MTH6102

Bayesian Statistical Methods

Jamie Griffin

Queen Mary, Autumn 2022

Contents

1 Likelihood 4

1.1 Maximum likelihood estimation . . . 5

1.2 Standard error . . . 8

2 Bayesian inference 9

2.1 Bayes' theorem . . . 9

2.2 Bayes' theorem and Bayesian inference . . . 10

2.3 Conjugate prior distributions . . . 12

2.4 Point estimates and credible intervals . . . 16

3 Specifying prior distributions 20

3.1 Informative prior distributions . . . 20

3.2 Less informative prior distributions . . . 21

4 Markov chain Monte Carlo methods 22

4.1 The Metropolis algorithm . . . 22

5 Predictive distributions 27

5.1 Simulating the posterior predictive distribution . . . 29

6 Hierarchical models 30

6.1 Inference for hierarchical models . . . 31

6.2 Advantages of hierarchical models . . . 34

7 Tests and model comparisions 35

Introduction

So far at Queen Mary, the statistics modules have taught the classical or frequentist approach which is based on the idea that probability represents a long run limiting frequency. In the Bayesian approach, any uncertain quantity is described by a probability distribution, and so probability represents a degree of belief in an event which is conditional on the knowledge of the person concerned.

This course will introduce you to Bayesian statistics. These notes are self-contained but you may want to read other accounts of Bayesian statistics as well. A useful introductory textbook is:

- Bayesian Statistics: An Introduction (4th ed.) by P M Lee. (This is available as an e-book from the library)

Parts of the following are also useful:

- Bayesian Inference for Statistical Analysis by G E P Box and G C Tiao.
- Bayesian Data Analysis by A Gelman, J B Carlin, H S Stern and D B Rubin.
- Probability and Statistics from a Bayesian viewpoint (Vol 2) by D V Lindley.

Chapter 1

Likelihood

First we review the concept of likelihood, which is essential for Bayesian theory, but can also be used in frequentist methods. Let y be the data that we observe, which is usually a vector. We assume that y was generated by some probability model which we can specify. Suppose that this probability model depends on one or more parameters θ, which we want to estimate.

Definition 1.1. If the components of y are continuous, then the likelihood is defined as the joint probability density function of y; if y is discrete, then the likelihood is defined as the joint probability mass function of y. In either case, we denote the likelihood as

$$p(y \mid \theta).$$

Example 1.1. Let $y = y_1, \ldots, y_n$ be a random sample from a normal distribution with unknown parameters μ and σ. Then θ is the vector (μ, σ), and the likelihood is the joint probability density function

$$p(y \mid \mu, \sigma) = \prod_{i=1}^{n} \phi(y_i \mid \mu, \sigma).$$

Note that $\prod$ here is the symbol for the product:

$$\prod_{i=1}^{n} a_i = a_1 \times a_2 \times \cdots \times a_n.$$

ϕ is the normal probability density function with parameters μ and σ

$$\phi(y_i \mid \mu, \sigma) = \frac{e^{-(y_i-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

Example 1.2. Suppose we observe k successes in n independent trials, where each trial has probability of success q. Now $\theta = q$, the observed data is k, and the likelihood is the binomial probability mass function

$$p(k \mid q) = \binom{n}{k} q^k (1-q)^{n-k}.$$

It is also possible to construct likelihoods which combine probabilities and probability density functions, for example if the observed data contains both discrete and continuous components. Alternatively, probabilities may appear in the likelihood if continuous data is only observed to lie within some interval.

Example 1.3. Assume that the time until failure for a certain type of light bulb is exponentially distributed with parameter λ, and we observe n bulbs, with failure times $t = t_1, \ldots, t_n$.

The likelihood contribution for a single observation t_i is the exponential probability density function

$$\lambda e^{-\lambda t_i}.$$

So the overall likelihood is the joint probability density function

$$p(t \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i}.$$

Suppose instead that we observe the failure time for the first m light bulbs with $m < n$, but for the remaining $n - m$ bulbs we only observe that they have not failed by time t_i. Then for $i \leq m$, the likelihood contributions are as before.

For $i > m$, the likelihood is the probability of what we have observed. Denoting the random variable for the failure time by T_i, we have observed that $T_i > t_i$, so the likelihood contribution is

$$p(T_i > t_i) = e^{-\lambda t_i}.$$

This is because the cumulative distribution function is $p(T_i \leq t_i) = 1 - e^{-\lambda t_i}$.

Hence the overall likelihood is now

$$p(t \mid \lambda) = \prod_{i=1}^{m} \lambda e^{-\lambda t_i} \prod_{i=m+1}^{n} e^{-\lambda t_i}.$$

1.1 Maximum likelihood estimation

In example 1.1, the parameters μ and σ of the normal distribution which generated the data are known as population parameters. An *estimator* is defined as a function of the observed data which we use as an estimate of a population parameter. For example the sample mean and variance may be used as estimators of the population mean and variance, respectively.

To use the likelihood to estimate parameters θ, we find the vector $\hat{\theta}$ which maximizes the likelihood $p(y \mid \theta)$, given the data x that we have observed. This is the method of maximum likelihood, and the estimator $\hat{\theta}$ is known as the maximum likelihood estimator, or MLE.

Usually the parameters θ are continuous, and those are the only examples we cover, but the idea of likelihood also makes sense if the unknown quantity is discrete.

When finding the MLE it is usually more convenient to work with the log of the likelihood: as the log is a monotonically increasing function, the same θ will maximize $p(y \mid \theta)$ and $\log(p(y \mid \theta))$. The log-likelihood is denoted by

$$\ell(\theta; y) = \log(p(y \mid \theta)).$$

Since the likelihood is typically a product of terms for independent observations, the log-likelihood is a sum of terms, so using the log greatly simplifies finding the derivatives in order to find the maximum.

Returning to the binomial example 1.2, the log-likelihood is

$$\begin{aligned}\ell(q; k) &= \log\left(\binom{n}{k} q^k (1-q)^{n-k}\right) \\ &= \log\binom{n}{k} + k\log(q) + (n-k)\log(1-q).\end{aligned}$$

To find the MLE, we differentiate with respect to q and set to zero

$$\frac{d\ell}{dq} = \frac{k}{q} - \frac{n-k}{1-q} = 0.$$

Rearranging gives the MLE $\hat{q}$ as

$$\hat{q} = \frac{k}{n}.$$

For the normal example 1.1, we have log-likelihood

$$\begin{aligned}\ell(\mu, \sigma; y) = \log\left(p(y \mid \mu, \sigma)\right) &= \log\left(\prod_{i=1}^{n} \phi(y_i \mid \mu, \sigma)\right) \\ &= \sum_{i=1}^{n} \log\left(\phi(y_i \mid \mu, \sigma)\right) \\ &= \sum_{i=1}^{n} \left(-\log(\sqrt{2\pi}) - \log(\sigma) - \frac{(y_i - \mu)^2}{2\sigma^2}\right) \\ &= -n\log(\sqrt{2\pi}) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \\ &= -n\log(\sqrt{2\pi}) - \frac{n}{2}\log(\gamma) - \frac{1}{2\gamma}\sum_{i=1}^{n}(y_i - \mu)^2\end{aligned}$$

where $\gamma = \sigma^2$.

Differentiating gives

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\gamma}\sum_{i=1}^{n}(y_i - \mu) = 0.$$

Hence the MLE for μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}.$$

$$\frac{\partial \ell}{\partial \gamma} = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Setting to zero, substituting our value for $\hat{\mu}$ and rearranging gives

$$\hat{\sigma}^2 = \hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Note that this is different from the usual sample variance estimator, which has $1/(n-1)$ instead of $1/n$.

Finally, for the exponential example 1.3, the likelihood for the second case, where we had both probability density functions and probabilities, was

$$p(t \mid \lambda) = \prod_{i=1}^{m} \lambda e^{-\lambda t_i} \prod_{i=m+1}^{n} e^{-\lambda t_i} = \lambda^m e^{-\lambda S}$$

where $S = \sum_{i=1}^{n} t_i$.

Hence the log-likelihood is

$$\ell(\lambda; t) = m\log(\lambda) - \lambda S.$$

Differentiating and setting to zero gives

$$\frac{d\ell}{d\lambda} = \frac{m}{\lambda} - S = 0.$$

Hence the MLE is

$$\hat{\lambda} = \frac{m}{S}.$$

In each example, we could calculate the second derivatives at the stationary points to verify that we have found a maximum.

1.2 Standard error

In frequentist statistics, an estimator $\hat{\psi}$ is a function of the data y which we use to estimate a population parameter ψ. To quantify the precision of $\hat{\psi}$, the *standard error* is defined as the standard deviation of $\hat{\psi}$ if the data y were repeatedly generated (given certain underlying population parameters), and for each realisation of y we calculate $\hat{\psi}$. So it quantifies the uncertainty in $\hat{\psi}$ due to random variation in the data we might have observed.

In the binomial example 1.2, we have log-likelihood

$$\ell(q;k) = \log\binom{n}{k} + k\log(q) + (n-k)\log(1-q).$$

We saw that the MLE is

$$\hat{q} = \frac{k}{n}$$

k is a binomal random variable with variance $Var(k) = nq(1-q)$, and so

$$Var(\hat{q}) = \frac{nq(1-q)}{n^2} = \frac{q(1-q)}{n}.$$

We would estimate this variance by substituting our estimate $\hat{q}$ for the unknown true value of q, so the estimated standard error is

$$Se(\hat{q}) = \sqrt{\frac{\hat{q}(1-\hat{q})}{n}} = \sqrt{\frac{k/n(1-k/n)}{n}} = \sqrt{\frac{k(n-k)}{n^3}}.$$

In the case of maximum likelihood estimation, there is a general approximate formula for the standard error based on the second derivative of the log-likelihood, which we do not cover in this module.

We also do not cover confidence intervals, but we do cover the Bayesian version, which are called credible intervals.

Chapter 2

Bayesian inference

2.1 Bayes' theorem

Bayes' theorem is a formula from probability theory that is central to Bayesian inference. It is named after Rev. Thomas Bayes, a nonconformist minister who lived in England in the first half of the eighteenth century. The theorem states that:

Theorem 2.1. Let Ω be a sample space and $A_1, A_2, \ldots, A_m$ be mutually exclusive and exhaustive events in Ω (i.e. $A_i \cap A_j = \emptyset, i \neq j, \cup_{i=1}^{k} A_i = \Omega$; the A_i form a partition of Ω.) Let B be any event with $p(B) > 0$. Then

$$p(A_i \mid B) = \frac{p(A_i)p(B \mid A_i)}{p(B)} = \frac{p(A_i)p(B \mid A_i)}{\sum_{j=1}^{m} p(A_j)p(B \mid A_j)}$$

The proof follows from the definition of conditional probabilities and the law of total probabilities.

Example 2.1. Suppose a test for an infection has 90% sensitivity and 95% specificity, and the proportion of the population with the infection is $q = 1/2000$. Sensitivity is the probability of detecting a genuine infection, and specificity is the probability of being correct about a non-infection. So $p(\text{+ve test} \mid \text{infected}) = 0.9$ and $p(\text{-ve test} \mid \text{not infected}) = 0.95$. What is the probability that someone who tests positive is infected?

Let the events be as follows:
B: test positive
A_1: is infected
A_2: is not infected

We want to find $p(A_1 \mid B)$. The probabilities we have are $p(A_1) = 1/2000$, $p(B \mid A_1) = 0.9$ and $p(B \mid A_2) = 1 - (B^C \mid A_2) = 1 - 0.95 = 0.05$.

Applying Bayes' theorem,

$$p(A_1 \mid B) = \frac{p(A_1)p(B \mid A_1)}{\sum_{j=1}^{2} p(A_j)p(B \mid A_j)}$$

$$p(A_1)p(B \mid A_1) = 1/2000 \times 0.9 = 0.00045$$

$$p(A_2)p(B \mid A_2) = (1 - 1/2000) \times 0.05 = .05$$

Hence

$$p(A_1 \mid B) = \frac{0.00045}{0.00045 + .05} = 0.0089$$

So there is a less than 1% chance that the person is infected if they test positive.

Bayes' Theorem is also applicable to probability densities.

Theorem 2.2. Let X, Y be two continuous r.v.'s (possibly multivariate) and let $f(x, y)$ be the joint probability density function (pdf), $f(x \mid y)$ the conditional pdf etc. then

$$f(x \mid y) = \frac{f(y \mid x)\, f(x)}{f(y)} = \frac{f(y \mid x)\, f(x)}{\int f(y \mid x')\, f(x')\, dx'}$$

Alternatively, Y may be discrete, in which case $f(y \mid x)$ and $f(y)$ are probability mass functions.

2.2 Bayes' theorem and Bayesian inference

In the Bayesian framework, all uncertainty is specified by probability distributions. This includes uncertainty about the unknown parameters. So we need to start with a probability distribution for the parameters $p(\theta)$.

We then update the probability distribution for θ using the observed data y. This updating is done using Bayes' theorem

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{p(y)}.$$

$p(y \mid \theta)$ is the likelihood for parameters θ given the observed data y.

We don't normally need to find the normalizing constant $p(y)$, which is given by

$$p(y) = \int p(\theta)\, p(y \mid \theta)\, d\theta \;\text{ or }\; \sum_{\theta} p(\theta)\, p(y \mid \theta)$$

So the procedure is as follows:

- Start with a distribution $p(\theta)$ - this is known as the *prior distribution.*
- Combine the prior distribution with the likelihood $p(y \mid \theta)$ using Bayes' theorem.
- The resulting probability distribution $p(\theta \mid y)$ is known as the *posterior distribution.*
- We base our inferences about θ on this posterior distribution.

The use of Bayes' theorem can be summarized as

Posterior distribution $\propto$ prior distribution $\times$ likelihood

Example 2.2. Suppose a biased coin has probability of heads q, and we observe k heads in n independent coin tosses. We saw the binomial likelihood for this problem:

$$p(k \mid q) = \binom{n}{k} q^k (1-q)^{n-k}$$

For Bayesian inference, we need to specify a prior distribution for q. As q is a continuous quantity between 0 and 1, the family of beta distributions is a reasonable choice.

If $X \sim \text{Beta}(\alpha, \beta)$, its probability density function is

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \ 0 \le x \le 1$$

where B is the beta function and $\alpha, \beta > 0$ are parameters.

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx, \text{ also } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The uniform distribution on $[0,1]$ is a special case of the beta distribution with $\alpha = 1, \beta = 1$.

Combining prior distribution and likelihood, we have the posterior distribution $p(q \mid k)$ given by

$$p(q \mid k) \propto p(q)\,p(k \mid q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)} \binom{n}{k} q^k (1-q)^{n-k} \propto q^{k+\alpha-1}(1-q)^{n-k+\beta-1}$$

We can recognize this posterior distribution as being proportional to the pdf of a $\text{Beta}(k+\alpha, n-k+\beta)$ distribution. Hence we do not need to explicitly work out the normalizing constant for $p(q \mid k)$, we can immediately see that $q \mid k$ has a $\text{Beta}(k+\alpha, n-k+\beta)$ distribution.

For a Bayesian point estimate for q, we summarize the posterior distribution, for example by its mean. A Beta(α, β) rv has expected value $\dfrac{\alpha}{\alpha+\beta}$. Hence our posterior mean for q is

$$E(q \mid k) = \frac{k+\alpha}{n+\alpha+\beta}.$$

Recall that the maximum likelihood estimator is

$$\hat{q} = \frac{k}{n}.$$

So for large values of k and n, the Bayesian estimate and MLE will be similar, whereas they differ more for smaller sample sizes.

In the special case that the prior distribution is uniform on $[0, 1]$, the posterior distribution is Beta$(k+1, n-k+1)$. and the posterior mean value for q is

$$E(q \mid k) = \frac{k+1}{n+2}.$$

2.3 Conjugate prior distributions

In the binomial example 2.2 with a beta prior distribution, we saw that the posterior distribution is also a beta distribution. When we have the same family of distributions for the prior and posterior for one or more parameters, this is known as a conjugate family of distributions. We say that the family of beta distributions is conjugate to the binomial likelihood.

The binomial likelihood is

$$p(k \mid q) = \binom{n}{k} q^k (1-q)^{n-k}$$

and the beta prior probability density is

$$p(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha, \beta)}$$

Considered as functions of q, the prior density function and the likelihood have the same functional form as each other (namely proportional to $q^r(1-q)^s$ for some r, s). Also, when we multiply them together, we still have the same form. This is what characterizes conjugate distributions.

Example 2.3. Consider the exponential example 1.3, in which we have observed the time to failure of n light bulbs as $t_1, \ldots, t_n$. The likelihood for the observed data is

$$p(t \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda S}$$

where $S = \sum_{i=1}^n t_i$.

The gamma distribution with parameters $\alpha, \beta > 0$ has probability density function

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)},\ x > 0$$

This looks like a candidate for a conjugate prior distribution for the parameter λ in the exponential model.

With a Gamma(α, β) prior for λ, the posterior distribution is

$$\begin{aligned} p(\lambda \mid t) \propto p(\lambda)\, p(t \mid \lambda) &= \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}\, \lambda^n e^{-\lambda S} \\ &\propto \lambda^{\alpha-1} e^{-\beta\lambda}\, \lambda^n e^{-\lambda S} = \lambda^{n+\alpha-1} e^{-\lambda(S+\beta)}. \end{aligned}$$

Hence the posterior density function is proportional to a Gamma$(n+\alpha, S+\beta)$ pdf, and so the posterior density for λ must be a gamma distribution with parameters $n+\alpha$ and $S+\beta$.

The mean of a Gamma(α, β) distribution is $\dfrac{\alpha}{\beta}$, and so the posterior mean for λ is

$$E(\lambda \mid t) = \frac{n+\alpha}{S+\beta}.$$

Recall that the maximum likelihood estimator (MLE) was

$$\hat{\lambda} = \frac{n}{S}.$$

As n and S increase, the posterior mean approaches the MLE.

Example 2.4. As another example, consider a random sample $y = y_1, \ldots, y_n$ from a normal distribution with parameters μ and σ, $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. For now, we take σ to be known, and so we are only looking for a prior and posterior distribution for μ.

The likelihood for the sample is

$$\begin{aligned} p(y \mid \mu) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right) \end{aligned}$$

where we have dropped terms that do not contain μ.

The sum in the exponent can be expanded as

$$\begin{aligned} \sum_{i=1}^n (y_i-\mu)^2 &= \sum_{i=1}^n y_i^2 - 2\mu\sum_{i=1}^n y_i + n\mu^2 \\ &= S_2 - 2n\mu\bar{y} + n\mu^2 \end{aligned}$$

where $S_2 = \sum_{i=1}^n y_i^2$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

Inside the exponent we can add or subtract terms not involving μ, because we only need to define the likelihood up to a constant of proportionality

$$\begin{aligned} p(y \mid \mu) &\propto \exp\bigg(-\frac{1}{2\sigma^2}(S_2 - 2n\mu\bar{y} + n\mu^2)\bigg) \\ &\propto \exp\bigg(-\frac{1}{2\sigma^2}(-2n\mu\bar{y} + n\mu^2)\bigg) \\ &\propto \exp\bigg(-\frac{1}{2\sigma^2}(n\bar{y}^2 - 2n\mu\bar{y} + n\mu^2)\bigg) \\ &= \exp\Big(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\Big) \end{aligned}$$

If we take the prior distribution for μ as normal, $\mu \sim N(\mu_0, \sigma_0^2)$, then the prior probability density is

$$p(\mu) \propto \exp\bigg(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\bigg).$$

Hence the posterior density after seeing the data y is

$$\begin{aligned} p(\mu \mid y) &\propto p(\mu)\; p(y \mid \mu) \\ &\propto \exp\bigg(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\bigg) \exp\bigg(-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\bigg) \\ &\propto \exp\bigg(-\frac{(\mu^2 - 2\mu_0\mu)}{2\sigma_0^2} - \frac{n(\mu^2 - 2\bar{y}\mu)}{2\sigma^2}\bigg) \\ &= \exp\bigg(-\frac{1}{2}\bigg[\mu^2\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right)\bigg]\bigg) \end{aligned}$$

where any factors not involving μ have been dropped at each stage. We can recognise $p(\mu \mid y)$ as being proportional to a normal density $N(\mu_1, \sigma_1^2)$ for some μ_1, σ_1. An $N(\mu_1, \sigma_1^2)$ pdf for μ is proportional to

$$\exp\bigg(-\frac{(\mu - \mu_1)^2}{2\sigma_1^2}\bigg) = \exp\bigg(-\frac{1}{2}\bigg[\mu^2\left(\frac{1}{\sigma_1^2}\right) - 2\mu\left(\frac{\mu_1}{\sigma_1^2}\right) + \mu_1^2\left(\frac{1}{\sigma_1^2}\right)\bigg]\bigg)$$

Equating the terms in μ^2 and μ with those in the expression for $p(\mu \mid y)$, we see that

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \quad \text{and} \quad \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}.$$

Rearranging these gives

$$\mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right) = \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right) \Big/ \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)$$

Hence the posterior distribution for μ is $N(\mu_1, \sigma_1^2)$ with these parameters.

The mean of the posterior distribution μ_1 can be written as a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$

$$\mu_1 = (1-w)\mu_0 + w\bar{y}, \text{ where } w = \frac{1}{1 + \dfrac{\sigma^2}{n\sigma_0^2}}.$$

The weight $w \to 1$ as $n \to \infty$ or $\sigma_0 \to \infty$, hence in either of these limits, the posterior mean approaches the sample mean.

Example 2.5. Now consider example 2.4, but this time suppose that μ is known and the standard deviation σ is unknown.

The likelihood for the sample $y_1, \dots, y_n$ is

$$p(y \mid \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

Let $\tau = 1/\sigma^2$. τ is referred to as the precision. It turns out that we can find a more standard form for a conjugate distribution for τ than for σ. In terms of τ, the likelihood is

$$p(y \mid \tau) \propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

As a function of τ, this likelihood has the same functional form as a gamma probability density function. Hence we consider $\tau \sim \text{Gamma}(\alpha, \beta)$ as a prior distribution. The prior pdf is

$$p(\tau) = \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)}, \; \tau > 0$$

The posterior distribution is

$$\begin{aligned} p(\tau \mid y) &\propto p(\tau)\, p(y \mid \tau) \propto \tau^{\alpha-1} e^{-\beta\tau}\, \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum_{i-1}^{n}(y_i-\mu)^2} \\ &= \tau^{\alpha+\frac{n}{2}-1} e^{-\left(\beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)\tau} \end{aligned}$$

Hence the posterior distribution is

$$\tau \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}\right)$$

and we can say that the family of gamma distributions is conjugate to the likelihood for the normal precision parameter, if the mean μ is known.

If μ and τ are both unknown then there is a bivariate distribution which is conjugate. This is constructed by taking as a marginal distribution

$$\tau \sim \text{Gamma}$$

and then the conditional distribution

$$\mu \mid \tau \sim \text{Normal.}$$

The joint prior distribution is the product of these two pdfs. Combining the prior with the likelihood for the sample, the posterior is of the same form as the prior, but we do not go into the details here.

2.4 Point estimates and credible intervals

Once we have found a posterior distribution $p(\theta \mid y)$, in the Bayesian framework all our inference are based on this distribution. This includes point estimates. For a one-dimensional parameter θ, we can summarize the posterior distribution just as we would normally summarize a distribution, using familiar summaries such as the mean, median or mode. The median or mean are the most commonly used.

The mode (the value of θ which maximizes the posterior density) may be used in cases where it is difficult to find the posterior distribution. With a uniform distribution as a prior, i.e. $p(\theta)$ a constant, the posterior distribution is proportional to the likelihood, and so the posterior mode in this case is the same as the maximum likelihood estimate.

For the examples we have seen so far, there is a simple formula for the posterior mean. For the beta or gamma posterior distributions, there is no exact formula for the median. However, computer packages, including R, allow us to easily find the median for common probability distributions.

For a continuous random variable Θ with cumulative distribution function $F(\theta)$, we have by definition

$$P(\Theta \leq \theta) = F(\theta)$$

The quantile function is defined as the inverse function to the cdf

$$Q = F^{-1}$$

If $p = F(\theta)$ for some $p \in [0, 1]$, then $Q(p) = \theta$. In particular, if $p = 0.5$ and m is the median of the distribution, then $F(m) = 0.5$ and $Q(0.5) = m$.

For common distributions, there is a set of four functions on R to calculate the density (or probability mass function), the cdf, the quantile function and to generate random variates. For example for the gamma distribution, these functions are named `dgamma, pgamma, qgamma, rgamma`, respectively.

2.4.1 Credible intervals

In frequentist inference, confidence intervals are used to express a range of uncertainty around a parameter estimate. Suppose random samples y are repeatedly generated. For each sample we can estimate the true parameter θ by $\hat{\theta}(y)$, and also construct an interval. A 95% confidence interval is an interval $(\theta_L(y), \theta_U(y))$ with

$$P\left(\theta_L(y) \leq \theta \leq \theta_U(y)\right) = 0.95$$

It is tempting to say the interval $(\theta_L(y), \theta_U(y))$ calculated from the current dataset y contains the true value θ with probability 0.95. But such statements are only possible in the Bayesian framework. With frequentist confidence intervals, the probability statement refers to the probability over repeatedly generating datasets y. Frequentists consider θ to be a fixed (but unknown) quantity, not a random variable.

In the Bayesian framework, we do have a probability distribution for θ. After seeing the data, this is the posterior distribution $p(\theta \mid y)$. Based on this distribution, we can make probability statements about θ.

For some $\alpha \in [0, 1]$, a $100(1-\alpha)\%$ credible interval is some (θ_L, θ_U) such that

$$P(\theta_L < \theta < \theta_U) = 1 - \alpha$$

For example, $\alpha = 0.05$ for a 95% credible interval.

There are two main ways of defining credible intervals, equal tail intervals and highest posterior density intervals.

Equal tail intervals have equal probability mass above and below the interval (θ_L, θ_U),

$$\begin{aligned} P(\theta < \theta_L) &= \alpha/2 \\ P(\theta > \theta_U) &= \alpha/2. \end{aligned}$$

Highest posterior density intervals are chosen so that the probability density function has equal height at either end of the interval, $p(\theta_L \mid y) = p(\theta_U \mid y)$.

The equal tail interval is usually easier to calculate in practice, and so is more widely reported. If the posterior distribution is among the well-known named distributions, then just as for the median, we can use the quantile functions in R such as `qgamma` and `qnorm` to calculate equal tail credible intervals.

2.4.2 Transformed parameters

Suppose we have found the posterior distribution $p(\theta \mid y)$ for a parameter θ. Let $\psi = g(\theta)$ be a monotonic, increasing transformation of θ, such as $\psi = \log(\theta)$ or $\sqrt{\theta}$.

Some posterior summaries are preserved by the transformation. For example, suppose that θ_M is the posterior median for θ

$$P(\theta \leq \theta_M) = 0.5.$$

Then

$$P(\psi \leq g(\theta_M)) = P(g(\theta) \leq g(\theta_M)) = P(\theta \leq \theta_M) = 0.5$$

and so $g(\theta_M)$ is the posterior median for ψ.

Similarly, let (θ_L, θ_U) be a 95% equal tail credible interval for θ.

$$P(\theta \leq \theta_L) = 0.025,\ P(\theta \geq \theta_U) = 0.025.$$

Adapting the argument for the median leads to

$$P(\psi \leq g(\theta_L)) = 0.025,\ P(\psi \geq g(\theta_U)) = 0.025.$$

and so $(g(\theta_L), g(\theta_U))$ is a 95% equal tail credible interval for ψ.

This shows that the posterior median and equal tail credible intervals are preserved by increasing, one-to-one transformations.

The posterior mean is not in general preserved by non-linear transformations. As an example, suppose that $\psi = \theta^2$.

$$E(\psi) = E(\theta^2) = Var(\theta) + E(\theta)^2$$

Since in general $Var(\theta) > 0$, it follows that $E(\psi) \neq E(\theta)^2$.

The shape of a probability density changes under non-linear transformations of the random variable. This change of shape means that in general the posterior mode is not preserved by such a transformation, and nor are the endpoints of highest posterior density credible intervals.

2.4.3 Multiple parameters

If θ, the unknown parameters(s), is a vector, then we still base our estimates on $p(\theta \mid y)$. If we are interested in predictions of future data, then we use the entire joint distribution.

For point estimates of individual parameters, we typically use the marginal distribution. For example, if $\theta = (\theta_1, \theta_2, \theta_3)$, then the marginal posterior distribution for θ_1 is

$$p(\theta_1 \mid y) = \int \int p(\theta_1, \theta_2, \theta_3 \mid y)\ d\theta_2\ d\theta_3$$

In practical Bayesian inference, Markov Chain Monte Carlo methods, covered later in these notes, are the most common means of approximating the posterior distribution. These methods produce an approximate sample from the joint posterior density, and once this is done the marginal distribution of each parameter is immediately available.

Example 2.6. Generating samples from the posterior distribution may also be helpful even if we can calculate the exact posterior distribution. Suppose that the data are the outcome of a clinical trial of two treatments for a serious illness, the number of deaths after each treatment. Let the data be k_i deaths out of n_i patients, $i = 1, 2$ for the two treatments, and the two unknown parameters are q_1 and q_2, the probability of death with each treatment. Assuming a binomial model for each outcome, and independent $\text{Beta}(\alpha_i, \beta_i)$ priors for each parameter, the posterior distribution is

$$q_i \mid k_i \sim \text{Beta}(k_i + \alpha_i, n_i - k_i + \beta_i),\ i = 1, 2.$$

We have independent prior distributions and likelihood, so the posterior distributions are also independent.

$$p(q_1, q_2 \mid k_1, k_2) = p(q_1 \mid k_1)\, p(q_2 \mid k_2) \propto p(k_1 \mid q_1)p(q_1)\, p(k_2 \mid q_2)p(q_2)$$

However, it is useful to think in terms of the joint posterior density for q_1 and q_2, as then we can make probability statements involving both parameters. In this case, one quantity of interest is the probability $P(q_2 < q_1)$, i.e. does the second treatment have a lower death rate than the first. To find this probability, we need to integrate the joint posterior density over the relevant region, which is not possible to do exactly when it is a product of beta pdfs.

We can approximate the probability by generating a sample of (q_1, q_2) pairs from the joint density. To generate the sample, we just need to generate each parameter from its beta posterior distribution, which can be done in R using the `rbeta` command. Then once we have the sample, we just count what proportion of pairs has $q_2 < q_1$ to estimate $P(q_2 < q_1)$.

Chapter 3

Specifying prior distributions

The posterior distribution depends on both the observed data via the likelihood, and also on the prior distribution. So far, we have taken the prior distribution as given, but now we look at how to specify a prior.

3.1 Informative prior distributions

An informative prior distribution is one in which the probability mass is concentrated in some subset of the possible range for the parameter(s), and is usually based on some specific information. There may be other data that is relevant, and we might want to use this information without including all previous data in our current model. In that case, we can use summaries of the data to find a prior distribution.

Example 3.1. In example 2.3, the data was the lifetimes of light bulbs, $t = (t_1, \dots, t_n)$, assumed to be exponentially distributed with parameter λ (the failure rate, reciprocal of the mean lifetime). The gamma distribution provides a conjugate prior distribution for λ. Suppose that we had information from several other similar types of light bulbs, which had observed failure rates $r_1, \dots, r_k$. Let the sample mean and variance of these rates be m and v.

A Gamma(α, β) distribution has mean $\dfrac{\alpha}{\beta}$ and variance $\dfrac{\alpha}{\beta^2}$. We can match these to m and v.

$$m = \frac{\alpha}{\beta}, v = \frac{\alpha}{\beta^2} = \frac{m}{\beta}$$

Rearranging gives

$$\beta = \frac{m}{v}, \alpha = \beta m = \frac{m^2}{v}.$$

Hence we can use these values of α and β as the parameters of our prior distribution.

Other examples of summary data include published estimates and their standard errors or confidence intervals. We can match these quantities to the mean (or median), standard deviation or percentiles of the prior distribution, as appropriate. In this kind of example there is doubt over whether the published estimate is really measuring the same quantity as the parameter in our current model represents, and so it may be a good idea to increase the uncertainty, by increasing the standard error or width of the confidence interval before matching to the prior distribution parameters.

3.2 Less informative prior distributions

If there is not specific numerical information upon which to base a prior, then one may aim to choose a prior distribution which conveys as little information as possible. However, there is no unique way of doing this.

We could choose a flat prior distribution, i.e. uniform over some range. If the full range of the parameter is unbounded, then we could assign a uniform distribution on some range that includes all plausible values for the parameter. For example suppose the parameter is a standard deviation σ. We might choose

$$p(\sigma) = 1/c, \ 0 < \sigma < c$$

for some large c. But the prior will not be uniform for transformations of σ such as $\log(\sigma)$ or σ^2.

There are more formal methods to attempt to specify uninformative prior distributions, but we do not cover them in this module.

3.2.1 Weakly informative prior distributions

Instead of trying to make the prior distribution completely uninformative, an alternative is to use the prior to convey some information about the plausible range of the parameters, but otherwise let the data speak for themselves.

In regression models, either linear or generalized linear models, after rescaling the covariates to have a standard deviation of 1, we could choose a prior distribution centred around zero, to convey the information that extremely large effects on the outcome are unlikely, for example a normal distribution $N(0, s^2)$. The scale parameter of the prior distribution (s here) would be based on what magnitude of effects had been found in the past in analyses of similar types of outcomes.

An alternative to a normal distribution is the Cauchy distribution, which may be preferred as it has heavier tails, so if the data do strongly suggest a large effect then this can be reflected in the posterior distribution.

Chapter 4

Markov chain Monte Carlo methods

The examples covered so far have been conjugate, in which case we can derive an exact formula for the posterior distribution, including the normalizing constant. However, this is only possible in a few simple cases. Bayesian methods are in theory applicable to complicated models with many parameters. Computational methods that have come into widespread use in the past three decades have made practical Bayesian inference possible for a much wider range of examples than was previously the case. The main computational methods used are various types of Markov chain Monte Carlo methods.

The term "Monte Carlo methods" refers to any method that generates random samples in order to do some calculation that cannot easily be done by non-random methods. Simple Monte Carlo methods are those where we can generate an independent sample from a probability distribution that we want to make statements about. In Bayesian inference this would be the posterior density $p(\theta \mid y)$. Example 2.6 was of this type, as it was possible to generate a random sample from the joint posterior distribution of q_1 and q_2. This sample could then be used to estimate $P(q_2 < q_1)$, which is not possible to calculate exactly.

4.1 The Metropolis algorithm

In general, it is not possible to generate an independent random sample of θ values from the posterior distribution $p(\theta \mid y)$, and so simple Monte Carlo cannot be used. Markov chain Monte Carlo methods generate a sample from a probability distribution that is not independent, rather each element in the sample is correlated with the previous element.

A Markov chain is a sequence $\theta_1, \theta_2, \ldots$ of random variables. The probability distribution of θ_i only depends on the previous value θ_{i-1}

$$p(\theta_i \mid \theta_1, \theta_2, \ldots, \theta_{i-1}) = p(\theta_i \mid \theta_{i-1}).$$

The Metropolis algorithm is a type of Markov chain Monte Carlo (MCMC). Let $h(\psi;\theta)$ be a probability density function for ψ which is symmetric in ψ and θ, for example the normal pdf $\phi(\psi;\theta,s)$. Let f be another pdf. The algorithm constructs a Markov chain $\theta_1, \theta_2, \theta_3, \ldots$, where the θ_i are continuous rvs (in our applications).

The algorithm constructs the Markov chain as follows:

- Start with θ_1, where $f(\theta_1) > 0$.
- For each $i > 1$, generate ψ_i from the distribution $h(\psi_i;\theta_{i-1})$.
- Define $r = \min\left(1, \dfrac{f(\theta_i)}{f(\theta_{i-1})}\right)$.
- Set $\theta_i = \begin{cases} \psi_i & \text{with probability } r \\ \theta_{i-1} & \text{with probability } 1-r. \end{cases}$

h is called the proposal distribution, and r is the acceptance probability. f is sometimes called the target distribution: this is what we are aiming for, i.e. we want to generate a sample with pdf f.

The sequence $\theta_1, \theta_2, \ldots$ has the property that if $\theta_{i-1} \sim f$, then $\theta_i \sim f$, in other words f is the equilibrium distribution of the chain. However, we don't start with $\theta_1 \sim f$, because if we could, then we would not need this algorithm. But for large enough i, if some technical conditions are met, then each $\theta_i \sim f$ approximately. (The notation $\theta \sim f$ is used here to mean that θ is distributed with pdf f.)

The algorithm eventually produces points distributed with pdf f, but it does not produce an independent random sample. Nevertheless, we can still use the sample to make inferences about f. Often, we discard some initial number of steps B in order to reduce the influence of the choice of starting value θ_1. B is known as the burn-in. Then, we would continue the chain for some large number M of additional steps, so that inferences about the sample are based on $\theta_{B+1}, \theta_{B+2}, \ldots, \theta_{B+M}$.

The proposal distribution h is often taken as a normal distribution centred on the current point

$$\psi_i \sim N(\theta_{i-1}, s^2).$$

The pdf h is symmetric in θ and ψ, as required by the algorithm

$$h(\psi;\theta) = \frac{1}{\sqrt{2\pi}s} e^{-\dfrac{(\psi-\theta)^2}{2s^2}} = \frac{1}{\sqrt{2\pi}s} e^{-\dfrac{(\theta-\psi)^2}{2s^2}} = h(\theta;\psi).$$

The acceptance probability r is

$$r = \min\left(1, \frac{f(\psi_i)}{f(\theta_{i-1})}\right).$$

If the proposal scale s (the standard deviation of h) is very small, then ψ_i is close to θ_{i-1}, and so $f(\psi_i)$ is close to $f(\theta_{i-1})$. Hence there is a high probability of accepting the proposal, but it is only a small step. If s is large, then ψ_i may be far from θ_{i-1}, and $f(\psi_i)$ may be much lower than $f(\theta_{i-1})$. Now there is a lower probability of accepting the proposal, but it is a larger step.

Ideally we would want large steps which are often accepted, to reduce the correlation between successive θ_i values, but in practice some compromise between taking large steps and having a high acceptance probability is needed. This means that some intermediate value for s tends to be best for reducing the correlation. The best value for s depends on the model and the data. There is a general rule of thumb, which is to aim for an acceptance probability $r = 0.23$. This number is calculated based on f being a multivariate normal density, but using simulation it has been shown to work well in a wider range of problems.

4.1.1 Relevance to Bayesian inference

The Metropolis algorithm produces an approximate sample from a general distribution with pdf f. In the algorithm, f appears in acceptance probability

$$r = \min\left(1, \frac{f(\psi_i)}{f(\theta_{i-1})}\right).$$

As f only appears in a ratio, we only need to know f up to a constant. If $f(\theta) = cg(\theta)$, where c does not involve θ, then we could use g instead of f.

In Bayesian inference, the posterior density is

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\; d\theta} \propto p(\theta)\, p(y \mid \theta).$$

It is difficult to find the normalizing constant $\int p(\theta)p(y \mid \theta)\; d\theta$. When using the Metropolis algorithm, there is no need to find this constant. We can put $g(\theta) = p(\theta)\, p(y \mid \theta)$, and use g in the algorithm for calculating the acceptance probability r, in place of f. This will generate a Markov chain $\theta_1, \theta_2, \ldots$ which is an approximate sample from $p(\theta \mid y)$.

4.1.2 Metropolis algorithm implementation

The following description of how the algorithm can be implemented for Bayesian inference is in terms of a single scalar parameter θ, but the general method can be used when θ is a vector of parameters.

The most common choice of proposal distribution is a normal distribution, and that is all we will consider in this module for the Metropolis algorithm.

Assume we have observed data y and have specified a probability model that depends on a parameter θ, so that we can compute the likelihood $p(y \mid \theta)$. The prior distribution is $p(\theta)$.

Define $g(\theta) = p(\theta)\,p(y \mid \theta)$, the non-normalized posterior density. The Metropolis algorithm proceeds as follows:

- Choose some $s > 0$.
- Start with θ_1, where $g(\theta_1) > 0$.
- For each $i > 1$:
    1. Generate $\psi_i \sim N(\theta_{i-1}, s^2)$.
    2. Define $r = \min\left(1, \dfrac{g(\psi_i)}{g(\theta_{i-1})}\right)$.
    3. Generate $u \sim Uniform(0,1)$, the continuous uniform distribution on $[0,1]$.
    4. Set $\theta_i = \begin{cases} \psi_i & \text{if } u \leq r \\ \theta_{i-1} & \text{if } u > r. \end{cases}$

Steps 1 to 4 are repeatedly carried out to generate a sequence $\theta_1, \theta_2, \theta_3, \ldots$ with some repeated values, i.e. $\theta_i = \theta_{i-1}$ if the proposed value ψ_i was rejected.

Step 1 is equivalent to setting $\psi_i = \theta_{i-1} + sz$, where $z \sim N(0,1)$.

Steps 3 and 4 are how in a computer implementation we would set $\theta_i = \psi_i$ with probability r, and $\theta_i = \theta_{i-1}$ otherwise.

Note that if $r^* = \dfrac{g(\psi_i)}{g(\theta_{i-1})} > 1$, then $u < r^*$ with probability 1, hence in steps 2 and 4 we could use r^* instead of r and the algorithm would not change.

Working on the log scale

In realistic applications, all computations are usually done using the log of the posterior density. The likelihood is typically a product of many terms

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta).$$

Due to the finite accuracy of computers, if we multiply the likelihood contributions $p(y_i \mid \theta)$ together for a large dataset, the result is inaccurate. So we calculate

$$\log\left(p(y \mid \theta)\right) = \sum_{i=1}^{n} \log\left(p(y_i \mid \theta)\right).$$

Define $\mathcal{L}(\theta) = \log\left(p(\theta)\,p(y \mid \theta)\right) = \log\left(p(\theta)\right) + \log\left(p(y \mid \theta)\right)$, the log of the posterior density (up to a constant).

To work on the log scale, the generation of the proposed value ψ_i is unchanged, but the part of the algorithm with the acceptance step changes.

For each $i > 1$:

1. Generate $\psi_i \sim N(\theta_{i-1}, s^2)$, just as before.
2. Define $\delta = \mathcal{L}(\psi_i) - \mathcal{L}(\theta_{i-1})$.
3. Generate $u \sim Uniform(0, 1)$, the continuous uniform distribution on $[0, 1]$.
4. Set $\theta_i = \begin{cases} \psi_i & \text{if } \log(u) \leq \delta \\ \theta_{i-1} & \text{if } \log(u) > \delta. \end{cases}$

Here, $\delta = \log(r^*)$ from section 4.1.2.

Summarizing the posterior distribution

The Metropolis algorithm generates an approximate sample from the posterior distribution $p(\theta \mid y)$, $\theta_1, \theta_2, \ldots, \theta_M$. This is not an independent sample, since each θ_i is correlated with the previous element in the sequence θ_{i-1}. But we can still summarize it like an independent sample. The sample mean can be used to estimate the posterior mean of θ, and similarly for the median. The 2.5th and 97.5th sample percentiles can be used as a 95% equal tail credible interval for θ.

It is common to discard the initial part of the sample, the first B values for some $B > 0$, to reduce the influence of the choice of starting value θ_1. B is called the burn-in.

Multiple parameters

Usually θ is a vector of parameters of length k. The Metropolis algorithm still proceeds in the same way in this case. We generate a Markov chain where each element of the sequence is a vector θ_i. The only change is to step 1, the proposal of candidate parameter values. For each $i > 1$, we can generate

$$\psi_i \sim N_k(\theta_{i-1}, \Sigma)$$

a multivariate normal proposal distribution with mean vector θ_{i-1} and covariance matrix Σ, where Σ is chosen at the start.

Chapter 5

Predictive distributions

In the Bayesian framework, all our inferences about θ are based on the posterior distribution $p(\theta \mid y)$. As well as the parameter(s) θ, we may be interested in new data x. Inference about new data is also based on $p(\theta \mid y)$.

Assume that x is independent of y (conditional on θ), but is generated by the same probability model. The posterior predictive distribution for x given y is

$$p(x \mid y) = \int p(x, \theta \mid y)\, d\theta.$$

From standard rules for conditional distributions

$$p(x, \theta \mid y) = p(x \mid \theta, y)\, p(\theta \mid y)$$

x is assumed to be independent of y (given θ), and so $p(x \mid \theta, y) = p(x \mid \theta)$.

Therefore the posterior predictive distribution is given by

$$p(x \mid y) = \int p(x \mid \theta)\, p(\theta \mid y)\, d\theta.$$

This is the probability distribution for unobserved or future data x. It takes into account the uncertainty remaining about θ after we have seen y as well as the random variation in x given any particular value for θ.

In conjugate examples, it is usually possible to derive an expression for the posterior predictive distribution

$$p(x \mid y) = \int p(x \mid \theta)\, p(\theta \mid y)\, d\theta.$$

However, this tends to be less simple than finding $p(\theta \mid y)$. Also, in most realistic problems there will not be a conjugate posterior distribution.

It is generally easier to find the mean and variance of $p(x \mid y)$ than deriving the full distribution. Suppose that X and W are general random variables. Then

$$E(X) = E(E(X \mid W))$$

and

$$Var(X) = Var(E(X \mid W)) + E(Var(X \mid W)).$$

In these identities, we replace W by the parameters and X by the new data we would like to predict.

Example 5.1. Suppose that the observed data is $k \sim \text{Bin}(n, q)$. With a $\text{Beta}(\alpha, \beta)$ prior for q, the posterior distribution is $\text{Beta}(a, b)$, where $a = k + \alpha,\ b = n - k + \beta$.

Suppose the data we want to predict is $x \sim \text{Bin}(m, q)$, where m is known. The posterior predictive distribution is

$$p(x \mid k) = \int_0^1 p(x \mid q)\, p(q \mid k)\, dq.$$

This integral is possible to do, but involves working with beta or gamma functions. We can find the mean and variance of x directly. Put

$$\mu = \frac{a}{a+b},\ v = \frac{ab}{(a+b)^2(a+b+1)},$$

the posterior mean and variance of q. The general formula for conditional expectations gives

$$E(x) = E(E(x \mid q)).$$

We want expectations in the posterior distribution, so conditioning on k we have

$$E(x \mid k) = E(E(x \mid q, k))$$

$$E(x \mid q, k) = E(x \mid q) = mq$$

hence the posterior predictive mean is

$$E(x \mid k) = E(mq \mid k) = mE(q \mid k) = m\mu.$$

For the variance,

$$Var(x \mid q, k) = Var(x \mid q) = mq(1 - q).$$

Hence the posterior predictive variance is

$$\begin{aligned}
Var(x \mid k) &= E(Var(x \mid q, k)) + Var(E(x \mid q, k)) \\
&= E(mq(1 - q) \mid k) + Var(mq \mid k) \\
&= mE(q \mid k) - mE(q^2 \mid k) + m^2 Var(q \mid k) \\
&= mE(q \mid k) - m\left(Var(q \mid k) + E(q \mid k)^2\right) + m^2 Var(q \mid k) \\
&= m\mu - m(v + \mu^2) + m^2 v.
\end{aligned}$$

The general procedure is to find the mean and variance of the predicted data conditional on a specific parameter or parameter vector, and then to find the mean and variance over the posterior distribution of parameters.

5.1 Simulating the posterior predictive distribution

Suppose that we have generated a sample $\theta_1, \theta_2, \ldots, \theta_M$ from the posterior distribution $p(\theta \mid y)$. Given this sample, we can simulate the posterior predictive distribution $p(x \mid y)$. We just generate random values

$$x_j \text{ from the distribution } p(x \mid \theta_j),\ j = 1, 2, \ldots, M.$$

Then

$$x_1, x_2, \ldots, x_M$$

is a sample from the posterior predictive distribution.

Since the usual computational method for Bayesian inference is to use a form of MCMC such as the Metropolis algorithm, we typically will have a sample of θ values.

Once we have a sample from the distribution $p(x \mid y)$, we can summarize the sample to estimate various quantities. For example, the sample mean of $x_1, x_2, \ldots, x_M$ is an estimate of the posterior predictive mean of x.

To estimate the posterior predictive probability that x is less than some value a, simply count what proportion of the sample is less than a. If the data is discrete, to estimate the posterior predictive probability that $x = 0$, count what proportion of the sample is equal to 0.

The sample quantiles of $x_1, \ldots, x_M$ can be used to form a prediction interval. For example, the 2.5th and 97.5th sample quantiles form a 95% posterior predictive interval. A new data-point x will be inside this interval with probability 0.95.

Chapter 6

Hierarchical models

So far we have considered a single parameter or vector of parameters θ, and the observed data $y_1, \ldots, y_n$ are independent given θ. One way of generalizing statistical models is to account for the fact that data are grouped in some way. In that case, we can add more levels to the model. In this module, we only consider two-level models, but more complicated structures follow along the same lines.

Suppose the data are in n groups, with m_i observations in group i. Now the data is

$$y = \{y_{ij} : i = 1, \ldots, n, j = 1, \ldots, m_i\}.$$

$y_{i1}, \ldots, y_{im_i}$ depend on a set of parameters $\psi_i, i = 1, \ldots, n$. These parameters $\psi_1, \ldots, \psi_n$ are assumed to be generated by a probability distribution with parameters θ. $p(\psi_i \mid \theta)$ acts like a prior distribution for ψ_i, but using a hierarchical model we can also estimate θ from the data.

The joint probability density for ψ and y, given θ, is

$$p(\psi, y \mid \theta) = \prod_{i=1}^{n} p(\psi_i \mid \theta) \prod_{j=1}^{m_i} p(y_{ij} \mid \theta, \psi_i) = p(\psi \mid \theta)\, p(y \mid \theta, \psi)$$

where $\psi = (\psi_1, \ldots, \psi_n)$.

In the Bayesian framework, we specify a prior distribution $p(\theta)$ but not for ψ, since we have the probability distribution $p(\psi \mid \theta)$.

For a one-level model, Bayes' theorem is

$$p(\theta \mid y) \propto p(\theta)\, p(y \mid \theta)$$

With a two-level model, we are estimating both θ and ψ, and so we have

$$p(\theta, \psi \mid y) \propto p(\theta, \psi)\, p(y \mid \theta, \psi).$$

$$p(\theta, \psi) = p(\theta)\ p(\psi \mid \theta)$$

and so

$$p(\theta, \psi \mid y) \propto p(\theta)\ p(\psi \mid \theta)\ p(y \mid \theta, \psi) = p(\theta)\ p(\psi, y \mid \theta).$$

$p(\psi, y \mid \theta)$ is taken as the likelihood, although only y and not ψ is the observed data.

Example 6.1. Assume the data are the times until failure for a batch of light-bulbs, $t_1, \ldots, t_m$. These times follow an exponential distribution with parameter λ, which is given a Gamma(α, β) prior distribution. In example 3.1, it was assumed that we have estimates of the failure rate from k previous studies, $r_1, \ldots, r_k$. The sample mean and variance of these estimates were used to find the two gamma distribution parameters.

Suppose that we had the data from the previous studies, $t_{i1}, \ldots, t_{im_i},\ i = 1, \ldots, k$. We could fit a single model to the entire set of data - a hierarchical model would make sense here.

The number of groups (batches of light-bulbs) is $n = k + 1$. Each group i has its own failure rate λ_i, which we can take as following a Gamma(α, β) distribution. Conditional on λ_i, the data $t_{i1}, \ldots, t_{im_i} \sim$ Exp(λ_i). We would also formulate a prior distribution for α and β. In the general notation of the previous section, the top-level parameter θ is the vector (α, β) here, and the group-level parameter ψ_i is λ_i here.

The hierarchical model is more coherent statistically than the approach of 3.1, as it treats each batch of light-bulbs the same. Also, it correctly accounts for the uncertainty in the values of α and β.

6.1 Inference for hierarchical models

We estimate the joint posterior distribution of θ and ψ, $p(\theta, \psi \mid y)$.

When using MCMC, this means generating a sample of all these quantities, $\theta, \psi_1, \ldots, \psi_n$. The simple Metropolis algorithm will be less efficient when n is large, as we need to sample many parameters. However there are other MCMC methods which can be used, such as Gibbs sampling. We are not covering those methods in this module.

Once we have a sample from the posterior density $p(\theta, \psi_1, \ldots, \psi_n \mid y)$, we can treat this just like any other joint posterior density. To obtain a sample from the marginal posterior density of any subset of parameters, just ignore the other elements of the sample. For example, to get a sample from $p(\theta \mid y)$, ignore $\psi_1, \ldots, \psi_n$.

6.1.1 Hierarchical model posterior predictions

Hierarchical models are useful for predictions. Now that there is more than one level to the model, we can estimate the posterior predictive distribution of:

- a new data-point in a group in the dataset;
- a new data-point in a group not in the dataset.

Suppose that we have a sample of size M from the posterior distribution $p(\theta, \psi_1, \ldots, \psi_n \mid y)$:

$$\begin{array}{l} \theta_1, \psi_{11}, \ldots, \psi_{n1} \\ \theta_2, \psi_{12}, \ldots, \psi_{n2} \\ \ldots \\ \theta_M, \psi_{1M}, \ldots, \psi_{nM} \end{array}$$

To predict a new data-point x in group i, generate $x_k \sim p(x \mid \psi_{ik}, \theta_k)$ for $k = 1, \ldots, M$.

To predict a new data-point z in a new group, generate $\tilde{\psi}_k \sim p(\psi \mid \theta_k)$, then $x_k \sim p(x \mid \tilde{\psi}_k, \theta_k)$ for $k = 1, \ldots, M$.

Example 6.2. Suppose the data are the observed numbers of cases of a certain disease

$$y = \{y_{ij} : i = 1, \ldots, n, j = 1, \ldots, m_i\}.$$

y_{ij} is the number of cases in district j within county i. The population of the district is N_{ij}. We assume that

$$y_{ij} \sim \text{ Poisson}(\lambda_i N_{ij}).$$

λ_i is the disease rate in county i, with $\lambda_i \sim$ Gamma(α, β), $i = 1, \ldots, n$.

α and β are given prior distributions, for example half-normal. The joint prior distribution is $p(\alpha, \beta)$.

The posterior density is

$$p(\lambda_1, \ldots, \lambda_n, \alpha, \beta \mid y) \propto p(\alpha, \beta) \prod_{i=1}^{n} \left[p(\lambda_i \mid \alpha, \beta) \prod_{j=1}^{m_i} p(y_{ij} \mid \lambda_i) \right].$$

$p(\lambda_i \mid \alpha, \beta)$ is the Gamma probability density function with outcome λ_i. $p(y_{ij} \mid \lambda_i)$ is the Poisson probability mass function for outcome y_{ij} with Poisson distribution parameter (mean) $\lambda_i N_{ij}$.

MCMC methods can be used to generate a sample of size M from $p(\lambda_1, \ldots, \lambda_n, \alpha, \beta \mid y)$:

$$\begin{array}{l} \lambda_{11}, \ldots, \lambda_{n1}, \alpha_1, \beta_1 \\ \lambda_{12}, \ldots, \lambda_{n2}, \alpha_2, \beta_2 \\ \ldots \\ \lambda_{1M}, \ldots, \lambda_{nM}, \alpha_M, \beta_M \end{array}$$

Then to make inferences about λ_1, we just take the first column, $\lambda_{11}, \ldots, \lambda_{1M}$, and this is a sample from the marginal posterior distribution $p(\lambda_1 \mid y)$.

To make inferences about $\mu = \alpha/\beta$, set

$$\mu_k = \frac{\alpha_k}{\beta_k}, \; k = 1, \ldots, M.$$

Then $\mu_1, \ldots, \mu_M$ is a sample from the marginal posterior distribution $p(\mu \mid y)$, and we can use the sample median and quantiles to estimate the posterior median and credible interval limits just as we have done previously for one-parameter models.

Two types of posterior predictions that we can make are:

1. The posterior predictive distribution for the number of cases x in a new district (i.e. a district not in our dataset), which is in county i, where county i does appear in our dataset. Let P be the population of this district.

   The procedure here is to generate

   $$x_k \sim \text{ Poisson}(\lambda_{ik} P), \; k = 1, \ldots, M.$$

   Then $x_1, \ldots, x_M$ is a sample from the posterior predictive distribution.

2. The posterior predictive distribution for the number of cases z in a new district, which is in a county that is not in our dataset.

   Now we need to generate a posterior predictive sample for the disease rate in the county as well. Let Q be the population of this district. The procedure is to generate

   $$\tilde{\lambda}_k \sim \text{ Gamma}(\alpha_k, \beta_k), \; k = 1, \ldots, M,$$

   and then

   $$z_k \sim \text{ Poisson}(\tilde{\lambda}_k Q), \; k = 1, \ldots, M.$$

   Now $z_1, \ldots, z_M$ is a sample from the posterior predictive distribution.

   The sample mean of $z_1, \ldots, z_M$ is an estimate of the posterior predictive mean for z, and the 2.5th and 97.5th percentiles of the sample form a 95% posterior predictive interval for z.

   To find the posterior predictive probability that z will be zero, we just count what proportion of the sample is equal to zero.

6.2 Advantages of hierarchical models

Pooling information

We can pool information while allowing some variation. Pooling information means using information from multiple groups to estimate the distribution $p(\psi \mid \theta)$. Then in groups with little data (e.g. in example 6.2, few districts or only districts with small populations), the posterior mean of ψ is close to the mean of $p(\psi \mid y)$. If the data in group i provides more information (e.g. many districts or some districts with large populations), then the posterior mean of ψ_i is mainly determined by $y_{11}, \ldots, y_{im_i}$.

Posterior predictions

We can make posterior predictions for more quantities than with a non-hierarchical model.

We could analyse data from each group separately. In that case, we can predict a new observation in the same group, as this situation is just a single-level model, estimating ψ_i from the sample $y_{i1}, \ldots, y_{im_i}$. But we can't predict a new observation in a group not in our dataset.

With a hierarchical model, the probability distribution $p(\psi \mid \theta)$ allows us to generalize to a new group.

Correctly accounting for uncertainty

If there is variation between groups in the parameter(s) ψ_i, this implies that the observations in the same group i are correlated with each other. Hence if we fitted a single-level model that assumed the data-points are independent, then we would tend to underestimate the uncertainty in the parameter estimates. So credible intervals (or confidence intervals in the frequentist framework) would be too narrow.

Reducing the influence of arbitrary prior distributions

The hierarchical structure moves the prior distribution further away from the inferences. In example 6.2, the disease rates in each county follow a gamma distribution

$$\lambda_i \sim \text{ Gamma}(\alpha, \beta),\ i = 1, \ldots, n.$$

If instead we fitted a non-hierarchical model, we could fix α and β as prior parameters for each λ_i.

In the hierarchical model we define a prior distribution $p(\alpha, \beta)$, and estimate α and β. So with the hierarchical model, the choice of prior distribution has less influence on our inferences about λ_i. The level at which we choose fixed prior parameters is moved one level further away from the λ_i.

Chapter 7

Tests and model comparisions

The final topic is a brief look at Bayesian versions of hypothesis tests and model comparisons.

Example 7.1. Consider the normal example 2.4, in which the data $y_1, \ldots, y_n \sim N(\mu, \sigma^2)$, and σ is known. Suppose that we are interested in deciding whether or not μ is zero.

In the frequentist framework, we have a null and an alternative hypothesis. In this case, the null hypothesis would be $H_0 : \mu = 0$. As a one-sided alternative hypothesis, take $H_1 : \mu > 0$. The null hypothesis is tested using the p-value. This is the probability of observing a statistic at least as extreme as the observed value, if H_0 is true.

Since σ is known, we can use the sample mean $\bar{Y}$ as a test statistic. If H_0 is true, then $\bar{Y} \sim N(0, \sigma^2/n)$. If we observe $\bar{Y} = \bar{y}$, then the one-sided p-value is

$$P(\bar{Y} \geq \bar{y} \mid \mu = 0) = P\left(\frac{\sqrt{n}\bar{Y}}{\sigma} \geq \frac{\sqrt{n}\bar{y}}{\sigma} \mid \mu = 0\right) = 1 - \Phi\left(\frac{\sqrt{n}\bar{y}}{\sigma}\right).$$

The Bayesian framework does not use p-values. Instead, probability statements are based on the posterior distribution $p(\mu \mid y)$. We can use this distribution to calculate posterior probabilities such as $P(\mu > 0 \mid y)$. With prior distribution $\mu \sim N(\mu_0, \sigma_0^2)$, the posterior distribution is

$$\mu \sim N(\mu_n, \sigma_n^2), \text{ where } \mu_n = \frac{\mu_0/\sigma_0^2 + n\bar{y}/\sigma_2}{1/\sigma_0^2 + n/\sigma_2}, \; \sigma_n^2 = \frac{1}{1/\sigma_0^2 + n/\sigma_2}.$$

For sufficiently large σ_0^2, i.e. for an uninformative prior distribution,

$$\mu_n \approx \bar{y}, \; \sigma_n^2 \approx \sigma^2/n.$$

Hence the posterior probability

$$P(\mu \leq 0 \mid y) = P\left(\frac{\sqrt{n}(\mu - \bar{y})}{\sigma} \leq -\frac{\sqrt{n}\bar{y}}{\sigma} \mid y\right) \approx \Phi\left(-\frac{\sqrt{n}\bar{y}}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}\bar{y}}{\sigma}\right).$$

This is the same numerical value as the p-value. Hence, the posterior probability and the p-value may be the same or similar, but they do not have to be, as they are probabilities of different quantities.

Example 7.2. For another example, suppose we want to know if a coin is fair, or if it is biased towards heads. The coin is tossed n times and we observe k heads. Let the probability of heads be q, and take $k \sim \text{Bin}(n, q)$. A frequentist approach would set

$$H_0 : q = 0.5$$
$$H_1 : q > 0.5$$

Label the random variable for the number of heads as X. The one-sided p-value for testing H_0 is

$$P(X \geq k \mid q = 0.5).$$

Suppose $n = 5$, and that we observed $k = n$. Then the p-value is $P(X = k \mid q = 0.5) = 0.5^5 = 0.0313$.

For a Bayesian version, we can take a uniform prior distribution for q. Then the posterior distribution is $\text{Beta}(k + 1, n - k + 1)$. If $k = n = 5$, the posterior is $\text{Beta}(6, 1)$. The normalized posterior density and cdf are

$$p(q \mid k) = 6q^5, \; F(q) = q^6.$$

The posterior probability that $q \leq 0.5$ is

$$P(q \leq 0.5 \mid k) = F(0.5) = 0.5^6 = 0.0156.$$

Here we have two different numbers from a frequentist and Bayesian approach, although in this case they are not hugely different.

In these two examples, we compared one-sided tests with posterior probabilities. With a continuous prior and posterior distribution, the posterior probability $P(\mu \neq 0 \mid y)$ is 1, since any single point has zero probability, and so there is no posterior probability to compare to the two-sided test.

Multiple models

So far, we have been using Bayes' theorem in the form

$$p(\theta \mid y) \propto p(\theta)\, p(y \mid \theta).$$

We have usually only needed $p(\theta \mid y)$ up to a constant of proportionality. With conjugate distributions, we can recognise the full posterior density, and when using the Metropolis algorithm, the normalizing constant cancels out.

Bayes' theorem in full is

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{p(y)}.$$

The denominator is

$$p(y) = \int p(\theta)\, p(y \mid \theta)\; d\theta.$$

This integral is usually harder to calculate compared to the methods for inference that we have covered so far, especially if θ is multi-dimensional. But it is used in one method for Bayesian model comparison.

Suppose that we have more than one candidate model that might fit the data. Label the models $M_1, M_2, \ldots, M_r$. We assume that one of these models generated the data. Each model has a vector of parameters $\theta_k,\ k = 1, \ldots, r$.

For example, in example 7.1, we could take models

$$\begin{aligned} M_1 &: \ y_i \sim N(0, \sigma_2) \\ M_2 &: \ y_i \sim N(\mu, \sigma_2) \end{aligned}$$

We expand the notation to make probabilities conditional upon a particular model. The prior distribution within model M_k becomes $p(\theta_k \mid M_k)$ and the likelihood is $p(y \mid \theta_k, M_k)$. Bayes' theorem for model M_k becomes

$$p(\theta_k \mid y, M_k) = \frac{p(\theta_k \mid M_k)\, p(y \mid \theta_k, M_k)}{p(y \mid M_k)}.$$

The term $p(y \mid M_k)$ can be used in Bayes' theorem in another way, for looking at probabilities of different models rather than of parameter values within one model. To do this, we need to specify prior probabilities for each model, $p(M_k),\ k = 1, \ldots, r$. We could choose a discrete uniform distribution

$$p(M_k) = \frac{1}{r},\ k = 1, \ldots, r.$$

However, we do not have to choose this distribution.

Then Bayes' theorem is applied to give

$$p(M_k \mid y) = \frac{p(M_k)\, p(y \mid M_k)}{p(y)}$$

where

$$p(y \mid M_k) = \int p(\theta_k \mid M_k)\, p(y \mid \theta_k, M_k)\, d\theta_k.$$

The denominator is

$$p(y) = \sum_{j=1}^{r} p(M_j)\, p(y \mid M_j).$$

$p(M_k \mid y)$ is the posterior probability that model M_k is the true model. This provides a a Bayesian method for choosing between models.

Bayes factors

Suppose that we are just considering two models, M_1 and M_2. The ratio of posterior probabilities is

$$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(M_1)\, p(y \mid M_1)}{p(M_2)\, p(y \mid M_2)}$$

In general, for a probability p, the term *odds* means $\dfrac{p}{1-p}$.

The prior odds of model M_1 vs M_2 is

$$\frac{p(M_1)}{p(M_2)} = \frac{p(M_1)}{1 - p(M_1)}$$

and the posterior odds of model M_1 vs M_2 is

$$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(M_1 \mid y)}{1 - p(M_1 \mid y)}.$$

These are the first two terms in the ratio of posterior probabilities. The third term is called the Bayes factor B_{12}:

$$\begin{aligned} B_{12} &= \frac{p(y \mid M_1)}{p(y \mid M_2)} \\ &= \frac{\int p(\theta_1 \mid M_1)\, p(y \mid \theta_1, M_1)\, d\theta_1}{\int p(\theta_2 \mid M_2)\, p(y \mid \theta_2, M_2)\, d\theta_2} \end{aligned}$$

We have:

$$\text{Posterior odds} = \text{prior odds} \times \text{Bayes factor}$$

$p(\theta_k \mid M_k)$ and $p(y \mid \theta_k, M_k)$ are the prior and likelihood for model M_k. Both of these needs to be properly normalized (unless the same constant appears in model M_1 and M_2).

A Bayes factor B_{12} greater than 1 supports model M_1, whereas a value less than 1 supports model M_2. Rules of thumb for the size of the Bayes factor have been suggested, for example:

Range of B_{12}			Evidence for M_1
1	to	$10^{\frac{1}{2}}$	slight
$10^{\frac{1}{2}}$	to	10	moderate
10	to	10^2	strong
	$>$	10^2	decisive

Example 7.3. Continuing the example of the fair coin 7.2, now let there be two models

$$\begin{aligned} M_1 &: \ k \sim \text{Bin}(n, 0.5) \\ M_2 &: \ k \sim \text{Bin}(n, q) \end{aligned}$$

Model M_1 has no unknown parameters, and model M_2 has one, namely q.

For model M_1, there are no parameters to integrate over, and so $p(k \mid M_1)$ is just the probability of the observed data if $q = 0.5$. If as before we observe $k = n$, this is

$$p(k \mid M_1) = 0.5^n.$$

For model M_2

$$p(k \mid M_2) = \int_0^1 p(q \mid M_2)\, p(k \mid q, M_2)\, dq.$$

$p(q \mid M_2)$ is the prior distribution for q and $p(k \mid q, M_2)$ is the likelihood. For a general Beta(α, β) prior distribution, we have

$$p(k \mid M_2) = \int_0^1 \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)} \binom{n}{k} q^k (1-q)^{n-k}\, dq$$

where $B(\alpha, \beta)$ is the Beta function. The constants $B(\alpha, \beta)$ and $\binom{n}{k}$ need to be included in the calculation of $p(k \mid M_2)$.

Suppose again that we take a uniform prior for q, and that $n = 5$, and we observe $k = 5$. Then $p(q \mid M_2) = 1$ and $p(k \mid q, M_2) = q^n$.

$$p(k \mid M_2) = \int_0^1 q^n\, dq = \left[\frac{q^{n+1}}{n+1}\right]_0^1 = \frac{1}{n+1}.$$

The Bayes factor is

$$B_{12} = \frac{p(k \mid M_1)}{p(k \mid M_2)} = (n+1)\, 0.5^n = \frac{n+1}{2^n} = 0.1875.$$

We previously calculated a one-sided p-value as 0.5^n. A two-sided p-value is more comparable to the Bayes factor. The two-sided p-value is

$$P(X = 0 \mid q = 0.5) + P(X = n \mid q = 0.5) = 2 \times 0.5^n = \frac{1}{2^{n-1}} = 0.0625.$$

Now take prior probabilities for each model as $p(M_1) = 1/2$ and $p(M_2) = 1/2$. The ratio of posterior probabilities is

$$\frac{p(M_1 \mid k)}{p(M_2 \mid k)} = \frac{p(M_1)}{p(M_2)}\, B_{12} = B_{12}.$$

$p(M_2 \mid k) = 1 - p(M_1 \mid k)$. Rearranging gives the posterior probability of model M_1 as

$$p(M_1 \mid k) = \frac{B_{12}}{1 + B_{12}} = 0.158.$$

Sensitivity to prior distributions

Suppose that we are comparing two models M_1 and M_2, and that model M_1 has a single parameter $\theta_1 \in \mathbb{R}$, with prior distribution $\theta_1 \sim N(0, \sigma_0^2)$.

$$p(y \mid M_1) = \int p(\theta_1 \mid M_1)\, p(y \mid \theta_1, M_1)\; d\theta_1$$

In typical problems, the likelihood $p(y \mid \theta_1, M_1)$ approaches zero for θ_1 outside some range $(-A, A)$. If we take σ_0 to be large enough (i.e. a flat, uninformative prior for θ_1), then

$$p(\theta_1 \mid M_1) \approx \frac{1}{\sqrt{2\pi}\sigma_0} \text{ for } -A < \theta_1 < A$$

Hence for large enough σ_0, the Bayes factor is

$$B_{12} \approx \frac{1}{\sqrt{2\pi}\sigma_0} \frac{\int p(y \mid \theta_1, M_1)\; d\theta_1}{\int p(\theta_2 \mid M_2)\, p(y \mid \theta_2, M_2)\; d\theta_2}$$

So if for example we replace a very large σ_0 by 100 σ_0, then B_{12} is divided by 100. However, the posterior distribution within model M_1 will hardly change, as the posterior is approximately proportional to the likelihood for large σ_0.

Because of this sensitivity of Bayes factors, and therefore of posterior model probabilities, to the prior distribution, many Bayesian statisticians prefer not to use these methods for comparing models.

There are alternatives for checking or comparing models which combine Bayesian and frequentist ideas, such as posterior predictive checks. We do not have time to cover these in this module.

Another option is to avoid choosing among different models, and instead construct a sufficiently flexible model. Models with many parameters can be easier to deal with in the Bayesian framework:

- conceptually, we can go from joint posterior to the marginal posterior distribution;
- having slightly informative prior distributions helps if there is not enough data to estimate all parameters using the current dataset alone5.