BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

# Bayesian Statistical Methods (BAY)

Semester 2, 2018

Prepared by:

Associate Professor Lyle Gurrin

Centre for Epidemiology and Biostatistics,

Melbourne School of Population and Global Health

The University of Melbourne



THE UNIVERSITY OF MELBOURNE

# Background

Bayesian inference is concerned with fitting full probability models to data and summarizing the results as a probability distribution for the parameters that define the model and any other unobserved quantities that can be predicted (prospectively or retrospectively) as a consequence of the model. The Bayesian approach allows the data analyst to make probabilistic statements about both quantities they observe and quantities about which they wish to learn, including model parameters, imputations for missing data and predictions for future observations. In contrast, the standard frequentist approach makes modelling assumptions only about observable quantities and cannot benefit from the explicit use of *subjective* probability to quantify uncertainty about inferences. Bayesian inference is importance in modern statistical practice, including biostatistics. This is in part due to a belief by some statisticians that it has philosophical and logical advantages. It is, however due much more to the ability of the Bayesian modelling process to tackle complex hierarchical problems and, most importantly, the availability of new computational techniques and user-friendly software tied to the ever-increasing processing power of desktop computers.

# Unit summary

This unit will provide a thorough introduction to the concepts and methods of modern Bayesian statistical methods with emphasis on practical applications in biostatistics. It begins with a discussion of the role of subjective probability in quantifying uncertainty in the scientific process. The concept of full probability modelling will be introduced and developed through single- and multi-parameter models with conjugate prior distributions. We explore the relationship between noninformative and informative prior distributions and their effect on posterior estimates, and explain the connection between Bayesian methods using noninformative priors and frequentist approaches. The application of Bayesian methods for fitting hierarchical models to correlated data structures will be developed. Computational techniques for use in Bayesian statistics, especially the use of simulation from posterior distributions using Markov chain Monte Carlo techniques (MCMC) will be covered with emphasis on the use of the *Stan* software package (accessed using either *R* or *Stata*) for the implementation of Bayesian analyses. The overall aim is to develop students' ability to perform and critically interpret Bayesian statistical analyses in health and medical research.

# Workload requirements

The expected workload for this unit is 8-12 hours per week on average, consisting of guided readings, discussion posts, independent study and completion of assessment tasks.

# Prerequisites

Epidemiology (EPI), B Mathematical Background for Biostatistics (MBB), Probability and Distribution Theory (PDT), Principles of Statistical Inference (PSI), Linear Models (LMR), Categorical Data & Generalised Linear Models (CDA)

# Learning Outcomes

At the completion of this unit students should be able to:

1. **Explain the logic of Bayesian statistical inference:**
   a. The use of full probability models to quantify uncertainty in statistical conclusions.
   b. The role of subjective probability in quantifying uncertainty about unknown parameters.
   c. The Bayesian updating process that combines a prior probability distribution with the data via the likelihood function to produce a posterior distribution that can be used to make direct probability statements about unknowns.
2. **Familiarity with standard Bayesian models:**
   a. Develop and analytically describe simple one- or two-parameter models with conjugate prior distributions and standard models containing two or more parameters including specifics for the normal location-scale model.
   b. The specification of the conjugate prior distribution as an informative prior distribution.
   c. Be able to program the above models in Microsoft Excel, R, Stata and, if necessary, *Stan* and provide an appropriate interpretation of the output on which inferences can be based.
3. Understand the relationship between **noninformative and informative prior distributions** and their effect on posterior estimates, and understand the connection between Bayesian methods using noninformative priors and frequentist approaches.
4. Recognise situations where both simple and more complex biostatistical data structures can be expressed as using Bayesian or **hierarchical Bayesian models**, and to be able to specify the technical details of such models including the simulation and inference of posterior estimates.
5. Explain and use the most common **computational techniques for Bayesian analysis**, especially the use of simulation from posterior distributions based on Markov Chain Monte Carlo (MCMC) methods, with emphasis on implementing each of the steps in an explicit strategy for posterior simulation.
6. Acquire skills to perform **practical Bayesian data analysis** relating to health research problems, and to be an advocate for full probability modelling when appropriate.

# Unit content

The unit is divided into 6 modules, summarized below. Each module will involve approximately 2 weeks of study and may include some or all of the following:

1. Module notes providing a guide around reading the relevant sections and examples of the prescribed textbook, including some additional text on special topics and illustrative examples.
2. Selected readings from published articles, technical reports and books, particularly the documentation from the *Stan* project.
3. Work to be completed by the student which may be:

   a. Assignment exercises of a more theoretical nature, similar to end-of-chapter exercises in the textbooks listed below.
   b. Simple simulations or computer exercises using Microsoft Excel, Stata or R.

c. Simple implementations of Bayesian analyses in *Stan* (using both co-ordinator-provided and student-generated computing code) to tackle cited examples and some simulation suggestions, again similar in style to end-of-chapter exercises in the textbooks.

# Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings to which they refer and working through the accompanying exercises. You will learn a lot more efficiently if you tackle the exercises systematically as you work through the notes. You are encouraged to post any content-related questions to *eLearning*, whether they relate directly to a given exercise, or are a request for clarification or further explanation of an area in the notes. You should also work through all of the computational examples in the notes for yourself on your own computer.

# Method of communication with co-ordinator

The co-ordinator will answer questions related to the module notes and practical exercises, and address any other issues that require clarification. I will be encouraging communication and interaction between students and between students and the co-ordinator via the discussions facility on *eLearning*, in addition to using the site for posting unit materials. Paper copies of course notes and any required reading will be sent out by old-fashioned post. It is assumed that all students taking this unit are familiar with BCA processes by now – in particular with how to contact Erica or Emily at BCA HQ in Sydney for administrative assistance, and how to access the (new!) *eLearning* system. If you need help with any of this please let us know.

# Module descriptions

**Module 1 – Background and introduction to Bayesian ideas**

- Philosophical aspects of Bayesian inference in public health and biostatistics. The scientific process and role of subjective probabilities in assessing scientific evidence. The role of Bayesian inference.
- Overview of Bayesian statistical inference
  (a) Set up full probability model
  (b) Condition on observed data to obtain posterior distribution
  (c) Evaluate fit of model and implications of results.
- Mechanics of Bayesian analysis and inference: Joint probability distribution, prior density, sampling model, Bayes' rule, likelihood and odds ratios.
- Probability as a quantitative measure of uncertainty.
- Introduction to the *Stan* software.

*Bayesian Data Analysis*

Chapter 1 – Probability and Inference, Sections 1.1 – 1.5.

*Stan Modeling Language - User's Guide and Reference Manual*

Chapter 1 (not 1.7) – Overview

Chapter 29 – Bayesian Data Analysis
Chapter 32 – Bayesian Point Estimation
Chapter 8 – Execution of a *Stan* Program
Chapter 69 – Stan Program Style Guide

**Module 2 – Single-parameter models**

- Estimating a probability from binomial data. Bernoulli trials and the binomial likelihood. Prediction in the binomial example for uniform prior. Posterior distribution as a compromise between data and prior information.
- The beta distribution and conjugate prior distributions in general.
- Bayesian analysis of the normal distribution. Likelihood of one and multiple points. Unknown mean and known variance. Shrinkage.
- Other standard single parameter models; exponential and Poisson.
- Noninformative prior distributions. Proper and improper prior distributions. Pivotal quantities and the special cases location and scale. Difficulties with noninformative (and precise) prior distributions.

*Bayesian Data Analysis*

Chapter 2 – Single-parameter models

**Module 3 – Multiparameter models and large sample inference**

- Bayesian approach to multiparameter problems: Integrate the joint posterior distribution over "nuisance" parameters to obtain the marginal posterior distribution. Marginal distribution as a mixture of the condition distributions; implications for simulation.
- Normal data with non-informative and semi-conjugate prior distribution.
- Large sample inference and frequency properties of Bayesian inference.
- Frequency properties of Bayesian inference, simple Bayesian rules using noninformative priors and the connection between frequentist and Bayesian inference.

*Bayesian Data Analysis*

Chapter 3 – Introduction to multiparameter models, Sections 3.1 – 3.3
Chapter 14 – Introduction to regression models, Sections 14.1 – 14.2, 14.8

*Stan Modeling Language - User's Guide and Reference Manual*

Chapters 9.1 – 9.5 (not 9.2) – Regression Models

**Module 4 – Computation for Bayesian data analysis**

- The importance in Bayesian inference of simulations from the posterior distribution – joint and marginal.
- Markov chain Monte Carlo (MCMC) (1) Gibbs sampler; (2) Metropolis and Metropolis-Hastings algorithms; and (3) Hamiltonian Monte Carlo and *Stan*.
- Diagnosing convergence of multiple chains
- Effective sample size

*Bayesian Data Analysis*

Chapter 11 – Basics of Markov chain simulation

*Stan Modeling Language - User's Guide and Reference Manual*

Chapter 30 – Markov chain Monte Carlo
Chapter 34 – Hamiltonian Monte Carlo

Appendix B – *Stan* for users of BUGS

**Module 5 – Hierarchical models**

- Hierarchical specification of probability models; parameters as a sample from a common population distribution.
- Exchangeability and setting up hierarchical models.
- Computation with hierarchical models.
- Full analytical treatment of the simple hierarchical normal model.

*Bayesian Data Analysis*

Chapter 5 – Hierarchical models

*Stan Modeling Language - User's Guide and Reference Manual*

Section 9.9 Hierarchical Logistic Regression
Section 9.10 Hierarchical Priors
Section 14.2 Meta-Analysis

**Module 6 – Hierarchical linear models**

- Random effects regression coefficients exchangeable and varying in batches
- Interpreting a normal prior distribution as additional data
- Random intercepts and slopes – random coefficient models

*Bayesian Data Analysis*

Chapter 15 – Hierarchical linear models, Sections 15.1 – 15.5

*Stan Modeling Language - User's Guide and Reference Manual*

Section 9.13 – Multivariate Priors for Hierarchical Models

# BAY Timetable

| Study Week | Week Commencing | Topic | Assessment |
|---|---|---|---|
| 1 | 30 July 2018 | Module 1 | |
| 2 | 6 August | | |
| 3 | 13 August | Module 2 | Assignment 1 due |
| 4 | 20 August | | |
| 5 | 27 August | Module 3 | Assignment 2 due |
| 6 | 3 September | | |
| 7 | 10 September | | |
| 8 | 17 September | Module 4 | Assignment 3 due |
| Mid-semester break | 24 September | | |
| 9 | 1 October | | |
| 10 | 8 October | Module 5 | Assignment 4 due |
| 11 | 15 October | | |
| 12 | 22 October | Module 6 | Assignment 5 due |
| 13 | 29 October | | |
| 14 | 5 November | | |
| 15 | 12 November | | Assignment 6 due |

# Assessment

Assessment will include 2 written projects worth 30% each following modules 3 and 6, to be made available at the beginning of these modules and completed within 3 weeks. In addition, students will be required to submit solutions to selected practical exercises (one assignment for each for modules 1, 2, 4 and 5), worth a total of 40%, by deadlines specified on the timetable below (2 weeks later for modules 1, 2 and 5, 3 weeks for Module 4 since this included the non-teaching week).

You are required to submit your work typed in Word or similar (e.g. using Microsoft's Equation Editor for algebraic work) and we strongly recommend that you become familiar with equation typesetting software such as this. If extensive algebraic work is involved you may submit neatly handwritten work, however please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. This handwritten work should be scanned and collated into a single pdf file and submitted via the *eLearning* site. See the BCA Assessment Guide document for specific guidelines on acceptable standards for assessable work.

The co-ordinator will avoid answering questions relating directly to the assessable material until after it has been submitted, but we encourage students to discuss the relevant parts of the notes among themselves, via *eLearning*. **Explicit solutions to assessable exercises should not be posted.** Each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source. **The due date for submission of the required exercises from each module is 11:59pm on the Monday following the two- or three-week period that module, as indicated below.**

# Submission of assessments and academic honesty policy

You should submit all your assessment material via *eLearning* unless otherwise advised. The use of *Turnitin* for submitting assessment items has been instigated within unit sites. For more detail please see pages 3-5 the BCA Assessment Guide.

The BCA pays great attention to academic honesty procedures. Please be sure to familiarise yourself with these procedures and policies at your university of enrolment. Links to these are available in the BCA Assessment Guide. When submitting assessments using *Turnitin* you will need to indicate your compliance with the plagiarism guidelines and policy at your university of enrolment before making the submission.

# Late submission of assessments and extension procedure

Assessment deadlines are important. Due to prerequisites, late results may preclude students from enrolling to study subsequent units. Different universities have different result submission deadlines. BCA results will be transmitted to universities the week before the earliest university deadline. Regarding granting of extensions or late submissions, we urge students to contact the co-ordinator well in advance (at least one week) if they anticipate requiring an extension for one of the assessable tasks due a clearly foreseeable reason. Late submissions will be penalised at the rate of 5% per day, including weekends, to a maximum penalty of 50% of the total mark.

# Feedback

Our feedback to you:

The types of feedback you can expect to receive in this unit are:

- Formal individual feedback on submitted exercises assignments
- Feedback from non-assessed online quizzes
- Responses to questions posted on *eLearning*

Your feedback to us:

One of the formal ways students have to provide feedback on teaching and their learning experience is through the BCA student evaluations at the end of each unit. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

# Learning resources

There are three textbooks that cover Bayesian data analysis using *Stan* which in this order may be thought of as introductory, intermediate and advanced:

1. Kruschke JK. Doing Bayesian Data Analysis: A Tutorial with R, JAGS and *Stan*. Academic Press / Elsevier, 2015.
2. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and *Stan*, CRC Press / Taylor and Francis / Chapman and Hall, 2016.
3. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis (3rd Edition). CRC Press / Taylor and Francis / Chapman and Hall, 2014.

**The prescribed textbook for this unit is Gelman *et al*. (2014).** Students should purchase a copy of this textbook.

A small number of additional required readings from journal articles, books and various sources for material on *Stan* will be provided to students. A number of additional reference books are recommended as further reading, with the list below in approximate order of difficulty; *Bayesian Data Analysis* above is approximately equivalent in difficulty to Spiegelhalter *et al*. (2004). Peter Congdon's second book Applied Bayesian Modelling focuses on the use of WinBUGS.

Iversen GR. Bayesian Statistical Inference. Series Number 07-043. Sage Publications, Beverly Hills, CA, 1985.

Berry DA. Statistics: A Bayesian Perspective. Wadsworth: Belmont, California, 1996.

Lee PM. Bayesian Statistics: An Introduction (3rd Edn). Edward Arnold: London, 2004.

Thompson J. Bayesian Analysis with Stata. Stata Press, 2014. **NOTE:** This book explains how to use Stata in-built Bayesian routines, not how to access *Stan* through Stata.

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

Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS book. CRC Press / Taylor and Francis / Chapman and Hall, 2013.

Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, 2007.

Congdon P. Bayesian Statistical Modelling. 2nd Edn. Wiley, 2006.

Congdon P. Applied Bayesian Modelling. 2nd Edn. Wiley, 2014.

Robert C. The Bayesian Choice. Springer, New York, 2001.

O'Hagan A. Kendall's Advanced Theory of Statistics Vol 2B: Bayesian Inference. Arnold: London, 1994.

Carlin BP, Louis TA. Bayes and Empirical Bayes Methods for Data Analysis. 2nd Edn. Chapman and Hall/CRC: Boca Raton, Florida, 2000.

Berger JO. Statistical Decision Theory & Bayesian Inference. Springer-Verlag: Berlin, 1985.

Bernardo JM, Smith AFM. Bayesian Theory. John Wiley and Sons, Ltd: Chichester, 1994.

# Software

For this unit you will need to have access to Microsoft Excel, Stata, or R for simple simulations and *Stan* (mc-stan.org) for model-fitting using MCMC routines. *Stan* can be accessed through R, Stata and various other platforms.

# Changes to BAY since last delivery, including changes in response to student evaluation

BAY was last delivered in 2016. The current structure of the unit does not acknowledge the ***New Bayes*** (a co-ordinator-invented term) which is that the discipline is now separated into three streams:

1) **Inference** (full probability modelling, probability as a measure of uncertainty)
2) **Computation** (approximations, simulation, Markov chain Monte Carlo and associated software – BUGS, JAGS and *Stan*)
3) **Applications and Data Analysis** (working with real data)

The WinBUGS software ironically no longer works reliably on Windows installation beyond about Windows 7.

Students had mixed feelings about the required textbook "Bayesian Data Analysis" (3rd Edition) since its technical complexity reaches well beyond what we expect to cover in this unit.

**Proposed action plan for this delivery:**

Consider using an alternative textbook (done – we'll stick with *Bayesian Data Analysis* since the alternatives are not suitable)

Decommission WinBUGS from use in BAY and switch to *Stan*.

Short Assessments to contain only one topic / data example, similar to the LCD model.