THE UNIVERSITY OF NEWCASTLE AUSTRALIA

**FACULTY OF SCIENCE AND INFORMATION TECHNOLOGY**
School of Mathematical and Physical Sciences
**Student Enquiries**
Telephone: (02) 4921-5000

# STAT3120: Applied Bayesian Methods
## Semester 2, 2016

# ASSIGNMENT 2

Due Friday 16 September 2016, 5pm

**Late assignments will not be accepted without prior written permission of the lecturer.**

## Total marks: 45

Marking Criteria:

1. Answers must be written in clear English with all appropriate working and/or supporting computer output shown.
2. Raw computer output without explanatory text is unacceptable.
3. Students are required to understand the content in Weeks 2-7 to answer these questions.
4. As part of your workings you should include applicable R-code.

Question 1 [Total: 12 marks]

(a) Consider a sample of $n$=15 data points, $y$, from the Normal distribution with unknown mean $\mu$ and variance $\sigma^2$. For this sample we wish to allow the 14th and 15th data points, ($y_{14}, y_{15}$), to have an inflated variance, namely $10\sigma^2$.

(i) Write down the likelihood $p(y \mid \mu, \sigma^2)$, including all constants. [2 marks]

To perform inference on $p(\mu, \sigma^2 \mid y)$ we partition using $p(\mu \mid y)$ and $p(\sigma^2 \mid y, \mu)$.

(ii) Mathematically derive the density $p(\sigma^2 \mid y, \mu)$ assuming the non-informative prior $p(\sigma^2) \propto 1/\sigma^2$. Derive $E(\sigma^2 \mid y, \mu)$ and $V(\sigma^2 \mid y, \mu)$. [4 marks]

(b) A biologist believes that lifetimes in days of fruitflies follow an exponential distribution. The density for a single observation y drawn from an exponential distribution is

$p(y; \lambda) = \lambda e^{-\lambda y}, \quad y > 0.$

For data, the biologist observes 100 randomly selected fruitflies and records how many days they live.

(i) Write the likelihood of $\lambda$ if there are 100 data values, represented symbolically as $y_1,...,y_{100}$. What parametric family is the conjugate prior for the exponential likelihood? (Just name the family.) [2 marks]

(ii) The biologist states that a prior distribution with mean 3 days and standard deviation 2 would reflect their previous knowledge about fruitfly lifetimes. What parameters should they specify in their prior? (numeric answer; show your work) [2 marks]

(iii) Based on the likelihood and the prior that you have specified, find the posterior density $p(\lambda \mid y_1,...,y_{100})$. Name it and give the numeric values of the parameters. [2 marks]

Question 2 [Total: 13 marks]

The table below is adapted from the text BDA. It consists of the numbers of fatal accidents and approximate numbers of passenger miles (in 100 millions) flown each year from 1976 to 1985.

| Year | Fatal accidents | Passenger miles |
|---|---|---|
| 1976 | 24 | 3863 |
| ‘77 | 25 | 4300 |
| ‘78 | 31 | 5027 |
| ‘79 | 31 | 5481 |
| ‘80 | 22 | 5814 |
| ‘81 | 21 | 6033 |
| ‘82 | 26 | 5877 |
| ‘83 | 20 | 6223 |
| ‘84 | 16 | 7433 |
| ‘85 | 22 | 7107 |

(a) Assume the numbers of fatal accidents are independently and identically distributed. Suggest an appropriate distribution to model the number of fatal accidents. [2 marks]

(b) Set a suitable prior distribution for the mean parameter in (a) and determine its posterior distribution based on the data in the table above. Plot this posterior density. [3 marks]

(c) Find a point estimate and 95% interval for the average yearly number of fatal accidents. [2 marks]

(d) Part (a) has ignored the information about passenger miles. Perhaps a better model here would take account that each fatal accident count is proportional to the number of miles flown. Set up a model of accident counts that accounts for the number of miles flown. [2 marks]

(e) Using this new model, set a prior distribution for the mean parameter and determine its posterior distribution. Plot this density and compare it with that in (b). [4 marks]

Question 3 [Total: 20 marks]

Consider the example on corn yields but with an extended data set, with sample sizes, sample means and sample variances as given below. We have five different corn growers giving individual yields for a particular new type of genetically engineered corn that has three seasons per year. The corn is distributed by a research station to the growers. Growers give yields for the most recent seasons in tons/hectare. Summary statistics for the data are given below.

| Grower | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $n_i$ | 16 | 19 | 14 | 12 | 8 |
| $\bar{y}_i$ | 15.3 | 16.2 | 16.4 | 13.2 | 13.5 |
| $s_i^2$ | 8.2 | 12.3 | 7.9 | 5.2 | 6.2 |

(a) By adapting code from the week 7 lecture, fit a hierarchical Normal model to this data, using the Empirical Bayes approach. Provide estimates and inference for each unknown true grower mean yield, including a plot showing all estimated posterior densities, one for each unknown mean. [7 marks]

(b) Plot the estimated parent density for the five grower means. At what percentile of this distribution does of each grower's estimated posterior mean yield lie? Do they all fit in with the prior distribution? [2 marks]

(c) What is the posterior probability that each of the five growers has the largest mean yield? What is your conclusion in this case? [2 marks]

(d) What is the posterior probability that, for each pair of growers (grower *i* and *i*+*k*), the mean yield for grower(*i*) exceeds the mean yield for grower (*i*+*k*)? [2 marks]

(e) What is the most likely ordering for the 5 mean grower yields in your MC sample? What are the 2nd, 3rd and 4th most likely orderings? [4 marks]

(f) Based on (a)-(d), and any other evidence you can provide, where do you believe significant differences exist among the 5 grower means (if any)? [3 marks]

## End of Assignment 2