Sta 360/601: Lab 8

In last week's lab some of you noticed that a log-normal distribution did not seem to fit the data very well. This lab explores one of the reasons for that.

Suppose that after talking to the EPA again, they give you a new data set which again contains 100 air pollution measurements, but this time also includes the date of each measurement. Your helpful assistant labels each measurement with the day of the week, giving you the data in the file data.txt. Here are the first few lines:

```
         pm       day
3 0.6480816 Wednesday
4 0.5406124  Thursday
5 0.6196524    Friday
6 0.3206031  Saturday
7 0.3183529    Sunday
8 0.4910938    Monday
```

You realize that since there are a lot more people driving, factories working, etc. on weekdays, perhaps it makes more sense to fit two separate lognormal distributions: one for weekdays, and one for weekends. You could go full-blown time-series analysis, but let's not do that today.

1. Write down a model where the pollution concentration y is Log-Normal with parameters (μ_1, σ_1^2) for weekdays, and (μ_2, σ_2^2) for weekends. Make sure you explicitly give your priors. **Your model should have prior probability 1 that $\mu_1 > \mu_2$.**

2. Using Gibbs sampling and/or Metropolis-Hasting, obtain at least 10,000 (post-burnin) draws from the posterior for your model. Provide traceplots for μ_1 and σ_1^2 showing that the sampler has converged.

3. Provide posterior point estimates and 95% confidence intervals for each of the four parameters.

4. What is the posterior probability that $\mu_1 > \mu_2$? What is the posterior probability that $\sigma_1 > \sigma_2$?

5. What is the posterior probability that the pollution level on a randomly chosen future Tuesday is higher than the pollution level on a randomly chosen future Saturday?