Journal of Statistics Education

Teaching Bayesian Statistics in a Health Research Methodology Program

Eleanor M. Pullenayegum
Lehana Thabane
McMaster University
St. Joseph's Healthcare

Journal of Statistics Education Volume 17, Number 3 (2009), www.amstat.org/publications/jse/v17n3/pullenayegum.html

---

Key Words: Bayesian methods; Priors; Convergence; Health sciences.

Abstract

Despite the appeal of Bayesian methods in health research, they are not widely used. This is partly due to a lack of courses in Bayesian methods at an appropriate level for non-statisticians in health research. Teaching such a course can be challenging because most statisticians have been taught Bayesian methods using a mathematical approach, and this must be adapted in order to communicate with non-statisticians. We describe some of the examples we used whilst teaching a course in Bayesian methods to a group of health research methodologists.

1. Introduction

Bayesian methods were introduced in the 18th century by Bayes [Bayes, 1763] and Laplace, and have become feasible in practical applications in the last ten years. Despite their advantages and a wealth of statistical methodological research on the subject (see e.g. Ashby [Ashby, 2006] for a review), their adoption in health research remains insular. We believe that this is due to a lack of biostatisticians trained in the practical implementation of the methods, a lack of clinicians and epidemiologists familiar with the methodology, and shortcomings in communication between statisticians and epidemiologists. The purpose of this paper is to share some examples that we found helpful in promoting discussion and building practical skills in a class of health research methodologists.

We teach in a Health Research Methodology program, admitting students to MSc and PhD degrees in one of five streams (biostatistics, clinical epidemiology, health technology assessment, health services research, population and public health). Courses are not stream-specific, and so students from different streams are taught in the same classes. Theoretical courses already exist in statistics programs, so we chose to keep our focus on concepts and practical implementation. The aims of the course were to familiarise students with Bayesian ideas, to highlight the differences between Bayesian and frequentist methods, and to equip students with the basic tools to design, conduct, analyse and report a simple Bayesian study. We covered the basics (Bayesian thinking, analyses using WinBUGS [Lunn et al, 2000], choosing priors, assessing convergence, reporting), then highlighted specific applications of Bayesian methodology to health research (trial design, meta-analysis, missing data, decision-making). A detailed syllabus is in Appendix A.

Our course textbook was "Bayesian Approaches to Clinical Trials and Health-Care Evaluation" by Spiegelhalter *et al* (2004). Before each class, students did some preparatory reading, then participated in a seminar and discussion to deal with the key concepts (e.g. How do you choose a prior? Doesn't the prior bias the results? What are the differences between a Bayesian and frequentist meta-analysis? How does Bayesian thinking impact the design and conduct of a clinical trial?). The following week, students applied their learning in a lab class, handing in their work the week of the next lab class. This allowed students to deal with concepts first, without becoming bogged down in computational problems, and ensured that students explored and experimented with the methodology whilst receiving programming support and feedback. The remainder of this paper describes some specific examples of key topics.

We now describe two examples of how we communicated the course material. The first describes some examples used to initiate class discussion on assessing convergence. The second example is a lab class on prior selection.

2. Convergence

The WinBUGS manual bears the warning "Beware – MCMC sampling can be dangerous!". One such danger is non-convergence of the Markov Chain Monte Carlo (MCMC) sampler. The output of the MCMC sampler should be used only when the sampled values are coming from the target posterior distribution. We begin this section with a brief overview of MCMC methods, then move to specific examples used to introduce students to the issues.

Background

The concept at the heart of Bayesian statistics is that the posterior distribution of a parameter θ given data, is proportional to the likelihood of the data multiplied by the prior distribution for θ, i.e. $f(\theta|data) \propto f(\theta)L(data|\theta)$. Whilst the concept is simple, its application quickly becomes complex, particularly when the unknown parameter θ is multi-dimensional. In these cases it is often impossible to derive the posterior distribution of θ analytically. Instead, computer-intensive methods for sampling from the posterior can be used. These sampling methods set up a Markov Chain which will eventually sample from the posterior distribution and thus provide most of the necessary information. This is called Markov Chain Monte Carlo (MCMC), of which the Gibbs sampler [Geman & Geman, 1984] is one example.

The Gibbs sampler is used when θ is multi-dimensional. If θ_k is the k^{th} element of the p-dimensional vector θ, and $\theta_{(k)}$ denotes the vector of all elements of θ except the k^{th}, then the k^{th} componentwise posterior is $f(\theta_k|data, \theta_{(k)})$. Although deriving the joint posterior distribution $f(\theta|data)$ may be difficult or impossible, it will often be easier to derive the componentwise posteriors. The Gibbs sampler begins with an initial estimate of the vector θ, denoted θ^0, then samples $\theta_1^1,\ldots,\theta_p^1$ in turn from the relevant componentwise posteriors $f(\theta_k|data, \theta_{(k)})$. This completes one cycle of the Gibbs sampler. We then proceed to take samples $\theta^2,\ldots,\theta^N$ by replacing θ^0 with our current estimate of θ.

The Gibbs sampler works because the estimates $\theta^0,\ldots,\theta^N$ form a Markov chain whose stationary distribution is the posterior distribution of interest, i.e. $f(\theta|data)$. Since (under regularity conditions) a Markov chain will converge to its stationary distribution, if we sample from the Gibbs sampler for long enough, we will eventually be sampling from the posterior distribution of interest. The question in assessing convergence is, "How long is 'long enough'?" It is a difficult question because the answer depends on the prior distributions, the likelihoods, the initial estimates, and the data, so that it is not possible to identify a single number of iterations for which all chains will have converged.

Teaching convergence was challenging, as none of our students had heard of a Markov Chain, and the majority had not seen a definition of convergence. To compound the problem, we were not able to find any literature on convergence at a suitable level. Rather than have students read technical papers as preparation for the class, we asked them to run some examples that were intended to raise questions about convergence issues. We did not direct students' attention to specific features of the examples, but rather asked them to comment on any unusual results, compare results between examples, and try to explain their observations. Although students had varying degrees of mathematical and statistical training, because they were all scientists, the process of observing and attempting to interpret findings was familiar to them. We now describe the examples used, and for each outline some of the questions it raises. Answers to the questions are given in Section 2.5.

2.1 How many updates?

The first example "How many updates?" used the Seeds dataset [Crowder, 1978] from the WinBUGS example sets. Briefly, the data records the proportion of seeds germinating from an experiment which used 21 plates and looked at two factors: type of seed (aegyptiaco 73 vs. 75) and type of root extract (cucumber vs. bean). The analysis is by a random effects logistic regression with intercept α_0, log odds ratio (OR) of seed type α_1, log OR of root extract α_2, and α_{12} the interaction between the two factors. The random effects for plates are b[i], which are assumed to come from a Normal distribution with precision tau. If p_{ij} is the probability of seed j on plate i germinating, the regression equation is thus

$$\text{logit}(P_{ij}) = \alpha_0 + \alpha_1 I(\text{type}_{ij} = \text{aegyptiaco 73}) + \alpha_2 I(\text{root}_{ij} = \text{cucumber}) + \alpha_{12} I(\text{type}_{ij} = \text{aegyptiaco 73})\, I(\text{root}_{ij} = \text{cucumber}) + b[i]$$

This connects with other topics in the course, as random effects logistic regression is also used in the context of cluster-randomized trials, multi-centre trials, and meta-analyses, which were discussed in a separate class (see Appendix; note that in our offering convergence preceded cluster-randomized trials due to the availability of the computer lab and instructors, but ideally, the two sessions would be switched). Using initial values of list(alpha0 = 0, alpha1 = 0, alpha2 = 0, alpha12 = 0, tau = 1, b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)), we asked students to compare summary statistics at 1000, 2000, 10000 and 20000 iterations. Since tau starts in the tails of the posterior distribution and autocorrelations are high, convergence is fairly slow.

Table 1: Summary statistics for tau, the precision of the random effects in the Seeds example

Iterations	*Mean*	*SD*	*MC Error*	*2.5%*	*Median*	*97.5%*
1000	114.2	547.0	52.0	2.339	16.32	806.6
2000	67.36	390.2	27.24	2.668	13.28	444.5
10,000	64.55	269.0	11.65	2.81	13.33	549.6
20,000	53.39	220.3	6.873	2.757	12.57	415.2

SD = Standard deviation; MC = Monte Carlo

The results in Table 1 show that the posterior distribution of the precision parameter tau changes as more iterations are run: the posterior mean decreases each time more iterations are added, as does the standard deviation. The posterior median in general decreases with more iterations, but not uniformly. A natural question to ask is, "How do you know when you have done enough iterations?". The answers to this question, and to further questions raised in Sections 2.2 – 2.4, will be deferred till Section 2.5 to allow for a fuller discussion.

2.2 Initial Values

To emphasize the point, students then ran the analysis again, this time with two sets of initial values: the set above and list(alpha0=1, alpha1=-0.1, alpha2=0.1,alpha12=-1,tau=0.01,b=c(10,10,10,10,10,10,10,10,10,10,10,10,-10,-10,-10,-10,-10,-10,-10,-10,-10,-10)). After running 1000 iterations, they were asked to look at the history plot (Figure 1) then to run another 9000 iterations and look at the autocorrelation and Brooks-Gelman-Rubin (BGR) plots (Figures 2 and 3).

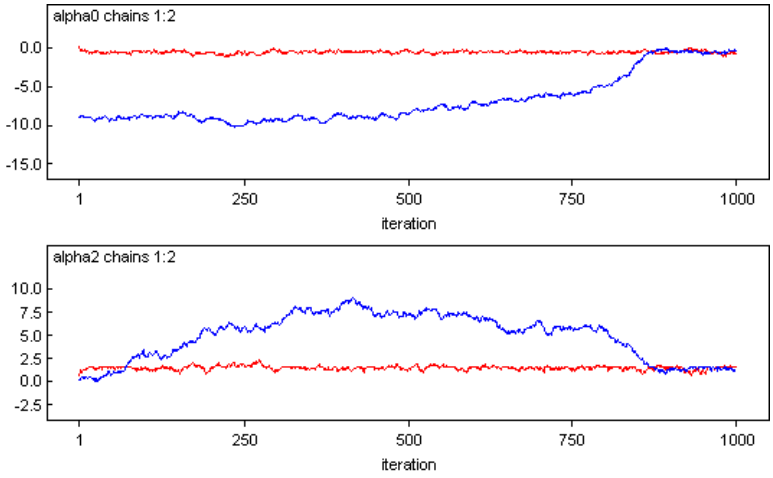

Figure 1: History plots for the Seeds example

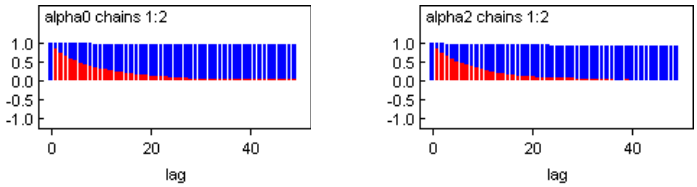

Figure 2: Autocorrelation plots for the Seeds example

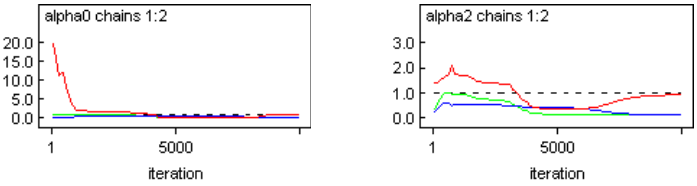

Figure 3: Brooks-Gelman-Rubin plots for the Seeds example

In Figure 1, it is evident that the two chains are giving very different estimates of alpha0 and alpha2 until around 900 iterations, at which point the second chain appears to sample from the same space as the first chain. One question that students asked was, "Why do the history plots for alpha 2 overlap and then stray apart again?"

Figures 2 and 3 show autocorrelation and Brooks-Gelman-Rubin plots, neither of which will be familiar to students. They will be most helpful when examined alongside the corresponding plots for the Normal example below, and so we postpone comment on these until Section 2.3.

2.3 A rapidly converging chain

As a comparison, students also ran a simple Normal example (see Table 2). In this example, the data consist of ten observations drawn from a Normal distribution of mean mu and precision tau. Mu is given a normal prior with mean 0 and precision 0.01, and tau is given a Gamma (0.001,0.001) prior. Students were asked to run 1000 iterations initially, then another 9000 iterations, and to look at history plots, autocorrelations, and BGR plots.

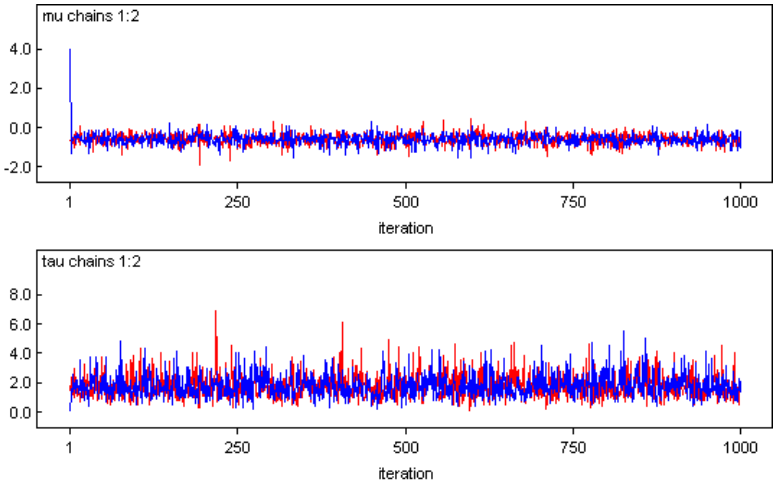

Figure 4: History plots for the Normal example

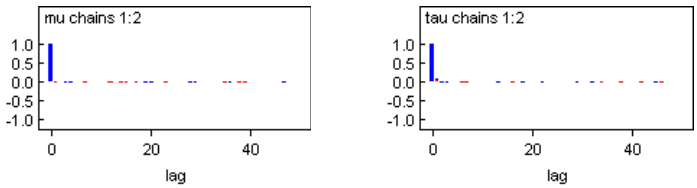

Figure 5: Autocorrelation plots for the Normal example

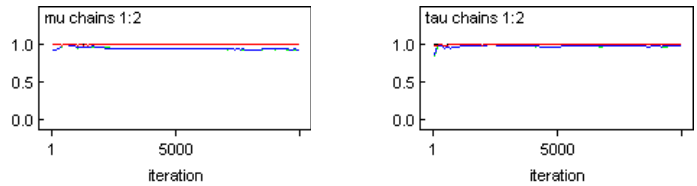

Figure 6: Brooks-Gelman-Rubin plots for the Normal example

```
model{
for(i in 1:10){ y[i]~dnorm(mu,tau)}
mu ~ dnorm(0,0.01)
tau ~ dgamma(0.001,0.001)
}

Data

list(y=c( -1.251818103,  0.200977229,  0.309874153, -1.270994629, -0.864859684,
 -1.201747282, -0.632211810,  0.306132427,  0.005017514 ,-1.623128936,
0.809304421, -0.187802948,  0.682484549,  0.770453876,  1.068758329,
0.267315595,  0.330954901,  0.343979499, -0.545634345, -0.268890128))
```

Table 2: Code for simple Normal example

The important feature of the history plot (Figure 4) is that the two chains overlap, even at very early iterations, despite starting far apart. This raises the question "Why do some chains converge faster than others?" Another interesting feature is that the plots are much more jagged than those for the Seeds example in Figure 1. This is a feature of chains with low autocorrelations.

Contrasting Figures 5 and 6 with Figures 2 and 3, we see that in the Seeds example the autocorrelations remain substantial at later lags, whereas in the Normal example they are minimal after lag 0. For the Brooks-Gelman-Rubin plots in the Normal example, all three sets of lines are very close to 1 and hold steady, whereas in the Seeds example for alpha0 the red line starts very high and then settles down to a value close to one, and for alpha2 all three lines are unstable until the last iterations. Students will not know what this signifies or why it is important, so questions emerging are "What is the autocorrelation plot measuring?", "Why do the autocorrelations for chain 2 in the Seeds example seem not to tail off?", and "What is a Brooks-Gelman-Rubin plot?". Although it may seem strange to ask students to look at plots of statistics they do not understand, students will be curious about what autocorrelations and BGR plots are, because these are buttons on the Sample Monitor Tool in WinBUGS (and students may have clicked on these buttons before).

In comparing Figures 1 and 4, and Figures 2 and 5, students will likely observe that the chains with the higher autocorrelations take longer to converge. In our class, students assumed that high autocorrelations indicate that the chain has not converged, and their original question had to be re-phrased as "Why do chains with higher autocorrelations take longer to converge?". A second important question is "Does autocorrelation affect inference?"

2.4 The stochastic nature of the Gibbs sampler

The final example was the Normal example in Table 2, but using two sets of identical initial values: list(mu=1,tau=0.01), and list (mu=1,tau=0.01). Students were asked to look at summary statistics for each chain separately after 10, 1000 and 10,000 updates, and to comment. The point of this example was to demonstrate that the sampled values at each iteration are stochastic. Results are shown in Table 3. The obvious question that emerges is, "How can two chains start with the same initial values and yet give different results?".

It is important to note that if using these results for inference, one would usually discard the initial burn-in phase when producing the summary statistics. In this case, however, the point is to demonstrate the randomness of each sample, rather than to make inference. Moreover, because the model is very simple, convergence occurs rapidly.

Table 3: Summary statistics for a simple Normal example with two sets of initial values

10 updates	Node	Mean	sd	MC error	2.5%	median	97.5%	start	sample
Chain 1	mu	-0.6853	0.277	0.08982	-1.211	-0.6036	-0.2657	1	10
Chain 1	tau	1.807	0.6514	0.2194	1.036	1.83	2.988	1	10
Chain 2	mu	-0.2435	1.42	0.3789	-1.311	-0.6153	3.953	1	10
Chain 2	tau	1.543	0.6354	0.2799	0.06088	1.559	2.522	1	10
1000 updates									
Chain 1	Mu	-0.6059	0.2749	0.008896	-1.154	-0.598	-0.0768	1	1000
Chain 1	tau	1.773	0.8523	0.03007	0.5025	1.666	3.849	1	1000
Chain 2	mu	-0.6159	0.3066	0.008566	-1.211	-0.6203	-0.09159	1	1000
Chain 2	tau	1.795	0.8254	0.03088	0.5742	1.655	3.799	1	1000
10,000 updates									
Chain 1	Mu	-0.6046	0.2649	0.002563	-1.145	-0.6035	-0.06871	1	10000
Chain 1	tau	1.797	0.849	0.009359	0.5518	1.658	3.819	1	10000
Chain 2	mu	-0.6007	0.2733	0.002705	-1.146	-0.6019	-0.06525	1	10000
Chain 2	tau	1.79	0.8389	0.008582	0.54	1.655	3.785	1	10000

2.5 Summary of student questions and suggested answers

At the end of the class, we asked students to write a paragraph outlining what they had learned. Here, and in their projects, students demonstrated that they realised assessing convergence was important, and that they could use history plots and BGR plots as checks for non-convergence. An alternative strategy to the free-flow paragraph for assessing learning would be to ask students to write down answers to the questions generated during discussion time. Instructors may choose to teach the material covering each question in turn, in which case a possible ordering, and two additional questions, are given below, together with sample answers to the questions and pointers to material that can be discussed in relation to each one.

1. Why do we need iterations at all? What is WinBUGS doing?

   This should cover the material in the "Background" section above: computer-intensive methods as opposed to analytical derivations, Markov Chains, Monte-Carlo simulation, stationary distributions, and the Gibbs sampler.

2. How can two chains start with the same initial values and yet give different results?

   MCMC uses random sampling. At each stage a new $\boldsymbol{\theta}$ is drawn at random from the current componentwise posterior distributions. Thus even if two chains start with the same initial value $\boldsymbol{\theta}^0$, the random draws of $\boldsymbol{\theta}^1$ will be different.

3. What is the autocorrelation plot measuring?

Autocorrelation measures how closely related successive samples of the same parameter are. Most students will be familiar with the concept of correlation, but will not have seen autocorrelation plots or variograms before. It is thus necessary to explain that the lag 1 autocorrelation for a univariate parameter α is estimated by the sample correlation of the pairs (α^1, α^2), (α^2, α^3),…,(α^{N-1}, α^N). Similarly, a lag k correlation is the correlation between the pairs (α^1, α^{k+1}),…,(α^{N-k}, α^N). Generally, the higher the autocorrelation between successive draws, the slower convergence will be.

4. Why do chains with larger autocorrelations take longer to converge?

   We need the chain to cover the whole parameter space. Chains with high autocorrelations show small changes in the sampled parameters from one iteration to the next. Essentially, they take small steps, and so take longer to cover the parameter space than chains with smaller autocorrelations, which take large steps. This is particularly apparent when the initial values are in the tails of the posterior distribution.

5. Does autocorrelation affect inference?

   Students will be used to the idea that correlation must be taken into account when dealing with non-independent data, so the question is how this should be dealt with in the Gibbs sampler. Importantly, the summary statistics provided by WinBUGS are not simple moment estimates, but rather use the theory of time series to account for the autocorrelation. A related issue that students may riase is thinning, as this is one of the options on the Sample Monitor Tool in WinBUGS. Thinning the chain to e.g. one in ten samples means keeping only every tenth iteration. This may seem appealing because the autocorrelation in the thinned chain will be smaller, however the only advantage of thinning is to reduce storage requirements: MacEachers & Berliner (1994) demonstrate that you always get more precise estimates if the entire chain is used.

6. Why do the autocorrelations for chain 2 in Example 2.2 seem not to tail off?

   This happens because the second chain has started far out in the tails of the target distribution. Correlations will tend to be large and positive for those portions of the chain that are far from the target distribution, and because WinBUGS by default calculates the correlations over the entire history of the chain, these high correlations from early on dominate the estimate of autocorrelation. If the early iterations are excluded from the autocorrelation plot, the correlations do tail off (see Figure 7). This example also raises the question of how to select initial values. Once the chain has converged, the choice of initial values does not matter, and in many cases it will be sufficient to allow WinBUGS to draw initial values at random. However in cases with highly autocorrelated chains, it can be helpful to choose initial values that are plausible under the posterior distribution, as this will speed up convergence. Since the posterior distribution is not known *a priori*, these values must be based upon an educated guess. For instance, most of the initial values in Section 2.1 were reasonable, with the exception of tau. Taking tau=1 means that we could be 95% sure that the plate effects would not affect the log odds by more than +/- 1.96, i.e. the odds ratios corresponding to plates would be between 0.14 and 7, 95% of the time. In fact, we likely expect plate-to-plate variability to be much smaller than this. If we took tau=100, the odds ratios corresponding to plates would be between 0.82 and 1.22, 95% of the time. We stress, however, that the subjectivity involved in choosing these initial values does not impact the validity of the results, provided that the chain has converged and an appropriate burn-in sample has been discarded.

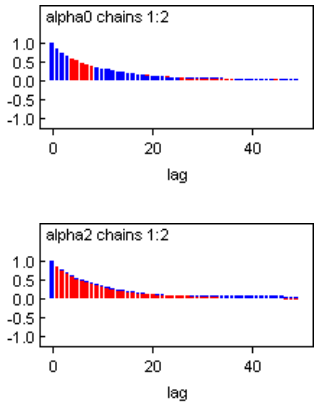

Figure 7: Autocorrelation plots for example 1.2, omitting the first 900 iterations

7. What is convergence?

   This is an opportunity to discuss Gibbs sampling and Markov Chains, as outlined in the "Background" section above. Importantly, we are not talking about convergence of a single number, but convergence in distribution (students will be familiar with the concept of convergence in distribution in the case of the Central Limit Theorem). Once the chain has converged we will be sampling from the same distribution no matter how many more iterations we run. For most students, it will not be helpful to define convergence in distribution in the mathematical way.

8. Why do some chains converge faster than others?

   In the cases observed in these examples, high autocorrelations result in slower convergence, and chains that start far from their target distributions take longer to converge. When there is a single unknown continuous parameter whose prior is conjugate to the likelihood, convergence will be instantaneous.

9. In the first example, why is it that the chains for alpha2 overlap then stray apart again?

   This is an opportunity to discuss convergence in distribution and multivariate convergence. Around iteration 50 it is true that the two chains overlap for alpha2, but this is not true for other parameters in the model, e.g. for alpha0 the chains are still very far apart. We have sampled from the conditional posterior $f(\text{alpha2}|\text{alpha0}^n,\text{alpha12}^n, \text{tau}^n,b^n)$ (for n close to 50) and have obtained a plausible value of alpha2 by chance, however since our estimates of alpha0, tau and b are still far in the tails, future samples of alpha2 stray away again. It is not until all the parameters are being drawn from their target values that the sampled values of alpha2 for the two chains overlap consistently.

10. What is a BGR plot?

    BGR stands for Brooks-Gelman-Rubin, and the plot is a plot of their convergence diagnostic [Brooks & Gelman, 1998]. The diagnostic requires the user to run two or more chains, whose initial values are overdispersed relative to the target posterior distribution (that is, the initial values are more spread out than would be expected had they been sampled from the posterior distribution). The variance within chains is then compared to the pooled variance amongst all sampled values (regardless of chain). This pooled variance is the sum of the within-chain variance and the between-chain variance. If the chains have not converged and are not "mixing" – heuristically, not overlapping when plotted simultaneously on a history plot – then there will be between-chain variance, meaning that the pooled variance will be larger than the within-chain variance. If the chains have converged, then the within-chain variance should be similar to the pooled variance. This is the same concept as ANOVA and so quite accessible to students. The diagnostic shows three lines: the red line is a plot of the pooled variance divided by the within-chain variance; the green line is the width of the central 80% interval of the pooled samples; and the blue line represents the mean width of the within-chain central 80% intervals. If the chain has converged, then the red line should be close to 1, and the blue and green lines should be holding steady at constant values.

11. How do you know when you have done enough iterations?

    We can never be certain. Convergence diagnostics test for specific aspects of non-convergence, but there is no global test. As a general guideline, start with several sets of initial values quite far apart, and run iterations until the history plots for the two chains overlap. Check the autocorrelation plots; if these show correlations that drop off very slowly, convergence may also be slow, conversely, if the autocorrelations drop off rapidly, convergence may be rapid. Following these informal diagnostics, more formal tests should be used. The general advice is to use more than one test, as each tests detects a particular aspect of non-convergence [Cowles and Carlin 1996]. If there are many parameters, it may not be necessary to monitor all of them. For example, if there are 100 patients and we are fitting a random effects model with grouping by patient, it would be adequate to monitor one or two of the random effects, rather than all 100 [Spiegelhalter *et al.* WinBUGS manual v 1.4.1]. Monitoring too many parameters will result in multiple testing issues, so that the probability of at least one convergence diagnostic giving statistically significant results by chance alone becomes quite large. For more formal tests of convergence, if using WinBUGS, the next step can be to check the Brooks-Gelman-Rubin plots. If these look reasonable, then the "coda" function can be used to save the history of the chain in a format that will allow further convergence diagnostics in another package (e.g. the R package BOA [Smith, 2005]). An overview of diagnostics may be found in Cowles and Carlin [1996].

    Whilst we demonstrated the use of more advanced tests in BOA, we did not ask that students learn the package – most students had never used R, and we felt that this would be too big a task to place on them given that they had already had to learn WinBUGS. Instead, we indicated what can be done within WinBUGS, and emphasized that if they were using WinBUGS to fit complex models, they would either need to do some additional convergence diagnostics themselves, or ask a statistician for help.

3. Priors

The steps of a Bayesian analysis can be laid out as (i) describe existing knowledge (this forms the prior); (ii) describe the additional knowledge (this is the likelihood); (iii) update the existing knowledge with the new knowledge (this forms the posterior distribution). This is appealing because it reflects the incremental way in which well developed health research programs build knowledge. Indeed, research proposals in the health sciences always begin with a literature review, and even frequentist studies use prior knowledge when calculating sample sizes.

When specifying a prior, researchers may have one of three goals. Some may wish to specify a distribution that describes the existing knowledge, perhaps by using pooled estimates from a meta-analysis. Others may note that previous work in the area was carried out on different populations or according to a different protocol, in which case the prior may describe the researcher's belief as to the relevance of the previous results. In this case there is some subjectivity, and sensitivity analysis may be appropriate. Other researchers may not wish prior knowledge to contribute at all. One instance where this is appropriate is a meta-analysis in which all the previous evidence is described in the studies included in the review, and will hence be contributing to the posterior through the likelihood. In these cases, an uninformative prior can be used. Such priors are so vague that, in comparison to the data, they contribute almost no information to the posterior distribution. The terms uninformative, vague and diffuse are often used interchangeably to describe these priors.

Prior specification can be problematic. For example, priors that are intended to be vague may be more informative than anticipated, or the distribution of the data may differ substantially from that described in the prior. This can lead to a prior-data conflict [Evans & Moshonov, 2006], where the prior places little or no mass on the range of values observed in the data. At times the conflict may be legitimate, for example if the results were genuinely surprising in a way that could not have been foreseen, and here no action need be taken. At other times the conflict may arise through different experimental conditions, changes in populations over time, or in an error in specifying the parameters of the prior. In this case, the conflict may render the resulting posterior impossible to interpret.

After a lecture discussing sources of information for prior specification (e.g. previous literature, pilot studies, expert opinion), and some standard

choices for prior distributions, there was a lab session in which students could practise choosing a prior and see the results of their choices. The initial part of the lab asked students to choose priors for some of the simple scenarios they might encounter in practice: priors for probabilities, means, and precision parameters. The second portion looked at some of the difficulties in selecting priors for precisions: the importance of choosing the upper bound of a uniform prior carefully, and the question of how vague an inverse gamma prior for a variance component is. The final portion of the lab looked at some of the ways priors can be used in Bayesian analysis, and was intended to illustrate that priors can be helpful, rather than just a nuisance to specify. We describe one example from each section of the lab, choosing those examples that produced the most discussion.

3.1 Beta priors for probabilities

When estimating proportions or comparing proportions between groups, students need to select a prior for a probability. To give them practice at this, we presented them with the following question:

Suppose we want to estimate the probability that a randomly selected student graduating from a McMaster bachelor's program will be in full-time employment within 6 months of completing the program. We plan to sample 1000 students.

Under each of the following scenarios, construct a prior for the probability of being in full-time employment within 6 months of graduating, and explain your reasoning:

a) You have no prior information on the probability of being in full-time employment.

b) You found a small study conducted at McMaster last year in which 10 students were sampled, 8 of whom were in full-time employment within 6 months of graduating.

c) The Career Centre at the University of Guelph shares information that amongst a random sample of 5000 of their students, 3500 were in full-time employment within 6 months of graduating.

Suppose we find that in our study, of the 1000 students sampled, 850 were in full-time employment within 6 months of graduating. What is the posterior distribution of the probability of being in full-time employment using each of your priors in a) to c)? Comment on your results. *Note: you may do this either using WinBUGS or analytically (or both). If you do it analytically, please derive the posterior distribution from the prior and the likelihood.*

Background

The beta density is a popular choice of prior for binomial problems, because it is conjugate to the binomial distribution. That is, if we wish to estimate the probability p of success and observe r successes amongst n Bernoulli trials, then if the prior for p is Beta(a,b), the posterior is Beta (a+r,b+n-r) so that the prior parameters a and b contribute the same amount of information as a successes and b failures. Other distributions can be used, and natural choices are those restricted to lie between 0 and 1, although other distributions can of course be truncated. Aside from the beta family, another option that is range-restricted is a Normal distribution on the log odds, and this can be helpful as an introduction to setting priors for logistic regression problems. For example, in scenario a) we may be 95% sure that the probability lies between 0.05 and 0.99, i.e. that logit(p) lies between -2.9 and 4.6. We could then set the prior for logit(p) to be Normal, mean 0.85, standard deviation 1.9. Similarly, the information in scenario b) suggests that a priori we can be 95% sure that the probability lies between 0.44 and 0.97, i.e. that logit(p) lies between -0.24 and 3.48, and so we might set the prior for logit(p) to be Normal, mean 1.62, sd=0.95. Whilst in our offering of the course, we chose to address this in a separate lab on logistic regression, this alternative prior specification could well be explored alongside the uniform and beta families.

The other piece of background information needed to answer this question is that Guelph and McMaster are neighbouring universities. Guelph is slightly smaller and is a leader in plant and animal life sciences. McMaster is well known for its medical school and has a strong Bachelor of Health sciences program. The two schools have similar admission averages, and both score well in university rankings. Thus while there are important differences between them, it is hard to predict which will have the higher employment rate, particularly since this will be driven by the proportion of students pursuing postgraduate degrees.

Students quickly noticed that although choosing a Beta(2,8) prior for b) was a reasonable choice, choosing a Beta(1500,3500) prior for c) was so informative that it swamped the data. This is a situation that often arises in health research, where there have been previous studies and meta-analyses providing a good deal of prior information. A decision then has to be made on what extent this prior information should contribute. This should be made before analysing the data, however most students chose the Beta(1500,3500) prior without considering the impact it would have. Thus an important message to communicate is that the prior should be formed by considering not just the amount of prior knowledge, but also the size of the sample to be collected and the degree to which the prior data applies to the problem at hand. In the above example, students need to argue their case for how much weight the Guelph results should receive. The Guelph information indicates that 70% is a good guess, but because of the differences between the two universities, there is still considerable uncertainty as to what the McMaster probability is. Most students chose to give the Guelph results a weight equivalent to a sample size of 50 or 100, thus allowing the prior information to give a reasonable estimate but allowing the data to drive the results. Graphs showing the prior, likelihood and data for students' original beta(1500, 3500) choice, and a downweighted beta(30,70) choice are shown in Figure 8. Posterior medians and 95% credible intervals for the various priors discussed are given in Table 4.

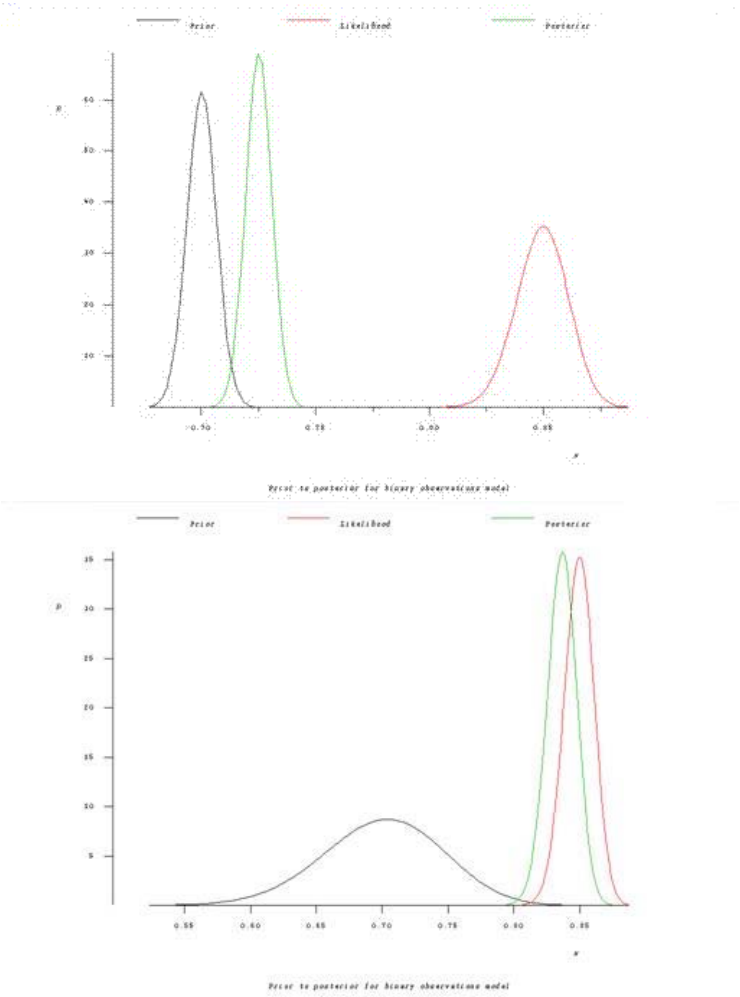

Figure 8: Plots of the prior, likelihood and posterior distributions for example 2.1, using students' original choice of a Beta(1500,3500), and a revised choice of Beta(30,70)

Table 4: Posterior medians and 95% credible intervals for candidate priors

Scenario	Prior	Posterior Median	95% Credible Interval
a)	Beta(1,1)	0.8497	0.8267 to 0.8709
a)	Logit(p) ~ Normal(0.85,sd=1.9)	0.8498	0.8271 to 0.8713
b)	Beta(2,8)	0.8497	0.8267 to 0.8705
b)	Log(p) ~ Normal(1.62, sd=0.95)	0.8499	0.8267 to 0.8710
c)	Beta(1500,3500)	0.7250	0.7137 to 0.7363
c)	Beta(30,70)	0.8002	0.7758 to 0.8235

3.2 Precision Priors

Many problems in healthcare research involve clustered data. The following problem deals with the difficult situation of specifying a prior for a variance component.

Consider a (fictional) cluster-randomized trial of a school-based intervention to investigate whether the use of computers in the elementary school classroom improves end-of-year test grades. Test grades are from a standardized test and vary from 0 to 100 with a mean of 60.

Four schools are randomized to receive either the intervention (computers) or control (standard teaching resources). There are 3 classes per school, and 30 students per class. When analysing this data, our model must account for correlation within classes and within schools. We do this using a variance components model using random effects. We let both classes and schools have random effects following a Normal distribution with mean 0 and unknown precision. We consider two options for setting uninformative priors on the variance parameters. The first option is to

use Gamma(0.001, 0.001) priors for the precision parameters, and the second is to use Uniform priors for the within-school and within-class standard deviations. Obtain the posterior distributions of the between-class, between-school and residual variances under each set of priors. Are the two sets of distributions similar? Are both sets of priors uninformative? Try to explain the results that you see.

Background

The Gamma(0.001,0.001) prior is the traditional choice for a vague prior for a precision, however Gelman [Gelman, 2006] recently showed that in a variance components model it is in fact informative when the true precision is close to zero (as evidenced by the fact that a Gamma (0.00001, 0.00001) prior will give different results). This example is intended to demonstrate this phenomenon. The model used is

$$\text{test score}_{ijk} = \alpha_0 + \alpha.\text{group x group}_k + \alpha.\text{class}_j + \alpha.\text{school}_k + \varepsilon_{ijk}$$
$$\alpha.\text{class}_j \sim N(0,\text{var.class})$$
$$\alpha.\text{school}_k \sim N(0,\text{var.school})$$
$$\varepsilon_{ijk} \sim N(0,\text{var.within})$$

where test score$_{ijk}$ represents the test score for child i in class j within school k, group$_k$ is a binary indicator equal to 1 if school k was randomized to intervention and 0 otherwise, α.class and α.school are random effects for classes within schools and schools respectively, and ε_{ijk} is a residual. The example compares two choices for priors on the variance parameters: the traditional inverse gamma distribution, where we take tau.class = 1/var.class, tau.school=1/var.school and then specify tau.class ~ Gamma(0.001,0.001) and tau.school ~ Gamma(0.001,0.001), or alternatively we take sd.class = sqrt(var.class), sd.class ~ U(0,100), sd.school = sqrt(var.school), sd.school ~ U(0,100). In both cases we use tau.within=1/var.within, tau.within ~ Gamma(0.001,0.001). The two distributions in question are depicted in Figures 9 and 10. Figure 10 shows that on the scale of the standard deviation, the Gamma prior has a large peak near zero. Figure 9 shows that the Uniform prior for the standard deviation has a large peak near zero on the scale of precision (though not as large a peak as for does the Gamma prior). This is worth pointing out, as students often like flat priors, but do not realise that no prior will be flat for both the standard deviation and the precision. The other point of note is that vague Uniform priors can over-estimate variances when the degrees of freedom are small (four or fewer) [Gelman, 2006], as is the case for the between-schools variance in this example.

When first introduced to WinBUGS, students will find it strange that the Normal distribution is parameterized in terms of its mean and precision, rather than mean and standard deviation. There is a statistical reason for this, namely that a gamma distribution for the precision is conjugate to the Normal distribution, however non-statisticians will not always find this enlightening. We found it helpful to explain conceptually why the reciprocal of the variance is an estimate of precision: larger standard deviations lead to wider confidence intervals and hence less precise estimates.

We provided students with the code for the model (see Appendix B), and gave them the data, which was generated with the residual variance equal to 225, within-school variance equal to 64, and within-class variance equal to zero. The summary statistics are given in Table 5, and the posterior densities for the variance due to class and the variance due to schools are given in Figure 11.

Table 5: Summary statistics for uniform and inverse-Gamma priors in a variance components example

Node	Mean	Sd	MC error	2.5%	Median	97.5%	Start	sample
Gamma								
alpha.group	7.442	3.534	0.05246	-0.1838	7.581	14.15	1	100000
alpha0	60.89	2.206	0.02207	56.27	60.92	65.33	1	100000
var.class	1.099	3.02	0.05302	9.126E-4	0.08963	9.122	1	100000
var.school	18.33	165.8	1.168	0.001946	4.43	112.0	1	100000
var.within	257.3	19.44	0.07105	222.1	256.3	298.3	1	100000
Uniform								
alpha.group	6.918	5.463	0.1046	-5.545	7.271	17.16	501	99500
alpha0	60.82	3.029	0.04088	54.34	60.89	66.84	501	99500
var.class	3.218	5.693	0.07867	0.002381	1.269	17.79	501	99500
var.school	76.7	283.2	3.672	0.2516	19.41	496.6	501	99500
var.within	256.1	19.33	0.06294	220.9	255.2	296.8	501	99500

Surprisingly, students did not have difficulty interpreting the variance components model, perhaps because the role of the random effects in a WinBUGS model is more explicit than other packages (e.g. SAS, SPSS), where the random effects are usually not written into the regression equation. Students were supposed to notice firstly that the results are different for the two choices of prior (and hence that the prior is contributing information to the results in at least one of the cases), and secondly that the results for the gamma prior are more precise for each of the summary statistics.

We found that students tended to look at the estimates of location (usually the mean, sometimes the medians) without looking at standard deviations, MC errors, or credible intervals. Also, not all students thought to look at the estimate of intervention effect (alpha.group), as the question did not explicitly ask for it. That the credible interval for the intervention effect is different for the two priors is of concern. After some guidance (prompting to look at the width of the credible intervals and the size of the standard deviations of the posterior distributions), students were able to see that the gamma prior gives less diffuse posteriors than the uniform prior, and results in posterior estimates for the class and school components that are closer to zero. They also understood that diffuse gamma priors for variance components near zero may be more informative than intended.

This example is quite difficult, as there are no standard choices for vague priors when the degrees of freedom associated with the variance parameter are small. In these cases sensitivity analysis is advised [Spiegelhalter, 2004], and using an informed prior can be helpful – neither of the vague priors used in this example actually represents sensible beliefs about the variance components. Since we were teaching at the graduate level, we felt that the level of difficulty was acceptable, however those wishing to simplify the problem could do so by eliminating one of the

variance components (e.g. looking at just schools, rather than classes within schools). It might also be helpful to ask students to try a Gamma (0.00001, 0.00001) and a Uniform(0,1000) prior for comparison.

In other examples, students who were not statisticians all chose to use uniform priors for standard deviations rather than gamma priors for precisions, most likely because the standard deviation is easier for them to interpret.

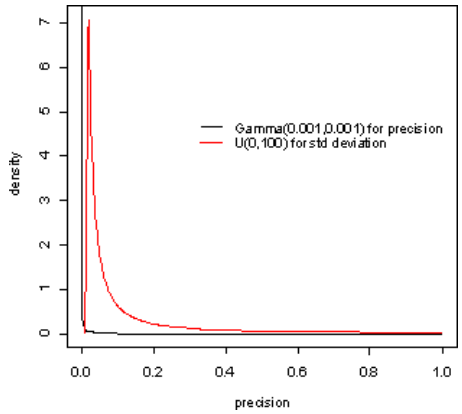

Figure 9: Comparison of the Gamma(0.001,0.001) prior density for precision and the precision density implied by a Uniform(0,100) prior for the standard deviation

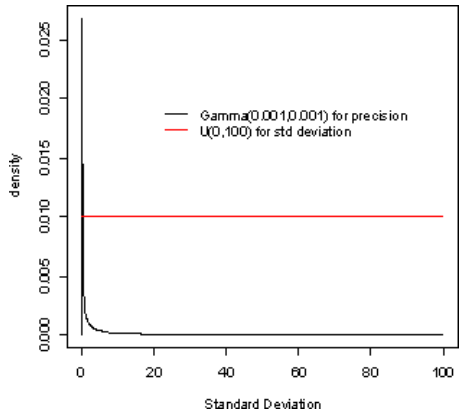

Figure 10: Comparison of the Uniform(0,100) prior for standard deviation and the prior standard deviation density implied by taking a Gamma (0.001,0.001) prior on the precision

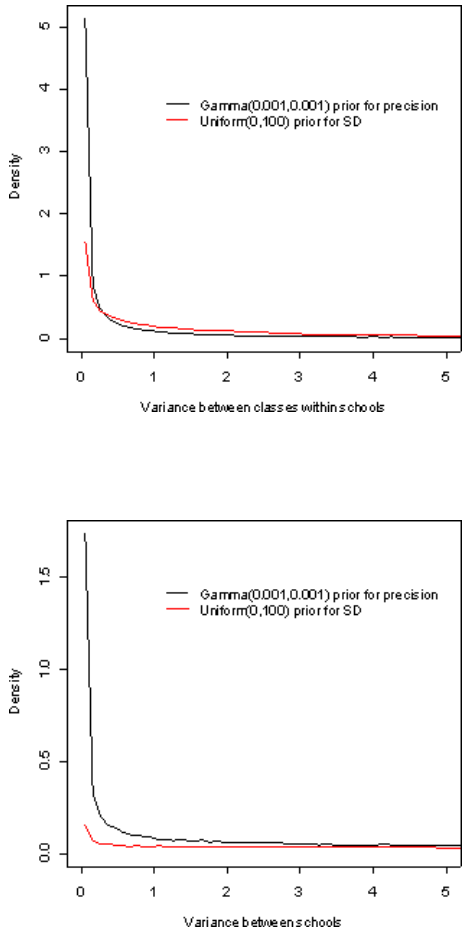

Figure 11: Posterior distributions for var.class and var.school for both the gamma and the uniform priors

3.3 Using Informed Priors

The last example referred students to an article by Senn [Senn, 2007] discussing a first-in-man study in which all six healthy volunteers who were given the experimental drug experienced a cytokine storm, whereas the two volunteers given placebo did not. A frequentist analysis of the data yields a result that is either not significant or just significant (depending on the choice of test), although it is obvious that something is wrong because a cytokine storm is a rare event. Bayesian analysis can incorporate the information on the rarity of the event using a prior distribution, and the article itself has suggestions for what a reasonable probability of an event might be. Whilst this example is not at all difficult, it does illustrate the value of being able to specify prior beliefs.

Students commented that this was one of the most helpful labs, and in their projects showed that they could choose an appropriate prior and discuss the consequences of their choice. A separate lab allowed students to practise setting priors for regression coefficients in linear and logistic regression problems.

The choice of prior can be a controversial topic in Bayesian analyses. Approaches include choosing a vague prior, basing the prior on previous evidence, eliciting expert opinion, or being purely subjective. Clearly the last option is of limited use in isolation, as the results will be relevant only to those who share the researcher's prior beliefs. Using a range of subjective priors can be useful as a sensitivity analysis, see for example Spiegelhalter et al [1994] for a discussion of sceptical and enthusiastic priors for monitoring a clinical trial. Sensitivity analysis can also be helpful when standard choices for vague priors are not well established, as for example in the variance component for schools in section 3.2.

Often researchers will choose vague priors, and students tend to favour these because then the Bayesian results will be similar to frequentist results: frequentist results are more familiar, and so are often assumed to be the "right" results. In fact choosing an informed prior can make logical sense, and can be appealing when coupled with careful sensitivity analysis. Using a vague prior or conducting a frequentist analysis amounts to looking at the evidence from the given study in isolation, rather than in the light of what is already known. This is precisely the point illustrated by the example in Section 3.3. Moreover, informed priors can be helpful for parameters on which the data contributes little information, e.g. for the cluster-randomised trial in Section 3.2 the four schools provide very little information on between-school variance, so it is difficult to construct a prior that is not informative.

4. Conclusions

We have discussed approaches to tackling two major topics in Bayesian methods: convergence and prior specification. For teaching convergence, we found that our examples-based approach meant that it was the students who were posing the questions, thus providing motivation for understanding a difficult topic whilst making use of the observational skills as scientists. We have also outlined suggested approaches to answering the students' questions. Our lab on prior specification allowed students to practice selecting priors, and highlighted the need to consider the relevance of prior information, potential difficulties in selecting vague priors, and the advantages offered by informed priors.

Feedback from the labs was overwhelmingly positive, and students appreciated the hands-on approach to learning with tutoring support. We had a small class (5 students), and since adequate tutoring was key to the success of the labs, a larger class would likely need additional TA support. As instructors, we found that discussion was very difficult in a conventional computer lab, where stations are arranged in rows and monitors obscure students' faces. If enough students owned laptops labs could be held in a classroom.

We have shared some examples that we found helpful in teaching Bayesian methods to a group of health research methodologists. By the end of the course, students had a good grasp of the key Bayesian concepts and could conduct a simple Bayesian analysis. We found that teaching Bayesian methods to students with a limited statistical background to be challenging, but worthwhile.

Appendix A

Topic	Learning Objectives – by the end of the class, students should be able to
1. Introduction to Bayesian Thinking and Statistics	Identify the differences between Bayesian and frequentist methods Describe what a distribution is Use Bayes' theorem to link the prior and posterior distributions Use the posterior distribution for inference
2. Lab 1 - Estimating and comparing means of Normal variables	Become familiar with the WinBUGS environment Use WinBUGS to estimate the posterior distribution of a mean parameter (given a prior) Use WinBUGS to compare the mean responses between two groups
3. Lab 2 – Binary outcomes	Use Bayesian methods to estimate and compare proportions Implement these models in WinBUGS
4. Lab 3 – Regression models	Formulate a linear regression as a Bayesian estimation problem. Formulate a logistic regression as a Bayesian estimation problem. Implement these models in WinBUGS
5. Prior distributions	Discuss the impact of the prior on the results, and the importance of selecting a suitable prior Describe methods for eliciting priors, and understand the strengths and weaknesses of each List some standard prior choices (conjugate priors, non-informative priors)
6. Bayesian Trial Design	Desribe the principles of a simple Bayesian sample size calculation Explain how Bayesian methods can be used for interim monitoring.
7. Lab 4 – Priors, Trial Design	Choose a prior for a probability and a prior for a mean. Describe the impact of their choice. Choose a prior for a precision parameter. Describe settings where a conjugate prior for a precision could be misleading.
8. Convergence and sampling diagnostics	Describe heuristically what WinBUGS is doing Explain why assessing convergence is important and describe some of the dangers of MCMC sampling Use simple convergence diagnostics (autocorrelation, history plots, Brooks-Gelman-Rubin diagonistics) to detect common settings when convergence has not have taken place.

9. Meta-analysis/ Multi-centre trials/ cluster-randomised trials	Formulate a random effects meta-analysis and explain each of the components Formulate a model for analysing a multicentre trial and explain each of the components Formulate a model for analysing a cluster-randomised trial and explain each of the components
10. Lab 6 – Meta-analysis	Implement a random effects meta-analysis and interpret the results
11. Missing Data (Lecture + Lab)	Describe how missing data is treated in Bayesian analysis, separately for responses and covariates Use WinBUGS to do multiple imputation
12. Reporting the Results	Describe and Bayesian methods in a paper Report the results of a Bayesian analysis
13. Decision-Making	Use Bayes factors can be used to summarise evidence, and explain what the resulting Bayes factor means. Describe how decision theory can be used to guide decisions (such as whether to accept or reject a null hypothesis, choose the sample size for a trial, or decide whether to stop a trial) Identify the Bayesian elements of a cost-effectiveness evaluation.

Appendix B

```
model{
for(i in 1:360){
        test[i] ~ dnorm(mu[i],tau)
        mu[i] <- alpha0 + alpha.class[class[i]]  + alpha.school[school[i]] + alpha.group*(group[i]-1)
}

alpha0 ~ dnorm(60,0.04)
for(i in 1:12){ alpha.class[i] ~ dnorm(0,tau.class)}
for(i in 1:4){ alpha.school[i] ~ dnorm(0,tau.school)}
alpha.group ~ dnorm(0,0.0044)

tau.class ~ dgamma(0.001,0.001)
tau.school ~ dgamma(0.001,0.001)
tau ~ dgamma(0.001,0.001)

var.class <- 1/tau.class
var.school <- 1/tau.school
var.within <- 1/tau

}
```

The data below were generated with true tau=1/225, tau.class = infinity
tau.school = 1/64 = 0.016. Obtain the posterior distributions of var.class,
var.school and var.within (you will need to set some initial values).

Now consider a set of Uniform priors:

```
model{
for(i in 1:360){
        test[i] ~ dnorm(mu[i],tau)
        mu[i] <- alpha0 + alpha.class[class[i]]  + alpha.school[school[i]] + alpha.group*(group[i]-1)
}

alpha0 ~ dnorm(60,0.04)
for(i in 1:12){ alpha.class[i] ~ dnorm(0,tau.class)}
for(i in 1:4){ alpha.school[i] ~ dnorm(0,tau.school)}
alpha.group ~ dnorm(0,0.0044)

tau.class <- 1/var.class
tau.school <- 1/var.school
var.class <- sd.class*sd.class
```

```
var.school <- sd.school*sd.school

sd.class ~ dunif(0,100)
sd.school ~ dunif(0,100)
tau ~ dgamma(0.001,0.001)

var.within <- 1/tau
```

References

Ashby D. Bayesian statistics in medicine: a 25-year review. Stat Med 2006; 25:3589–3631.

Bayes T. Essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London (1763).

Brooks SP, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics 7: 434-455 (1998).

Cowles MK, Carlin BP. Markov Chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 1996; 91:883-904.

Crowder, M J (1978) Beta-binomial Anova for proportions. *Applied Statistics*. **27**, 34-37.

Evans M, Moshonov H. Checking for Prior-Data Conflict. Bayesian Analysis 2006; 1(4): 893-914.

Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1(3): 515-533 (2006)

Geman S, Geman D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741

Haynes RB. Forming Research Questions. *J Clin Epidemiology* 2006 59:881-886.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325--337.

Smith, B.J., 2005. *Bayesian Output Analysis Program (BOA), Version 1.1.5* [online]. The University of Iowa. Available from http://www.public-health.uiowa.edu/boa [accessed March 23, 2005]

Senn S. Safety first? *Significance* 2007 June 4(2): 79-80.

Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). Journal of the Royal Statistical Society, Series A, 157:357{416, 1994.

Spiegelhalter DJ, Abrams KR and Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Wiley 2004.

Spiegelhalter DJ. Comments on Guidance for the use of Bayesian Statistics in Medical Device Clinical Trials. FDA, 2006.

Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. J Clin Epidemiol. 2005 Mar;58(3):261-8.

Eleanor M. Pullenayegum
Department of Clinical Epidemiology and Biostatistics
McMaster University
Hamilton, ON, Canada
and
Biostatistics Unit/Centre for Evaluation of Medicines
Father Sean O'Sullivan Research Centre
St Joseph's Healthcare
Hamilton, ON, Canada
E-mail:pullena@mcmaster.ca

Lehana Thabane
Department of Clinical Epidemiology and Biostatistics
McMaster University
Hamilton, ON, Canada
and

Biostatistics Unit/Centre for Evaluation of Medicines
Father Sean O'Sullivan Research Centre
St Joseph's Healthcare
Hamilton, ON, Canada
E-mail:ThabanL@mcmaster.ca

Volume 17 (2009) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications