251 Project

Derek Walton & Bryce Martin

Project idea: we want to compare the quality of the two closest malls to BYU: Provo City Center & University Parkway. We will be gathering sample ratings by randomly picking stores from each and getting their ratings off of google maps. Our parameters of interest are the average rating for the stores in each location (our 'populations'), as well as the standard deviation of the ratings. Another thing we could look into is any correlation between number of ratings and overall rating.

We'll likely model the ratings with a beta distribution (adjusting the 5 star to be '100%' and 1 star to be '0%') and model the count data with a normal distribution (taking the log to transform the data).

The rating data is easy to (manually) get from google maps, and we'll randomly pick stores based on the malls' respective websites with all their stores listed.

We hope that at the end of this analysis we will be able to determine the overall quality of both malls, how close the stores tend to be of similar quality, and possibly even if the the more 'niche' stores (stores with lower counts of ratings) have a different rating on average than the high rating count stores. This can help local residents make informed decisions on which mall to go to when they are looking for consistent quality or fewer people.

Data Prep

Given the output of the confidence intervals, we can not confidently say that the log(n) has any significant correlative effect on the ratings.

We are therefore skip trying to estimate that as a parameter and estimate only two things for each population: the true ratings for each and the standard deviation of those ratings.

Because there isn't a known posterior distribution for our rating data, we use a Monte Carlo approximation to analyze our data. Our prior is an uninformative uniform distribution because we wanted to ensure a prior that would least effect our posterior distribution. For our likelihood, we assume a good approximation may be the beta distribution with alpha = 3 and beta = 1.5, as we imagine the data is more left skewed. We acknowledge that the beta distribution doesn't allow for the endpoints, so we slightly change the data on the endpoints (5 star to

4.99, 1 star to 1.01). Since we aren't predicting how any individual would rate the store but the stores overall rating, we can confidently say that a given true store's rating is not equal to exactly 1 or 5 stars. This allows us to use the beta distribution for our likelihood.

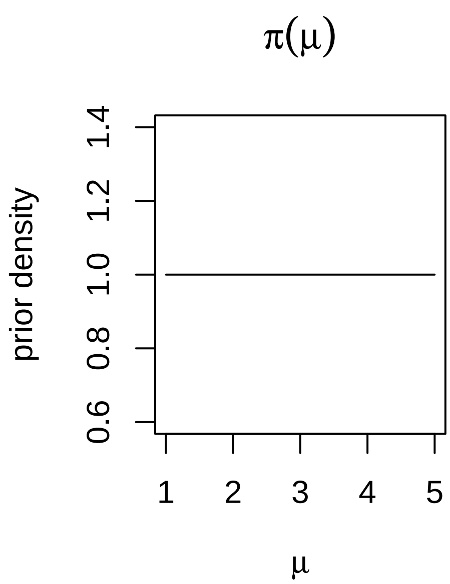
Our model for a single store would be:

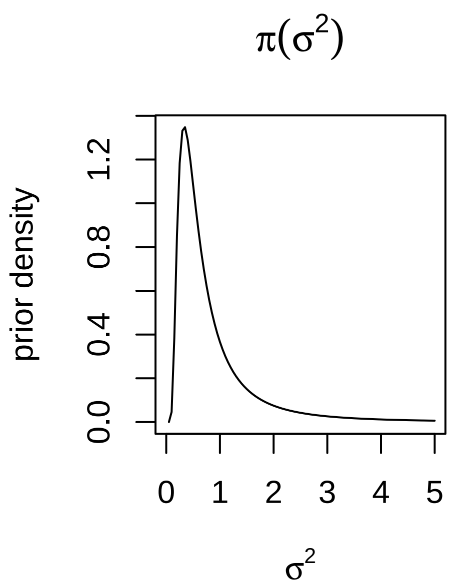
$$x_i = \text{Average rating for the ith store at a specific mall}$$

$$\text{Data} = x_1, x_2, ..., x_n$$

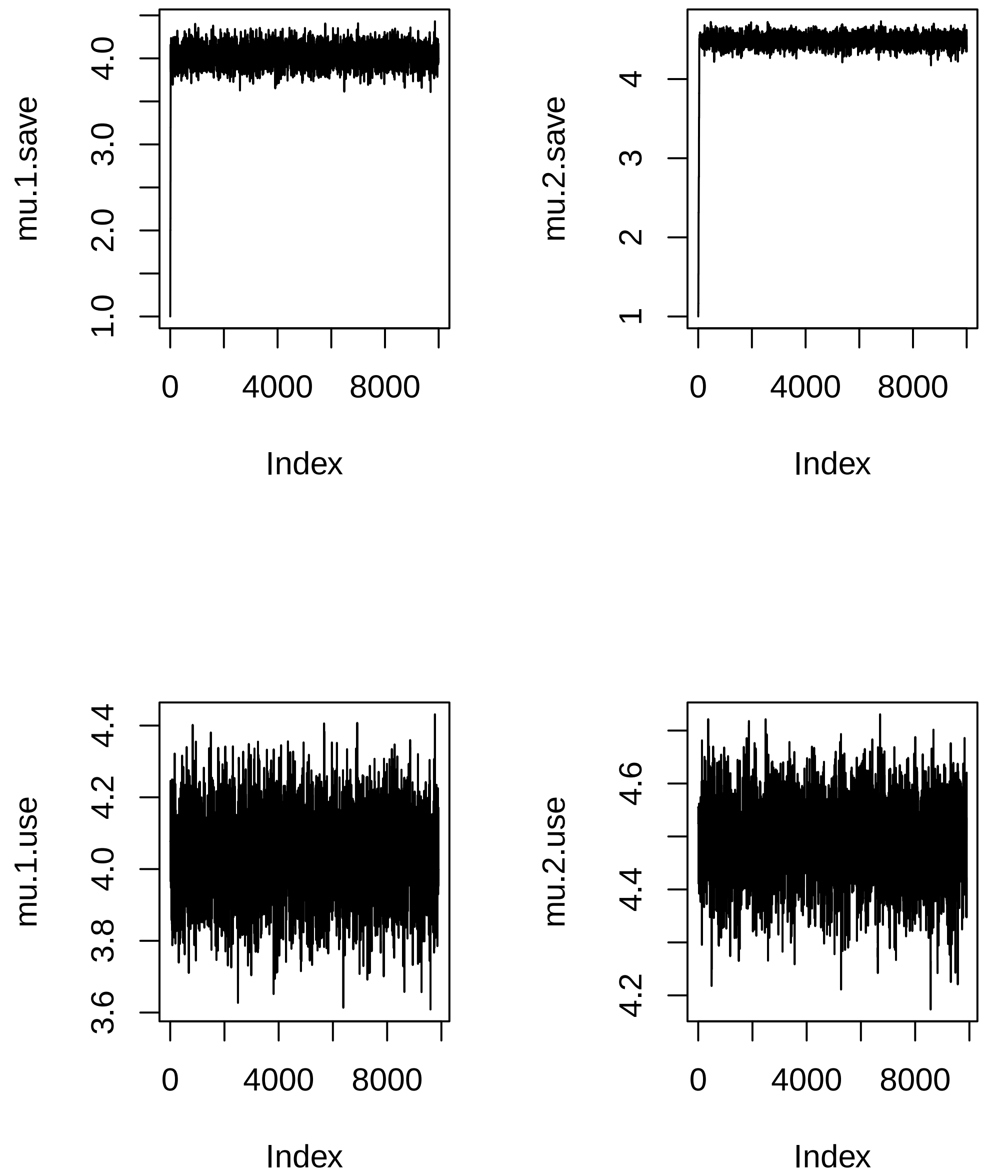
$$\text{Prior: } X \sim Unif(0, 1)$$

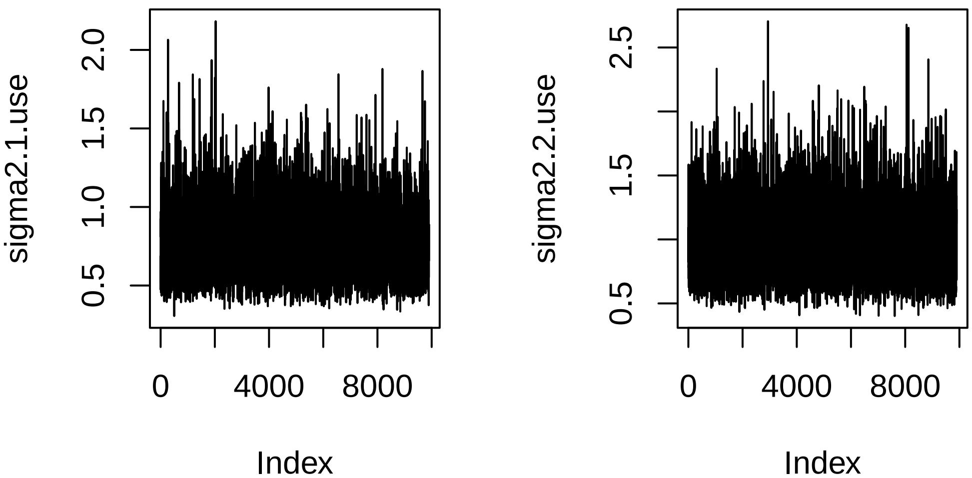
$$\text{Likelihood: } f(Data|\mu, \sigma^2) \sim Beta(3, 1.5)$$

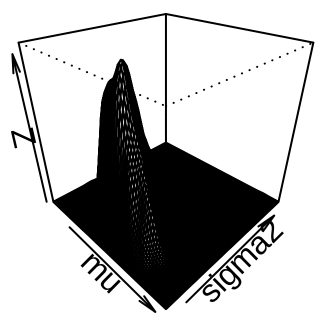


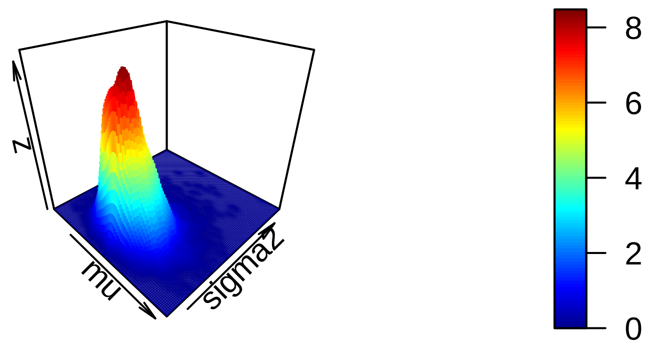


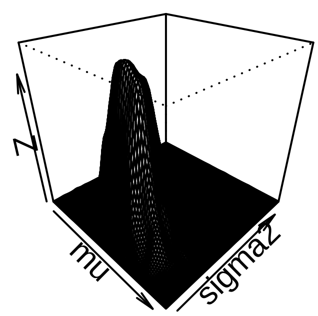
Because we're trying to estimate two parameters of interest, we will use Gibbs Sampling to estimate both the average rating as well as the variance of the ratings.

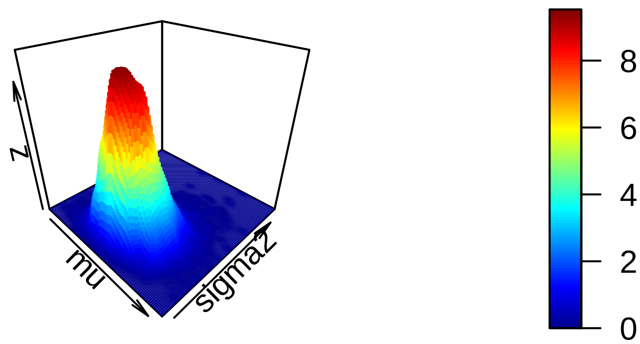

mu.1.save
1.0
2.0
3.0
4.0
0
4000
8000
Index
mu.2.save
1
2
3
4
0
4000
8000
Index
mu.1.use
3.6
3.8
4.0
4.2
4.4
0
4000
8000
Index
mu.2.use
4.2
4.4
4.6
0
4000
8000
Index

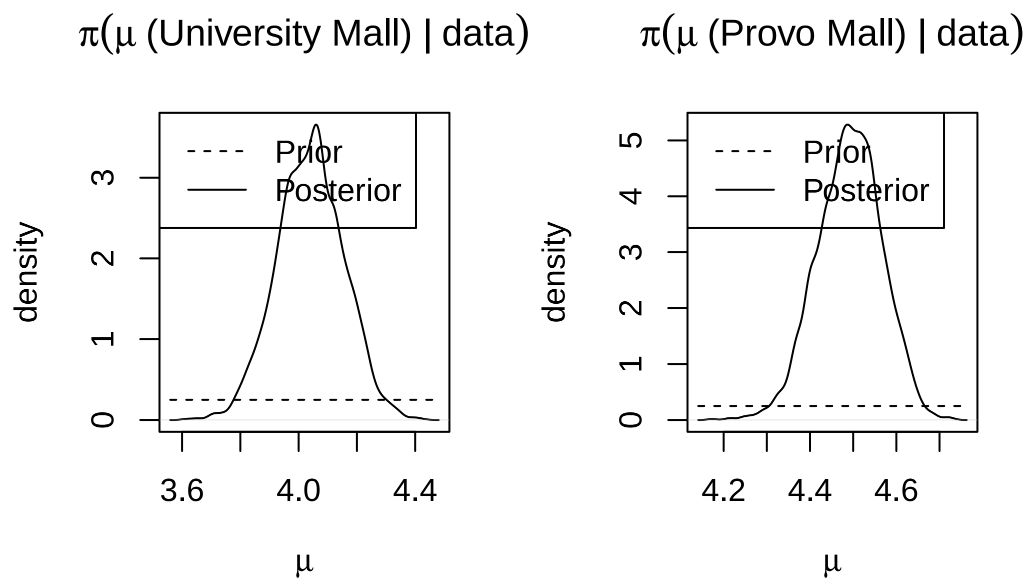

sigma2.1.use
0.5
1.0
1.5
2.0
0
4000
8000
Index
sigma2.2.use
0.5
1.5
2.5
0
4000
8000
Index


Z
mu
sigma2


z
mu
sigma2
8
6
4
2
0

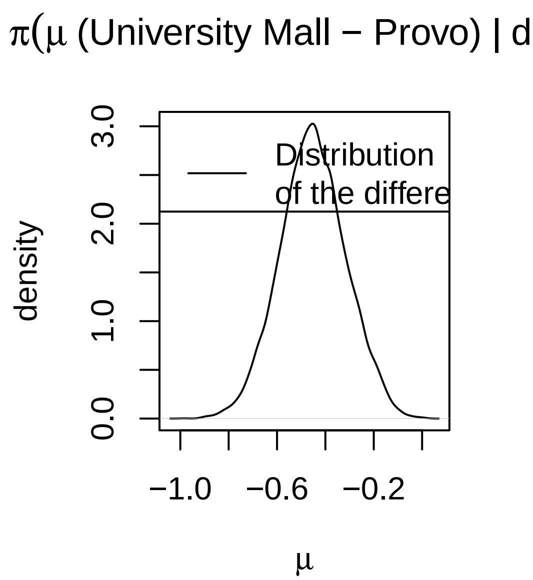

z
mu
sigma2

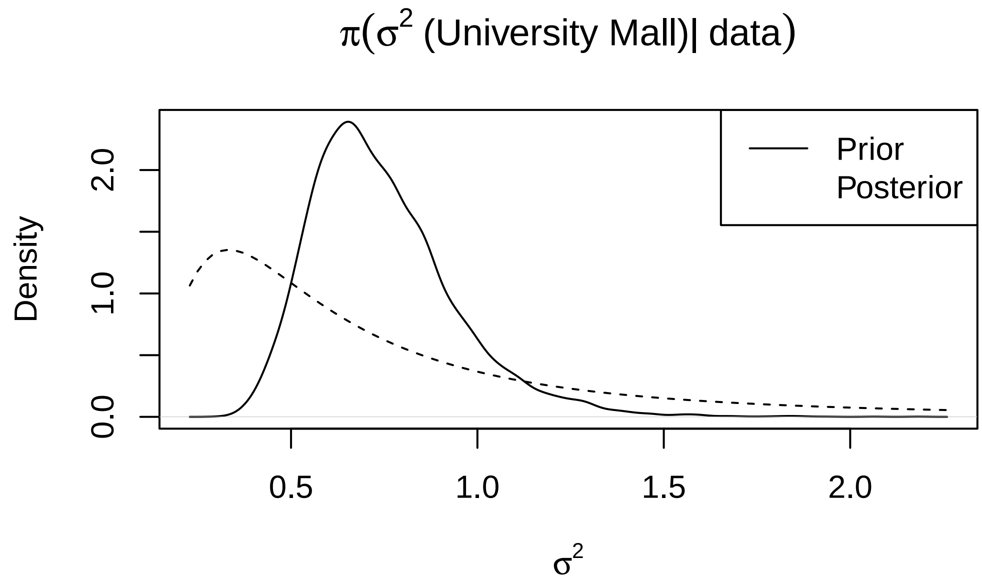


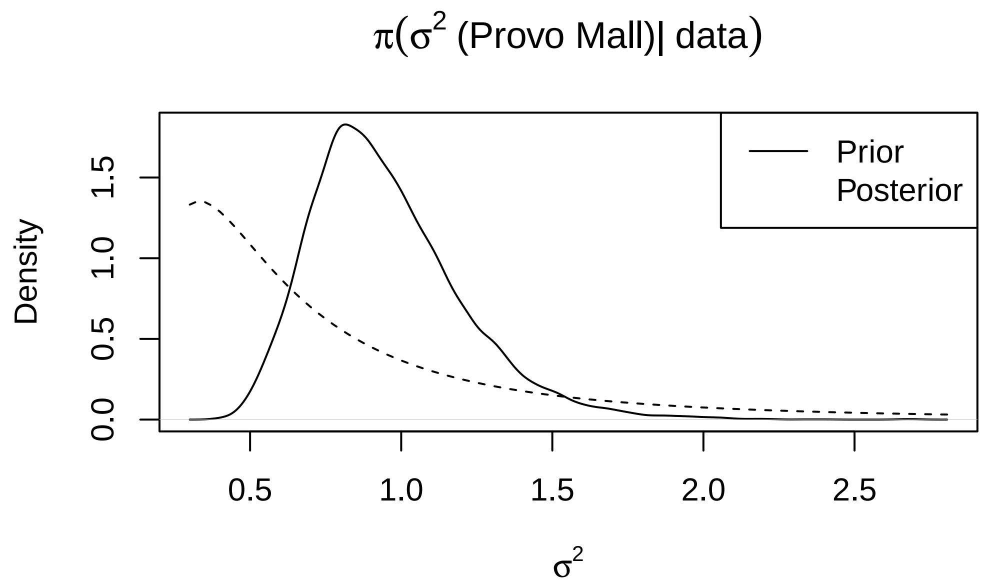



```
     2.5%     97.5%
3.813840 4.259602

     2.5%     97.5%
```

```
4.340887 4.629902

[1] 4.039997

[1] 4.491395

[1] 0.7451857

[1] 0.9517535
```








```
     2.5%     97.5%
0.4562967 1.2193081

       1%       99%
```

```
0.6501306 1.1560063

     2.5%     97.5%
0.5669308 1.5443336

       1%       99%
0.7238638 1.3134391
```

Appendix (all code)