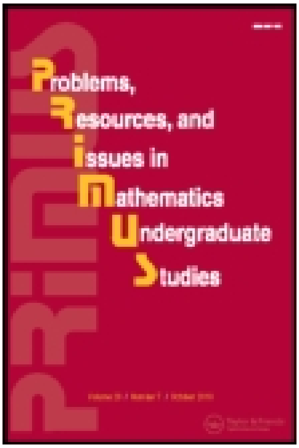


PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/upri20

Teaching Markov Chain Monte Carlo: Revealing the Basic Ideas Behind the Algorithm

Wayne Stewart & Sepideh Stewart

Published online: 07 Dec 2013.

To cite this article: Wayne Stewart & Sepideh Stewart (2014) Teaching Markov Chain Monte Carlo: Revealing the Basic Ideas Behind the Algorithm, PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies, 24:1, 25-45, DOI: 10.1080/10511970.2013.824054

To link to this article: http://dx.doi.org/10.1080/10511970.2013.824054

PRIMUS, 24(1): 25–45, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1051-1970 print / 1935-4053 online
DOI: 10.1080/10511970.2013.824054

Teaching Markov Chain Monte Carlo: Revealing the Basic Ideas Behind the Algorithm

Wayne Stewart and Sepideh Stewart

Abstract: For many scientists, researchers and students Markov chain Monte Carlo (MCMC) simulation is an important and necessary tool to perform Bayesian analyses. The simulation is often presented as a mathematical algorithm and then translated into an appropriate computer program. However, this can result in overlooking the fundamental and deeper conceptual ideas that are necessary for an effective diagnosis of MCMC output. In this paper we discuss MCMC simulation conceptually in the context of a Bayesian paradigm without revealing the formal algorithm first. We propose a tactile simulation method with a two-state discrete parameter where a coin supplies the proposal values and given the acceptance sets, the die value determines whether be not to accept the proposal.

Keywords: Markov chain Monte carlo, simulation, Bayesian inference, Bayes' box, undergraduate teaching.

1. INTRODUCTION

In recent years an increasing number of researchers from all areas of science, medicine, bioinformatics, engineering, economics, and many more use Markov chain Monte Carlo (MCMC) simulation. In our experience Metropolis–Hastings (M–H) and Gibbs sampling (a special case of M–H) are the most common forms of MCMC theory used in practice. Casella and George [5] describe a simple case of how and why the Gibbs sampler works and provide insight for more complicated cases.

Despite having a useful and well-developed algorithm, in the authors' experience, many learners often do not have a good conceptual understanding

Address correspondence to Wayne Stewart, Department of Mathematics, University of Oklahoma Norman, OK 73019-3103, USA. E-mail: wayne.s.stewart@ou.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/upri of MCMC. Although in some cases learners are capable of translating the algorithm into a program, they are often unaware of its more basic conceptual ideas which are central and useful in explaining MCMC output.

While there is some literature on teaching and learning introductory Bayesian statistics available [1–3, 10], there have been no studies done on how to teach the MCMC algorithm. In this paper we discuss the pedagogical issues surrounding the teaching of MCMC simulation and step-by-step construction of the Bayes' box as a necessary foundation for MCMC sampling. Moreover, we present an innovative tactile way of teaching a special case of the M–H algorithm within a Bayesian context without compromising the essence of the algorithm.

2. CONSTRUCTION OF THE BAYES' BOX

For simplicity let us first consider Bayes' formula in the discrete case:

$$p(\theta|x) = \frac{p(\theta) f(x|\theta)}{\sum_{\Theta} p(\theta) f(x|\theta)}. \tag{1}$$

The prior $p(\theta)$ is a probability function expressing prior beliefs about the parameter of interest. As its name suggests it is what the modeler believes about the parameter prior to incorporating information from the data. The likelihood $f(x|\theta)$ is the way information from the data enters the Bayesian update formula which combines the likelihood with the prior to form the posterior $p(\theta|x)$ which is simply an updated prior.

A methodical way of showing the calculation of the posterior using (1) would be through constructing the Bayes' box [1–3] where the parameter, prior, likelihood, and posterior are all clearly displayed.

As an example of constructing the Bayes' box we will use a uniform prior, a binomial likelihood with $n = 10$ trials and $x = 4$ successes. The probability of success (θ) would naturally be a real number between zero and one, however, in this case we will approximate using the following discrete values: $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. This can be done by hand, which has learning benefits at the initial stages of concept making or by using the following suggested *R* function:

```
bayesbox<-function(theta,prior,lik){
h<-prior*lik
post<-h/sum(h)
bbox.mat=matrix(c(theta,prior,lik,h,post),
byrow=FALSE,nr=length(theta),nc=5,
dimnames=list(NULL,c(“theta”,“p(theta)”,“f(x|theta)”,
“h(theta)”,“p(theta|x)”)))
totals<-c(NA,sum(prior),NA,sum(h),sum(post))
return(list(bayesbox=rbind(bbox.mat,totals),h=h))}
```

Table 1. Bayes' box created using R with $n = 10$ trials and $x = 4$ successes

State	θ	$p(\theta)$	$f(x\|\theta)$	$h(\theta)$	$p(\theta\|x)$
1	0.1000	0.1111	0.0112	0.0012	0.0123
2	0.2000	0.1111	0.0881	0.0098	0.0969
3	0.3000	0.1111	0.2001	0.0222	0.2201
4	0.4000	0.1111	0.2508	0.0279	0.2759
5	0.5000	0.1111	0.2051	0.0228	0.2256
6	0.6000	0.1111	0.1115	0.0124	0.1226
7	0.7000	0.1111	0.0368	0.0041	0.0404
8	0.8000	0.1111	0.0055	0.0006	0.0061
9	0.9000	0.1111	0.0001	0.0000	0.0002
Total		1.0000		0.1010	1.0000

The following R code illustrates how to create the Bayes' box (see Table 1) for the above example with a uniform prior:

```
>bayesbox(theta=seq(0.1,0.9,length=9),
  prior=rep(1,9)/9,
lik =dbinom(x=4,size=10,prob=seq(0.1,0.9,length=9)))
```

Table 1 is comprised of six columns. The first column is called "State" and refers to each of the different states of θ. The second column labeled θ is the parameter of interest which arises from the problem under investigation. The third column is the prior and is the prior probability of θ. The fourth column is the likelihood which for discrete x (as here) is the probability of x for a given θ written $f(x|\theta)$. The fifth column is calculated by multiplying the prior by the likelihood and is denoted by $h(\theta) = p(\theta)f(x|\theta)$. The last column is the posterior and is calculated using Bayes' formula in (1), or simply $h(\theta)/\sum h(\theta)$.

2.1. The Usefulness of the Bayes' Box in Teaching

As shown in Table 1, although the prior and posterior columns each add up to one (since they are probability functions of θ), in general the likelihood column does not. This comes from the fact that $f(x|\theta)$ is a probability function for x when θ is constant, while in the Bayes' box it is the other way round (θ is varying and x is constant), hence the name "likelihood function." If one was to sum over x then $\sum_X f(x|\theta) = 1$. However, when f is a likelihood function, x is constant and the column will be summed over θ, thus $\sum_\Theta f(x|\theta) \neq 1$ where $\Theta = \{\theta_i\}$.

Students often find this concept difficult to understand. However, the construction of the likelihood function can be clearly visualized and explained through the Bayes' box. In the case of the example shown in Table 1 with

$x = 4$, it is easy to see that θ is the variable and x is the constant, hence the likelihood function values will not add up to one. In addition, summary statistics can be calculated from the Bayes' box. For example, the posterior mean $E(\theta|x) = \sum \theta_i p(\theta_i|x)$.

2.2. Why Teach MCMC

Since the original development of the M–H algorithm various authors have published this work in different representations (see Table 2 for a sample of these algorithms and their corresponding authors). Even though they have different notations, number of steps, and in some cases slightly different approaches, they are all logically equivalent.

Instructors often present MCMC in a summary fashion out of its context either as a "black box" with no mathematical algorithm shown or a formal representation of the algorithm with some possible description. In the case of the former, Jackman [2009, xxi] is concerned that: MCMC techniques for exploring posterior densities are often regarded in a 'black box' way, with an overly casual concern for what these methods are actually doing, or how well they do it.

> In the latter case the algorithm is immediately translated into a computer program without giving the learner a chance to see the underlying concepts.

The essential reason for teaching MCMC is that the practitioner will more likely succeed in diagnosing and interpreting MCMC output. This is necessary in order that accurate estimates of parameters are made. For example, what is meant by a chain? What is burn-in and how does it occur? Why is auto-correlation an issue and how does it come about with parameters that are conditionally independent within the model? What should be seen in a trace plot? Why do some chains take longer to converge than others and how does this make a difference in the treatment of chains? These questions and more are answered with confidence once a deeper understanding of the algorithm has been attained. Thus, the learner will have a more critical view of MCMC output produced and take the WinBUGS (Windows Bayesian inference Using Gibbs Sampling) manual warning "Beware -MCMC sampling can be dangerous!", more seriously.

3. DIFFICULTIES TEACHING MCMC SIMULATION

Introduction to Bayesian Statistics (STATS 331) is the only undergraduate Bayesian statistics course offered by the statistics department at The University

Table 2. Different ways of representing MCMC algorithm

Given $x^{(f)}$ 1. Generate $Y_e \sim q\left(y \mid x^{(t)}\right)$. 2. Take $X^{(i+1))} = \left\{ \begin{array}{c} Y_i \text{ with probability } \rho\left(x^{(t)}, Y_i\right) \\ x^{(t)} \text{ with probability } 1 - \rho\left(x^{(t)}, Y_t\right) \end{array} \right\}$ Where $\rho(x, y) = min\left\{\frac{f(y)\, q(x \mid y)}{f(x)\, q(y \mid x)}, 1\right\}$ Robert and Casella [98, p. 233]	1. Pick the initial value $x_1 \in R^n$ and set $k = 1$ 2. Draw $y \in R^n$ from the proposal distribution q(x_k,y) and calculate the acceptance ratio $\alpha(x_k, y) = min\left(1, \frac{\pi(y)q\,(y, x_k)}{\pi(y)q\,(x_k, y)}\right)$ 3. Draw $t \in (0, 1)$ from uniform probability density. 4. If $\alpha\,(x_k, y) \geq t$ set $x_{k+1} = y_2$ else $x_{k\div 1} = k_x$ When k = K, the desired sample size, stop, else increase $k \rightarrow k + 1$ and go to step 2. Kaipio and Somersalo [8, p. 96]
A more general form of the Metropolis- Hastings algorithm is as follows Given a current value $x^{(s)}$ of x, 1. Generate x^* from $J_s\left(x^* \mid x^{(s)}\right)$; 2. Complete the acceptance ratio $r = \frac{p_0\,(x^*)}{p_0\left(x^{(s)}\right)} \times \frac{J_s\left(x^{(s)} \mid x^*\right)}{J_s\left(x^{(*)} \mid x^{(s)}\right)};$ 3. Sample n $\sim$ *uniform*(0,1), If u < r set $x^{(s+1)} = x^*$, else set $x^{(s+1)} = x^{(s)}$. Jackman [7, pp. 201–202]	1. Sample $\theta^{\#}$ from n ‘proposal’ or jumpier distribution $J_i(\theta_t^{-}\theta^{(t-1)})$ 2. $r \leftarrow \frac{p(\theta^* \mid y)J_t(\theta^*, \theta^{(t+1)})}{p\left(\theta^{(t+1)} \mid y\right) J_t\left(\theta^{(t-1)}, \theta^*\right)}$ 3. $\alpha \leftarrow \min(r, 1)$ 4. sample U $\sim$ Uni f(0, 1) 5. if U $\leq \alpha$ then 6. $\theta^{(t)} \leftarrow \theta$ 7. else 8. $\theta^{(t)} \leftarrow \theta^{(1-t)}$ 9. end if Hoff, [6, p. 184]

of Auckland. The course was proposed and developed in 2009 by the first named author to provide undergraduates the opportunity of being exposed to the other major paradigm of statistics and was designed as a stepping stone to an existing postgraduate Bayesian course (STATS 731). The students are third-year statistics majors who have completed at least two courses on classic statistics and are encountering Bayesian statistics for the first time. Their assumed knowledge includes: high school mathematics; basic understanding of *R*; an introductory understanding of probability density functions; a basic understanding of the chi-squared test; one and two sample *t*-test; paired sample *t*-test; regression; one-way and two-way ANOVA.

In the past MCMC simulation was introduced later in the course to summarize a continuous posterior distribution as MCMC is largely redundant within the discrete context. This unfortunately coincided with the introduction or use of a number of the following new or complex ideas.

1. Integral calculus: Probabilities are calculated through integrals and not sums, for example $m(x) = \int_{\Theta} p(\theta)f(x|\theta)d\theta$ instead of $m(x) = \sum_{\Theta} p(\theta)f(x|\theta)$.
2. WinBUGS: Implemented to handle the modeling and simulation.
3. R2WinBUGS: Due to graphical shortcomings and limited options for data manipulation in WinBUGS extra packages such as R2WinBUGS are needed to do a thorough analysis with proper interpretation.
4. Complex models - Most real-world problems are modeled with continuous parameters.

The mathematical ideas together with introduction of new software and complex models are all vital and set the scene for the teaching of the Bayesian paradigm with continuous parameters. However, this sudden escalation of complexities takes the focus off the Bayesian paradigm. To motivate a possible solution to this situation the following learning theories were considered.

4. LEARNING THEORIES INFLUENCING THE LECTURER'S PRIOR BELIEFS

In a developing theory, Tall [11–13] has introduced a framework for mathematical thinking based on three worlds of mathematics, the *conceptual embodiment*, *operational symbolism*, and *axiomatic formalism*. The world of conceptual embodiment is based on:

> our operation as biological creatures, with gestures that convey meaning, perception of objects that recognise properties and patterns. . .and other forms of figures and diagrams (Tall, [13, p. 22]).

In this world, we think about the things around us in the physical world, and it:

> includes not only our mental perceptions of real-world objects, but also our internal conceptions that involve visuo-spatial imagery (Tall, [11, p. 30]).

The world of operational symbolism is the world of practicing sequences of actions which can be achieved effortlessly and accurately. Finally the world of axiomatic formalism:

> builds from lists of axioms expressed formally through sequences of theorems proved deductively with the intention of building a coherent formal knowledge structure" (Tall, [13, p. 22]).

On the other hand, there are some things that words cannot quite express and must be performed to gain better understanding. This is the heart of Bruner's [4] enactive representation (e.g., learning a sport or driving a car [4]). According to his theory there are three ways to make experience a model of the world: action, visual or other sensory information, and language. These forms of representation are known as enactive, iconic, and symbolic respectively.

In line with Bruner's ideas the first named author introduced the MCMC algorithm earlier using a tactile method with a discrete parameter to give students a hands-on experience. Moreover, the teaching of the formal algorithm (Tall's formal world) was delayed until the tactile version of the algorithm was taught. To reduce the concentration of new concepts and ideas later in the course, an implementation of WinBUGS and R2WinBUGS (this interfaces *R* with WinBUGS) for simple cases was carried out.

5. DESCRIPTION OF TEACHING THE M–H ALGORITHM

In this section we will give a detailed description of the order and way the MCMC algorithm is introduced. The first step is to discuss the underlying conceptual ideas on how proposed values are accepted/rejected in the algorithm. This is followed by a tactile demonstration of the algorithm. To prove that the tactile method would give a sample from the posterior, Markov Chain theory is employed to produce the steady state solution. This solution is then compared with the exact posterior displayed from the Bayes' box, which are indeed the same. Finally, we show how any number of states can be viewed as multiple applications of the two state problem.

5.1. A Conceptual Description of the Algorithm

Consider a proposal distribution that randomly produces states of θ from the parameter space. On the basis of an acceptance probability these states can be accepted or rejected. Those that are accepted form the posterior sample and will have a relative frequency close to the posterior probability $p(\theta|x)$. This is accomplished by using a proposal distribution $q(\theta)$ and changing it into the posterior by accepting or rejecting what it proposes based on natural properties of $h(\theta)$ (see Table 1) which is proportional to the posterior probability and is always present in Bayesian analyses.

One key idea is that a proposal distribution will be changed (filtered) into the posterior, one proposed θ at a time. An ideal proposal to start with is the uniform distribution since it gives equal weighting to all states in the parameter space.

Suppose we have a binomial experiment with n trials and x successes. We will limit θ (the probability of success) to two possible states namely θ_1 and θ_2 and a uniform proposal $q(\theta)$, which means $q(\theta_1) = q(\theta_2) = 0.5$. In this case it would be better to use a non uniform prior, to prevent confusion with the proposal. As part of any simulation technique, an initial value must be assigned. The candidates would then be $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \ldots, \theta^{(N)}$, where N is the number of candidates proposed and $\theta^{(t)} \in \{\theta_1, \theta_2\}$.

Figure 1 shows $h(\theta)$ for the given two states. The heights of h corresponding to the first and second states are $h(\theta_1) = h_1$ and $h(\theta_2) = h_2$, respectively. The initial value is set manually (irrespective of the heights h_1 and h_2). Suppose in this case $\theta^{(0)} = \theta_2$. The uniform distribution $(q(\theta))$ will then randomly supply another value of θ say θ_1. This will be accepted with less probability than θ_2 since h_1 is smaller than h_2. We do not wish to always reject θ_1 whenever proposed, since it has a non-zero chance of occurring in the posterior distribution though with less frequency than θ_2 .

θ_1 will be accepted with probability h_1/h_2, if this is not accepted it will be rejected and the sampler will remain at the previous value. Suppose θ_1 is accepted. If the candidate distribution now proposes θ_2 it will be accepted with probability one.

An important point relative to Markov chains is the fact that each new simulated value depends on where the sampler is and not where it has been. The

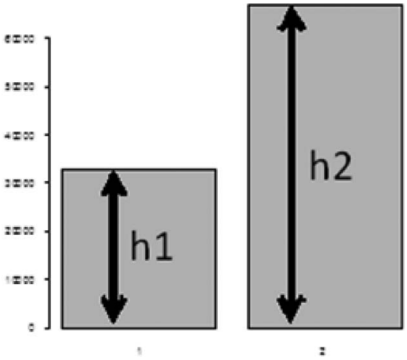


Figure 1. h function proportional to the posterior of θ.

above acceptance and rejection process can be summarized in the following formula

$$\alpha_{i,j} = \min\left\{1, \frac{h_j}{h_i}\right\}, \tag{2}$$

where $\alpha_{i,j}$ is the probability of accepting state j given that the sampler is at state i. Note that the uniform distribution is supplying candidate $\theta\prime$ values independently, therefore from a stochastic perspective the candidate values are independent, hence, $q(\theta^{(t)}|\theta^{(t-1)}) = q(\theta^{(t)}) = 0.5$.

5.2. A Tactile Method of Teaching the M–H Algorithm

The tactile method starts with creating an artificial two-state Bayes' box (see Table 3) having the following properties: (I) prior distribution $p(\theta)$ different from the proposal $q(\theta)$; and (II) the smallest ratio of h values will give the only non-unary acceptance probability (see Appendix A for the *R* code). The prior is constructed so that the required h ratios will agree with probabilities generated from the throw of a fair die.

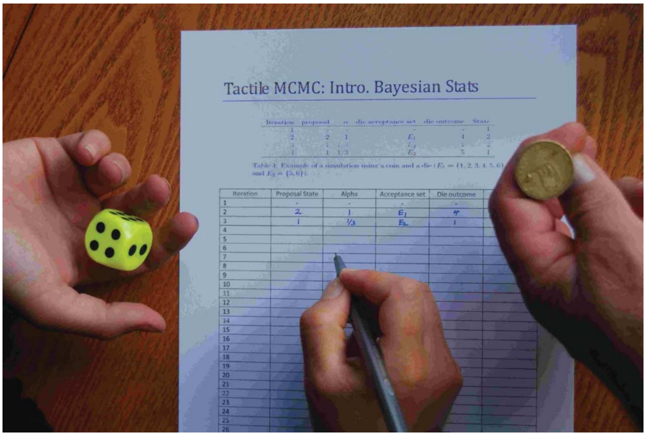
The simulation can be performed with a coin (the proposal) and a die (the acceptance) where students work co-operatively in groups of three as shown in Figure 2. The instructor can create a barplot from the frequencies of the two states generated by the participants (see Appendix B for a copy of the worksheet).

To start the tactile simulation an initial state must be assigned (either state 1 or state 2). By default this will always be accepted. State 1 represented by a head has parameter value 0.4 and state 2 represented by a tail has parameter value 0.8.

From column h in Table 3 the smallest ratio of the h values is $1/3$. This means that $\alpha_{i,j} \in \{1, 1\backslash 3\}$ ($\alpha_{1,2} = 1$ and $\alpha_{2,1} = 1\backslash 3$), which can be simulated with the roll of a die. Since there are only two probabilities to emulate for the acceptance, we need to assign die outcomes to them. For probability one (with probability one the proposed state is accepted), we must have the event $E_1 = \{1, 2, 3, 4, 5, 6\}$ obtaining any one of the six possible sides. For probability

Table 3. A Bayes' box designed for a tactile simulation

State	θ	p(θ)	f(x| θ)	h(θ)	P(θ|x)
1	0.4000	0.0420	0.2007	0.0084	0.2500
2	0.8000	0.9580	0.0264	0.0253	0.7500
Totals		1.0000		0.0338	1.0000



Tactile MCMC: Intro. Bayesian Stats

Iteration	proposal	α	die acceptance set	die outcome	State
1	-	-	-	-	1
2	2	1	E_1	4	2
3	1	1/3	E_2	4	2
4	1	1/3	E_2	5	1

Table 4: Example of a simulation using a coin and a die ($E_1 = \{1, 2, 3, 4, 5, 6\}$ and $E_2 = \{5, 6\}$).

Iteration	Proposal State	Alpha	Acceptance set	Die outcome
1	-	-	-	-
2	2	1	E_1	4
3	1	1/3	E_2	1
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				

Figure 2. Tactile simulation in action.

1/3, we will arbitrarily assign the event $E_2 = \{5, 6\}$ obtaining a five or a six. An example of a possible sequence of the simulation is presented in see Figure 3 .

In practice the simulation shown Figure 3 (recorded in Table 4) can be repeated by making reference only to a bar plot of the proposal and the h function (see Figure 4). The procedure can be sped up by noticing that there is no need to roll the die when the acceptance probability is one (going from low to higher h values) hence, whatever is proposed in such circumstances is accepted. Notice how the posterior is proportional to h. Looking at the plot of the h function and starting initially with state 1, suppose state 2 is proposed, then we will accept this because it has a higher (relative to the current state) posterior probability. Now starting from state 2, suppose state 1 is proposed which has a lower posterior probability (smaller h), hence accepted with probability ($h_1 \backslash h_2 = 1/3$). If this was rejected then remain at state 2. Next, suppose state 1 is proposed again, then the acceptance probability will be 1/3. If this is accepted then the current state will be state 1, hence the sampler produced states 1,2,2,1 thus far.

Table 4 can be made dynamic and used for teaching purposes by a powerpoint slide (see Figure 5) where H represents candidate state 1 and T state 2.

A summary of the simulation done in class can be included in the Bayes' box (see Table 5). The last column refers to the relative frequency of the simulated states.

5.3. A Sketch Proof of the Simulation

Will the method of accepting and rejecting proposals in the above example eventually give the stationary distribution? Since the simulation is a Markov

1. Set initial state: State 1;
2. Throw the coin and record the outcome (Head=state 1; Tail=state 2): State 2;
3. Calculate alpha: 1 in this case;
4. Choose the appropriate die acceptance set: In this case it will be E_1;
5. Throw the die and record the outcome, if the number is in the acceptance set then accept the proposal: Suppose in this case the outcome is 4. Since $4 \in E_1$, we accept the proposal, thus the sampler moves to state 2;
6. Throw the coin: State 1 this time;
7. Calculate alpha: This time alpha becomes $\frac{1}{3}$ (because we are at state 2 and proposal is at state 1);
8. Choose the appropriate die acceptance set: This time E_2;
9. Throw the die and get a 1: Therefore reject the proposal (because $1 \notin E_2$) and remain in state 2;
10. Throw the coin: Say state 1 again;
11. Calculate alpha: alpha is $\frac{1}{3}$ again (because we are at state 2 and proposal is at state 1);
12. Choose the appropriate die acceptance set: E_2 again;
13. Throw the die and get a 5: Therefore accept the proposal (because $5 \in E_2$) and move to state 1.

Figure 3. A possible tactile simulation sequence.

Table 4. Example of a simulation using a coin and a die ($E_1 = \{1, 2, 3, 4, 5, 6\}$ and $E_2 = \{5, 6\}$)

Iteration	Proposal	α	Die acceptance set	Die outcome	State
1	–	–	–	–	1
2	2	1	E_1	4	2
3	1	1/3	E_2	1	2
4	1	1/3	E_2	5	1
⋮	⋮	⋮	⋮	⋮	⋮

chain process we can make a transition matrix from the sampler probabilities α_{ij} and q_j. The transition probabilities are calculated by using the following formula

$$p_{ij} = \alpha_{ij} q_j, \; i \neq j,$$

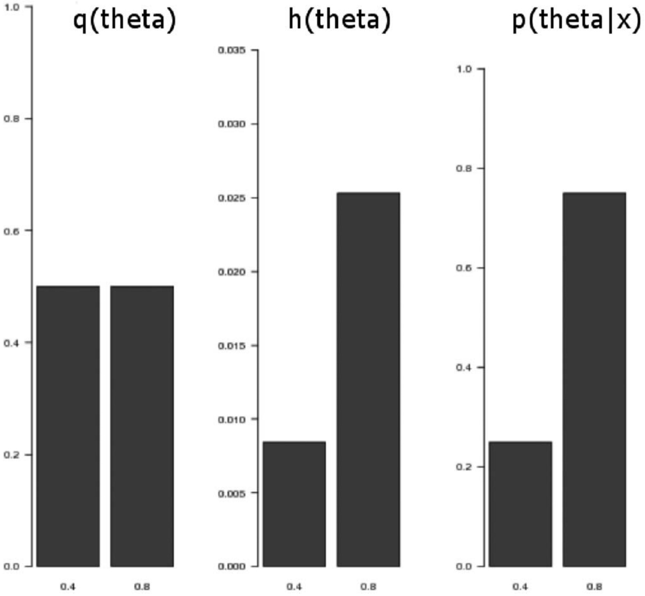
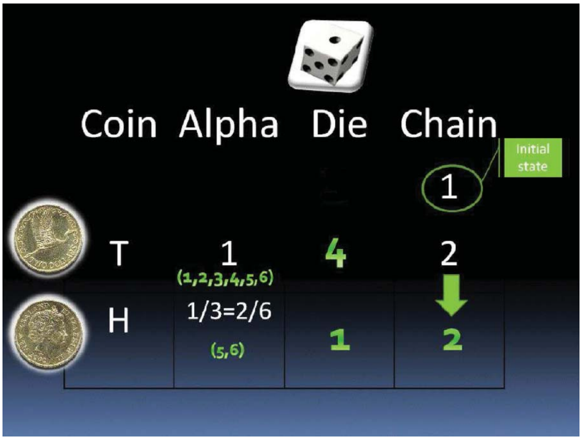


Figure 4. Barplots of the proposal, the function *h*, and the exact posterior.



Coin	Alpha	Die	Chain
			1
T	1 (1,2,3,4,5,6)	4	2
H	1/3=2/6 (5,6)	1	2

Figure 5. A useful Powerpoint slide for teaching the simulation.

where p_{ij} is the transition probability of going from state i to j. Hence, the transition matrix

$$P = \begin{pmatrix} p_{11} & p_{12} = 1 \times 1/2 \\ p_{21} = 1/3 \times 1/2 & p_{22} \end{pmatrix},$$

Table 5. Simulated posterior added to the Bayes' box

State	θ	$p(\theta)$	$f(x\|\theta)$	$h(\theta)$	Postexact	Sim
1	0.4000	0.0420	0.2007	0.0084	0.2500	0.2513
2	0.8000	0.9580	0.0264	0.0253	0.7500	0.7487
Totals		1.0000		0.0338	1.0000	1.0000

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/6 & 5/6 \end{pmatrix}.$$

A corresponding stochastic eigenvector v gives the long-term (and steady-state) state vector

$$v\mathrm{P} = v.$$

As the entries of P are positive, the Markov chain is regular and hence there is a unique probability vector v. To find v we will solve the system $\mathrm{v(I - P) = 0}$ as follows

$$\mathrm{v(I - P) = 0},$$

$$\begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} 1/2 & -1/2 \\ -1/6 & 1/6 \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}.$$

Where I is the dentity matrix. The general solution of the system is

$$v = \begin{pmatrix} v_1 & v_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3}v_2 & v_2 \end{pmatrix}.$$

For v to be a probability vector,
$1 = v_1 + v_2 = \frac{1}{3}v_2 + v_2 = \frac{4}{3}v_2$ which gives, $v_2 = \frac{3}{4}$ and thus $v_1 = \frac{1}{4}$.
Hence

$$v = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

Notice that this is the exact posterior as shown in Table 5.

Table 6. Three states

State	θ	$p(\theta)$	$f(x\|\theta)$	$h(\theta)$	$p(\theta\|x)$
1	0.2000	0.3333	0.0881	0.0294	0.2949
2	0.5000	0.3333	0.2051	0.0684	0.6867
3	0.8000	0.3333	0.0055	0.0018	0.0184
		1.0000		0.0996	1.0000

6. MULTIPLE STATES

The two-state MCMC simulation is a logical place to start in terms of moving to multiple states since the sampler will always be comparing a proposal with a previously accepted state. Take for example a binomial experiment in which there are $n = 10$ trials and $x = 4$ successes and we are concerned with obtaining the posterior distribution of θ, the probability of a success. The Bayes' box given in Table 6 is a summary of the calculation of the posterior $p(\theta|x)$ for the three states of the parameter $\theta \in \{0.2, 0.5, 0.8\}$. Notice that the prior distribution was chosen to be uniform, which means that for each of the three states $p(\theta_i) = 1/3$.

In order to simulate the posterior, MCMC was performed not by throwing a coin and die (since the acceptance probabilities cannot in general be created through throws of a die) but by using the formula in (2) and writing an *R* function to accept a proposal with probability $\alpha_{i,j}$. A code snippet which performs the required action is given below:

```
for(i in 2:n){ # starts at 2 because initial
value was previously assigned
sample(1:length(h),1,replace=TRUE)->prop[i] # q,
prop=proposal state (1,2)
alpha[i]=min(1,h[prop[i]]/h[post[i-1]])#h is a
vector
u[i]=runif(1) # a random number between 0 and 1
if(u[i]<=alpha[i]){post[i]<-prop[i]} # if u[i]
is less than alpha, accept proposal
else{post[i]=post[i-1]} # otherwise retain the
previous post value
}
```

In Table 7 the first 20 (20/20000) iterates are displayed. Notice that $\alpha_{1,2} = \alpha_{3,2} = \alpha_{3,1} = \alpha_{i,i} = 1$. The only interesting acceptance probabilities relate to transitions from states that have high posterior probability to those with lower probabilities, hence from Table 7, $\alpha_{2,1} = 0.4295$, $\alpha_{1,3} = 0.0625$, and $\alpha_{2,3} = 0.0268$.

Table 7. First 20 iterates of the MCMC sampler for three states

Iteration	Prop	U	Alpha	Post
1	1.0000			1.0000
2	1.0000	0.8640	1.0000	1.0000
3	2.0000	0.5169	1.0000	2.0000
4	2.0000	0.4019	1.0000	2.0000
5	1.0000	0.6313	0.4295	2.0000
6	2.0000	0.3237	1.0000	2.0000
7	2.0000	0.6505	1.0000	2.0000
8	2.0000	0.6063	1.0000	2.0000
9	2.0000	0.4484	1.0000	2.0000
10	3.0000	0.4519	0.0268	2.0000
11	2.0000	0.3654	1.0000	2.0000
12	2.0000	0.5190	1.0000	2.0000
13	2.0000	0.3559	1.0000	2.0000
14	2.0000	0.1047	1.0000	2.0000
15	1.0000	0.7313	0.4295	2.0000
16	1.0000	0.2823	0.4295	1.0000
17	3.0000	0.0095	0.0625	3.0000
18	3.0000	0.8564	1.0000	3.0000
19	3.0000	0.4508	1.0000	3.0000
20	1.0000	0.3310	1.0000	1.0000

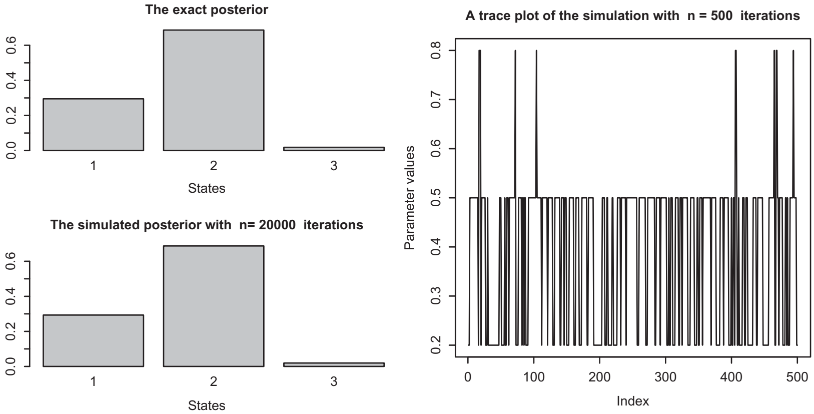


Figure 6. Simulating the posterior for three states.

The relative frequencies of the three states give simulated posterior probabilities, these are plotted in Figure 6 along with the exact posterior barplot and a trace of the movement of the states through 500 iterations. The same was repeated for 10 states and the plots given in Figure 7.

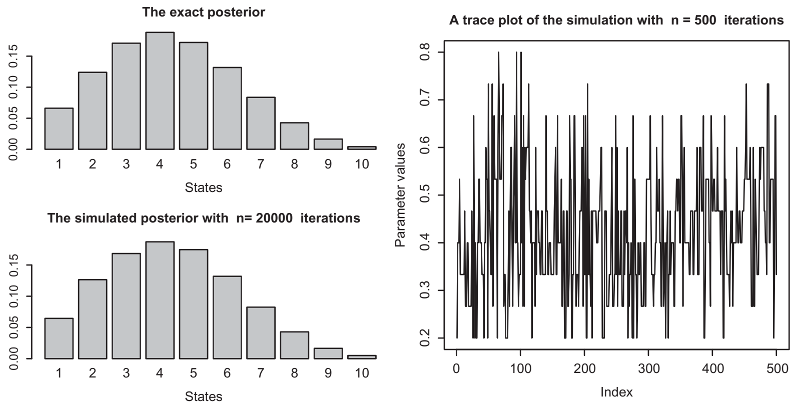


Figure 7. Simulation for 10 states.

7. CONTINUOUS PARAMETERS

The development thus far has been to approximate a continuous parameter with a discrete one. This enabled a systematic teaching of MCMC from Bayes' boxes and coin-die flips to an *R* function that runs the simulation. Now that this has been established it takes very little to move to the completely continuous case. Since by now the students have come to use and appreciate a computer function that runs the simulation by itself, we need only adjust the code with which they are already familiar. The important point for the students to understand is the fact that in the case of the binomial parameter θ, we now have an infinite number of values from $\{0, \ldots, 1\}$ (none of which will repeat). Therefore it makes little sense to associate a state number with each parameter value but to deal only with the parameter values and use a histogram instead of a barplot to display the posterior. Below is an R code snippet showing the minor changes needed to accomplish this:

```
for(i in 2:n){ # starts at 2 because initial value
   assigned
prop[i]=runif(1) # prop and post are vectors
   containing
parameter values
alpha[i]=min(1,h(prop[i])/h(post[i-1]))# h is a
function
u[i]=runif(1)
if(u[i]<=alpha[i]){post[i]<-prop[i]}
else {post[i]<-post[i-1]}
}
```

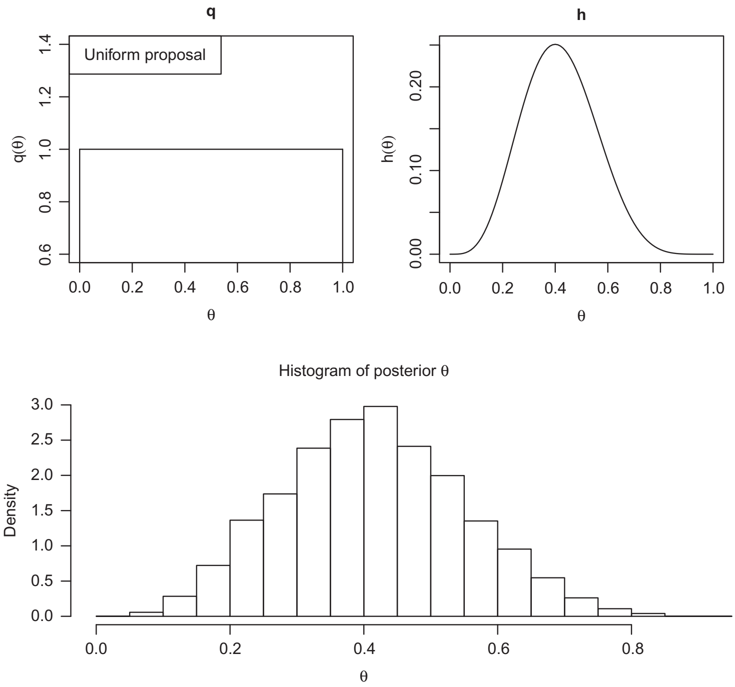


Figure 8. Posterior histogram for the same Binomial experiment ($n = 10, x = 4$).

A plot of $q(\theta)$, $h(\theta)$ and a histogram of the posterior θ values for the same binomial experiment discussed above is shown in Figure 8.

8. CONCLUSION

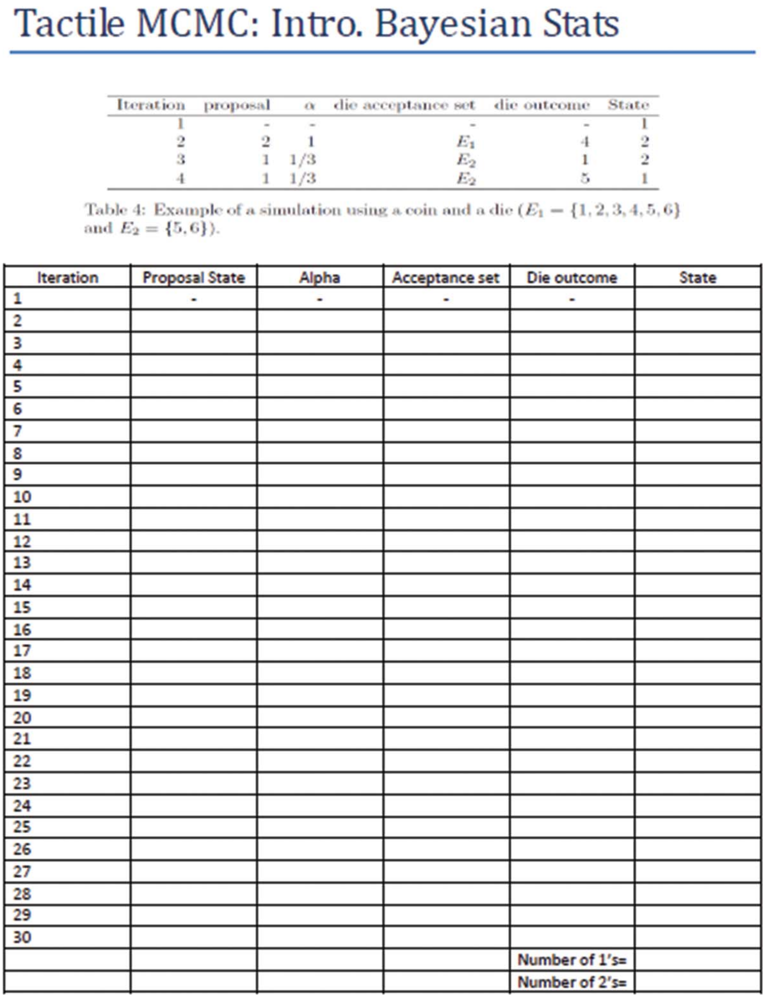
In the authors' experience teaching and learning MCMC is often difficult. Although, most learners use the algorithm mechanically, their conceptual understanding is often limited. In many text books the algorithm is typically presented formally (Tall's formal world) with much assumed knowledge such as: "accept proposal with probability 1/3". The two-state tactile technique clarifies the heart of MCMC simulation. The consideration of what state the sampler is in (not where it has been) when deciding acceptance probabilities is the Markov Chain part and throwing the coin and die relates to the Monte Carlo part. The tactile simulation teaches MCMC in a way that would be almost impossible otherwise. The learners not only see a symbolic representation but also feel the simulation through the coin and die (Bruner's iconic and enactive). The method addresses the *how* question by its use of the coin and the die and the *why* question is answered by linear algebra. This simulation does not replace the formal algorithm but can be helpful as an introduction to it and progress to the formal world with more understanding.

The simulation discussed in this article is two state but can be extended to a larger number and finally into an infinite number for a continuous parameter. The example discussed here used a simplified form of M–H algorithm with a uniform proposal, a further development would be to use a non-uniform proposal. The two-state method as outlined in this article has been presented at statistics and mathematics conferences (USCOTS 2011 and New Zealand Mathematics Society (NZMS) 2010) by the authors. The method was also demonstrated recently at a 2-day Bayesian course to a group of scientists and lecturers organized by the New Zealand Social Statistics Network (NZSSN 2011). Our observation thus far is that students and researchers welcome this hands-on method and have satisfying experiences.

APPENDIX A. *R* CODE USED TO MAKE TWO-STATE BAYES' BOXES

```
bb2<-function(k=1,lik,theta){ # K=1...6, K/6
(acceptance prob)
library(xtable)
lik1<-lik[1]
lik2<-lik[2]
pi1<-(k/6*lik2/lik1)/(1+k/6*lik2/lik1)
prior=c(pi1,1-pi1)
h<-prior*lik
post=h/sum(h)
mat<-cbind(theta,prior,lik,h,post)
rownames(mat)<-1:length(lik)
Totals=c(NA,sum(prior),NA,sum(h),sum(post))
mat2=rbind(mat,Totals)
layout(matrix(c(1,2,3,4),nr=1,nc=4,byrow=TRUE))
barplot(matrix(prior,nc=2,nr=1,byrow=TRUE,dimnames=
  list(c(“Coin”),theta)),
ylim=c(0,1),las=1,main=“Prior NOT the proposal”)
barplot(matrix(c(0.5,0.5),nc=2,nr=1,byrow=TRUE,
  dimnames=list(c(“Coin”),theta)),
ylim=c(0,1),las=1,main=“Proposal/n Uniform”)
barplot(matrix(h,nc=2,nr=1,byrow=TRUE,dimnames=
  list(c(“Coin”),theta)),
ylim=c(0,max(h)+0.5*max(h)),las=1,main=“Proportional
   to target/n h”)
barplot(matrix(post,nc=2,nr=1,byrow=TRUE,dimnames=
  list(c(“Coin”),theta)),
ylim=c(0,max(post)+0.5*max(post)),las=1,main=
  “Posterior target/n post”)
return(list(bbox=mat2,latex=xtable(mat2,digits=4),
  mat=mat,h=h))
}
e.g. bb2(k=2,lik=dbinom(x=5,size=10,prob=
  c(0.4,0.8)),theta=c(0.4,0.8))
```

APPENDIX B. WORKSHEET FOR THE TACTILE SIMULATION



Tactile MCMC: Intro. Bayesian Stats

Iteration	proposal	α	die acceptance set	die outcome	State
1	-	-	-	-	1
2	2	1	E_1	4	2
3	1	1/3	E_2	1	2
4	1	1/3	E_2	5	1

Table 4: Example of a simulation using a coin and a die ($E_1 = \{1,2,3,4,5,6\}$ and $E_2 = \{5,6\}$).

Iteration	Proposal State	Alpha	Acceptance set	Die outcome	State
1	-	-	-	-	
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
				Number of 1's=	
				Number of 2's=	

REFERENCES

1. Albert, J. 2002. Teaching introductory statistics from a Bayesian perspective. In *The Sixth International Conference on Teaching Statistics (ICOTS6)*, pp. 1–6. Cape Town, South Africa: Brian Phillips.
2. Berry, D. A. 1996. *Statistics: A Bayesian Perspective*. Belmont, CA: Duxbury Press.
3. Bolstad, W. M. 2002. Teaching Bayesian statistics to undergraduates: Who, what, where, when, why and how, In *The Sixth International Conference on Teaching Statistics (ICOTS6)*, pp. 1–6. Cape Town, South Africa: Brian Phillips.
4. Bruner, J. S. 1966. *Toward a Theory of Instruction*. New York: Norton.
5. Casella, G. and E. I. George. 1992. Explaining the Gibbs sampler. *The American Statistician*. 46(3): 167–174.
6. Hoff, P. D. 2009. *A First Course in Bayesian Statistical Methods*. New York: Springer.
7. Jackman, S. 2009. *Bayesian Analysis for the Social Sciences*. Chichester, UK: Wiley.
8. Kaipoi, Y. and E. Somersalo. 2005. *Statistical and Computational Inverse Problems*. New York: Springer.
9. Robert, C. P. and G. Casella. 1998. *Monte Carlo Statistical Methods*. New York: Springer.
10. Stewart, W. 2007. Introducing Bayesian statistics to undergraduates. In *Proceedings of the sixth Southern Hemisphere Conference on Mathematics and Statistics Teaching and Learning (DELTA'07)*, pp. 151–159. El Calafate, Argentina: D'Arcy-Warmington, Anne and Luaces, Victor Martinez and Oates, Greg and Varsavsky, Cristina.
11. Tall, D. O. 2004. Building theories: The three worlds of mathematics. *For the Learning of Mathematics*. 24(1): 29–32.
12. Tall, D. O. 2008. The transition to formal thinking in mathematics. *Mathematics Education Research Journal*. 20(2): 5–24.
13. Tall, D. O. 2010. Perceptions operations and proof in undergraduate mathematics. *Community for Undergraduate Learning in the Mathematical Sciences (CULMS) Newsletter*. 2: 21–28.

BIOGRAPHICAL SKETCHES

Wayne Stewart received his Ph.D. in Bayesian statistics at The University of Auckland, New Zealand with a thesis about the local sensitivity of the posterior to inputs. He is currently teaching statistics courses at the Mathematics Department, University of Oklahoma. His primary research interests are in teaching and learning Bayesian statistics and advanced statistical thinking.

Sepideh Stewart received her Ph.D. in mathematics education at The University of Auckland, New Zealand. She is currently an Assistant Professor at the Department of Mathematics, University of Oklahoma. Her primary research interests are in teaching and learning linear algebra, Bayesian statistics and advanced mathematical and statistical thinking.