STAT 532 Midterm I - Fall 2015
Due: Monday, October 5 at 10:00 (beginning of class)

Work independently using only your brain, BDA3 (the class textbook), *Willful Ignorance* book, *your* notes/handouts, *your* homeworks, and R.

Show all work neatly (and in order). Plots should appear with the corresponding exercise, and only include computer code/output necessary to completely answer a question (i.e. before including R-code ask yourself - would I want to see this if I was grading this?). Use **11 pt font** and you can single space your answers to save paper. If you have questions, please email me or talk to me first before posting anything on D2L. (100 points total)

1. (4 pts) Suppose you are in a discussion with a researcher in another department (never take probability theory or mathematical statistics) who wants to use Bayesian methods to estimate a single parameter in their research. You ask what prior they are considering and they respond with "I'm going to use the non-informative prior." In one paragraph, respond politely to the researcher. You may include questions to the researcher in your response, but you should also do some explaining. (Max 1/2 page)

2. (4 pts) Consider the following statement from Gelman et al.:

   > If so few data are available that choice of "uninformative" prior matters, then we should put relevant information into the prior!

   Argue *for* this position. (Max 1/4 page)

3. (8 pts) A researcher in another discipline would like you to explain what a likelihood function really is and how it differs from a probability distribution. She has only taken a semester of Intro Stats and it was 8 years ago, though she uses multiple linear regression a lot in her own research. Part of her question stemmed from hearing that the natural logarithm of the likelihood function is often used instead of the likelihood function and she doesn't understand why this works. She also doesn't understand how a posterior distribution differs from a likelihood function, because in the Bayesian analyses she has seen the posterior mean was very close to the MLE, and reported intervals were also very similar. Be sure to address these two specific questions in your response. (Max 1 page).

4. (4 pts) The Pareto distribution can be used to model situations similar to the exponential or log-normal models. The probability density function is given by

$$p(y|\alpha, y_m) = \alpha \frac{y_m^\alpha}{y^{\alpha+1}} I(y \geq y_m).$$

   Show this distribution can be used as a conjugate prior distribution if $p(y|\theta) = Uniform(0, \theta)$.

5. (5 pts) Consider the following quote from *Willful Ignorance* (pg 369), and discuss how it might relate to the assumptions of exchangeability we will use in our models.

   > Mathematical probabilities and the statistics used to estimate them depend on willful ignorance. It is only by ignoring certain specific considerations pertaining to the event or statement in question that a mathematical probability becomes possible. This mathematical probability applies to individuals *en masse*, but must overlook some of what Keynes called the "vague though more important" individual circumstances. The vagueness of what must be ignored may preclude formulation of uncertainty as a numerical probability.

6. In the 1949, researchers did thorough surveys of a sample of 18 of the Galapogos islands and recorded all bird species detected for each island. Researchers returned to the same islands and repeated the surveys in 1959. With the data, they are first interested in estimating the mean number of birds going locally extinct on Galpogos islands over the decade. They realize not seeing the species in 1959 is not proof of extinction on that island, but they are willing to assume their surveys are thorough enough that this is a good surrogate. They also had other evidence from field observations that species were going extinct (in fact this was motivation to complete this study and analysis).

   (a) (3 pts) Choose a common distribution for the data generating model, and write out the pdf or pmf. Discuss the assumptions of the model relative to the information you have about the study. (You may proceed with using a model even if you are uncomfortable with assumptions, as long as you point this out and discuss it).

   (b) Consider the following five ways of choosing prior distributions and briefly *justify your specific choice within each*

      i. (2 pts) Choose a proper, conjugate prior distribution meant to reflect little prior knowledge in the parameter of interest (a "non-informative" prior)

      ii. (2 pts) Choose a proper, conjugate prior distribution meant to be weakly informative.

      iii. (2 pts) Choose an improper prior distribution, but try to relate it to the conjugate prior distribution.

      iv. (3 pts) Choose an informative prior using the conjugate distribution. You may ask me questions for this part (pretending I'm the researcher) or just state what you specific knowledge you are assuming to have in order to create the prior. If you ask me, include record of the questions you asked me and the answers I gave.

      v. (4 pts) Specify an informative prior by splitting the range of the parameter into intervals and assigning a probability to each interval (The prior should look like a step-function or histogram). You should try to capture the same prior information in this part as you did in part (v).

   (c) (1 pt) Plot all five priors together.

   (d) (2 pts) The observed data can be found in `Extinctions.csv`. Plot the raw data and plot the likelihood function.

   (e) (12 pts) Obtain the posterior distributions for each prior. Clearly label them by the (i) - (v) in part (b) and show your work for obtaining each one. Show relevant code for obtaining the posterior when using prior (v).

   (f) (4 pts) For each posterior, calculate and display 99% posterior intervals for the parameter of interest, as well as the probability the mean number of extinctions is greater than 5. Make it easy to compare results from the different priors.

   (g) (3 pts) Discuss sensitivity of the posterior distribution to the priors.

   (h) (4 pts) Using the prior from (v), approximate the posterior predictive distribution computationally (i.e. obtain draws to approximate it). You do not have to include the actual R-code, but write out the steps you used in the code to produce the distribution. Plot it.

   (i) (5 pts) Using the posterior predictive distribution, draw many samples of the same size as the original data set and calculate the ratio of the sample variance to the sample average for each one. Display in a histogram and overlay the ratio obtained from the observed data. Comment on the results and why this might be a useful check in this modeling situation.

7. Researchers are interested in estimating the probability of laying an egg (in a day) for Buff Orpington hens between the ages of 24 months and 30 months. They are specifically interested in whether they are justified in claiming that the probability is greater than or equal to 0.5 because this is something they can use to possibly help sell this breed of hen. They know there is individual heterogeneity in egg laying rates, but they need an overall rate to release to people considering buying such hens for egg production. They collect data on 60 individually housed hens fed the same food and cared for in the same way. They think of the study as three 2-week sessions of data collection because they will switch technicians collected data every two weeks. They record whether each hen lays an egg each day, but you are initially just given the total eggs laid in two weeks for each hen.

   (a) (3 pts) Write out a reasonable model for the data generating process. Briefly discuss assumptions and the degree to which you think they are reasonable. (You may proceed with using a model even if you are uncomfortable with assumptions, as long as you point this out and discuss it).

   (b) (3 pts) From pilot data and years of informal observations researchers believe that on average hens lay an egg every other day. This information was collected without knowledge of individual hens (they were just counting number of eggs laid divided by number of hens in the group). With this information, decide upon a prior to use. Justify your choice and plot it.

   (c) (3 pts) The data recorded by the researchers for the first 2 week period are available in `EggData1.csv`. Plot the data and then obtain the posterior distribution (or an approximation to it) using the prior in (b). Plot the likelihood, prior, and posterior distribution on the same plot.

   (d) (3 pts) The data recorded by the researchers for the second are available in `EggData2.csv`. Use the posterior distribution from (c) as the prior distribution for analyzing this second period of data. Do you think this prior is well justified? Why or why not? Plot the posterior on the same plot as the posterior from (c).

   (e) (3 pts) Now use the posterior for part (d) as the prior for analyzing the data available in `EggData3.csv`. Graphically show the progression of knowledge about the probability of laying an egg each day starting from the prior you chose in part (b).

   (f) (6 pts). Now, repeat parts (d) and (e), but instead of directly using the posterior distribution as the prior, use it to construct a "weaker" prior distribution. Justify your choices (but still pretend you have not seen the future data when constructing the prior – i.e. you construct the prior for use with `EggData2.csv` after doing the analysis with `EggData1.csv`, but before seeing the data from the second data collection period).

   (g) (2 pts) For part (f), Graphically show the progression of knowledge about the probability of laying an egg each day starting from the prior you chose in part (b).

   (h) (3 pts) The researchers talk to a statistician who suggests they should just combine all the data from the 3 time periods into a single data set (summing over all three periods) and use the original prior from part (b). Do this and compare the resulting posterior distribution to the other two obtained after all 3 periods of data collection. The researchers would like to know which analysis to report. Give them your opinion.

   (i) (4 pts) Now, assume in talking to the researchers that egg production can depend on how much the birds eat and whether they are molting. How would you account for such covariates in estimation of the probability? You do not have to actually do it, but write out a reasonable model as a starting point (priors included) and justify your priors as much as possible.