STAT 532 Assignment 4 -2015
Due: Friday, Sept. 25th (by 3:00 pm)

Show all work **neatly** and **in order** for full credit (I shouldn't have to search for pages). When plots are included they should be nicely scaled and supplement your discussion. In the body of the report, only include computer code that is explicitly asked for and output that is necessary to completely answer a question. Other well organized code and output can be included in the appendix so that I can check your work and provide comments if needed. I do not want to have to search through pages to find your answer, plots, or proof you did something. Homework assignments provide an opportunity to practice synthesizing information into reports as you will likely have to do for future jobs, so practice displaying things in a professional manner.

NOTE: I expect that you will not refer to the solutions of previous students or search for solutions on-line until after assignments are completed and handed in. If you feel you need a hint, please email me or post on D2L and then I can share the information with everyone. I would like to continue to give weight to homeworks for this class. If you have someone's binder from past years, I would prefer that you just give it back to him/her. Thanks!

Problem 1: Refer back to Problem 5 from Assignment 3. For parts (g)-(j) we used the analytical solution for the posterior distribution. Now, let's investigate the use of sampling to approximate the quantities obtained.

1. (5 pts) First, we will use the `rgamma()` function to obtain draws from the posterior distribution. Use those draws to approximate the quantities we are after (HW3 Problem 5 parts (h)-(j)). Try this for different numbers of draws and discuss what you see and learn from this. You should also be comparing to the analytical results from Assignment 3.

2. (8 pts) Suppose you don't know how to obtain draws from a canned function or calculate areas for the posterior distribution. Simulate draws from the posterior distribution using the grid approximation method. Play around with the coarseness/fineness of the grid, and use plots to show differences, and briefly discuss what you learned. Include a nicely organized and clear chunk of your code demonstrating how you implemented a grid approximation for sampling (I do not want your whole script).

3. (3 pts) Use plots and summary measures to compare the calculations in (1) to those using the draws from part (2). Briefly discuss the comparison.

4. (2 pts) Gelman suggests we only really need 100 independent draws to characterize a distribution. Do you agree with this statement? As always, justify your answer.

5. (3 pts) Using the draws from (1) or (2), display the posterior distributions of $\lambda^2/(1-\lambda)$ and $log(\lambda)$. Approximate the expected values and variances of these functions of λ. Considering all the work you did in doing transformations of random variables in 501/502, why should you be excited about this problem?

6. (3 pts) Derive the posterior predictive distribution $p(\tilde{y}|y)$ analytically using the posterior distribution and fake data you generated in Assignment 3.

7. (2 pts) What assumption(s) do you need to make about the future observations for this to be useful?

8. (2 pts) Obtain 10000 draws of $\tilde{y}$ from the posterior predictive distribution and plot them.

9. (2 pts) Graphically compare the approximate posterior predictive distribution obtained via simulation to the true posterior predictive distribution.

10. (5 pts) Compare the mean and variance of the posterior predictive distribution to the mean and variance of the posterior distribution. Discuss and explain why the results from the comparison make sense. Your audience should be another grad student who is *not* taking this class (you can even try it out on one if you would like - let me know if you do). You should try to tie your explanation to other things you should have covered in a non-Bayesian course regarding prediction of new responses.

11. (4 pts) Now, ignoring the observed data (or supposing you don't yet have any), obtain 10000 draws from the *prior* predictive distribution and plot them. Show a little work, and briefly compare the posterior and prior predictive distributions.

12. (4 pts) Now, think about generating multiple data sets like yours using the posterior predictive distribution. Draw 1000 different samples of size 20 from the posterior predictive distribution. Store them in a matrix. Show 9 of them in a 3×3 panel plot.

13. (5 pts) Construct an approximate posterior predictive distribution for the sample standard deviation. To do this, calculate the sample standard deviation for each one of your 1000 posterior predictive samples (use the `apply()` function!) and then plot them. Add a vertical line for sample standard deviation for your "observed" data (the data you generated in Assignment 3) and discuss how it compares.

14. (4 pts) Repeat the previous problem using the sample maximum.

Problem 2 (5 pts) In the first paragraph of Section 2.9, Gelman et al. provide an example about how to think about a weakly informative for regression models on the logarithmic or logistic scale. Convince me that you have convinced yourself that his example is correct and makes sense (show me some work and thoughts).

Problem 3 (5 pts) Assume $y|\theta$ has an exponential distribution with rate parameter θ, and that you want to use a $Gamma(\alpha, \beta)$ prior on θ.

1. (4 pts) Suppose you observe that $y \geq 100$, but never observed the actual value of y because the data are censored. What is the posterior distribution of θ given that $y \geq 100$, as a function of α and β? Figure out the posterior mean and posterior variance.

2. (3 pts) Now, suppose you find out that $y = 100$. What is the posterior distribution of θ and compare the posterior mean and variance to that from part (a).

3. (2 pts) Do these results surprise you given that we observe more information when we know $y = 100$, as opposed to just knowing $y \geq 100$?

Problem 4 (3 pts) Read the last paragraph on page 261 in Willful Ignorance. Briefly discuss how the information and opinions in the paragraph relate to your experiences with how statistical inference is used in research in other disciplines.