# Lab 3

Stephen R. Proulx

## Bayesian Statistical Modeling Winter 2022

## Lab Exercise, Week 3

*When is this lab due?* Labs are due on the Thursday after they are assigned. However, in many cases you can complete them during the lab period itself. This assignment is due on Thursday, 2/1/2024. Submit Rmd and pdf to gradescope.

### Medicago height data

We'll use publicly available data from: https://figshare.com/articles/dataset/Medicago_truncatula_plant_height_data/8018873 The data we use are heights of the plant *Medicago truncatula*. The experiment in question was looking at the effect of plants on caterpillars, but we will just use the plant height data. It comes in multiple replicate observations, and we'll fit the mean and variance, using a normal distribution for the likelihood, for each replicate separately.

Load the data. The code below is written assuming this Rmd file is saved in the same folder as the data file *MedicagoHeights.Rds*. You can copy the MedicagoHeights data to the same folder as your Rmd file. (Or, alternatively, you can give the path to the data in the load function.)

```
load("MedicagoHeights.Rds")
view(plant.heights)
```

### Visualize and summarize the data

Plot histograms or density plots of height for each replicate and summarize the means and standard deviations of each replicate. You may use the `filter` function to pick data from specific replicates, or add `fill=Replicate` to the `aes` directive in ggplot. like this:

```
ggplot(  plant.heights    ,aes(x=height,fill=Replicate))  +
  geom_density(alpha=0.5)
```

### Summarize the replicates

Summaries can be done using dplyr or by first subsetting the data to get each replicate and then using precis or mean and variance commands.

```
plant.heights %>%
  group_by(Replicate) %>%
  summarise(mean=mean(height),  sd=sd(height))
```

### Fit the different replicates and see how they compare

We'll use a grid approximation to fit the means and standard deviations for each replicate separately. To get you started we'll define the log likelihood function. It requires that the height data be in a dataframe called `dataset` and a column called `height`.

```
height_loglik_f <- function(mu_input,sigma_input){
  sum(dnorm(
  dataset$height ,
  mean=mu_input,
  sd=sigma_input,
  log=TRUE ))
}


height_loglik_f_vec <- Vectorize(height_loglik_f)
```

Now define a grid of $\mu$ and $\sigma$ values and then go through the same steps that we did in class to calculate the posterior probability for each $\mu$ and $\sigma$ pair. Do this seperately for each replicate in the dataset. In order to use the first replicate you would do:

```
dataset <- filter(plant.heights, Replicate == "Rep1")  # dataset is replicate 1
```

### Quantify the posterior distribution

Use the methods you have available to quantify the posterior distribution.

### Posterior simulations

Use the fitted model from replicate 1 to create a posterior simulation. This means that you will sample the values of $\mu$ and $\sigma$, and then produce an observation of 98 plant heights (this is how many observations there are of each replicate), and find the mean of this simulated dataset. Simulate this 1,000 times, and compare the distribution of this mean to the actual mean of replicate 2. Describe what you see, is it what you expected?

### Extra credit

Do the same procedure for the other replicates. What can you say is similar or different about the replicates? Are there any anomalies?