STA 410/2102, Spring 2004 — Assignment #4

Due at **start** of class on April 8. Worth 15% of the final mark.

Note that this assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own.

New variant Creutzfeldt-Jakob disease (nvCJD) is thought to be caused by human consumption of beef from cattle infected with Bovine Spongiform Encephalopathy (BSE). Not much is known for sure about nvCJD, however, so we might be interested in confirming that eating beef from infected cattle really is a cause of nvCJD, and that it is the only cause.

Since nvCJD is rare even among those who have eaten beef from BSE-infected cattle, it may be reasonable to model the number of cases of nvCJD in a community by a Poisson distribution. Suppose we have data on n communities, which have nearly equal populations, and which were supplied with beef from pretty much the same sources. We model the number of cases of nvCJD in community i, y_i, as being a random value from a Poisson distribution with mean λ_i, with the y_i for different communities being independent given the λ_i. Three hypothesis might be considered for how λ_i is related to how many million kilograms of beef, x_i, were consumed by people in the ith community:

H_1 : The mean is a positive constant that does not depend on the amount of beef consumed: ie, $\lambda_i = \alpha$, for some positive constant α.

H_2 : The mean is proportional to the amount of beef consumed: ie, $\lambda_i = \beta x_i$, for some positive parameter β.

H_3 : The mean is a positive constant plus an amount proportional to the amount of beef consumed: ie, $\lambda_i = \alpha + \beta x_i$.

If H_1 is true, we might conclude that eating beef from cattle infected with BSE is not the cause of nvCJD. If H_2 is true, we might conclude that eating beef from cattle infected with BSE is the only cause of nvCJD. If H_3 is true, we might conclude that eating beef from cattle infected with BSE is a cause of nvCJD, but that there is also some other cause, since the mean is positive even when x_i is zero.

The Bayesian approach to judging how plausible these three models/hypotheses are is based on the "marginal likelihood" for each model, which is the prior probability of the observed numbers of cases, $y_1, \ldots, y_n$, under each model. (The amounts of beef eaten, $x_1, \ldots, x_n$, are regarded as fixed quantities, which are not being modeled.)

In general, the marginal likelihood of data y with respect to a model, H, with parameters θ is defined to be

$$P(y|H) \;=\; \int P(\theta|H)P(y|\theta,H)d\theta$$

Here, $P(\theta|H)$ is the prior density for the parameter of model H, and $P(y|\theta,H)$ is the likelihood for model H, based on the observed data. To obtain the posterior probabilities

for a set of models (assumed to be the only possibilities), we would multiply the marginal likelihood of each model by our judgement of its prior probability, and then normalize these numbers to sum to one.

To apply this method to the three hypotheses above, we need to define prior distributions for the parameters of each. Let's suppose that experts on nvCJD and BSE have selected the following priors:

H_1 : $\alpha \sim \text{Uniform}(0, 10)$.

H_2 : $\beta \sim \text{Uniform}(0, 5)$.

H_3 : $\alpha \sim \text{Uniform}(0, 4)$ and $\beta \sim \text{Uniform}(0, 5)$, with α and β independent in the prior.

Data is available on six communities, as follows:

i	x_i	y_i
1	0.5	4
2	2.2	8
3	1.2	2
4	2.0	6
5	0.7	2
6	0.2	1

You should write R functions to evaluate the marginal likelihoods for H_1, H_2, and H_3, based on this data. Using each of these models, you should also find the posterior mean for λ for a community with $x = 1$. (In other words, you find the posterior mean of α for H_1, of β for H_2, and of $\alpha + \beta$ for H_3.)

You should do the integrations required using a function that you write for integrating a one-dimensional function using Simpson's Rule, with some specified number of intervals. You should try using increasing numbers of intervals for the integration until you reach a point where the approximations to the integrals appear to have converged.

You should hand in a listing of your R functions, properly formatted and commented, the results you obtained for the marginal likelihoods and for the posterior expected values (including results for at least two numbers of intervals for Simpson's Rule, to demonstrate convergence), and a brief discussion of the results, commenting on what they mean in terms of the scientific problem being solved.