STAT314/461 Bayesian Inference

Patrick Graham
patrick.graham@canterbury.ac.nz
room tba

September, 2021

Things to discuss

- lab this week
- office hours
- Week 6

What we will cover this term

not necessarily in exactly this order

- Bayesian computation - from direct simulation to Markov Chain Monte Carlo methods.
  - rejection sampling
  - importance sampling
  - Metropolis-Hastings algorithm
  - Gibbs Sampler
- Hierarchical Models
- Problems of the "missing data type"

Computation will be in R

Assignments

- 20/09 - due 1/10 max 10% for STAT314
- 4/10 - due 15/10, max 10% for STAT314
- exact dates may change depending on how fast we go.

Best 4 of 5 assignments will count for your grade.

Things we need to know about aside from Bayesian Inference

- logistic regression
- multivariate normal distribution (see dmvnorm() in the mvtnorm package)
- R programming:
    1. For loops
    2. While loops
    3. functions - using and writing?
- setting priors

Recap on Bayesian Inference

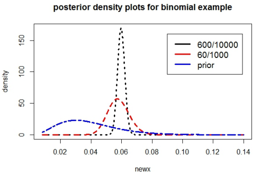
Recap on Bayesian Inference: Some big Bayesian ideas

What is Bayesian modelling?

- An approach to statistical modelling which makes explicit use of probability models for
  - observables (data)
  - parameters of models
- Uses the data actually observed to update the prior distribution (pre-data) for the parameters to a posterior distribution (post-data); $p(\theta|\mathrm{data})$
- "Statistics is the study of uncertainty; Uncertainty should be measured by probability." (Denis Lindley, 2000)

Review some basics

The aim is always $p(\mathrm{unknown}|\mathrm{data})$



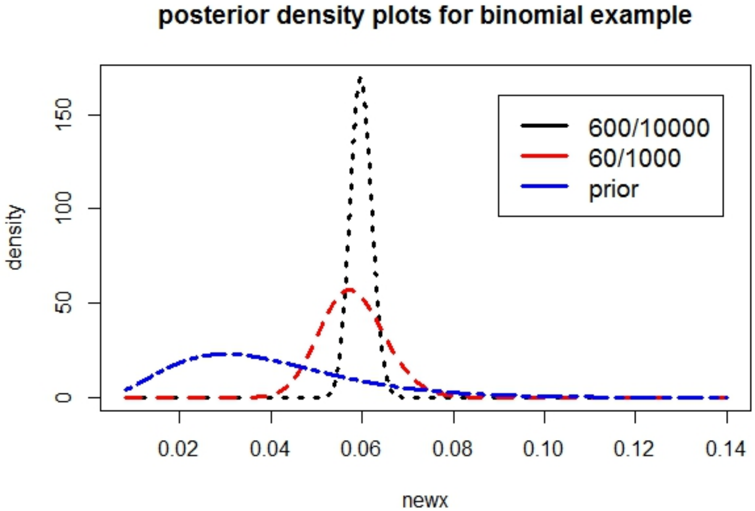
- aka the posterior distribution – think "post-data" distribution
- computation of the posterior requires:
  - a model for the data
  - a model for the parameters of the data model – the "prior" or "prior model"
  - (model $\equiv$ statistical model (recognises uncertainty))
- the prior (think "pre-data" or "pre - current data) allows information about the problem that is not in the data to be brought to the analysis

Some features of Bayesian modelling and inference

- A full distribution is available for inference; not just point estimates, standard errors and intervals.
- So its easy to make inferences like $\Pr(\theta > c|\mathrm{data}) = 0.1$.
- Given a posterior distribution for model parameters, inferences for observables not yet seen follows straightforwardly from the rules of probability theory.

The idea of using a distribution for inference

Suppose we are interested in estimating a binomial proportion; e.g proportion unemployed:



Bayesian Inference: More formally

Suppose Y denotes some variable of interest, e.g. income
Y_i - value of Y for i^{th} individual
$\mathbf{Y}$ collection of Y values, $\mathbf{Y} = (Y_1, Y_2, \ldots)$
We model $\mathbf{Y}$ as a random variable with distribution

$$p(\mathbf{Y}|\theta) = f_{\mathbf{Y}}(\mathbf{Y}|\theta) \tag{1}$$

Usually assume conditional independence so

$$p(\mathbf{Y}|\theta) = \prod_i p(Y_i|\theta) = \prod f_Y(Y_i|\theta) \tag{2}$$

We observe $\mathbf{Y}^{obs} = (Y_1, Y_2, \ldots, Y_n)$
The first job for a Bayesian is to compute

$$\begin{aligned} p(\theta|\mathbf{Y}^{obs}) &= \frac{p(\mathbf{Y}^{obs}|\theta)p(\theta)}{\int p(\mathbf{Y}^{obs}|\theta)p(\theta)\,\mathrm{d}\theta} \\ &\propto p(\mathbf{Y}^{obs}|\theta)p(\theta) \end{aligned} \tag{3}$$

Bayesian Inference: Likelihood and Prior

$$p(\theta|\mathbf{Y}^{obs}) \propto p(\mathbf{Y}^{obs}|\theta)p(\theta) \tag{4}$$

$$\text{Likelihood} \times \text{Prior}$$

Likelihood - probability density for observing the data actually seen, under the assumed model; a function of θ. Compute it using the model $p(\mathbf{Y}|\theta) = f_{\mathbf{Y}}(\mathbf{Y}|\theta)$.
Under conditional independence $p(\mathbf{Y}|\theta) = \prod_{i=1}^{i=n} f_Y(Y_i|\theta)$.
e.g. suppose $\mathbf{Y}^{obs} = (400, 1012, 961)$; assuming conditional independence (given θ), the likelihood function is

$$p(\mathbf{Y}^{obs}|\theta) = f_Y(400|\theta) \times f_Y(1012|\theta) \times f_Y(961|\theta) \tag{5}$$

Prior - probability model for θ which represents information about θ that is not contained in the observed data.

Bayesian Inference for observables not yet seen

e.g. for

- future observations
- missing observations

Just use probability theory to compute

$$p(\mathbf{Y}^{new}|\mathbf{Y}^{obs}) = \int p(\mathbf{Y}^{new}|\mathbf{Y}^{obs}, \theta) p(\theta|Y^{obs}) \, \mathrm{d}\theta \tag{6}$$

Can think of integration as Monte Carlo simulation

- draw θ from $p(\theta|\mathbf{Y}^{obs})$
- draw $\mathbf{Y}^{new}$ from $p(\mathbf{Y}^{new}|\mathbf{Y}^{obs}, \theta)$. Under conditional independence this amounts making independent draws from the data model $p(Y|\theta) = f_Y(Y|\theta)$.

Likelihood, prior and Bayesian computation

$$p(\theta|\mathbf{Y}^{obs}) = \frac{p(\mathbf{Y}^{obs}|\theta)p(\theta)}{\int p(\mathbf{Y}^{obs}|\theta)p(\theta)\,\mathrm{d}\theta} \tag{7}$$

- But notice we have been emphasising the numerator of (7), sometimes called the unnormalised posterior $q(\theta|\mathbf{Y^{obs}}) = p(\mathbf{Y}^{obs}|\theta)p(\theta)$.
- The Monte Carlo methods we will study generally only need the unnormalised posterior.
- This is a good thing:
  1. *We* specify the prior; *we* derive the likelihood from the model for the data that *we* specify. Hence the unnormalised posterior can always be written down and computed.
  2. In realistic applications the integration in the denominator can be high-dimensional and difficult.

Introduction to Monte Carlo methods for Bayesian Computation

Some basic ideas in the use of Monte Carlo methods for posterior computation (1)

- Suppose we can simulate from $p(\theta|\mathbf{Y}^{obs})$
- Let $h(\theta)$ be some function of θ
- We can approximate $E(h(\theta)|\mathbf{Y}^{obs}) = \int h(\theta)p(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta$ using the following Monte Carlo (MC) algorithm
  (i) for i in $1 : \mathrm{n_{sim}}$
  draw $\theta_{(i)}$ from $p(\theta|\mathbf{Y^{obs}})$
  compute $h_{(i)} = h(\theta_{(i)})$
  store $\theta_{(i)}, \ h_{(i)}$
  (ii) set

$$\hat{E}(h(\theta)|\mathbf{Y^{obs}}) = \frac{\sum_{i=1}^{n_{sim}} h_{(i)}}{n_{sim}}$$

Basic ideas in Monte Carlo computation (2)

Many things we are interested in are integrals and can be written as expectations of $h(\theta)$ for some choice of $h(\theta)$:

Expectation: $h(\theta) = \theta$

$$E(\theta|\mathbf{Y}^{obs}) = \int \theta p(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta$$

Variance: $h(\theta) = (\theta - E(\theta|\mathbf{Y}^{obs}))^2$

$$V(\theta)|\mathbf{Y}^{obs}) = \int (\theta - E(\theta|\mathbf{Y}^{obs}))^2 p(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta$$

Tail probability e.g. $h(\theta) = I_c(\theta)$, where $I_c(\theta) = 1$ if $\theta \leq c$; $I_c(\theta) = 0$ if $\theta > c$.

$$\begin{aligned}\Pr(\theta < c|\mathbf{Y}^{obs}) &= \int_{\infty}^{c} p(\theta|\mathbf{Y}^{obs})\mathrm{d}\theta \\ &= \int I_c(\theta) p(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta\end{aligned}$$

Basic ideas in Monte Carlo computation:(2b)

Interval Probability $h(\theta) = I_{(a,b)}(\theta) = 1$ if $a \leq \theta \leq b$; $I_{(a,b)} = 0$, otherwise.

$$\Pr(a \leq \theta \leq b|\mathbf{Y}^{obs}) = \int_a^b p(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta \tag{8}$$

$$\int I_{(a,b)}(\theta)p(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta \tag{9}$$

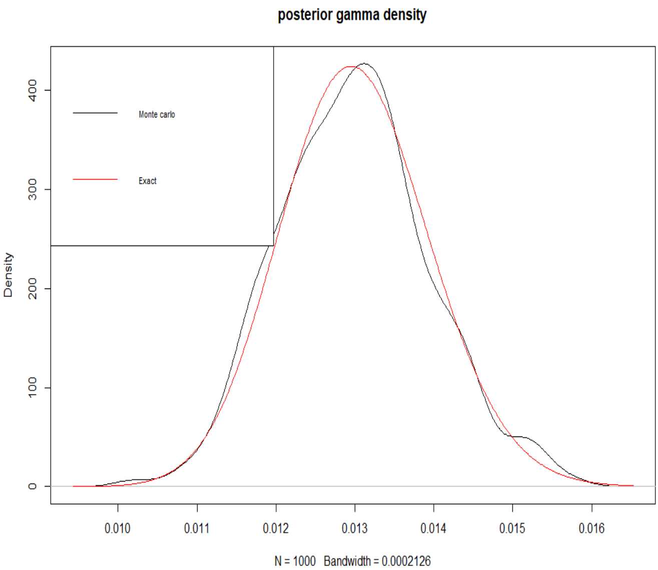
So, generate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ from the posterior, count up the number of theta values falling in (a, b) and divide by M. As $b \to a$ this gets pretty close to density estimation.

e.g.

```
plot(density(rgamma(n=1000,shape=alpha+sumY,
rate=beta+sumN))
```

General idea, in practice, is to draw values of θ from $p(\theta|\mathbf{Y}^{obs})$ and approximate features of the posterior by the corresponding features of the sample of θ values.

Compare Monte Carle density estimate with exact density



Basic ideas in Monte Carlo computation (3)

Recall:

$$p(\theta|\mathbf{Y}^{obs}) = \frac{p(\mathbf{Y}^{obs}|\theta)p(\theta)}{\int p(\mathbf{Y}^{obs}|\theta)p(\theta)\,\mathrm{d}\theta} \tag{10}$$

Setting $h(\theta) = p(\mathbf{Y}^{obs}|\theta)$, i.e the likelihood, we can see that the normalising constant (denominator) in the posterior also has the form of an expectation, but over the prior. So this is a bit different from the other examples where integration is over the posterior for θ. In principle we could apply the standard MC algorithm to approximate $K(\mathbf{Y}^{obs}) = \int p(\mathbf{Y}^{obs}|\theta)p(\theta)\,\mathrm{d}\theta$ by repeatedly sampling values from $p(\theta)$, computing $p(\mathbf{Y}^{obs}|\theta) = \prod_{i=1}^{n} p(Y_i|\theta)$ for each generated theta, and taking the average.

But $\prod_{i=1}^{n} p(Y_i|\theta)$ rapidly gets very small, which leads to numerical problems. We need to be a bit more clever to evaluate the normalizing constant (fortunately we, mostly, don't need to explicitly)

Aside: Laplace approximation to the normalising constant

$$p(\theta|\mathbf{Y}^{obs}) = \frac{p(\mathbf{Y}^{obs}|\theta)p(\theta)}{\int p(\mathbf{Y}^{obs}|\theta)p(\theta)\,\mathrm{d}\theta} = \frac{q(\theta|\mathbf{Y}^{obs})}{\int q(\theta|\mathbf{Y}^{obs})} \tag{11}$$

Let $u(\theta) = \log(q(\theta))$ and suppose $u(\theta)$ is maximised at $\theta = \hat{\theta}$, so $\hat{\theta}$ is the posterior mode. Let $u''(\theta) = \frac{d^2}{du(\theta)^2}$ and Let r denote the dimension of θ

$$K(\mathbf{Y^{obs}}) = \int (q(\theta|\mathbf{Y}^{obs}))\,\mathrm{d}\theta \approx q(\hat{\theta}|\mathbf{Y}^{obs})(2\pi)^{r/2}|-u''(\hat{\theta})|^{-1/2} \tag{12}$$

In practice we would compute

$$\log(K(\mathbf{Y^{obs}})) \approx u(\hat{\theta}|\mathbf{Y^{obs}}) + (r/2)\log(2\pi) - 0.5\log|-u''(\hat{\theta})| \tag{13}$$

$$= \sum_{i=1}^{n} \log(p(Y_i|\hat{\theta})) + \log(p(\hat{\theta})) + (r/2)\log(2\pi) - 0.5\log|-u''(\hat{\theta}| \tag{14}$$

can often then work with
$\log(p(\theta|\mathbf{Y}^{obs})) \approx \log(q(\theta|\mathbf{Y}^{obs})) - \log(\hat{K}(\mathbf{Y}^{obs}))$ where $\log(\hat{K}(\mathbf{Y}^{obs}))$ is rhs of (14)

Basic ideas of Monte Carlo computation (4): Work on the log-scale for intermediate calculations

- To avoid numerical problems we will usually work on the $\log_e$ scale when evaluating likelihoods and posterior distributions; leaving exponentiating as late as possible.
- For example, in the Metropolis-Hastings algorithm, to be considered later, it is necessary to evaluate the a ratio of posterior densities computed at different values of θ. Instead of working with $r_{MH} = p(\theta_1|\mathbf{Y})/p(\theta_2|\mathbf{Y})$ directly we work with $\log(r_{MH}) = \log(p(\theta_1|\mathbf{Y})) - \log(p(\theta_2|\mathbf{Y}))$ e.g. Instead of determining $(r_{MH} > 1)$ we determine $\log(r_{MH} > 0)$

Basic ideas in Monte Carlo computation (5): Monte Carlo error

- Yes, Monte Carlo methods provide *approximations* to posterior distributions.
- Monte Carlo error is under the control of the analyst - reduce error by increasing the Monte Carlo sample size!
- We can get a sense of the Monte Carlo error.
- For simple Monte Carlo methods this is fairly straightforward, at least for the expected value (of θ or $h(\theta)$).
- if $\boldsymbol{\theta}_{n_{sim}} = (\theta_{(1)}, \theta_{(2)}, \ldots \theta_{n_{sim}})$ are n_{sim} independent draws from the posterior and $s(\boldsymbol{\theta}_{n_{sim}})$ is the standard deviation of the draws, then $s(\boldsymbol{\theta}_{n_{sim}})/\sqrt{n_{sim}}$ is a reasonable approximation to the Monte Carlo error, for the expectation.

Basic ideas in Monte Carlo computation (6): More on Monte Carlo error

- $s(\boldsymbol{\theta}_{n_{sim}})$ is an approximation to the posterior standard deviation, so $V(\boldsymbol{\theta}_{n_{sim}}) = s(\boldsymbol{\theta}_{n_{sim}})^2$ is an approximation to the posterior variance.
- So, in a loose sense, we can think of

$$V^{\mathrm{tot}}(\boldsymbol{\theta}_{n_{sim}}) = V(\boldsymbol{\theta}_{n_{sim}}) + \frac{V(\boldsymbol{\theta}_{n_{sim}})}{n_{sim}} \tag{15}$$

$$= V(\boldsymbol{\theta}_{n_{sim}})(1 + \frac{1}{n_{sim}}) \tag{16}$$

  as a measure of total uncertainty about θ
- For $n_{sim} = 100$, total uncertainty is about 1% greater than posterior uncertainty; Total standard deviation is about $\sqrt{1 + 1/100} \approx 0.5\%$ greater than posterior standard deviation.
- Monte Carlo error can be often be a small contributor to total uncertainty.

Basic ideas in Monte Carlo computation (7): Some caveats

- Caveat 1: - approximating extreme posterior quantiles requires a bigger Monte Carlo sample size than approximating the posterior expectation,
- Caveat 2: The above arguments are based on the premise that we can draw $n_s im$ values independently from the posterior. The more complex Monte Carlo methods we will look at later generate *correlated* draws and approximation of the Monte Carlo standard error in this case is more difficult. For a given n_{sim}, the Monte Carlo standard error will be greater for correlated draws than for independent draws.

Applications of direct Monte carlo simulation

Direct simulation of the posterior of a function of random variables Ex (1)

Consider a simple two-group study: people randomly allocated to group A or B

- Group A - 10 people given drug A; 7 successes (respond to treatment)
- Group B - 10 people given drug B; 3 successes

A model for these data:
Conditional Independence:

$$p(Y_A, Y_B|\theta_A, \theta_B, N_A, N_B) = p(Y_A|\theta_A, N_A)p(Y_B|\theta_B, N_B)$$

$$Y_A \sim \mathrm{Binomial}(\theta_A, Na)$$
$$Y_B \sim \mathrm{Binomial}(\theta_B, Nb)$$

Suppose $p(\theta_A, \theta_B) = p(\theta_A)p(\theta_B)$ and $\theta_A \sim \mathrm{Beta}(1, 1)$, $\theta_B \sim \mathrm{Beta}(1, 1)$
Suppose further that we are interested in the relative risk, $\mathrm{rr} = \theta_A/\theta_B$.

Simulating the posterior of a function of random variables (2)

- Ultimately we want $p(\mathrm{rr}|\mathrm{data})$ where here data is (Y_A, Y_B) and we regard as N_A, N_B as known constants, fixed by the investigator.
- - first step is to obtain the posterior for the model parameters

$$
\begin{aligned}
p(\theta_A, \theta_B|\mathrm{data}) &\propto p(Y_A, Y_B|\theta_A, \theta_B, N_A, N_B)p(\theta_A)p(\theta_B) \\
&= [p(Y_A|\theta_A, N_A)p(\theta_A)] \times [p(Y_B|\theta_B, N_B)p(\theta_B)] \\
&= [\mathrm{Binomial}(Y_A|\theta_A, N_A)][\mathrm{Beta}(\theta_A|1, 1)] \\
&\times [\mathrm{Binomial}(Y_B|\theta_B, N_B)][\mathrm{Beta}(\theta_B|1, 1)] \\
&= [\mathrm{Beta}(\theta_A|Y_A + 1, N_A - Y_A + 1)]\times \\
&\times [\mathrm{Beta}(\theta_B|Y_B + 1, N_B - Y_B + 1)]
\end{aligned}
$$

Simulating the posterior of a function of random variables (3)

$$p(\theta_A, \theta_B|\mathrm{data}) = [\mathrm{Beta}(\theta_A|Y_A+1, N_A-Y_A+1)]\times \\ \times [\mathrm{Beta}(\theta_B|Y_B+1, N_B-Y_B+1)] \quad (17)$$

It is easy to simulate the joint posterior for (θ_A, θ_B). Just draw values independently from the appropriate Beta distributions. Almost as easy to simulate the posterior of $\mathrm{rr} = \theta_A/\theta_B$:

for (i in 1:n_{sim}) {
draw $\theta_A^{(i)}$ from $\mathrm{Beta}(Y_A+1, N_A-Y_A+1)$
draw $\theta_B^{(i)}$ from $\mathrm{Beta}(Y_B+1, N_B-Y_B+1)$
set $rr^{(i)} = \theta_A^{(i)}/\theta_B^{(i)}$
store $rr^{(i)}$
}

Summarise stored rr values (see R script `posterior_rr.r`)

Direct simulation of function of parameters: Ex. 2 Cancer rates by DHB

see R markdown file `cancer_example.pdf` and /or `cancer_example.rmd`

- If λ_i is the underlying rate of tongue cancer for the i^{th} DHB we can, using Monte Carlo simulation, compute quantites like $\Pr(\lambda_i > \lambda_j, \forall j \neq i | \mathbf{Y})$ i.e the posterior probability that the i^{th} DHB has the largest underlying cancer rate among all DHBs. We can compute this probability for each DHB.
- We can also compute the posterior distribution for the rank-order of DHBs, wrt tongue cancer rates.
- These types of inferences are difficult to obtain analytically

Useful probability decompositions for Monte Carlo simulation (and Bayesian Inference in general)

Direct simulation for a vector of random variables

We can always write a joint distribution as a product of a marginal and conditional distribution, e.g.

$$p(Y_1, Y_2) = p(Y_1)p(Y_2|Y_1)$$

Extends automatically to vectors or arbitrary length:

$$p(Y_1, Y_2, \ldots, Y_k) = p(Y_1)p(Y_2|Y_1)p(Y_3|Y_2, Y_1), \ldots, p(Y_k|Y_{k-1}, \ldots, Y_1)$$

And, of course this holds for any vector of r.v's, not just observables, so we might sometimes model a parameter vector as

$$p(\theta_1, \theta_2, \ldots, \theta_k) = p(\theta_1)p(\theta_2|\theta_1)p(\theta_3|\theta_2, \theta_1), \ldots, p(\theta_k|\theta_{k-1}, \ldots, \theta_1)$$

The point is we are breaking a multivariate distribution into a sequence of univariate (more generally, low-dimensional) distributions - often easier to model and simulate. If you are lucky you can take advantage of conditional independencies to simplify further.

Sequence of conditionals approach also applies to a combination of parameters and observables; and holds conditionally

$$p(\theta, Y_{n+1}|Y_1, \ldots, Y_n) = p(\theta|Y_1, \ldots, Y_n)p(Y_{n+1}|\theta, Y_1 \ldots, Y_n) \tag{18}$$

If $Y_1, \ldots, Y_n$ are the observed data we will of often write expressions such as (18) as

$$p(\theta, Y_{n+1}|\mathrm{data}) = p(\theta|\mathrm{data})p(Y_{n+1}|\theta, \mathrm{data}). \tag{19}$$

If the $Y's$ are modelled as conditionally independent given θ :

$$p(Y_1, Y_2, \ldots|\theta) = \prod_i p(Y_i|\theta) \tag{20}$$

(19) simplifies to

$$p(\theta, Y_{n+1}|\mathrm{data}) = p(\theta|\mathrm{data})p(Y_{n+1}|\theta) \tag{21}$$