Introduction to Bayesian Modeling

Homework Assignment #2

Due 24 March 2020 by the start of class.

1. The Aché tribe of Paraguay are part-time hunter-gatherers and have been in contact with Paraguayan society only since the mid-1970s. Part of Aché life is spent away from the village on extended forest treks. While trekking, the Aché subsist exclusively on foods that they collect on a given day. Armadillos comprise the vast majority of food calories consumed by the Aché and it is of interest to quantify the typical number of armadillos killed in a day.

   Our data involve observations on $n = 38$ Aché men. Let y_i be the number of armadillos killed by the i'th man on a given day. The Poisson distribution provides a natural model for the number of events occurring haphazardly over a given amount of time, so we assume

   $$y_1, \ldots, y_n | \theta \overset{\text{iid}}{\sim} \text{Pois}(\theta).$$

   We refer to θ as the *kill rate*. The broader Aché data include many hunting days for each man. For this analysis one day has been randomly selected for each man.

   Dr. Garnett McMillan, an expert on Aché hunting practices, believes that Aché men typically kill an armadillo every other day. This provides a "best guess" for θ of 0.5 armadillos. Dr. McMillan is 95% sure that the mean daily number of kills is no greater than 2 armadillos.

   In fact, our data show that across the randomly selected hunting days for our $n = 38$ Aché men, a total of 10 armadillos were killed.

   1. Using the expert information from Dr. McMillan, find a prior distribution for the kill rate. Assume that Dr. McMillan's "best guess" gives the mode of the prior distribution, and use Table 2.2 and the discussion of prior-finding on page 27 of your textbook. (HINT: Find α as a function of β, then use computer software to find β such that the resulting Gamma distribution accords with Dr. McMillan's beliefs.)

   2. Find the posterior distribution for the kill rate, based on your prior and the reported armadillo hunting data.

   3. Find the expected number of armadillos that will be killed if these 38 Aché hunters each hunt for one more day. Explain how you found this.

   4. Give a probability interval for how many armadillos will be killed if these 38 Aché hunters each hunt for one more day. Explain how you found this interval. There are many ways you might approach this problem, and I'm interested in seeing how you'll try to do it. Make sure to report the probability (approximate or exact) you've used for defining your interval.

- If you haven't yet done it, download and install OpenBUGS (http://www.openbugs.net/w/Downloads). You can find guidance on how to use WinBUGS, which is almost the same thing, in Section 3.2 of the textbook.

  If you're running a Mac OS, you may need to Google information on installing/running OpenBUGS. https://www.r-bloggers.com/running-r2winbugs-on-a-mac-running-osx/ might be helpful to you on this. Alternatively you could run JAGS, which is very similar, but I don't know JAGS well enough yet to provide much support for it.

- Read Example 3.1.3 in the textbook. Perform this example in OpenBUGS with $y_1 \sim \text{Bin}(80, \theta_1)$, $y_2 \sim \text{Bin}(100, \theta_2)$, $\theta_1 \sim \text{Beta}(1, 1)$, $\theta_2 \sim \text{Beta}(2, 1)$ with observations $y_1 = 32$ and $y_2 = 35$. Put each term in the model on a separate line. There should still be only two list statements with entries separated by commas. See Exercises 3.6 and 3.7 in the textbook for WinBUGS syntax.

2. Suppose n cities were sampled and for each city i the number y_i of deaths from ALS were recorded for a period of one year. We expect the numbers to be Poisson distributed, but the size of the city is a factor. Let M_i be the known population for city i (in thousands of people) and let

$$y_i \mid \theta \overset{ind}{\sim} \text{Pois}(\theta M_i), \quad i = 1, \ldots, n,$$

   where $\theta > 0$ is an unknown parameter measuring the common death rate per 1000 people for all cities. Given θ, the expected number of ALS deaths for city i is θM_i, so θ is expected to be small. Assume that independent scientific information can be obtained about θ in the form of a gamma distribution, say $\text{Gamma}(\alpha, \beta)$. Show that the posterior distribution for $\theta \mid y_1, ..., y_n$ is a Gamma distribution in this case, and find the value of its parameters.

3. Extending Problem 1, two cities are allowed different death rates. Let $y_i \overset{ind}{\sim} \text{Pois}(\theta_i M_i)$, $i = 1, 2$, where the M_is are known constants. Let knowledge about θ_i be reflected by independent gamma distributions, namely $\theta_i \sim \text{Gamma}(\alpha_i, \beta_i)$. Derive the joint posterior for (θ_1, θ_2). Characterize the joint distribution in a fashion similar to the textbook's Example 3.1.3 (where it is done for sampling data from two independent binomials). Think of θ_i as the rate of events per thousand people in city i. For independent priors $\theta_i \sim \text{Gamma}(1, 0.1)$, give the exact joint posterior with $y_1 = 500, y_2 = 800$ in cities with populations of 100 thousand and 200 thousand, respectively.

4. Perform a data analysis for the model in Problem 2 using the data $y_1 = 500, y_2 = 800, M_1 = 100, M_2 = 200$, and using independent $\text{Gamma}(1, 0.1)$ priors for the θ_is. Make OpenBUGS-based inferences for all parameters and functions of parameters discussed there using a Monte Carlo sample size of 10,000 and a burn-in of 1,000. This may involve an excursion into the "Help" menu to find the syntax for Poisson and Gamma distributions. Compare the posterior means for θ_1 and θ_2 based on the OpenBUGS output to the exact values from the Gamma posteriors that you obtained in Problem 2.

5. Read Example 5.1.4 in the textbook—the Reye's Syndrome (RS) example.

A situation that can arise in case-control sampling is where there is prior information on the odds ratio (OR), perhaps based on a previous case-control study that is similar to the one at hand. We can place a prior on the parameters by placing, say, a normal prior on $\delta \equiv \log(OR)$ and an independent Beta prior on θ_2. These induce a prior on (θ_1, θ_2). To apply this, we need to solve for θ_1 in terms of δ and θ_2. Some algebra gives

$$\theta_1 = \frac{e^{\delta}\theta_2}{1 - \theta_2(1 - e^{\delta})}.$$

To elicit a prior on δ we think about OR. If our best guess is that $OR = 3$, then we take the mean of the normal distribution for δ to be $\log(3) = 1.1$. Moreover, if we are, say, 90% sure that the OR is at least 0.8, then we are also 90% sure that $\log(OR)$ is at least $\log(0.8) = -0.22$. We need to find a normal distribution with a mean of 1.1 and a 10th percentile of -0.22. We know (or can look up) that the 10th percentile of a normal is -1.28 standard deviations below the mean, so we set $1.1 - 1.28\,\sigma = -0.22$ and solve for $\sigma = 1.03$. Our prior on δ is $N(1.1, (1.03)^2)$.

With respect to the RS data, an expert has no belief that there is or is not an effect of aspirin on RS, but that if there is one, it won't be "huge" in either direction. They place a $\mathrm{N}(0, 2)$ prior on δ, so their best guess for the OR is $e^0 = 1$. The prior also indicates that they are 95% sure that the OR is in the interval $e^{(-1.96*1.414, 1.96*1.414)} = (0.063, 16.0)$. The interval is "centered" on 1 and allows for a broad range of possibilities in both directions. Place the Jeffreys' Beta(0.5,0.5) prior on θ_2.

Analyze the RS data by adding to the following OpenBUGS code.

```
model{
 for(i in 1:2){ y[i] ~ dbin(theta[i],n[i]) }
 theta[2] ~ dbeta(a,b)
 delta ~ dnorm(mu, prec)
 theta[1] <- exp(delta)*theta[2]/(1-theta[2]*(1-exp(delta)))
 OR <- theta[1]/(1-theta[1])/(theta[2]/(1-theta[2]))
}
```

Examine the sensitivity of the results to the choice of prior. Try a prior that reflects much more skepticism about whether there is any effect and a prior that suggests that any effect will be a positive one. Also consider a $U[\log(0.02), \log(50)]$ prior for δ and a $\mathrm{Beta}(1, 1)$ prior for θ_2 to see what impact that has on the results. Be sure to calculate the posterior probability that $OR > 1$, and possibly some other posterior probabilities, like the probability that $OR > 2$.