CHL 5223H Applied Bayesian Methods

Homework 1
Monday January 16, 2017
DUE: MONDAY February 6, 2017

For the below homework, please show your work. You should be showing at least some examples of your computer code (R code for example) if you use the computer to compute some things. When you are asked to explains something, you should not be afraid write.

1. Preamble: You have been hired to predict the results of an election. The election is for town select-person. There are three select-persons who make up the the town council and who run the town. There is a division between the purple party and the brown party. If one side has a majority, they will control the issues in the town. Each select-person is elected in one of three town districts and both parties have candidates running in each district. If a candidate wins a majority of votes in a district, that candidate will be the select-person for that district. Also, assume that there will be 5001 citizens voting in each district.

   (a) Prior distributions. Create two different analyses. One based on a "non-informative prior" and the other on an informative prior. Justify your answer. For the informative prior consider the following information. In the past, the two parties have been quite balanced. In the past, each party would get between 40% to 60% of the votes. Use this information to create an informative prior. *(To recieve full credit for your specification of your informative prior, you need to present a prior which has approximately 95% of its mass between .40 and .60.)*

   (b) Posterior distribution for the voting percentage for each district. For each district, a simple random sample of the citizen of each district is asked whom they plan to vote for. In district one: 85 said purple and 65 said brown; in district two: 70 said purple and 80 said brown; and in district three: 50 said purple and 100 said brown. Calculate the posterior probability for the percent who will vote for purple in each district. In reporting your results, provide the distribution of the posterior and the value of the parameters of the parameters and also report appropriate statistics for these distribution (which includes the posterior mean, standard deviation, and some kind of 95% interval). (Note: do this for each of the two priors specified in the first part.)

   (c) Given the above information, provide the probability that the purple party will have a majority in the town council. Also, provide the probability that purple will win each of the districts. Do this by simulation. That is, simulate 10000 elections. For each simulated election, generate a probability, θ_i, which is the probability of a citizen voting for the purple party in district i. Then, generate the number of citizens voting for purple in each district. From there, one can elect either the purple candidate or the brown candidate for each district for each simulated election. Finally, one can see which party had a majority in the town council in the simulated election. (Note: do this for each of the two priors specified in the first part.) *(For full credit on this simulation, you need to show that for each simulated election you sample a θ_i and the number of citizens who vote in each district. Also, no, I don't want to see the actual list of 10000 sampled values.)*

2. (Some basic Bayesian questions) Let's calculate the posterior distribution of the average height (in inches) of a group of people. Suppose that we have two different datasets. The first dataset is: D1=(50, 47, 65, 74, 59, 64), and the second is: D2=(30, 25, 35, 45, 23, 33). Consider the following three priors for this data and then answer the question below.

   Prior 1: Assume the heights X_i of people from a particular population follows a normal distribution with μ and precision τ. Also, assume that the parameter τ is known and that μ has a normal prior distribution with mean μ_0 and precision τ_0. Assume that it is believed that the "average" height is 66 inches (about 182cm). Also, assume that it is generally thought that the 95% of the X_i's are between 54 and 78 inches. Then the posterior standard deviation is about one forth of that range and so the standard deviation of the X_i's given μ is about 6 inches, so the prior belief is that the fixed parameter τ is about 1/36. For the value of μ_0 it is believed that the true value μ is between 63 inches and 69 inches (with about 95% probability). So, the value of τ_0 is about 4/9.

Prior 2: Assume the heights, X_i, are distributed with mean μ and precision τ. Also, assume that τ has gamma prior distribution with parameters (shape and rate) of α and β, and μ has a normal distribution given that $\tau = t$ with mean μ_0 and precision $\theta\tau$. As with prior 1, we let μ_0 equal 66 inches. Also, from the reasoning in prior 1, we will assume that the prior mean of τ is 1/36. Since the mean of this gamma distribution is α/β, then one possibility is to have α=1 and β=36. We can use R or Splus to check the distribution of of the standard deviation of X_i given μ which we figured should be about 6. The following R/Splus command can be used to check the definition:

```
> quantile(1/sqrt(rgamma(10000,1,36)),probs=c(0,.025,.25,.50,.75,.975,1))
    0%    2.5%     25%      50%       75%     97.5%     100%
  1.83    3.13    5.07     7.17     11.46     36.95   449.41
```

We see that their is very good support in the neighbor of 6 for the prior distribution. To finish this prior, we need a value for the parameter θ. Now it could be argued that the values X_i given μ should be much more spread out than the values of μ around its mean. So, we will keep θ greater than one. Here, let us use θ to be about 4 and that leads to the following summary statistics for the prior distribution of the standard deviation of μ given mu_0:

```
>quantile(1/sqrt(4*rgamma(10000,1,36)),probs=c(0,.025,.25,.50,.75,.975,1))
   0%       2.5%     25%     50%     75%  97.5%     100%
   0.88     1.58    2.56    3.63    5.67  18.75   407.89
```

Prior 3: This prior uses the same basic model as prior as was used in Prior 2, but μ_0 equals 66, θ equals 0.1, α=.001, and β equals .001.

Do the following for question 2:

(a) For each of the three priors and for the two data sets (so there are 6 combinations in all), calculate the posterior mean, standard deviation, and a 95% credible region for average height. (Note: the standard precision of the t-distribution is not 1/variance.) You may calculate these values by finding the exact values using analytical methods (algebra) or by performing simulations. (You can also do it both ways if you wish.)

(b) For the six different posterior densities from the previous question, give the density plots and put all six on the same plot. (So, I want just one plot). Those densities can be either from the true, analytical distribution or they can be estimated from samples from the posterior distributions. For clarity, use a method to distinguish which is which. For example, you might consider the same color for the densities from the same prior and a different line type for the two different priors.

(c) Use data D1 and prior 3, find the predictive distribution for new observations. That is, using the posterior distribution, get the distribution of a new person from this population. To describe this predictive distribution, provide the mean, standard deviation, a 95% credible region and a density plot. These values can be obtained from either analytical methods, from sampling or a combination of analytical and sampling techniques. Don't forget to provide a description as to how you did this. (Either the formula or the sampling algorithm.)

(d) Please comment on the differences in the results obtained by using the different priors. To be more specific, try and imagine what kind of personal beliefs each of the three priors represent. So, each of the priors assumes that the “a prior” average height is 66 inches and then we are imagining that the person sees two different types of data. So, is there a person who might be describes as “not believing” the data when they see one set of data. So, perhaps that person might revise their estimate of the average height and perhaps have 95% credible region of the average which might not even contain any of the observed values. Also, note how their prior beliefs are reflected in the size of the posterior credible regions after seeing either of the two data sets. So, to answer this question, describe the beliefs of the three different people who would have each of the three different priors and state how they react to seeing each of the two data sets. (So, perhaps they were surprised by the data, they might have been suspicious of the data, or they might have seen the data as “confirming” their belief, etc.)

More specifically, describe when you might or might not consider using the different priors. When do you think that they are ”valid” or ”invalid” in terms of representing your beliefs.

3. (A basic, bare-bones, fixed effect meta analysis). Preamble: You are a scientist who is interested in the effect a class of drugs on the control of diabetes. One measure of diabetic control is called HbA1c. For diabetics who don’t have good control, the value of HbA1c is too high. It is hoped that the drug lowers the value of HbA1c. For the purpose of this question, we will make a series of simplifying assumptions. They may not be realistic, but these assumptions make the problem manageable for a homework problem. At the end of this question, there are some comments which would be relevant to applying this question to a real world application.

Assumptions: (i) It will be assumed that there is one, universal value, θ, which we are estimating. That is, the average amount HbA1c is lowered for all drugs in the class, for all dose levels, and for all populations. (ii) From a randomized trial i, one gets the average difference, $\bar{Y}_i$, and the standard error, S_i. The difference is between the value of the HbA1c at the end of the trial compared to the value at the beginning of the trial. The standard error is an estimate of the square root of the within trial variance of $\bar{Y}_i$. It will be assumed that the observed average difference, $\bar{Y}_i$, in trial i is conditionally normally distributed with mean θ and *variance* S_i^2 (given the value of θ and S_i). Note: here we are assuming that the sampled standard error is the true standard error. (So, measured without error). (iii) Also, before any experimental results are know, it is assumed that everyone’s prior distribution for θ is the vague, improper prior which is normal with mean zero and precision zero. (That is, μ_0 and τ_0 are both zero.)

Answer the following questions. Be sure to show the formula’s that you used, don’t just give numeric answers.

(a) You do a randomized clinical trial with 209 subjects and you get a value for $(\bar{Y}_1, S_1)$ of $(-1.82, .21)$. What is your (posterior) belief in θ which is the average amount that the drug lowers the value of HbA1c? (This should be expressed as a distribution.)

(b) After you ran your experiment, you learn that a colleague across the country has also run a clinical trial to measure the value of θ. Your colleague had 79 subjects and got a value for $(\bar{Y}_2, S_2)$ of $(-1.02, .28)$. Starting with the prior belief that you had from your own clinical trial, update your belief using the new data. (That is, use the information from the previous question as your prior distribution and then use this new information as data to get a new posterior distribution.) What is your new posterior belief on the values of θ?

(c) Now consider the problem from your colleagues point of view. When she first collects her data without seeing your data, what is her posterior belief for the value of θ? After she finds out about your data and she updates her belief, what is her new belief in the value of θ?

(d) A few months later, you go to a conference and find several other labs who have study the problem and ran clinical trials. You learn that they got values for $(n, \bar{Y}_i, S_i)$ of the following:

$(19, -1.9, 0.945)$, $(100, -2.00, 0.285)$, and $(20, -1.21, 0.545)$. (Note: n is the number of subjects for each trial.) Pooling all this information, what is your new belief in the value of θ? Provide the general formula for combining this information. (Note: since you can easily get the formula by googling up “fixed effect meta analysis”, there will be little weight to the actual formula. You will be mostly graded on explaining how the formula follows from the principles/formulas presented in the class material.)

Some closing notes. These notes are not part of the homework question. The data presented here is a subset of the data presented by Hirst et al, 2013, Diabetologia, 56:973-984. They considered a random effect meta analysis model. That model can be fit with some of the techniques that we will learn in a few weeks. For the purpose of this questions, the fixed effect model is simpler and highlights the connection between a simple Bayesian model and the main formula used in meta analysis. In reality, the random effect is probably more applicable in general. Otherwise, one is assuming that drug effect is constant over a wide range of conditions such as the different specific drugs in the drug family, the different doses used in a particular study, the different illness severity in the trial population as well as other trial population effects such as different gender mix, culture, social economic status, genetic mix. Also, as for assumptions used in this problem, the assumption that the trial standard error could be treated as the true standard error in the model is probably okay if there is a sufficiently large trial population. When doing a meta analysis, there are other considerations that one should consider such as publication bias. There have been a number of guidelines developed to perform a proper meta analysis study. If one is going to one for real, then one should consult the literature for these guidelines.