Course 52558: Problem Set 1
Due March 31st, 2009

SOFTWARE NOTE: in some of the questions you will need to use Matlab or R for computations and plotting. Examples in the course will generally be in Matlab but these can easily be "translated" into R. If you prefer R (can be installed on your own computer at no cost), you can definitely use it. To get started

- Install R and WinBUGS (that we will use later in the course) by following the instructions in http://www.stat.columbia.edu/ gelman/bugsR (we will use WinBUGS later so don't worry if you don't know what it is).
- Follow one of the many R tutorial that can be found on the web (e.g. from http://www.cyclismo.org/tutorial/R), at least until the example below does not look like Chinese

Now, take a look at the matlab example (NormalPriorPost.m) or R example (NormalPriorPost.txt) that plots the normal prior and posterior. You should understand what this code does before attempting the tasks below

1. **A Thought Experiment**. Let θ be the true proportion of men in Israel over the age of 40 with hyper-tension.
   (a) Though you may have little or no expertise in this area, use your social knowledge and common sense to give an initial point estimate (single value) of θ.
   (b) Now suppose that in a properly designed survey, of the first 5 randomly selection men, 4 are hypertensive. How does this information effect your initial estimate of θ?.
   (c) Finally suppose that at the survey's completion, 400 of 1000 men have emerged as hypertensive. Now what is your estimate of θ?
   (d) What guidelines for statistical inference do your answers suggest?
2. **Relationship between posterior and prior mean and variance**.
   (a) Show that for any continuous random variables X and Y
   $$E(X) = E(E(X \mid Y))$$
   (Note that a similar proof can be used for the discrete case)
   (b) Show that for any random variables X and Y
   $$\text{var}(X) = E(var(X \mid Y)) + var(E(X \mid Y))$$

(c) Let y denote the observed data. We assume y was generated from $p(y \mid \theta)$, where θ, the parameters governing the sampling of y are random and distributed according to $p(\theta)$. Use the above and describe (i.e. understand the equation and then put into words) the relationship between the mean and variance of the prior $p(\theta)$ and and the posterior $p(\theta \mid y)$.

3. **Posterior of a Poisson Distribution**. Suppose that X is the number of pregnant woman arriving at a particular hospital to deliver babies in a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest adopting a Poisson likelihood

$$p(x \mid \theta) = \frac{e^{-\theta}\theta^x}{x!}, x \in \{0, 1, 2, \ldots\}, \theta > 0$$

To provide support on the positive real line and reasonable flexibility we suggest a Gamma $G(\alpha, \beta)$ distribution prior

$$p(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \theta > 0, \alpha > 0, \beta > 0$$

where $\Gamma()$ is a continuous generalization of the factorial function so that $\Gamma(c) = c\Gamma(c-1)$. α, β are the parameters of this prior, or the hyper-parameters of the model. The Gamma distribution has mean $\alpha\beta$ and variance $\alpha\beta^2$.

Show that the posterior distribution $p(\theta \mid x)$ is also Gamma distributed. Determine its parameters α and β.

4. **Posterior of the Poisson Model**.

In this question we will use Matlab/R to explore the Poisson model with the Gamma prior considered above.

(a) The Matlab (GammaPrior.m) R (GammaPrior.txt) files in the Code directory of the course web page can be used to plot the Gamma prior. Use this code to investigate different values for α and β. Describe the qualitative behavior of this prior as a function of these parameters and try to explain why they are called 'shape' and 'scale' parameters, respectively.

(b) Continuing the previous question involving births, assume that in December 2008 we observed $x = 42$ moms arriving at the hospital to deliver babies, and suppose we adopt a Gamma(5,6), which has mean 30 and variance 180, reflecting the hospital's total for the two preceding years. Use Matlab/R to plot the posterior distribution of θ next to its prior. What are your conclusions?

(c) Repeat the above for different values of x. What are your conclusions.

5. **Extinction of Species**. Paleobotanists estimate the moment in the remote past when a given species became extinct by taking cylindrical, vertical core samples well below the earths surface and looking for the last occurrence of the species in the fossil record, measured in meters above the point P at which the species was known to have first emerged. Letting $\{y_1, \ldots, y_n\}$ denote a sample of such distances above P at a random set of locations, the model

$$(y_i|\theta) \sim \text{Unif}(0, \theta)$$

emerges from simple and plausible assumptions. In this model the unknown $\theta > 0$ can be used, through carbon dating, to estimate the species extinction time. This problem is about Bayesian inference for θ, and it will be seen that some of our usual intuitions do not quite hold in this case.

(a) Show that the likelihood may be written as

$$l(\theta : y) = \theta^{-n} I(\theta \geq max(y_1, \ldots, y_n))$$

where $I(A) = 1$ if A is true and 0 otherwise.

(b) The Pareto(α, β) distribution has density

$$p(\theta) = \begin{cases} \alpha\beta^\alpha \theta^{-(\alpha+1)} & \theta \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha, \beta > 0$. The Pareto distribution has mean $\frac{\alpha\beta}{\alpha-1}$ for $\alpha > 1$ and a variance of $\frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$ for $alpha > 2$.
With the likelihood viewed as a constant multiple of a density for θ, show that the likelihood corresponds to the Pareto(n 1,m) distribution. Now let the prior for θ be taken to be Pareto(α, β) and derive the posterior distribution $p(\theta|y)$. Is the Pareto conjugate to the uniform?

(c) In an experiment in the Antarctic in the 1980s to study a particular species of fossil ammonite, the following is a linearly rescaled version of the data obtained: $y = (0.4, 1.0, 1.5, 1.7, 2.0, 2.1, 3.1, 3.7, 4.3, 4.9)$. Prior information equivalent to a Pareto prior with $(\alpha, \beta) = (2.5, 4)$ was available. Plot the prior, likelihood, and posterior distributions arising from this data set on the same graph, and briefly discuss what this picture implies about the updating of information from prior to posterior in this case.

(d) Make a table summarizing the mean and standard deviation for the prior, likelihood and posterior distributions, using the (α, β) choices and the data in part (d) above. In Bayesian updating the posterior mean is often a weighted average of the prior mean and the likelihood mean (with positive weights), and the posterior standard deviation is typically smaller than either the prior or likelihood standard deviations. Are each of these behaviors true in this case? Explain briefly.