Introduction to Bayesian Statistics

Practicum 4

Using WinBugs to estimate a single Mean and Variance

**Background**

We have so far evaluated the posterior median and 95% BCI of a dichotomous outcome. Here is an example with a continuous outcome. As you saw earlier, the Bayesian way of dealing with analyzing continuous data is quite more complex than what needs to be done for dichotomous data. So, in the following example, we will try to become familiar with implementing such an approach for estimation of the posterior mean and variance of a continuous variable.

**Description of the data**

The data has been saved in the excel file “class dcc data.xlxs”. It reports the contamination of hands of children and educators with fecal coliforms in 103 classes of 52 day care centers in Québec, Canada. For this example, we will assume that all classes are independent from one another. The variables are described in the table below

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Coding |
| dcc | Day care center ID | NA |
| fc\_kid | Log (base 10) of the number of fecal coliforms on the hands of children per child-mL | Contiuous |
| fc\_ed | Log (base 10) of the number of fecal coliforms on the hands of educator per educator-mL | Contiuous |

The data that we will be using in this practicum has been exported to a concatenate file for easy use in Bugs. This data is called “class data short.txt”.

Write a model in WinBugs to estimate the posterior mean and variance of the number of fecal coliforms on the hands of children per child-ml (on the log 10 scale). To do this, you will need to have some prior information about the mean and variance of fecal coliforms on the hands of children. We will assume a very vague prior for the precision (1/variance), namely a gamma distribution with parameters 0.01, 0.01 (remember that the data is on the log of base 10 and per kid-ml, so this is quite a bad precision). For the mean, we will assume a mean of 0 (which is effectively 1 colony of fecal coliform per mL on the log base 10 scale) with a precision of 0.1 (still on the log of base 10 scale), which is quite vague.

1. Report and interpret the results

2. Run the available diagnosis in WinBugs and assess if the model converged well.

A common alternative way of specifying the variance, and which is often easier to visualize, and widely used in the literature, is to specify a prior on the standard deviation rather than the precision term. With this approach, it is assumed that the standard deviation of the parameter of interest follows a uniform distribution, which is a lot easier to visualize than the gamma distribution. In WinBUGS, we would write code similar to the following:

model

{

for (i in 1:n)

{

data[i]~dnorm(mu,tau)

}

mu~ dnorm(0,0.002)

tau<-1/(sd\*sd)

sd~dunif(0,20)

var<-sd\*sd

}

Of course, the values of the priors would need to be adjusted to the problem at hand. For example, a sd of 20 would result in a variance of 400, which is quite wide for the FC counts, but could be reasonable in other problems. Similarly, the prior on mu should be adapted to the problem at hand. Here, the precision of mu is 0.002 with an equivalent variance of 500.

So go ahead and run this alternative model using a uniform between 0 and 10 for the SD and a mean of 0 and precision of 0.1 for the mean.

1. Does it modify your results?

Next, we will see how changing priors may impact the results. We will re-run the analysis using two other priors and compare them to the results of question 3.

1. MATCH vague: For this prior, use the MATCH software we discussed earlier together with any pertinent prior information on the mean and variance of the number of fecal coliforms on the hands of children (on the log of base 10 per kid-ml). Since we don’t have much info on the variability, the priors on both the mean and precision should be quite wide. It may be easier to think about the prior variance in terms of the standard deviation instead of the precision (recall that we can specify the standard deviation in the form of a uniform distribution). Write down the distribution you identify from MATCH along with any parameter values.
2. MATCH strong: For this prior, For this prior, use the MATCH software we discussed earlier together with any pertinent prior information on the mean and variance of the number of fecal coliforms on the hands of children (on the log of base 10 per kid-ml). Let’s assume that the expert you consult to assess this prior is quite certain about the number of fecal coliforms on the hands of children and on how much that values varies in a population of toddlers. In this case the prior distributions on both the mean and precision should be quite narrow. It may be easier to think about the prior variance in terms of the standard deviation instead of the precision (recall that we can specify the standard deviation in the form of a uniform distribution). Write down the distribution you identify from MATCH along with any parameter values.

4. Report the results in the Table below

|  |  |  |
| --- | --- | --- |
| Prior type | Mean | Variance (or precision) |
|  | Prior Distribution (parameters) | Median | 95% Credible Interval | Prior Distribution (parameters) | Median  | 95% Credible Interval |
| Vague | Normal (0,0.1) |  |  | Gamma (0.01, 0.01) |  |  |
| MATCH Diffuse \_\_\_\_\_\_\_\_\_\_\_\_ |  |  |  |  |  |  |
| MATCH Strong \_\_\_\_\_\_\_\_\_\_\_\_ |  |  |  |  |  |  |

5. Compare the median estimates and credible intervals using the three options for priors in the Table above. Comment on their similitudes / differences.

6. As before, which would best approximate the frequentist estimate?

**Dealing with missing values of the outcome**

In large field epidemiological research, it is not rare to end up with several subjects with missing values. Since Bayesians view all parameters as having a probability distribution, nothing is ever truly exact, which facilitates the ability to deal with missing values in the Bayesian framework, as long as they are missing at random and are not associated with a selection bias of some kind. And even then, if someone could model the probabilities of selection, values of the missing values not missing at random could be imputed. In the data for the the number of fecal coliforms on the educators (teachers)’ hands, there was one teacher that was absent at the time of sampling, and therefore we have two missing values. Because we are aiming at estimating the posterior distribution of the mean of the number of fecal coliforms, we can use this distribution to predict the value of the missing educator. So it is really extremely straightforward. So what needs to be done? Nothing because we are indeed specifying the distribution and as long as it can be assumed that the value for that educator was missing at random (which was the case, there was not systematic reason for this person to be absent that day), then we can use the posterior generated from the likelihood of the OTHER data and the prior to sample a value for the missing data at each iteration.

However, when data are missing like this, it is strongly recommended to create sets of initial values corresponding to the missing values. That is, each missing values becomes a stochastic node (parameter) which needs to be estimated starting at a certain value. The “generate inits” command in WinBugs often generates values that may result in the program crashing. However, for our example, with only 1 value missing, WinBugs should be able to deal with it, especially that the model is very simple at the moment.

7. Write a model to estimate the posterior mean and variance of the number of fecal coliform on the hands of the educators per educator-mL (on the log 10 scale). Using the same priors are those for the children’s hands in question 1 (with the uniform priors on the SD) obtain predicted values for the missing educator data.

Report and interpret the results and assess if the model has converged.