammography for breast cancer screening, a randomised clinical trial of a vaccine for HIV in young Thai adults, and the frequency of influenza pandemics.

2.1 Bayes' rule

Bayes' rule, or theorem, is well known and widely used by all kinds of statisticians, as well as probabilists. Its simplicity belies its power and utility. The rule states that

$$\mathrm{p}(\tilde{A} = A|\tilde{B} = B) = \frac{\mathrm{p}(\tilde{B} = B|\tilde{A} = A)\mathrm{p}(\tilde{A} = A)}{\mathrm{p}(\tilde{B} = B)} \tag{2.1}$$

or

$$\mathrm{p}(\tilde{A} = A|\tilde{B} = B) = \frac{\mathrm{p}(\tilde{B} = B|\tilde{A} = A)\mathrm{p}(\tilde{A} = A)}{\int \mathrm{p}(\tilde{B} = B|\tilde{A} = a)\mathrm{p}(\tilde{A} = a)\,\mathrm{d}a}. \tag{2.2}$$

(A note about notation: I will frequently abuse notation in this course in the aim of simplicity. The notation above uses $\tilde{A}$ for the random variable and A for the realisation, p() for both probability mass and density, and integral signs even for random variables with discrete support [you may mentally

substitute summations where appropriate]. Where the random variable is [to my mind] obvious, I will often omit it, so that, say, $\mathrm{p}(A)$ is to be taken to mean $\mathrm{p}(\tilde{A} = A)$.)

Although the equation appears innocous, it is, in reality, profound, for it provides a way to invert conditional probabilities. Because *all* probabilities are, actually, conditional (though the conditions are sometimes supressed, the cause of much confusion), Bayes' rule provides a way to manipulate probabilities and move from one to another, in the above, from $\mathrm{p}(B|A)$ to $\mathrm{p}(A|B)$. We shall see why this is important in the following example.

2.1.1 Example: breast cancer screening in Germany

Gerd Gigerenzer is a psychologist of medicine, who heads the prestigious Max Planck Institute for Human Development in Germany. In his quite wonderful book, Understanding Uncertainty (ref. [4], which should be required reading for all statistics students), he describes an experiment he performed on German physicians. The 24 participating physicians were all volunteers, and had a variety of experience levels and backgrounds: some were newly qualified, others headed practices or departments, some worked in teaching hospitals, others in private practice. They were given the following information and task (paraphrased from memory):

> Imagine you are using mammography to screen asymptomatic women for cancer. The probability that an asymptomatic woman in her 50s in your area attending regular screening has breast cancer is 0.8%. If she actually has breast cancer, the probability is 90% that her mammogram will be positive. If she does not have breast cancer, the probability is 7% that she will, nevertheless, test postive. A woman sits in front of you with a positive mammogram. What is the probability that she actually has breast cancer?

Gigerenzer describes how some physicians were nervous, or agitated, in answering this question; others stormed away without answering it, some requested help from their children. The answers provided straddled almost two orders of magnitude: four thought the probability was around 1%, eight that it was about 90%. Only two got the correct answer—that there was about a 10% chance the woman actually had breast cancer. Let us use Bayes' theorem to show that this is the correct probability.

Let $B = 1$ if the woman has **breast** cancer and 0 otherwise. Let $M = 1$ if her **mammogram** is positive and 0 if not. Let $A = 1$ if the woman is

asymptomatic, lives in the **area**, is **aged** in her 50s, and is **attending** for regular screening, and 0 if not. The information Gigerenzer gave the participants is thus:

$$\begin{aligned} \mathrm{p}(B=1|A=1) &= 0.008, && (2.3)\\ \mathrm{p}(M=1|B=1,A=1) &= 0.9, && (2.4)\\ \mathrm{p}(M=1|B=0,A=1) &= 0.07. && (2.5) \end{aligned}$$

Note that the information he gave relates to women satisfying certain conditions, which is why for all equations we have conditioned on $A = 1$. We want the probability of breast cancer given the information we already had (i.e. $A = 1$) *and* the new information that the mammogram reveals (i.e. $M = 1$). Using Bayes' theorem we have:

$$\mathrm{p}(B=1|M=1,A=1) = \frac{\mathrm{p}(M=1|B=1,A=1)\mathrm{p}(B=1|A=1)}{\mathrm{p}(M=1|A=1)}. \quad (2.6)$$

The numerator follows directly from the information provided. Obtaining the denominator requires (ever so slightly) more work:

$$\mathrm{p}(M=1|A=1) = \sum_{b=0}^{1} \mathrm{p}(M=1|B=b,A=1)\mathrm{p}(B=b|A=1). \quad (2.7)$$

Using the numbers given in the problem statement gives

$$\mathrm{p}(B=1|M=1,A=1) = \frac{0.9\times 0.008}{0.9\times 0.008 + 0.07\times 0.992} = 9.4\%. \quad (2.8)$$

The very high false positive rate of mammography is one reason it has many detractors (see for example the references in [4]).

Note that there are *two* probabilities of having breast cancer in the equation above: a probability based on the risk group A only, i.e. *prior* to obtaining the addition information from the mammogram, $\mathrm{p}(B = 1|A = 1)$; and a probability that accounts additionally for the state of knowledge *after* obtaining this information, $\mathrm{p}(B = 1|M = 1, A = 1)$. These are called prior and posterior probabilities, and Bayes' rule provides the calculus by which the prior probability is modified by the new evidence, $\mathrm{p}(M = 1|B = 1, A = 1)$, to obtain the posterior. You will realise that even the prior probability is in fact a posterior in the sense that it represents the information after A is observed.

Note also that this analysis makes a strong (if justified) assumption: that these probabilities all apply to the individual woman sitting in front of you,

the physician. Let's give her a name, Marianne, and some other details: she's divorced, has two daughters, drives a red Skoda, plays tennis. The probability of 0.8% is, presumably, derived from the overall prevalence in her age group. Do you think it is appropriate to apply that probability to Marianne? Some people would argue, yes, without additional relevant information to condition on, this is appropriate (note that having had two children does modify her risk of breast cancer, however). Others would argue that it is inappropriate unless she had somehow been randomly selected (equiprobably) from the population of women with condition $A = 1$ before coming for screening. If other relevant information were provided (e.g. it would be sensible for the physician to ask if there is a family history of breast cancer, or to check for a palpable lump), this additional information should really be incorporated in the probability calculation by conditioning on it.

Finally, observe how Bayes' rule allows you to flip the conditioning round, from the probability of something known given something unknown (the test result given the underlying cancer status) to the probability of something unknown and uncertain (cancer status) given the knowledge that you have (the test result).

2.1.2 Historical background

Bayes' rule was developed by an English minister and amateur mathematician, Thomas Bayes. It was published [5] in 1763, some time after his death, by a self-declared friend, Richard Price, who appears to have edited the manuscript heavily, and heavy handedly. The paper is hard to read due to the archaic prose, but rewards the careful reader: confidence intervals are imagined about 200 years before Neyman and Pearson (*fils*) described them, as is the beta conjugate prior for the binomial, described later in this chapter.

2.2 Inference for a probability

The breast cancer example above illustrates a fairly non-controversial use of Bayes' theorem. Almost all statisticians would be happy with the calculations provided and the interpretation of the final result (exceptions make the world a more interesting place). However, traditionally the statistical world was split into two camps according to how else this formula could be used. One camp, which I will usually call *frequentists*, though they are also called *classical* statisticians, argued that Bayes' rule could only be used to make probabilistic statements about *events*. The other group, *Bayesians*, argued that Bayes' rule could be used to make probabilistic statements about *any-*

thing, including unknowns such as *parameters*. (Nowadays, the culture wars between frequentists and Bayesians are much rarer as most statisticians are more pragmatic than once they were.) Let us see how Bayesians might tackle one of the simplest, and most important, tasks that a statistician may face: estimating a proportion from a binomial sample.

Let us enliven the task by using a concrete example. Rerks-Ngarm et al [6] describe interim results from a high profile vaccine trial carried out in Thailand that hit the main media sources worldwide. The trial had a good, standard design: eligible participants were randomly split into two arms, one given the putative vaccine, the other a placebo, and they were followed up for 3.5 years, at which point they were invited back to take an HIV test. The endpoint is HIV conversion. There were some complications due to drop out and the decision of which analysis to use, and I will present the simplest one (the data used in the modified intention-to-treat analysis). Consider (for now) only the vaccine arm. Of 8197 eligible participants, 51 were infected during the trial period (so straight away we know the vaccine is not perfect). The question we will answer is, what is the "underlying" probability, p_v, of infection over this time window for those vaccinated?

Before we answer that, let us ensure we know what this probability represents. Participants in trials are, of course, not randomly selected from a population and forced to participate. They are referred to the trial, or volunteer, and must meet eligibility requirements (not being pregnant, for example, in the Thai trail) and must give written informed consent. This means they are not really representative of the population to which they belong. So p_v doesn't represent the probability of infection among vaccinees in the whole population. And the risk of infection is different in, say, Thailand and Singapore, so p_v doesn't represent the probability of infection in trial participants in other countries. It represents something more metaphysical/airy-fairy, the probability of infection in a hypothetical second trial in the same participants. Usually the airy-fairiness of this is brushed off by hoping that the ratio of p_v to p_u, the proportion infected in unvaccinated participants, generalises to the population as a whole, though there is no deductive reasoning that leads to that conclusion. But let us assume that p_v represents something of interest and try to estimate it nevertheless.

It seems appropriate to assume that $X_v \sim \text{Bin}(N_v, p_v)$, where X_v is the number of vaccinees infected (i.e. 51) and N_v the number of vaccinees (i.e. 8197). Theoretically this would be inappropriate if the infection of one participant influenced the risk of infection of another (if they were sexual partners or shared needles to inject drugs, for example) but the effect of such non-independence is probably small since the number of participants is small relative to the population as a whole. The probability (mass) function for

the data is then

$$\mathrm{p}(X_v = 51 | N_v = 8197, p_v) \propto p_v^{X_v}(1-p_v)^{N_v - X_v}, \tag{2.9}$$

(being a bit loose with notation by omitting the $\tilde{p}_v$).

We would like a point estimate to summarise the data, an interval estimate to summarise uncertainty, and possibly (later) a measure of the evidence that the vaccine was effective. You should know how to do all these already, but it may be useful to phrase the solution in our notation.

2.2.1 The frequentist approach: refresher

The traditional approach to estimating p_v is to find the value of p_v that maximises the likelihood of the data given that hypothetical value were the true value. This maximum likelihood estimate can be obtained using calculus (for very simple problems like this one) or numerically (using Newton–Raphson, simulated annealing, cross entropy, etc.). Since we can do the calculus quite easily here, let us do that. Differentiation for many problems is simplified by working with the log-likelihood (which, more importantly, dramatically minimises overflow issues in numerical methods, see later),

$$\log[\mathrm{p}(X_v | N_v, p_v)] = c_1 + X_v \log p_v + (N_v - X_v)\log(1-p_v). \tag{2.10}$$

Differentiating with respect to the argument we wish to maximise over,

$$\frac{\mathrm{d}\log[\mathrm{p}(X_v | N_v, p_v)]}{\mathrm{d}p_v} = \frac{X_v}{p_v} - \frac{N_v - X_v}{1-p_v}, \tag{2.11}$$

replacing p_v by $\hat{p}_v$ on setting the derivative equal to 0, and solving, gives $p_v = X_v/N_v = 51/8197 = 0.62\%$ which, reassuring to non-statisticians, is just the empirical proportion infected.

The method you probably know best to quantify the uncertainty in this estimate is to derive a 95% (confidence) interval using the standard error of $\hat{p}_v$,

$$SE(\hat{p}_v) = \sqrt{\frac{\hat{p}_v(1-\hat{p}_v)}{N_v}}. \tag{2.12}$$

Note that in deriving this, you have to cheat, by assuming $p_v = \hat{p}_v$, and wave your hands around, by assuming that the sample size is large enough to warrant using asymptotics in an actual application. Here, this is okay, because the sample size is actually pretty big, so both assumptions are reasonable. In small samples, though, uncertainty intervals derived from this standard error can be very misleading [7].

Anyway, the standard error for these data is 0.09%, so a 95% interval can be derived as $0.62\% \pm 1.96 \times 0.09\%$, i.e. (0.45,0.79)%. The likelihood, maximum likelihood estimate, and 95% interval are presented in figure 2.1.

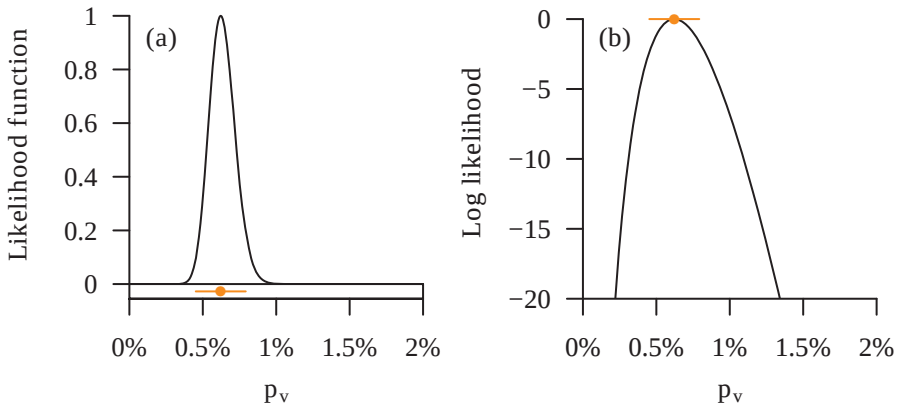

Figure 2.1: **Likelihood (a) and log-likelihood (b) functions for Thai HIV vaccine trial.** The maximum likelihood estimate is indicated by the dot, along with a 95% confidence interval based on the asymptotically Gaussian distribution of the maximum likelihood estimate. If you look closely enough, you might notice that the curve is slightly asymmetric (more readily observed on a log scale)—that it is only a little out means that asymptotic arguments used to derive the uncertainty interval are reasonable.

Let us briefly dwell upon the interpretation of these two quantities. The maximum likelihood estimate of p_v is *not* the most likely value of p_v, no matter how much it sounds like it should be. It is the value of p_v that makes the data most likely. Classical statisticians are not allowed to make probabilistic statements about parameters, so if you shared the notion that the maximum likelihood estimate was the most probable one, please disabuse yourself of it. Second, for similar reasons, there is not a 95% probability that p_v lies within the interval *blah*, because, to a classical statistician, it does not make sense to talk about the probability a parameter lies between two numbers (unless you mean the probability is 0 or 1). Rather, 95% of such intervals you make over your lifetime—in situations wherein the model is appropriate, there is no systematic error, and the sample size is "large"—will contain the true value, but, for the current interval, you cannot claim any probability. In other words, do not interpret either quantity the way you probably find most intuitive. At least, in frequentist statistics.

2.2.2 Tackling it Bayesianly

How does the Bayesian derive point and interval estimates for this problem? Bayes' rule is used to describe the probability density of the parameter, p_v, given the data, X_v and N_v, in exactly the same way as it would be to describe the probability of cancer given a positive test, i.e. the Bayesian is interested in

$$\mathrm{p}(p_v|X_v, N_v) = \frac{\mathrm{p}(X_v|p_v, N_v)\mathrm{p}(p_v|N_v)}{\int_0^1 \mathrm{p}(X_v|\pi, N_v)\mathrm{p}(\pi|N_v)\,\mathrm{d}\pi} \tag{2.13}$$

over the range of p_v, which is $[0, 1]$. The term $\mathrm{p}(p_v|X_v, N_v)$ is the posterior for p_v, and should be seen as a function of p_v. On the numerator are $\mathrm{p}(X_v|p_v, N_v)$, the likelihood function, and $\mathrm{p}(p_v|N_v)$, the prior for p_v. The denominator includes a dummy variable π in place of p_v. Note that the likelihood function here is just the regular likelihood function introduced by Fisher in the 1920s and described above, i.e. it is $\mathrm{p}(X_v|p_v, N_v) \propto p_v^{X_v}(1 - p_v)^{N_v - X_v}$. What is the prior distribution? Well, there is no *the* prior distribution: just as you choose one of many possible models for the data, must you choose one of many possible models for the parameters. The prior represents the information on the parameter—the proportion of vaccinees who would be infected by HIV during the trial—before any data are observed. One justifiable approach would be to assume *all* probabilities on $[0, 1]$ are equally likely until the trial is performed, which can be written as a formula thus

$$\mathrm{p}(p_v|N_v) = \mathrm{p}(p_v) = \mathbf{1}\{p_v \in [0, 1]\} \tag{2.14}$$

where $\mathbf{1}\{A\}$ is the indicator function equal to 1 if A is true and 0 otherwise. This is equivalent to the formulæ

$$p_v \sim \mathrm{U}(0, 1) \text{ or} \tag{2.15}$$
$$p_v \sim \mathrm{Be}(1, 1). \tag{2.16}$$

Note that we drop N_v from the condition because this prior is under the assumption that the sample size and probability of infection are independent.

The posterior under this model for data and parameter is thus

$$\mathrm{p}(p_v|X_v, N_v) = C p_v^{X_v}(1 - p_v)^{N_v - X_v} \tag{2.17}$$

over the range $[0, 1]$, where C is some constant. There are two ways to proceed: a smart way, which I'll describe later in the chapter, and a dumb way, which we'll use now. The latter involves taking a grid of values for p_v, spaced close to each other, evaluating the function $f_1(p_v) = p_v^{X_v}(1-p_v)^{N_v - X_v}$, approximating the integral by the sum of f over this grid times the spacing

between successive grid values, and exploiting the fact that the integral of $\mathrm{p}(p_v|X_v, N_v)$ is unity (because it is a probability density) to approximate C. Using `R` code, this can be implemented as follows.

```
pv    = seq(0.00001,0.05,0.00001)
xv    = 51; nv = 8197
logf1 = xv*log(pv) + (nv-xv)*log(1-pv)
f2    = exp(logf1-max(logf1))
intf2 = sum(f2)*(pv[2]-pv[1])
post  = f2/intf2
```

This is plotted in figure 2.2. Note that, of course, the posterior may take values more than one.

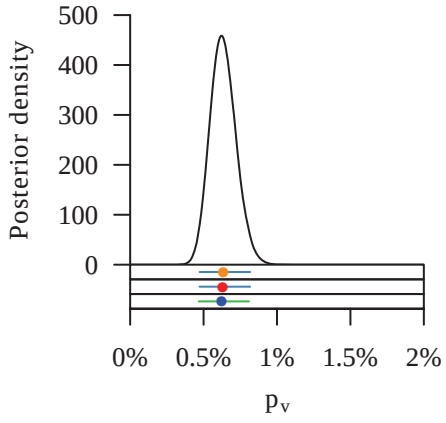

Figure 2.2: **Posterior density of the probability of infection among vaccinees in the Thai HIV vaccine trial.** The posterior mean is represented by an orange dot, the posterior median by a red dot, and the posterior mode by a blue dot; equal-tailed 95% intervals are represented as (two) light blue lines and the highest posterior density 95% interval by the green line. If you look closely enough, you might notice that the curve is slightly asymmetric; it is not assumed to be symmetric.

2.2.3 Point and interval estimates

The previous subsection shows how a posterior distribution for p_v might be derived. From this, it is easy to derive a point or interval estimate. There are several choices of point estimate to report for distributions of parameters, just as there are several that could be reported for distributions of data, and the choice of which to report is made similarly.

- The posterior mean, $E(p_v|X_v, N_v)$, is the most usual choice, and is readily calculated from most numerical methods to obtain the posterior.
- The posterior median is p_v such that $\int_{-\infty}^{p_v} \mathrm{p}(\tilde{p}_v = \pi|X_v, N_v) = 1/2$. It is more representative of the distribution than the mean is, if the distribution is skewed.
- The posterior mode is $p_v = \arg\max \mathrm{p}(p_v|X_v, N_v)$. It represents the most likely value of the parameter (contrast to the definition of the maximum likelihood estimate above).

To obtain these three estimates using R and following on from the code above:

```
pmean = sum(pv*post)/sum(post)
pcdf  = cumsum(post)/sum(post)
pmedian = 0.5*(max(pv[pcdf<0.5])+min(pv[pcdf>0.5]))
pmode = pv[which.max(post)]
```

For the Thai HIV trial example, these estimates are 0.63%, 0.63% and 0.62%, respectively. Note how close they are to each other and to the maximum likelihood estimate.

To obtain an uncertainty interval (usually called a credible interval, but I shall call them just "intervals" for short) there are two frequently used approaches. One is to use the quantiles of the posterior distribution; for a 95% interval these would usually be the 2.5%ile and the 97.5%ile, though any interval spanning 95% could be taken. Because the posterior distribution represents the distribution of the parameter after accounting for the data, the interpretation of this equal-tailed interval is that there is a 95% chance that the parameter lies in the interval (again, compare to the definition of the frequentist confidence interval above). Note that because the posterior distribution is not necessarily symmetric, there may be some parameter values that are outside the interval that have higher probability than some parameter values inside the interval. If this is a concern, a *highest posterior density* interval can be obtained, with a little more work. To do this, start with the abscissa of the posterior mode, gradually reduce the abscissa, including any parameter with posterior higher than this in the interval, until the interval first spans 95% (or whatever coverage is sought).

To obtain these interval estimates in R, you may do the following:

```
CI1 = c(max(pv[pcdf<0.025]),min(pv[pcdf>0.975]))
threshold = max(post)
coverage = 0
```

```
for(i in seq(0.999,0.001,-0.001))
{
  threshold = i*max(post)
  within    = which(post>=threshold)
  coverage  = pcdf[max(within)]-pcdf[min(within)]
  if(coverage>=0.95)break()
}
CI2 = pv[range(within)]
```

These two interval estimates are, respectively, $(0.47, 0.82)\%$ and $(0.47, 0.81)\%$. Not only are the close to each other, but they are close to the classical 95% confidence interval. This illustrates one very important point, namely that *in some situations* it really doesn't matter whether you use a classical or Bayesian approach, because both will give effectively the same answers. The situations in which the classical estimate is as good as the Bayesian are those when (i) the sample size is large, so classical results which rely on asymptotics are a good approximation to the actual sampling distribution of statistics such as estimators, (ii) there is no real prior information to be incorporated in the analysis, and (iii) when someone has already developed a classical numerical routine that you can use. Bayesian methods come into their own when any of these conditions are not met.

A further point worth making, is that the close correspondence between classical and Bayesian estimates in situations in which both can be used means that one can be used as an approximation to the other. If your philosophical proclivities are not Bayesianly-aligned, you can still use Bayesian methods as if they were classical and interpret them thus. More relevant for this course, you can use classical estimates from the literature as if they were Bayesian in deriving prior distributions, and can arguably interpret classical point and interval estimates the way you wish to, as if they were Bayesian.

2.3 Prior distributions

We have heretofore not directly addressed the question of what actually is a prior distribution and how is one chosen? Under the Bayesian paradigm, one represents uncertainty by a probability distribution. What is the probability a woman has cancer given she has a positive mammogram? What is the distribution of risk of getting infected by HIV given that we observed 51 infections out of 8197 participants? The posterior distribution encapsulates the residual uncertainty after analysing your data set. The prior represents the uncertainty before.

Just as you, the statistician analysing a data set, must choose a model for the data in terms of unknown parameters, so too must you, the Bayesian statistician analysing the data set, choose a model for the parameters of that model. So you may think of the likelihood as representing the model for the data and the prior the model for the parameters. Just as you must justify the former, so must you justify the choice of prior.

The prior distribution should have correct support for the parameters it represents. If p is a probability then taking a normal distribution as the prior for p would be silly, for it would give support to $p > 1$ or < 0. Similarly, a regression coefficient β can take values on the real line, so an exponential prior that forced β to be positive might not be considered appropriate.

For models with multiple parameters, a joint prior distribution must be specified. This often assumes that the joint prior is the product of one distribution for each parameter, e.g. $\mathrm{p}(a, b) = \mathrm{p}(a)\mathrm{p}(b)$, i.e. the parameters are *a priori* independent. However, priors need not be taken to be independent, and there are some fairly common situations in which they are not: for instance, if a posterior distribution from dataset 1 is used as a prior for dataset 2, then any correlations in the first posterior should be accounted for in the second prior.

2.3.1 Informative and non-informative priors

An informative prior for a parameter (or parameters) encapsulates information beyond that solely available in the data directly at hand. For instance, perhaps someone has estimated the same parameter in a previous study and reported a point and interval estimate for it in the literature: you might take these and use them to create a normal distribution with an appropriate mean and variance to act as a prior for your study. We will see an example of this early in the next chapter.

A non-informative prior is the opposite: it is a distribution that is either flat or almost flat on the part of the parameter space with reasonably high likelihood. For example, the uniform prior for p_v on the range $[0, 1]$ we used in the Thai vaccine trial is non-informative because over the range around 0.5–1.5%, it is flat (as it is all over the range). For a parameter that has support on the real line , a uniform distribution on $[-1\,000\,000, 1\,000\,000]$, or a $\mathrm{N}(0, 1000^2)$, *might* (depending on the data and model) be effectively flat over the effective parameter space. See figure 2.3 for a depiction.

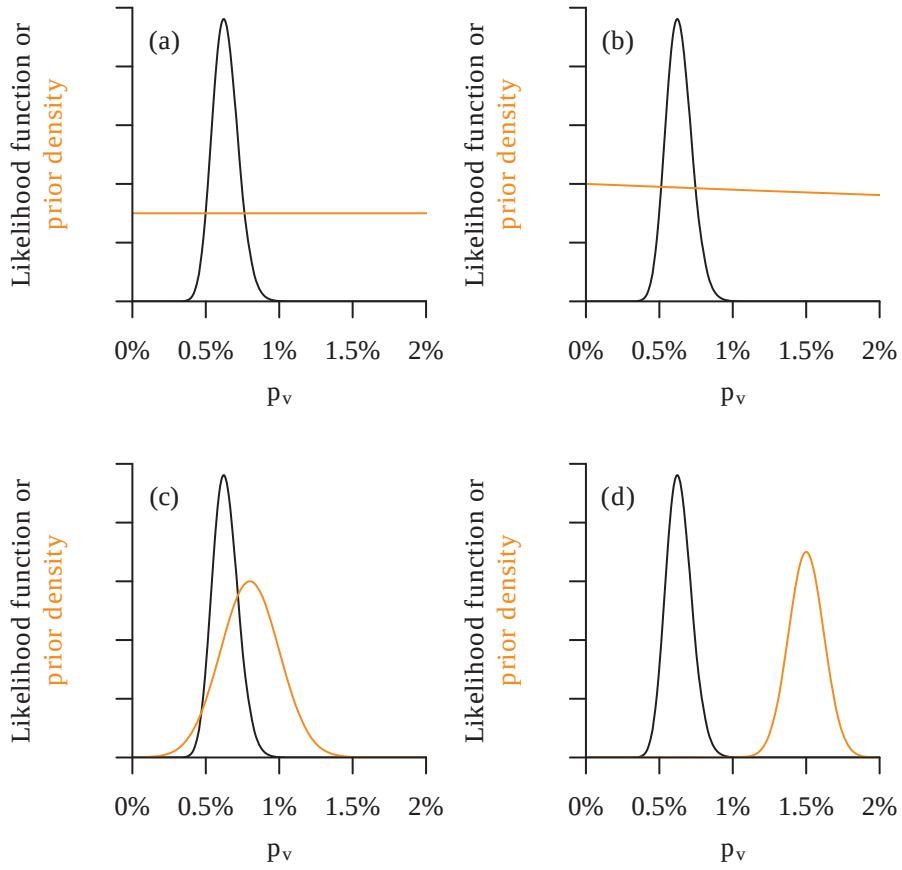

Figure 2.3: **Examples of non-informative (a and b) and informative prior distributions (c and d).** Black lines represent likelihood functions, orange lines the prior density. (a) a uniform prior. (b) an exponential prior with small rate relative to the spread of values supported by the likelihood. (c) a normal distribution, consistent with the data. (d) a normal distribution that clashes with the information from the data.

2.3.2 Proper and improper priors

Recall that a distribution has integral one, and prior distributions really ought to, too. A proper prior distribution is one that integrate to one, $\int_\Theta \mathrm{p}(\theta)\,\mathrm{d}\theta = 1$, where θ is the parameter and Θ its support. Sometimes, you might choose to use an improper prior, i.e. one in which $\int_\Theta \mathrm{p}(\theta)\,\mathrm{d}\theta$ is not finite. Examples include assuming $\theta \sim \mathrm{U}(-\infty, \infty)$, i.e. $\mathrm{p}(\theta) \propto 1$; $\theta \sim \mathrm{U}(0, \infty)$, i.e. $\mathrm{p}(\theta) \propto \mathbf{1}\{\theta > 0\}$; or $\theta \propto \mathbf{1}\{\theta > 0\}/\theta$.

Posteriors, too, can be proper or improper. Improper posteriors are a problem. Improper priors are not necessarily bad, though, for depending on the data and model, some improper priors lead to fully proper posteriors. Proper priors beget proper posteriors; improper priors may beget proper or improper posteriors. An improper posterior may result if the likelihood is badly behaved, for example, by having an asymptotic non-zero value as a parameter tends to infinity. In such situations, though, not only is a proper prior called for, an informative prior is vital to substitute for the lack of information in the dataset.

2.3.3 Conjugate priors

A final kind of prior, always proper, but which can be either informative or non-informative, is the conjugate prior. If a parameter for a model belongs to a specific family *a priori* and the same family *a posteriori*, then that family is said to be conjugate to the model and the prior a conjugate prior.

Let us use the Thai HIV trial as an example again. Above we chose to take a uniform prior for p_v, and I wrote this as a $\mathrm{Be}(1, 1)$, which may have seemed odd. The posterior we found had form

$$\mathrm{p}(p_v | X_v, N_v) \propto p_v^{X_v}(1 - p_v)^{N_v - X_v}. \tag{2.18}$$

Those of you who are particularly *au fait* with your distributions might have recognised this as proportional to a $\mathrm{Be}(1 + X_v, 1 + N_v - X_v)$ distribution. (If you did, congratulations. I didn't the first time.) But the posterior is a density so if it's proportional to a particular beta distribution it must *equal* that particular beta distribution (as both have integral one). So here we moved from a prior in which p_v was beta to a posterior in which p_v is beta. Ergo the beta distribution is conjugate to the binomial.

In general, if $p \sim \mathrm{Be}(\alpha, \beta)$, $X \sim \mathrm{Bin}(N, p)$ and $Y = N - X$, then the

posterior is:

$$\begin{aligned} \mathrm{p}(p|X,N) &\propto p^{\alpha-1}(1-p)^{\beta-1}p^X(1-p)^Y && (2.19)\\ &= p^{X+\alpha-1}(1-p)^{Y+\beta-1} && (2.20)\\ &\propto \frac{\Gamma(\alpha+X+\beta+Y)}{\Gamma(\alpha+X)\Gamma(\beta+Y)}p^{X+\alpha-1}(1-p)^{Y+\beta-1} && (2.21) \end{aligned}$$

where $\Gamma()$ is the gamma function. Thus $p|X,N \sim \mathrm{Be}(\alpha+X,\beta+Y)$.

A few other models have conjugate priors available to them and we shall encounter some as we move through the course. Most real problems do not have a suitable conjugate prior however. In situations in which a conjugate prior exists, it can be useful to exploit it, as analysis can be more analytic. For the HIV trial example, once we realise that the posterior for p_v is beta, we can calculate posterior properties directly. For example:

```
pmean = (1+xv)/(1+nv-xv)
pmedian = qbeta(0.5,1+xv,1+nv-xv)
pmode = xv/nv
CI1 = qbeta(c(.025,.975),1+xv,1+nv-xv)
```

2.4 Sequential updating

Imagine an experimentor (experimentor 1) trying to estimate a probability p given a series of Bernoulli trials, $x_i \in \{0,1\}$ for $i = 1,2,\ldots$, with $y_i = \sum_{j=1}^{i} x_j$. She has assigned a $\mathrm{Be}(\alpha,\beta)$ prior to p. Rather than wait for all the data to come in, every time a new data point arrives, she recalculates the posterior from scratch based on the information available thus far. Exploiting conjugacy, for an hypothetical dataset, her posterior would change as in table 2.1.

Now imagine a second experimentor (experimentor 2) observing the same data, but taking the posterior from the last iteration to be the prior and the data to be only the new data point. His posterior would change as in table 2.2.

As you can see, the posteriors our two experimentors end up with at each stage are the same, despite the (marginally) different calculations. Generally, if a dataset comes to you piecemeal, it does not matter whether you analyse the data once at the end, or sequentially update your prior as you proceed (in principle, though in practice you may find it convenient to do one or the other: for example, if the posterior early in the data collection is not well approximated by an analytic form (e.g. multivariate normal) that can

Table 2.1: Experimentor 1's analysis.

i	x_i	y_i	Prior	Posterior	Data used
1	0	0	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha, \beta + 1)$	0 / 1
2	1	1	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + 1, \beta + 1)$	1 / 2
3	1	2	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + 2, \beta + 1)$	2 / 3
4	0	2	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + 2, \beta + 2)$	2 / 4
5	0	2	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + 2, \beta + 3)$	2 / 5
6	1	3	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + 3, \beta + 3)$	3 / 6
7	0	3	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + 3, \beta + 4)$	3 / 7
8	1	4	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + 4, \beta + 4)$	4 / 8
i	x_i	y_i	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha + y_i, \beta + n_i - y_i)$	y_i / i

Table 2.2: Experimentor 2's analysis.

i	x_i	y_i	Prior	Posterior	Data used
1	0	0	$\text{Be}(\alpha, \beta)$	$\text{Be}(\alpha, \beta + 1)$	0 / 1
2	1	1	$\text{Be}(\alpha, \beta + 1)$	$\text{Be}(\alpha + 1, \beta + 1)$	1 / 1
3	1	2	$\text{Be}(\alpha + 1, \beta + 1)$	$\text{Be}(\alpha + 2, \beta + 1)$	1 / 1
4	0	2	$\text{Be}(\alpha + 2, \beta + 1)$	$\text{Be}(\alpha + 2, \beta + 2)$	0 / 1
5	0	2	$\text{Be}(\alpha + 2, \beta + 2)$	$\text{Be}(\alpha + 2, \beta + 3)$	0 / 1
6	1	3	$\text{Be}(\alpha + 2, \beta + 3)$	$\text{Be}(\alpha + 3, \beta + 3)$	1 / 1
7	0	3	$\text{Be}(\alpha + 3, \beta + 3)$	$\text{Be}(\alpha + 3, \beta + 4)$	0 / 1
8	1	4	$\text{Be}(\alpha + 3, \beta + 4)$	$\text{Be}(\alpha + 4, \beta + 4)$	1 / 1
i	x_i	y_i	$\text{Be}(\alpha + y_{i-1}, \beta + n_{i-1} - y_{i-1})$	$\text{Be}(\alpha + y_i, \beta + n_i - y_i)$	x_i / 1

be used in subsequent calculations, it is better to analyse the data all in one go.

The idea of sequential updating means you can always treat an existing posterior as a prior for the next analysis.

2.5 Example: timing of influenza pandemics

Let us consider one final example in this chapter. Influenza is a common infection caused by various influenza viruses. Infection leads to a spectrum of disease expressions, from no symptoms at all, to symptoms that are not differentiable from other common respiratory viruses, such as a sore throat, fatigue, or runny nose, to febrile symptoms that are the traditional hallmark of influenza infection, to severe infections that can lead to hospitalisation, pneumonia, or death. Multiple infection over one's lifetime is possible because influenza viruses in humans have a high mutation rate that allow them to evade the host's immune response after sufficient time. In addition, at a frequency of about three times a century, a substantially different influenza virus may evolve, to which very few individuals have any immunity; this causes a large influenza pandemic. The influenza pandemic of 1918–20 is thought to have killed between 20 and 100 million people worldwide (which, for reference, is more than died in the entire first world war, the end of which it precipitated).

Table 2.3: Years (CE) of influenza pandemics, as provided by REF Potter 2001 (C. W. Potter. A history of influenza. J. Appl. Microbiol., 91:5729, 2001.), plus the 2009 pandemic.

1729	1781	1830	1898	1918	1957	1968	2009

The emergence of influenza pandemics might be appropriately modelled as a Poisson process (if they are memoryless—which might not be the case). If event times follow a Poisson process, then the hazard rate between events is constant, the number of events within a given time interval is Poisson, and the time between events is exponential. If pandemics emerged as a Poisson process with rate $\lambda = 1/\mu$ then the likelihood function is given by:

$$f(\mathbf{t}|\lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda(t_i - t_{i-1})\} \times \exp\{-\lambda(T - t_n)\} \tag{2.22}$$

where we might justifiably set t_0 to be 1700, and T is the present (at the time of writing, $T = 2012$). We might choose a uniform prior for μ, the average time between outbreaks, in the absence of additional external data. This model we can fit by considering a grid for μ over a plausible range using the following R code:

```
loglikelihood=function(mu)
{
  tpan = c(1729,1781,1830,1898,1918,1957,1968,2009)
  tnow = 2012
  tstart = 1700
  tdeltas = c(tpan,tnow)-c(tstart,tpan)
  output = sum(dexp(tdeltas[1:8],rate=1/mu,log=TRUE))
         + pexp(tdeltas[9],rate=1/mu,log=TRUE,lower.tail=FALSE)
  output
}
mu = 1:100
logposterior = 0*mu
for(i in 1:length(mu))
   logposterior[i] = loglikelihood(mu[i])
                   + dunif(mu[i],0,1000)
logposterior = logposterior-max(logposterior)
posterior = exp(logposterior)
posterior = posterior/(sum(posterior)*(mu[2]-mu[1]))
```

A plot of the posterior is presented in figure 2.4.

The posterior mean average inter-pandemic interval is 52y (95%I 23–110y), in other words, roughly one pandemic per two human generations. Note that the distribution is asymmetric as a result of the small sample size, and so an asymptotic classical confidence interval would be inappropriate.

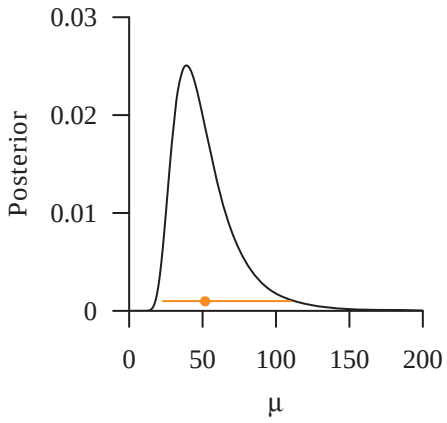

Figure 2.4: **Posterior distribution of mean inter-pandemic interval, Poisson process model.** The black line represents the posterior density, while the orange line and point represent a 95% interval and posterior mean respectively.