Bayesian inference and MCMC

CINN tutorials

2018

Outline

Frequentist inference

- Probability model $p(D|\theta)$ for data $D = \{Y_i\}_{i=1}^n$.
- Unknown parameters, θ, assumed fixed but unknown.
- Statistical inference: estimate the value of θ from data D:
  - procedures such as maximum likelihood estimation (MLE) give estimates $\widehat{\theta}$ of θ as a function of D;
  - MLE: value of $\widehat{\theta}$ which maximises $p(D|\theta) = L(\theta|y)$.
- Uncertainty about $\widehat{\theta}$ is estimated from its sampling distribution.
- This forms the basis of CI calculations etc.
- Frequentist approach: estimate the fixed unknown state of nature using observed data.

Example

- $n =$ number of independent trials (e.g. locations on an island).
- $\theta =$ probability of success for each trial (e.g. finding a species).
- $y =$ number of successes observed (e.g. where species seen).
- $Y|\theta \sim Bin(\theta, n)$, so, for $y = 0, 1, ..., n$

$$p(y|\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}$$

- The MLE for θ is $\widehat{\theta} = y/n$ and the uncertainty in $\widehat{\theta}$ is estimated through the sampling distribution of $\widehat{\theta}$:

$$y/n \sim \mathcal{N}(\theta, \theta(1-\theta)/n).$$

- From this we get 95% confidence intervals.
- Example: If $n = 20$ and $y = 9$, then $\widehat{\theta} = 0.45$.

Bayesian inference

- Probability model $p(D|\theta)$ for data $D = \{Y_i\}_{i=1}^n$.
- Unknown parameters θ are also random variables, rather than fixed unknown quantities.
- Uncertainty about θ modelled by probability distribution $p(\theta)$, the prior distribution of θ.
- $p(\theta)$ captures current knowledge of θ before observing data.
- Statistical inference: obtain $p(\theta|D)$ the posterior distribution of θ given the data D and our prior $p(\theta)$ using Bayes' rule.
- All inferences about θ derived from $p(\theta|D)$...
- Bayesian approach: update our current knowledge of the state of nature using evidence/data.

Bayes' rule

- Main inferential tool in Bayesian inference is Bayes' rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

- $p(\theta)$ is the prior distribution of θ;
- $p(D|\theta)$ is the distribution of the data given a fixed value of θ (known as the likelihood function, when a function of θ);
- $p(D)$ is the marginal likelihood;
- $p(\theta|D)$ is the posterior distribution of θ.

Result of Bayesian inference

- The posterior distribution $p(\theta|y)$ for θ given the data D contains all of the information of interest.
- All inferences about θ are derived from $p(\theta|y)$...
  ... can use summaries: posterior mean, median or mode, percentiles, etc. to describe the posterior.

Choice of prior

- The prior for θ should capture our knowledge before observing data.
- Possible sources:
  - scientific knowledge of the background;
  - previous studies;
  - expert judgments (see e.g. O'Hagan, 1998, and other papers in the same issue).

Conjugate priors

- For simple problems, a convenient option is to choose from a family of conjugate priors.
- A family of distributions is conjugate if when the prior is chosen from the family, the posterior is also a member of the family.
- Mathematically convenient, and widely used.

Example

- n independent trials with $\theta =$ probability of success.
- $y =$ number of observed successes. y has a binomial distribution:

$$p(y|\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}$$

- θ is a probability so it can only take values between 0 and 1.
- Choose shape for distribution to reflect knowledge of θ - e.g.
  - No preference for any one value of θ over another - uniform distribution - non-informative prior.
  - Mode is $\theta = 0.5$ and $p(\theta < 0.1) = p(\theta > 0.9) \approx 0$
  - Mode is $\theta = 0.25$ and with $p(\theta \leq 0.8) = 0.95$.
  - $\mathbb{E}[\theta] = 0.7$, $\mathrm{Var}[\theta] = 0.2$.

The beta distribution

- Conjugate prior for θ is a beta distribution:

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}.$$

- The expectation and variance are:

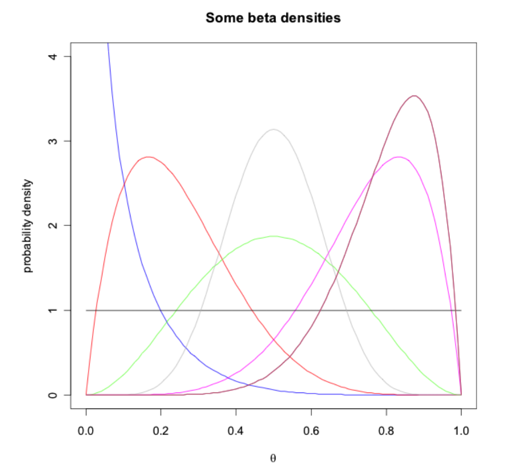
$$\mathbb{E}[\theta] = \frac{a}{a+b}$$

$$\mathrm{Var}[\theta] = \frac{ab}{(a+b)^2(a+b+1)}.$$

- The mode is

$$\mathrm{mode}[\theta] = \frac{a-1}{a+b-2}.$$

Examples of beta pdfs



Use of conjugate prior

- Ignoring terms not involving θ

$$p(y|\theta) \propto \theta^y(1-\theta)^{n-y}$$
$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

- So, using Bayes' theorem:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1}. \end{aligned}$$

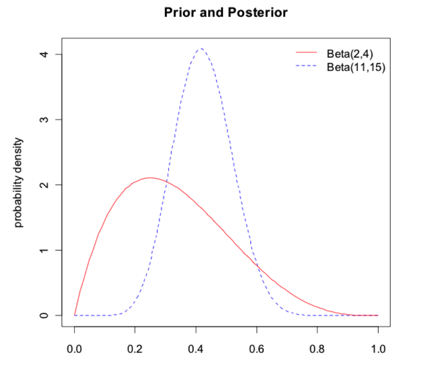
- We note that $p(\theta|y)$ is therefore a beta distribution:

$$\theta|y \sim \text{Beta}(a+y, b+n-y).$$

- The marginal likelihood does not need to be calculated (although it is available through finding the normalising constant of the beta distribution).

Specific example

- Example, $y = 9$ successes observed in $n = 20$ trials:
  - If the prior for θ is $\theta \sim \text{Beta}(2,4)$ so that $\mathbb{E}[\theta] = 2/6 = 0.333$;
  - then the posterior is $\theta|y \sim \text{Beta}(11,15)$ so $\mathbb{E}[\theta|y] = 11/26 = 0.423$.



Bayesian inference for Gaussians: case 1

- **Case 1:** σ^2 is known, and we wish to estimate the mean μ.
- μ is continuous and can take any number on the real line. The conjugate prior is the normal distribution, with pdf given by

$$
\begin{aligned}
p(\mu|\mu_0,\sigma_0^2) &= \frac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0)\right).
\end{aligned}
$$

Posterior derivation

- Ignoring terms not involving μ in $p(D|\mu)$, we obtain

$$\begin{aligned} p(D|\mu) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \\ &\propto \exp\left(-\frac{2n\overline{y}\mu - n\mu^2}{2\sigma^2}\right), \end{aligned}$$

  where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- After some algebra, we obtain $\mu|D \sim \mathcal{N}(\mu_n, \sigma_n^2)$, where $\mu_n = \frac{\sigma_0^{-2}\mu_0 + n\sigma^{-2}\overline{y}}{\sigma_0^{-2}+n\sigma^{-2}}$ and $\sigma_n^{-2} = \sigma_0^{-2} + n\sigma^{-2}$.

Summary

- If the variance is known, a normal distribution is conjugate for estimating the mean of normal data.
- i.e. normal prior for the mean leads to a normal posterior...
  - posterior standard deviation $\sigma_n < \sigma_0$ prior standard deviation;
  - posterior mean is a weighted average of the prior mean μ_0 and the data mean $\bar{y}$.
- If the prior variance σ_0^2 is very large (i.e. assuming a vague prior), the posterior of μ is approximately

$$\mu | D \sim \mathcal{N}(\bar{y}, \frac{\sigma^2}{n}). \tag{1}$$

Posterior point summaries

- Strictly, the posterior distribution of θ is our inference about θ.
- But summary statistics are often useful...
- Posterior expectation: $\mathbb{E}[\theta|D] = \int_\theta \theta p(\theta|D) d\theta$.
- Binomial example where $\theta|y \sim \text{Beta}(a+y, b+n-y)$:
    - Posterior expectation is $\mathbb{E}[\theta|y] = \frac{a+y}{a+b+n}$.
    - Compare with the prior expectation $\frac{a}{a+b}$ and the data mean $\frac{y}{n}$.
    - Can show that $\frac{a+y}{a+b+n}$ always lies between $\frac{a}{a+b}$ and the data mean $\frac{y}{n}$.
- In a general sense, the posterior is a compromise between the prior and data.

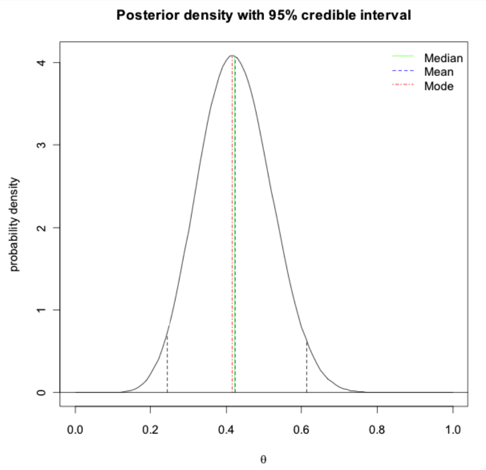
MAP estimate

- The maximum *a posteriori* estimate is the posterior mode (the value when the posterior is at its maximum).
- Binomial example where $\theta|y \sim \text{Beta}(a+y, b+n-y)$:
  - MAP estimate is $\frac{a+y-1}{a+b+n-2}$.
- Binomial example with uniform prior ($a = b = 1$):
  - The posterior density is $\text{Beta}(y+1, n-y+1)$.
  - Posterior mean is $\mathbb{E}[\theta|y] = \frac{y+1}{n+2}$ and MAP estimate $\frac{y}{n}$.
- So, with a uniform prior, the MAP estimate is the same as the MLE in a frequentist setting.

Posterior interval summaries

- A central $100(1-\alpha)\%$ posterior interval is given by the range of θ values above and below which $100(\alpha/2)\%$ of the posterior probability lie.
- This is called a $100(1-\alpha)\%$ credible interval for θ.
- For example, a 95% credible interval is the range of values of θ between the 2.5% and 97.5% percentiles of the posterior density.
- Another interval is the highest posterior density (HPD) region:
  - the set of values of θ that contains $100(1-\alpha)\%$ of the posterior probability, and such that the density within the region is always higher than the density outside.

Example of summaries



WinBUGS

- WinBUGS is a piece of software that automatically implements (variants of) the algorithm above. It provides:
    - a language for specifying a likelihood function and prior distributions;
    - a tool for describing models in graphical form;
    - flexible MCMC computations on the full joint distribution of all quantities (observed or parameters);
    - some MCMC diagnostic procedures;
    - marginal posterior densities and summaries;
    - a mechanism for exporting results to other programs (in particular, to R).

What WinBUGS is not

- WinBUGS does not provide:
  - tools for data manipulation and management;
  - facilities for exploratory data analysis;
  - joint posterior densities.
- All of these shortcomings can be dealt with by using WinBUGS together with R (or SAS, Genstat or some other general purpose statistics software).
- R is particularly suitable because:
  - it is a flexible object-oriented language for data analysis;
  - there are several software developments linking WinBUGS with R;
  - like WinBUGS, it is free.

Steps in using WinBUGS

1. Specify the **model** (likelihood and priors).
2. **Check** the model syntax.
3. Load the **data**.
4. **Compile** the model and specify the number of chains (for MCMC diagnostics).
5. Load **initial values** to start the chain.
6. Select **nodes** (parameters) whose posterior distributions are to be sampled.
7. Simulate the chain(s) - this is called **updating**.
8. Perform MCMC **diagnostics** - check for convergence.
9. **Examine** posterior densities and summaries.

Example 1: binomial with conjugate prior

- In $n = 20$ independent Bernoulli trials, $y = 7$ successes are observed
$$y \sim \text{Bin}(\theta, n)$$
and we assume a conjugate prior for θ:
$$\theta \sim \text{Beta}(\alpha, \beta).$$
- Suppose that from prior knowledge we can assume $\alpha = 3$, $\beta = 2$.
- Of course, we know the posterior distribution is
$$\theta | y \sim \text{Beta}(\alpha + 7, \beta + 13)$$
and no simulation is necessary.
- But just to illustrate WinBUGS...

BUGS code

Write BUGS code to specify the model ...

```
model
{
    y ~ dbin(theta, n)          ← likelihood
    theta ~ dbeta(3 ,2)         ← prior
}
```

This says ...

`y` has a binomial distribution with parameters `theta` and `n`,

and

`theta` has a beta distribution with parameters 3 and 2.

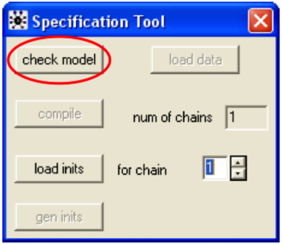
Checking the model

Check the model syntax ...

Select *Specification*... from the *Model* menu.

Click on *check model*.

```
model
{
```



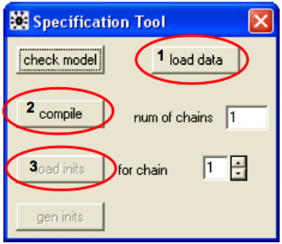
Load the data (1) ...

```
list(n=20, y=7)
```

... and compile (2).

Then load initial values (3) ...

```
list(theta=0.5)
```

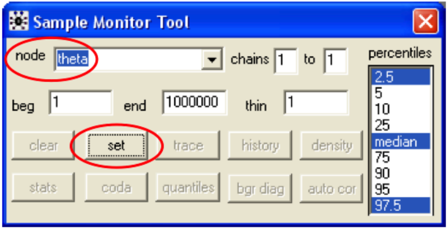


Monitoring

Select *Samples* from the *Inference* menu.

Select the nodes (variables) to monitor (just one in this case) ...

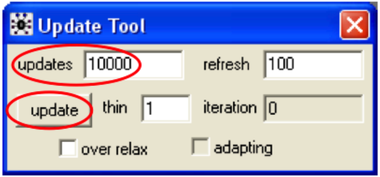
Type the node name and click *set* for each node.



Set the trace for all selected nodes (*) to monitor the MCMC simulations (optional) ...

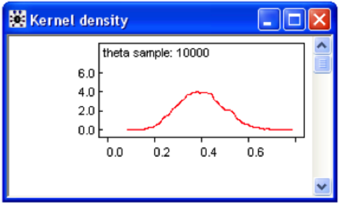
Results

Using the Update tool (in the *Model* menu), select 10,000 simulations …



Results:

Click *density* in the *Sample Monitor Tool* to get the posterior densities …



… and click *stats* for summary statistics …

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta	0.3982	0.09627	9.43E-4	0.2239	0.3952	0.5937	1	10000

Example 2: binomial with non-conjugate prior

- As in Example 1, $y \sim \text{Bin}(\theta, n)$ and $y = 7$ successes observed out of $n = 20$ trials.
- But now assume a different prior.
- Logit transform of θ: $\phi = \log\left(\frac{\theta}{1-\theta}\right)$, so that $-\infty < \phi < \infty$ (this is the *link function*).
- Assume a normal prior for ϕ: $\phi \sim \mathcal{N}(\mu, \tau)$ and choose $\mu = 0$ and $\tau = 0.01$ (i.e. a vague prior).
- This is a non-conjugate prior with no simple form for the posterior.
- Straightforward in WinBUGS.

BUGS code

BUGS code for the model is ...

```
model
{
    y ~ dbin(theta, n)          ← likelihood
    logit(theta) <- phi         ← link function
    phi ~ dnorm(0.0, 0.01)      ← prior

}
```

Data:

```
list(n=20, y=7)
```

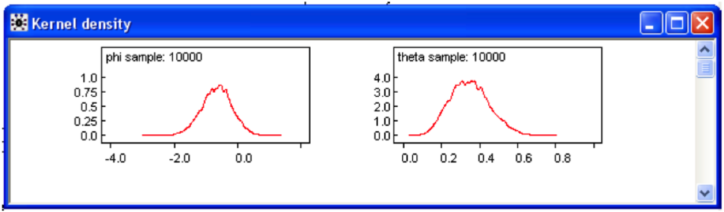
Initial values:

```
list(phi=0)
```

Model data and inits entered and checked as before

Results

Results with `theta` and `phi` both selected for monitoring ...



Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
phi	-0.6559	0.4811	0.00497	-1.623	-0.6423	0.2659	1	10000
theta	0.3493	0.1036	0.001081	0.1648	0.3447	0.5661	1	10000

Generating initial conditions

We need data or initial values for all random variables:

`y` – data available (`y=7`)

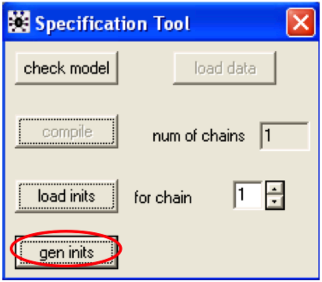
`phi` – set initial values (`phi=0`)

`y.new` – need initial values.

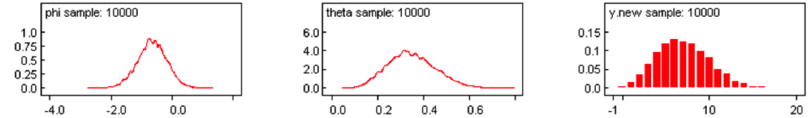
Rather than setting them we can generate them using the **gen.inits** button.

gen.inits simulates a value from the distribution `y.new ~ dbin(theta, n)` using the initial value of theta = 0.5 and observed value of n=20

$y.new \sim Bin(0.5, 20)$



Results



node	mean	sd	MC error	2.5%	median	97.5%
phi	-0.653	0.484	0.0051	-1.627	-0.652	0.281
theta	0.35	0.105	0.0011	0.164	0.343	0.570
y.new	6.998	2.96	0.0314	2.0	7.0	13.0

Predicted value is 7.0 (sd=2.96) with credible interval (2, 13)

Non-standard posteriors

- We can always find the equation for the posterior up to proportionality: $\pi(\theta|D) \propto p(D|\theta)p(\theta)$.
- When using conjugate priors (and sometimes for the non-informative priors), we used the covenient mathematical form of the posterior to:
  - normalise the posterior, so that we had an exact expression for its pdf;
  - summarise the posterior, using expectation, variance, intervals, etc.
- If we do not have conjugate priors, we can't do this:
  - both of these types of calculation involve integration, which is intractable in general.

The Metropolis algorithm by example

- Suppose $\theta \sim \text{Triang}(0.4, 0.6)$ and that $p(\theta)$ can be calculated up to a constant of proportionality

$$p(\theta) = 10\{1 - 10|\theta - 0.5|\} \propto \{1 - 10|\theta - 0.5|\} = f(\theta).$$

- Set an initial starting value $\theta^{(0)}$, say $\theta^{(0)} = 0.45$.
- **Step 1:** Generate a proposed value θ^*.
- In rejection sampling θ^* was drawn from a distribution $g(\theta)$.
- In the Metropolis algorithm, θ^* is drawn from a (symmetric) proposal distribution $q(.|\theta^{(p-1)})$ that is conditional on the previous value of θ, $\theta^{(p-1)}$, e.g.

$$\theta^*|\theta^{(p-1)} \sim \mathcal{N}(\theta^{(p-1)}, 100)$$

$$\theta^*|\theta^{(p-1)} \sim \mathcal{N}(0.45, 100).$$

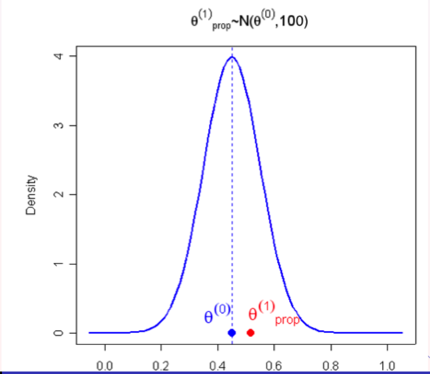
First step

Supppse using

$$\theta^*|\theta^{(p-1)} \sim \mathcal{N}(0.45, 100)$$

gives

$$\theta^* = 0.518.$$



Acceptance probability

- **Step 2.** Calculate the probability of accepting θ^*.

$$\alpha(\theta^{(i-1)}, \theta^*) = \min\left\{1, \frac{p(\theta^*)}{p(\theta^{(p-1)})}\right\} = \min\left\{1, \frac{f(\theta^*)}{f(\theta^{(p-1)})}\right\}.$$

- The ratio evaluates whether our proposed value θ^* is more or less likely to belong to $p(\theta)$ than our previous value $\theta^{(p-1)}$.
- In this case $\frac{f(\theta^*)}{f(\theta^{(p-1)})} = \frac{f(0.518)}{f(0.45)} = \frac{8.2}{5} = 1.64$.
- $\theta^* = 0.518$ is more likely than our previous value $\theta^{(0)} = 0.45$.
- θ^* is accepted with $\alpha(\theta^{(i-1)}, \theta^*) = \min(1, 1.64) = 1$.

Accept/reject step

Step

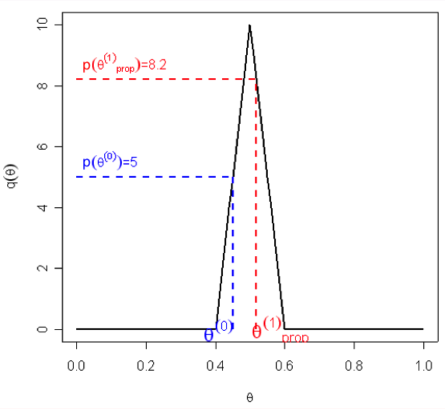
3. Accept θ^* with probability $\alpha(\theta^{(p-1)}, \theta^*)$. We accept θ^* with probability 1 so that

$$\theta^{(0)} = 0.45,$$

$$\theta^{(1)} = 0.518.$$

If

we had rejected the move, then we would have $\theta^{(1)} = \theta^{(0)}$.



Second iteration

- **Step 1.** Generate proposed value.
  - Use the proposal distribution:

$$\theta^*|\theta^{(p-1)} \sim \mathcal{N}(\theta^{(1)}, 100) = \mathcal{N}(0.518, 100)$$

  to generate a value $\theta^* = 0.596$.

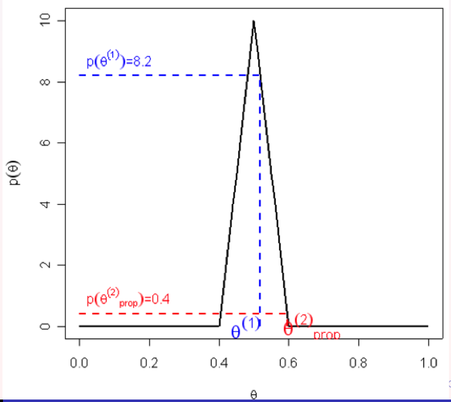
- **Step 2.** Calculate the acceptance probability

$$\alpha(\theta^{(1)}, \theta^*) = \min\left\{1, \frac{f(0.596)}{f(0.518)}\right\} = \min\left\{1, \frac{0.4}{8.2}\right\} = 0.049.$$

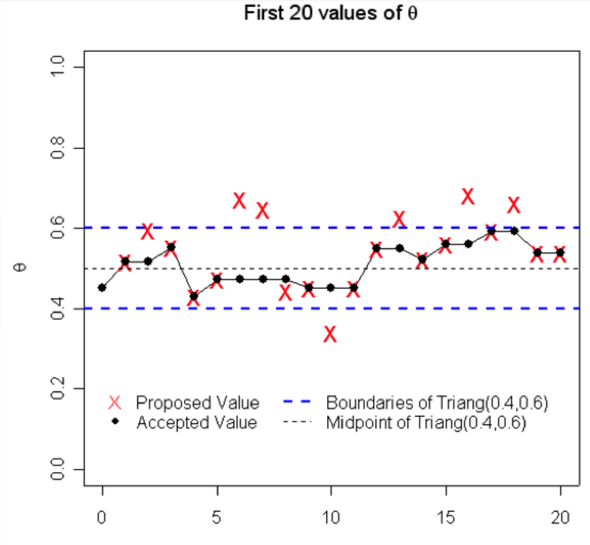
i.e. θ^* is much less likely to belong to $p(\theta)$ than $\theta^{(1)}$.

Second iteration accept/reject

Step 3. Drawing a value from a uniform distribution Unif$(0,1)$ gives $u = 0.68$, so we reject θ^* in favour of $\theta^{(1)}$ and $\theta^{(2)} = \theta^{(1)} = 0.518$. The realisation of the chain so far consists of the values 0.45, 0.518, 0.518.

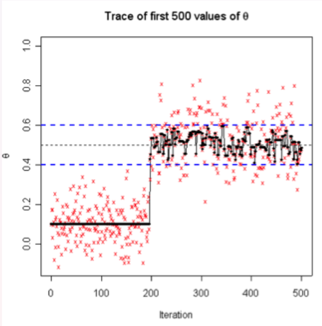


Trace plot


First 20 values of θ
θ
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Proposed Value
Accepted Value
Boundaries of Triang(0.4,0.6)
Midpoint of Triang(0.4,0.6)

Choosing different starting values

Suppose in the Metropolis example above we choose a very different starting value, e.g. $\theta^{(0)} = 0.1$. Initially θ sticks at around 0.1. Eventually a proposed value of between 0.4 and 0.6 is obtained and so the chain reaches some kind of equilibrium. Here we would need to ignore the first 200+ values (to be safe) - this is the burn-in period.



Note we want the chain to converge to the distribution $p(\theta)$, not a single value of θ, when it reaches equilibrium.

Metropolis-Hastings for posterior simulation

- When we have prior $p(\theta)$ and likelihood $f(D|\theta)$, the Metropolis-Hastings algorithm is implemented as follows.

Returns a dependent sample $\{(\theta^{(p)},) \mid 1 \leq p \leq P\}$ from $p(\theta|D)$.

- For p=1:P
    - Simulate $\theta^* \sim q(.|\theta^{(p-1)})$
    - Simulate $u \sim \mathscr{U}[0,1]$
    - if $u < \min\left\{1, \frac{p(\theta^*)f(D|\theta^*)q(\theta^{(p-1)}|\theta^*)}{p(\theta^{(p-1)})f(D|\theta^{(p-1)})q(\theta^*|\theta^{(p-1)})}\right\}$
        - $\theta^{(p)} = \theta^*$
    - else
        - $\theta^{(p)} = \theta^{(p-1)}$