Math 457 Exam 2
Due March 16, 2018

The Rules: This is an open book, open notes exam. You may talk to me or other students about any homework problem (or problems) that are similar to any part of this problem. You may even talk to other students to determine which homework problem(s) you think this exam is like. However, you may not speak about any specifics of this problem, including how to modify analyses you may have already done to fit the specifics of this problem.

The Problem: With this email, you will have received a datafile called something like Exam2dataA.dat (but the "A" will be a number). All people in the class have the same problem, but the datasets are different. The file contains the heights (in inches) of a random sample of 1500 individuals from some species. Height in this species is primarily governed by one gene with two possible alleles, but there is also a certain amount of random variation. More specifically, if Y_i is the height of individual i, then

$$Y_i \sim \begin{cases} ccN(\theta_{AA}, \sigma^2_{AA}), & \text{if } i\text{'s genotype is AA} \\ N(\theta_{Aa}, \sigma^2_{Aa}), & \text{if } i\text{'s genotype is Aa} \\ N(\theta_{aa}, \sigma^2_{aa}), & \text{if } i\text{'s genotype is aa} \end{cases}$$

You may assume that, if p_A is the population frequency of the A allele, then the genotype frequencies are given in the following table:

Genotype	AA	Aa	aa
Relative Frequency	p_A^2	$2p_A(1-p_A)$	$(1-p_A)^2$

Your task is to analyze the data and write up a report. Your report should be written as a nice narrative, and should include:

- A clear description of the model you ran, and why you made the model choices you did. This includes explaining what your prior distributions were, and how you chose parameters for those prior distributions.
- An explanation of how you ran the model. For example, if you did Gibbs sampling, I want to see a derivation of the full conditional distributions you used. You should attach a copy of your code as an appendix to your report. Note: Your write-up should be clear enough that I could re-create your analysis even if you didn't attach your code.
- A section in which you convince the reader that your Gibbs sampler (or other MCMC program) ran long enough, if applicable.
- A section in which you convince me that your results are reasonable. One way to do this would be something like getting a sample from the posterior predictive distribution and see how it compares to the actual data.
- Answers to the following questions:
  - What are the mean heights and standard deviation of heights for the three genotypes? Maybe give some kind of confidence intervals for these.

  - What is the (relative) frequency of the A allele in the population? What are the frequencies of the three genotypes?
  - Individual 100 has a height of 51.83806 inches. What is the probability she has the AA genotype? the Aa genotype? the aa genotype?
  - If AA is the genotype that yields the largest heights (on average), and aa yields the lowest heights (on average), which genotype has the highest variance of its genotypes? What probability would you attach to your response? (For example, you might say something to the effect that there is an 88% probability that σ^2_{AA} is the largest of the three variances.)

- At least 4 carefully chosen plots to illustrate your analysis.