Part 4 Section 1: Bayesian Inference – A Powerful Idea

Danail Obreschkow

2025-08-29

Prerequisites

Load libraries and set RNG seed for this chapter:

```
library(magicaxis)
library(pracma)
library(foreach)
set.seed(1)
```

Foreword

This part of the course gets to heart of modern statistics. It deals with the topic of statistical *inference*, which, in the most general sense, is the art of finding a theoretical model describing some empirical observations. More specifically, statistical inference is a mathematical way to distinguish better from worse models and to determine the free parameters of a model, given some data.

In general, statistical inference is an *inductive* process, since it relies on a *finite* sample of observations to determine a model that can normally describe *infinitely* many cases. It is therefore unavoidable to make some axiomatic assumptions to constrain the problem enough that deductive logic can be applied.

A good example for the inductive nature of inference is the search for physical laws: all known physical laws rely on a finite set of observations, yet apply to a *continuity* of situations. For instance, Johannes Kepler proposed his revolutionary laws of planetary motion based on a number of discrete position measurements of Mars, using only very limited additional data from the other five ancient planets (including Earth), collected by Tycho Brahe over a couple of decades. Despite this finite set of data, Kepler's laws are formulated, as if they applied to *any* conceivable planet in the solar system at *any* time in the past and future. This is, of course, a leap that defies strict logic: in principle, the laws of physics could suddenly change any day or behave differently for not-yet-discovered planets. Similarly, Brahe's data were not good enough to distinguish between $R^3 \propto T^2$ (Kepler's third law) and $R^3 \propto T^{1.99999}$. These considerations make it clear that the general character and exact formulation of physical laws do *not* fully follow from the data, but require *metaphysical* principles, such as the idea that the Universe obeys some *simplicity* (see Occam's razor), that it is built upon special *symmetries* (e.g. homogeneity and isotropy of space-time) and that it follows a mathematical *beauty* (e.g. integer exponents in Kepler's case).

The example above illustrates that statistical inference can rarely determine the 'true' model behind some data in a fully deductive way. A powerful mathematical framework for inference should, however, be able to distinguish between models that make different predictions regarding the actually observed data; and it should be able to account for prior knowledge about the 'true' model. The mathematical framework that most directly combines these two features is the 'Bayesian' framework, generally attributed to Thomas Bayes and Pierre-Simon Laplace. This framework allows us to assign *probabilities* to models and model parameters, given the data. Incidentally, Bayesian inference explains closely how humans and presumably most animals learn about their social and physical environment and interpret the world.

This part of the course is entirely dedicated to the Bayesian framework and associated tools, such as maximum likelihood estimation, maximum posterior estimation, and Laplace approximation.

Bayes' theorem

The core idea of the Bayesian framework is to infer a model M (or testing a hypothesis H) from a data set D (also known as evidence E) by identifying the model and data with A and B, respectively, in Bayes' theorem, $P(A|B)P(B) = P(B|A)P(A)$ (see part 3). In this small step lies the enormous power of modern statistical inference!

We can thus write

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \tag{1}$$

where

- $P(D|M)$ is the conditional probability of the data given the model, known as the *likelihood*;
- $P(M)$ is the *prior probability* (often just called the *prior*), encoding our knowledge about the model without considering the data;
- $P(M|D)$ is the *posterior probability* (often just called the *posterior*), i.e. the probability of the model, accounting for the data *and* prior knowledge;
- $P(D)$ is called the *marginal likelihood*, because it is obtained by integrating ('marginalising') $P(D|M)$ over all models.

Since $P(D)$ does not depend on the model, it acts merely as a normalisation factor that ensures that $P(M|D) + P(M^c|D) = 1$, where M^c is the complement of M, i.e. model M^c is true if M is false and vice versa. This normalisation implies that

$$P(D) = P(D|M)P(M) + P(D|M^c)P(M^c) = P(D|M)P(M) + P(D|M^c)[1 - P(M)] \tag{2}$$

Example: rare disease

As an example, let us re-visit the "Rare Disease" problem from part 3 of this course: 0.1% of the population are infected by the rare disease Frequentitus. A test is available that is 99.9% (99%) accurate at correctly determining sick (healthy) patients. Your test is positive. What is the probability that you have the disease?

To formalise this question, we can identify the model M with you being "sick", and the data D with the test being "positive". Using Bayes' theorem:

$$P(\text{sick}|\text{pos}) = \frac{P(\text{pos}|\text{sick})P(\text{sick})}{P(\text{pos}|\text{sick})P(\text{sick}) + P(\text{pos}|\text{healthy})P(\text{healthy})} = \frac{0.999 \cdot 0.001}{0.999 \cdot 0.001 + 0.01 \cdot 0.999} = \frac{1}{11} \approx 9\%$$

This solution is identical to that obtained part 3 using probability trees, but the Bayesian method is more efficient and more easily extends to more complex problems.

Updating priors

Imagine that you take some data D_1 and use it to determine the posterior probability $P(M|D_1)$ of a model M. Later you take some more data D_2, which are statistically *independent* from D_1. You are naturally interested in updating the posterior of M, such that it accounts for both data samples, i.e. you would like to compute the full posterior $P(M|D)$, where $D = \{D_1, D_2\}$ is the combined data. Given that the two data sets are independent, their joint probability is just $P(D) = P(D_1)P(D_2)$ and, likewise, $P(D|M) = P(D_1|M)P(D_2|M)$. Applying Bayes rule we get,

$$P(M|D) = \frac{P(D_1|M)P(D_2|M)P(M)}{P(D_1)P(D_2)} = \frac{P(D_2|M)}{P(D_2)} \cdot \underbrace{\frac{P(D_1|M)P(M)}{P(D_1)}}_{P(M|D_1)}.$$

This simple arithmetic shows that the full posterior can be obtained by using the first posterior $P(M|D_1)$ as the new prior for the second experiment and then update our belief using the new data $P(D_2|M)$. This is one of the fundamental theorems of Bayesian inference: *If we get additional, statistically independent data, the posterior can be updated by using the previous posterior as the new prior.* This process can be repeated ad infinitum.

Example: rare disease additional test

As a follow up from the previous example, let us now assume that you take another, statistically independent test for the disease Frequentitus. This test is 90% (95%) correct at determining sick (healthy patients). You test positive again. What is the overall probability that you have the disease?

It now suffices to use the previous posterior (1/11) as the new prior and apply Bayes rule again:

$$P(\text{sick}|\text{pos}) = \frac{P(\text{pos}|\text{sick})P(\text{sick})}{P(\text{pos}|\text{sick})P(\text{sick}) + P(\text{pos}|\text{healthy})P(\text{healthy})} = \frac{0.9 \cdot \frac{1}{11}}{0.9 \cdot \frac{1}{11} + 0.05 \cdot \frac{10}{11}} = \frac{9}{14} \approx 64\%$$

Parameterized models

In many practical cases, the model M to be inferred from the data D belongs to a discrete or continuous family of models, specified by one or multiple parameters. For instance, we might like to fit a normal distribution with parameters μ (mean) and σ (standard deviation) to some data points. Throughout this chapter, we will use the symbol θ to refer to such model parameters; e.g. $\theta = (\mu, \sigma)$. Where necessary, we will use $\bar{\theta}$ to denote the *true* model parameters that describe the *population* and $\hat{\theta}$ for a *point estimator*, i.e. a single guess of the true model parameters based on a *sample* drawn from that population.

In the case of parameterized models, it is common to rewrite the likelihood as

$$\mathcal{L}(\theta; D) \equiv P(D|\theta)$$

and consider $\mathcal{L}$ a function of θ at fixed D rather than the other way around. As a consequence, $\mathcal{L}$ is not generally normalised, i.e. $\int \mathcal{L}\, d\theta \neq 1$, unlike standard probability density functions (PDFs); hence the symbol $\mathcal{L}$ instead of P.

Let us emphasize one more time that the likelihood function *is the probability of the data D assuming that the model is specified by the parameters θ*, which is *not* to be confused with the probability of θ given D (i.e. the *posterior*). This distinction is the reason that we use the semi-colon notation $(\theta; D)$ for the likelihood instead of the vertical bar $(\theta|D)$, normally reserved for conditional probabilities. Of course, in the scientific literature, all sorts of notations are used, so be prepared to get confused! It is also quite common to drop the data vector and simply write $\mathcal{L}(\theta)$.

Following Bayes' theorem (Eq. 1) the posterior is related to the likelihood via,

$$P(\theta|D) \propto \mathcal{L}(\theta; D)P(\theta). \tag{3}$$

Note that this equation uses a proportionality symbol instead of an equal symbol, because we have dropped the θ-independent marginalized likelihood $P(D)$. It is understood that $P(\theta|D)$ has to be normalized, such that $\int P(\theta|D)\, d\theta = 1$, where the integral goes over the whole allowed domain of the parameter space.

Eq. 3 implies that $P(\theta|D) \propto \mathcal{L}(\theta; D)$, if and only if the prior is *uniform* (= *flat*), i.e. if $P(\theta) = \text{constant}$.

Incidentally, the data D often takes the form of n data points x_i, $i = 1, ..., n$, each of which can be a single value or a vector, provided with some measurement uncertainties. We can collect these points in a data vector $\mathbf{x} = (x_1, ..., x_n)$. Given these notations, the general inference problem then translates to determining the *posterior probability density* $P(\theta|\mathbf{x})$; and the likelihood associated with this problem is $\mathcal{L}(\theta; \mathbf{x})$.

Example: coin tossing

As an introductory example of likelihood functions, let us reconsider the coin toss experiment, discussed earlier in the context of binomial distributions. Our coin has an unknown probability $\bar{p}$ of landing heads up and thus a probability $(1 - \bar{p})$ of landing tails up. Hence, our model has a single real parameter $\theta = p$. Its true value $\bar{p}$ can lie between 0 and 1, and a fair coin has $\bar{p} = 0.5$.

Let us now assume that we flip the coin once and find the result to be heads up. This can be formalised by writing the data-vector $\mathbf{x} = (H)$. The likelihood is thus

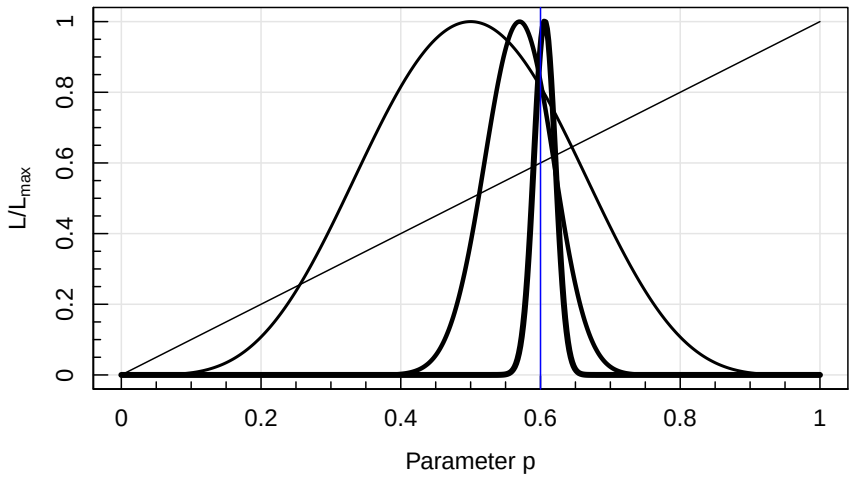
$$\mathcal{L}(p; H) = P(H|p) = p.$$

We continue tossing the coin n times, getting k heads and $(n-k)$ tails. Assuming that each successive coin flip is *independent and identically distributed* (IID), the likelihood function then becomes a standard binomial PDF (see part 3),

$$\mathcal{L}(p;k,n) = P(k|n,p) = \binom{n}{k} P(x_1|p) \cdot \ldots \cdot P(x_n|p) = \binom{n}{k} p^k (1-p)^{(n-k)}. \tag{4}$$

In **R**, this function is implemented as **dbinom(k,n,p)**. We can simulate a coin tossing experiment, by associating heads and tails with the binary numbers 1 and 0. A single toss can be simulated by drawing a uniform real random number $r \in [0,1]$ and setting the tossing result equal to $x = 1$ if $r < \bar{p}$ and $x = 0$ otherwise. (Alternatively, we could also use the **rbinom** function.) Let us assume that our coin is heads-biased with $\bar{p} = 0.6$ (called `p.true` in **R**). We 'toss' this coin 1000-times in **R** and plot the updated likelihood function at $n = 1, 10, 100, 1000$ using increasingly thick lines.

```
p.true = 0.6
x = as.numeric(runif(1e3)<p.true) # 1=heads, 0=tails
L = function(p,n) {
  k = sum(x[1:n])
  return(dbinom(k,n,p))
}
magplot(NA,xlim=c(0,1),ylim=c(0,1),
        xlab='Parameter p',
        ylab=expression('L/L' ['max']))
for (i in seq(0,3)) curve(L(x,10^i)/max(L(x,10^i)),lwd=i+1,n=500,add=T)
abline(v=p.true,col='blue')
```



The likelihood becomes more narrow and centred around the true value $\bar{p}$ (blue line) as n increases.

Let us now ask: Given k heads in n tosses and assuming a flat prior on p, i.e. $P(p) = 1$, what is the probability of the coin being tail-biased, i.e. $\bar{p} < 0.5$? From Eq. 3, it follows that

$$P(\bar{p} < 0.5|k,n) = \frac{\int_0^{0.5} \mathcal{L}(p;k,n)dp}{\int_0^1 \mathcal{L}(p;k,n)dp}.$$

Numerically, the integral can be evaluated using the **integrate** function (see part 2 of this course). The tail-bias probabilities corresponding to our $n = 1, 10, 100, 1000$ tosses are:

```
for (i in seq(0,3)) {
  n = 10^i
  phalf = integrate(L,0,0.5,n)$value/integrate(L,0,1,n)$value
  cat(sprintf('Number of tosses = %4d; P(p<0.5) = %.3f\n',n,phalf))
}
```

```
## Number of tosses =    1; P(p<0.5) = 0.250
## Number of tosses =   10; P(p<0.5) = 0.500
```

```
## Number of tosses =  100; P(p<0.5) = 0.082
## Number of tosses = 1000; P(p<0.5) = 0.000
```

As we can see, the values decrease to zero as $n \to \infty$, as expected, since the our simulated coins are head-biased, not tail-biased. The decrease does not need to be monotonic due to statistical fluctuations.

Example: spaceship emergency

To train our minds, let us discuss an analogous example to the coin experiment, in a different context.

Imagine you are on a spaceship and you need to land this spaceship on a planet, as a matter of emergency. Only two planets (A and B) are within reach and for the landing site to be hard enough, it has to rain on less than 50% of the days, on average, on the planet. You need to pick the best planet based on very limited data:

- On planet A it rained on 2 of 3 randomly observed independent days.
- On planet B it rained on 6 of 10 randomly observed independent days.

Which planet to you pick?

One might naively intuit that planet B is the smarter choice, since the data suggest that it rains more frequently on planet A than on planet B ($2/3 \approx 0.667$ vs $6/10 = 0.6$). However this answer is too quick. What we really need to compute are the probabilities for either planet to have rain on less than 50% of all days. To do so let us first determine the likelihoods as a function of the fraction r of rainy days:

- $\mathcal{L}_A(r) = P(2 \text{ of } 3 \text{ rainy days}|r) = \binom{3}{2} r^2(1-r)$
- $\mathcal{L}_B(r) = P(6 \text{ of } 10 \text{ rainy days}|r) = \binom{10}{6} r^6(1-r)^4$

Assuming a flat prior on r, that is $P(r) = 1 \ \forall r \in [0,1]$, Bayes' theorem implies that the posterior probability density for the planets to have r rainy days is

- $P_A(r|D) = \frac{\mathcal{L}_A(r)}{\int_0^1 \mathcal{L}_A(r')dr'}$
- $P_B(r|D) = \frac{\mathcal{L}_B(r)}{\int_0^1 \mathcal{L}_B(r')dr'}$

Hence, the posterior probabilities for the planets to be 'good' (r<0.5) are given by

- $P(A \text{ good}|D) = \int_0^{1/2} P_A(r|D)dr = \frac{\int_0^{1/2} r^2(1-r)dr}{\int_0^1 r^2(1-r)dr} = \frac{5}{16} = 0.3125$
- $P(B \text{ good}|D) = \int_0^{1/2} P_B(r|D)dr = \frac{\int_0^{1/2} r^6(1-r)^4dr}{\int_0^1 r^6(1-r)^4dr} = \frac{281}{1024} = 0.2744141$

(Note the analogy to the posterior probability of the coin being tail-biased.) Neither planet gives us particularly good chances of survival! However, given this analysis, it is clear that **planet A** has to be preferred, given no other information.

In **R**, we can forgo the analytical calculation and do everything numerically:

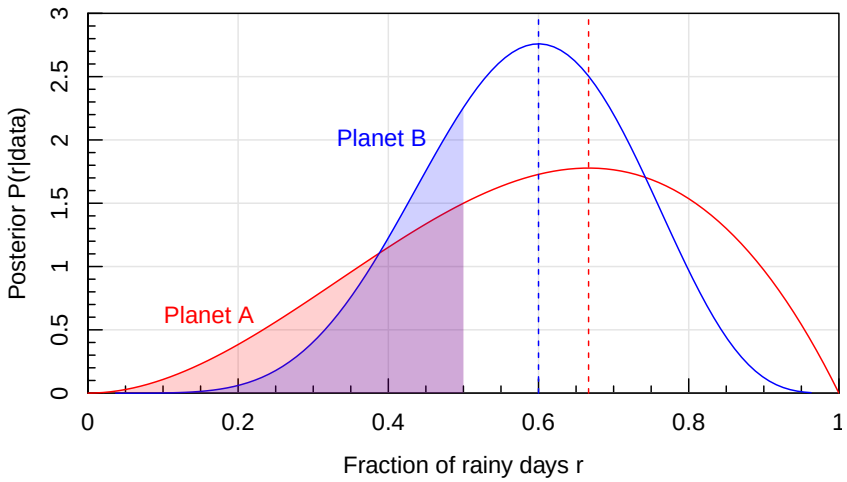
```
likelihood.A = function(r) dbinom(2,3,r)
likelihood.B = function(r) dbinom(6,10,r)
posterior.A = function(r) likelihood.A(r)/integral(likelihood.A,0,1)
posterior.B = function(r) likelihood.B(r)/integral(likelihood.B,0,1)
P.A.good = integral(posterior.A,0,0.5)
P.B.good = integral(posterior.B,0,0.5)
print(c(P.A.good,P.B.good))
```

```
## [1] 0.3125000 0.2744141
```

To sharpen our intuition for this problem let us plot the posteriors $P_A(r|D)$ and $P_B(r|D)$.

```
magcurve(posterior.A(x),xlim=c(0,1),ylim=c(0,3),xaxs='i',yaxs='i',
         xlab='Fraction of rainy days r',ylab='Posterior P(r|data)',col='red')
curve(posterior.B(x),col='blue',add=T)
abline(v=c(2/3,6/10),col=c('red','blue'),lty=2)
r = seq(0,0.5,length=100)
```

```
polygon(c(r,rev(r)),c(posterior.A(r),r*0),col='#ff000030',border=NA)
polygon(c(r,rev(r)),c(posterior.B(r),r*0),col='#0000ff30',border=NA)
text(0.24,0.6,'Planet A',pos=2,col='red')
text(0.47,2,'Planet B',pos=2,col='blue')
```



We see that while the mode of the posterior for planet B ($r = 6/10 = 0.6$) is lower than for planet A ($r = 2/3 \approx 0.667$), the integrated probability of good days ($r < 0.5$) is nonetheless higher for planet A. Graphically, this can be seen from the fact that the shaded red region has a larger surface area than the shaded blue region.

Maximum a posteriori and maximum likelihood estimation

Often one is interested in a *point estimator*, that is a single model parameter $\hat{\theta}$ which optimally describes the data (given a certain form of the model). Following the parameterized notation of Bayes' theorem, the most natural point estimator is the parameter θ, which maximises the posterior probability distribution $P(\theta|D)$. This parameter is known as the *maximum a posteriori (MAP)* estimator,

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} \mathcal{L}(\theta; D)P(\theta). \tag{5}$$

Quite frequently, one is interested in what model the data favour without accounting for prior information or one simply has no idea about a prior (apart, perhaps, from certain limits). In this case, it is sometimes justified to take the prior $P(\theta)$ to be constant. In statistical language this is normally called a *flat* prior. The MAP solution can then be found by maximising the likelihood. The latter parameter solution is known as the *maximum likelihood estimator (MLE)*,

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \mathcal{L}(\theta; D). \tag{6}$$

In the following sections, we will mainly discuss MLE rather than MAP estimation. Note, however, that the extension from the former to the latter is normally trivial, as it suffices to substitute $\mathcal{L}(\theta; D)$ for $\mathcal{L}(\theta; D)P(\theta)$ (for instance in the so-called Laplace approximation discussed in part 4, section 2).

Log-likelihood

Instead of working directly with the likelihood, it is often much more convenient to work with its natural logarithm, the *log-likelihood* function

$$\ell(\theta; D) \equiv \ln \mathcal{L}(\theta; D). \tag{7}$$

Note that sometimes the log-likelihood is abbreviated as LL.

Since the logarithm is a monotonic function, a maximum of $\mathcal{L}$ is also a maximum of ℓ and vice versa. Hence, $\hat{\theta}_{\mathrm{MLE}}$ can be obtained either through maximising $\mathcal{L}$ or ℓ.

There are several reason for ℓ to be used instead of $\mathcal{L}$:

- Firstly, the likelihood function is not normalised, i.e. $\int \mathcal{L}\, d\theta \neq 1$. Usually, $\mathcal{L}$ vanishes asymptotically or diverges as the number of data point n increases. In practice, this often implies extremely small or large values that can easily reach beyond the typical range of floating point numbers. Furthermore, the enormous dynamic range of $\mathcal{L}$ as a function of θ can be a challenge for numerical optimisers. For computational purposes log-likelihoods are therefore much more suitable.
- The likelihood is often a product of probabilities or probability densities, which themselves often involve exponential functions or power laws. Taking the logarithm turns products into sums and powers into products, making it much more easy to obtain closed-form solutions.
- Many applications of likelihood functions, such as the Laplace approximation, the Fisher Information and Jeffreys priors (all discussed in part 4, sect. 2) rely on the log-likelihood function.

Example: coin tossing (continued)

In the coin toss example above we shall now ask: Given n coin flips with k heads, what is the intrinsic probability p of the coin to produce heads? Assuming a flat prior $P(p) = 1$ (which is probably a bad assumption for a real coin!), the most probable value of p given the data is the MLE solution. Let us compute this solution after $n = 1, 10, 100, 1000$ tosses (see likelihood functions in the first figure):

```
for (i in seq(0,3)) {
  n = 10^i
  mle = optimise(L,c(0,1),n,maximum=TRUE)$maximum
  cat(sprintf('Number of tosses = %4d; MLE = %.2f\n',n,mle))
}
```

```
## Number of tosses =    1; MLE = 1.00
## Number of tosses =   10; MLE = 0.50
## Number of tosses =  100; MLE = 0.57
## Number of tosses = 1000; MLE = 0.61
```

As we see, the MLE gets closer to the true population parameter $\bar{p} = 0.6$ as n increases.

Let us now try to find the MLE solution analytically. To maximise $\mathcal{L}(p; k, n)$ (given in Eq. 4) as a function of p, we first take the logarithm:

$$\ell(p; k, n) = \ln \binom{n}{k} + \ln P(x_1|p) + ... + \ln P(x_n|k) = \ln \binom{n}{k} + k \ln(p) + (n-k) \ln(1-p). \tag{8}$$

It is easy to take the derivative of this log-likelihood:

$$\frac{d\ell}{dp} = \frac{k}{p} - \frac{n-k}{1-p} \tag{9}$$

The ML estimator $\hat{p}$ is obtained by setting this derivative to 0 and solving for p. This results in

$$\hat{p}(k, n) = \frac{k}{n}. \tag{10}$$

In other words, if a fraction $f_\mathrm{h} = k/n$ of the tosses results in heads, then the MLE method simply states that the coin is most likely to produce heads with a probability $\hat{p} = f_\mathrm{h}$. This is exactly what common sense would suggest, if all values of p are a priori equally likely – the implicit assumption of the MLE method.

Comparing competing models or families of models

Odds ratio

A particular situation is the case where we want to choose between only two competing models M_1 and M_2. In this case, the ratio between the posterior probabilities of M_1 and M_2 is called the *odds ratio* R (also known as *posterior odds*). According to Bayes' theorem,

$$R \equiv \frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} \cdot \frac{P(M_1)}{P(M_2)}. \tag{11}$$

Conveniently, the marginalised likelihood $P(D)$ has cancelled out.

Assuming that there are only two possible models M_1 and M_2, their respective posterior probabilities are,

$$P(M_1|D) = \frac{R}{1+R}, \quad P(M_2|D) = \frac{1}{1+R}. \tag{12}$$

This equation is, of course, just a restatement of Eqs. (1) and (2), obtained by substituting M with M_1 and its complement M^c with M_2.

Example: rare disease (again)

As a quick example, let us solve again the "rare disease" problem from the beginning of this chapter:

$$R = \frac{P(\text{pos}|\text{sick})P(\text{sick})}{P(\text{pos}|\text{healthy})P(\text{healthy})} = \frac{0.999 \cdot 0.001}{0.01 \cdot 0.999} = \frac{1}{10} \quad \Rightarrow \quad P(\text{sick}|\text{pos}) = \frac{R}{1+R} = \frac{1}{11}.$$

This solution is, of course, identical to the earlier one.

Bayes factor

If both models are a priori (i.e. before considering the data) equally likely, $P(M_1) = P(M_2)$, the odds ratio is just given by the likelihood ratio, generally known as the *Bayes factor*,

$$B \equiv \frac{P(D|M_1)}{P(D|M_2)}. \tag{13}$$

Even in the presence of non-equal priors for M_1 and M_2, the Bayes factor can be of some value as it quantifies the *additional* evidence that the data provide in favour of model M_1, beyond what was already known. Harold Jeffreys published a frequently used table to guide the interpretation of Bayes factors (see Table 1).

Table 1: Harold Jeffreys' evidence scale

$\|\log_{10}(\text{B})\|$	Strength of evidence
0.0 to 0.5	Marginal (barely worth mentioning)
0.5 to 1.0	Substantial
1.0 to 1.5	Strong
1.5 to 2.0	Very strong
>2.0	Decisive

Extension to parameterized models

In comparing two models M_1 and M_2, each of these models (or just one) can be a class of models, described by a set of parameters θ_1 and θ_2. In this case the likelihoods $P(D|M_1)$ and $P(D|M_2)$ are the average likelihoods, *marginalised* over θ_1 and θ_2, respectively. Explicitly, for both models $i \in \{1, 2\}$,

$$P(D|M_i) = \int P(D|\theta_i; M_i)P(\theta_i; M_i)d\theta_i. \tag{14}$$

If θ_i is a vector of k parameters, this is a k-dimensional integral over the whole allowed parameter domain. The likelihoods $P(D|M_i)$ can be called *marginal likelihoods*, but bear in mind that these are *model-specific* marginal likelihoods, not be confused with the global marginal likelihood $P(D) = P(D|M_1)P(M_1) + P(D|M_2)P(M_2)$. To avoid this confusion when comparing models, it is therefore common to refer to $P(D|M_i)$ as the *model evidence* or *Bayesian evidence* (for model M_i).

Following Eq. (14), the Bayes factor of two parameterized models becomes

$$B \equiv \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\theta_1; M_1)P(\theta_1; M_1)d\theta_1}{\int P(D|\theta_2; M_2)P(\theta_2; M_2)d\theta_2}. \tag{15}$$

Extension to multiple models

Following the same logic, we can, of course, compare any number N of models M_i, $i = 1, ..., N$, against each other, by comparing their model evidences $P(D|M_i)$. Some of these models may be parameterized, in which case the model evidence involves an integral over the respective parameters (see eq. 14), while others may have no parameters, in which case no such integral is needed. The actual posterior probability of each model M_i, assuming that one has to be true, is then given by

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j=1}^{N} P(D|M_j)P(M_j)}. \tag{16}$$

The denominator of this equation is just the global marginal likelihood for the N models.

Interestingly, the Bayes factor (and, equivalently, the model evidence) automatically penalises models that have too much substructure (e.g. too many parameters in the vector θ). It thus guards against over-fitting. We will expand on this topic later in the context of so-called *information criteria*.

In summary, the model evidence $P(D|M_i)$ quantifies how much we can *believe* in a model M_i, given the data D, irrespective of prior knowledge.

Example: which distribution?

Step 1: parameter-free models

Question: Were the random numbers $\{0.03, 0.13, 0.19\}$ *generated using 'runif(3)' or 'rnorm(3)'? Give the odds ratio and probabilities.*

There are two models to be compared: let us call them M_u (for **runif**) and M_n (for **rnorm**), respectively. Nothing is said about which model is a priori more likely. Thus we can assume that their priors are identical. Hence, the odds ratio is equal to the Bayes factor,

$$R = B = \frac{P(D|M_u)}{P(D|M_n)}.$$

The two likelihoods are given by the probability density functions associated with **runif** and **rnorm**:

- $P(D|M_u) = \prod_{i=1}^{3} \rho_u(x_i) = 1$, where $\rho_u(x) = 1$ if $x \in [0, 1]$ and $\rho_u(x) = 0$ otherwise
- $P(D|M_n) = \prod_{i=1}^{3} \rho_n(x_i)$, where $\rho_n(x) = (2\pi)^{-1/2} e^{-x^2/2}$

So let us compute this:

```
x = c(0.03,0.13,0.19) # data
likelihood.u = prod(dunif(x))
likelihood.n = prod(dnorm(x))
R = likelihood.u/likelihood.n
sprintf('R = %.2f',R)
```

```
## [1] "R = 16.18"
```

Hence, there is a probability $P = R/(1 + R) = 94.2\%$ that the three numbers were generated using **runif(3)** (which is indeed how I got them). In terms of Jeffreys' evidence scale (Table 1), this would be called "strong" evidence, since $\log_{10}(R) = 1.21$.

Step 2: models with bounded parameters

Question: I draw four random numbers a, b, μ, σ, *uniformly from the intervals* $a \in [-2, 0]$, $b \in [0, 2]$, $\mu \in [-2, 2]$, $\sigma \in [0, 5]$. *Then, I flip a fair coin to decide whether to pick a uniform distribution (heads) or a normal distribution (tail) of the form:*

- $\rho_u(x|a, b) = 1/(b - a)$, if $x \in [a, b]$, 0 otherwise
- $\rho_n(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$

Then, I use this distribution to produce the five random numbers $\{-0.12, 0.70, 0.88, 0.81, 0.47\}$. *What is the probability that my coin is showing heads?*

Note that, unlike in the previous question, we do not know the parameters $\{a, b\}$ and $\{\mu, \sigma\}$ of the uniform and the normal distributions, respectively. We therefore have to consider the continuum of distributions in the range $a \in [-2, 0]$, $b \in [0, 2]$, $\mu \in [-2, 2]$, $\sigma \in [0, 5]$.

The likelihoods for the two models M_u and M_n become

- $P(D|a, b; M_u) = \prod_{i=1}^{5} \rho_u(x_i|a, b)$
- $P(D|\mu, \sigma; M_n) = \prod_{i=1}^{5} \rho_n(x_i|\mu, \sigma)$

Since the coin is fair, both models are a priori equally probable. Hence, the odds ratio is again equal to the Bayes factor, but this time we have to use the Bayes factor for parameterized models given in Eq. (15):

$$R = B = \frac{\int_{-2}^{0} da \int_{0}^{2} db\ P(D|a, b; M_u) P(a, b; M_u)}{\int_{-2}^{2} d\mu \int_{0}^{5} d\sigma\ P(D|\mu, \sigma; M_n) P(\mu, \sigma; M_n)}.$$

This equation includes priors on the model parameters a, b, μ, σ, which are all flat priors due to the statement that they have been sampled "uniformly" from their intervals. The normalisation condition $\int_{-2}^{0} da \int_{0}^{2} db\ P(a, b; M_u) = 1$ implies that $P(a, b; M_u) = 1/4$; and likewise $P(\mu, \sigma; M_n) = 1/20$. Hence,

$$R = 5 \cdot \frac{\int_{-2}^{0} da \int_{0}^{2} db\ P(D|a, b; M_u)}{\int_{-2}^{2} d\mu \int_{0}^{5} d\sigma\ P(D|\mu, \sigma; M_n)}.$$

Let us compute this:

```
x = c(-0.12,0.70,0.88,0.81,0.47) # data
integrand.u = function(a,b) prod(dunif(x,a,b))
integrand.n = function(mu,sigma) prod(dnorm(x,mu,sigma))
evidence.u = integral2(integrand.u,-2,0,0,2,vectorized=F)$Q
evidence.n = integral2(integrand.n,-2,2,0,5,vectorized=F)$Q
R = evidence.u/evidence.n
sprintf('R = %.2f',R)
```

```
## [1] "R = 2.75"
```

Hence, there is a probability $P = R/(1 + R) = 73.4\%$ that the five numbers were drawn from a uniform distribution, i.e. that my coin showed heads (which was indeed the case). However, in terms of Jeffreys' evidence scale (Table 1), this would only be called "marginal" evidence, since $\log_{10}(R) = 0.44$.

Step 3: models with unbounded parameters

Question: Were the random numbers $\{8.096, 9.779, 4.475, 4.067, -1.594, 6.413, 4.730, 3.761, 12.278, 2.333\}$ *drawn from a uniform or a normal distribution? Give the odds ratio and probabilities.*

This question is very similar to the previous one, except that there is no indication on the parameters a, b, μ, σ. In principle, any real values are possible, as long as $a < b$ and $\sigma > 0$. One might be tempted to think that this problem can thus be solved by letting the integration boundaries go to infinity:

$$R = \frac{\int_{-\infty}^{\infty} da \int_{a}^{\infty} db\ P(D|a, b; M_u) P(a, b; M_u)}{\int_{-\infty}^{\infty} d\mu \int_{0}^{\infty} d\sigma\ P(D|\mu, \sigma; M_n) P(\mu, \sigma; M_n)}.$$

However, for the priors $P(a, b; M_u)$ and $P(\mu, \sigma; M_n)$ to be normalized, they would have to vanish asymptotically, resulting in an ill-defined $R = 0/0$. Technically speaking $P(a, b; M_u)$ and $P(\mu, \sigma; M_n)$ are *improper* priors, i.e. priors which *cannot* be normalised. Such improper priors can only be expressed up to an arbitrary normalization constant. One might think that two such arbitrary constants in the numerator and denominator of R cancel out, but there is no reason to assign the same constant to $P(a, b; M_u)$ and $P(\mu, \sigma; M_n)$. In fact, a re-parameterisation, say $a' = a/2$, would change the value of the integral if the prior $P(a, b; M_u)$ is not changed accordingly.

Interestingly, the issue about the ill-defined proportionality constants of improper priors can be avoided by using the same improper prior on *common* parameters in the models to be compared. In the present case, we could choose to parameterise both distributions using just the mean μ and the standard deviation σ.

Using the fact that the variance of a uniform distribution is equal to $(b-a)^2/12$, we find that $a = \mu - \sqrt{3}\sigma$ and $b = \mu + \sqrt{3}\sigma$. In this way, the numerator and denominator in R use the same improper prior $P(\mu, \sigma; M_u) = P(\mu, \sigma; M_n)$. If this is a flat prior, it can be factored out of the integrals and cancelled out. Thus,

$$R = \frac{\int_{-\infty}^{\infty} d\mu \int_0^{\infty} d\sigma \; P(D|a = \mu - \sqrt{3}\sigma, b = \mu + \sqrt{3}\sigma; M_u)}{\int_{-\infty}^{\infty} d\mu \int_0^{\infty} d\sigma \; P(D|\mu, \sigma; M_n)}. \tag{17}$$

Let me stress that, in the absence of explicit information on priors, flat priors are not necessarily the best choice! We will return to this topic later and discuss ways to determine the right prior in a more "objective" way (see Jeffreys priors at the end of part 4, sect 2).

Evaluating the double integrals with infinite boundaries in Eq. (17) is numerically rather challenging. We use the **integrate2** function of the **pracma** package with finite boundaries h instead of ∞ and then let h increase until the result converges. Even so, the numerical integral can sometimes produce bugs, which we catch with a **try** function. The following code should be sufficiently self-explanatory.

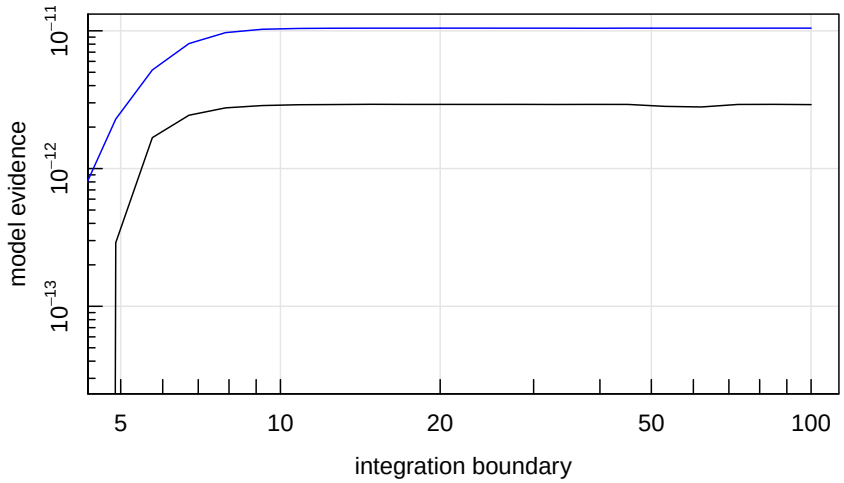
```
x = c(8.096,9.779,4.475,4.067,-1.594,6.413,4.730,3.761,12.278,2.333) # data

# integrands
integrand.u = Vectorize(function(mu,s) prod(dunif(x,mu-sqrt(3)*s,mu+sqrt(3)*s)))
integrand.n = Vectorize(function(mu,s) prod(dnorm(x,mu,s)))

# functions to evaluate marginal likelihoods (= model evidence),
# using h instead of infinity as the boundary of the 2D integrals
evidence.u = function(h) {
  # make boundaries [smin,smax] for s, which span the subinterval of [0,h] on which the
  # integrand integrand.u is non-zero (this allows a more stable numerical integration)
  smin = Vectorize(function(mu) min(h,max(abs(x-mu))*2/sqrt(12)))
  smax = Vectorize(function(mu) min(h,smin(mu)+h))
  # evaluate 2D integral, while catching errors
  out = try(integral2(integrand.u,-h,h,smin,smax)$Q,silent=TRUE)
  return(ifelse(is.numeric(out),out,0))
}

evidence.n = function(h) {
  # evaluate 2D integral, while catching errors
  out = try(integral2(integrand.n,-h,h,0,h)$Q,silent=TRUE)
  return(ifelse(is.numeric(out),out,0))
}

# check how the integrals converge as h increases
h = 10^seq(0,2,length=30)
f = foreach(i=1:30,.combine='c')%do%c(evidence.u(h[i]),evidence.n(h[i]))
f = matrix(f,nrow=2)
magplot(h,f[1,]+1e-99,log='xy',type='l',ylim=c(min(max(f[1,]),max(f[2,]))*0.01,max(f)),
        xlab='integration boundary',ylab='model evidence')
lines(h,f[2,]+1e-99,col='blue')
```




```
# odds ratio
R = max(f[1,])/max(f[2,])
sprintf('R = %.2f',R)
```

```
## [1] "R = 0.28"
```

Hence, there is a probability $P = R/(1+R) = 21.9\%$ that the three numbers were generated using *some* uniform distribution and a probability $P = 1/(1+R) = 78.1\%$ that they were generated using *some* normal distribution. (It was indeed a normal distribution of mean 3.2 and standard deviation 1.7.)

We caution that this solution is not necessarily unique. For instance, we could choose a different way to match up the free parameters of the uniform distribution to those of the normal distribution: instead of using their standard deviations, we could have used their full-width-half-max (FWHM) values as the common free parameter. This would lead to a slightly different prior ratio. This illustrates the general difficulty of making sensible choices for improper priors.

Just for later reference, using the Jeffreys prior ($P(\mu,\sigma) = \sigma^{-2}$, for normal distributions) instead of a flat one would change the odds ratio to $R = 0.20$ in this example.