251 Project

Derek Walton & Bryce Martin

Introduction

It's a Friday night in Provo and you want to go to a mall. But you realize there are two nearby. How do you decide which to visit? Both the Provo Towne Centre and University Place Mall seem to have a variety of stores and capabilities, making the decision even harder. Our research question thus follows: On average, which mall has better store ratings and which has more consistent ratings? Without having a specific store in mind, one way we can measure which mall is better is by exploring the average store ratings and the variance for the averages at each mall. That would be two parameters for each mall, totalling 4 parameters. This would tell us which mall, on average, is better, while also letting us know how consistent we can expect our experience across each mall to be. Additionally, we can also see if there is a relationship between the store ratings and number of ratings, helping us know whether having more ratings indicates an overall lower or higher average. This relationship being the slope in a linear regression would be our final, fifth parameter. We randomly selected 30 stores from each malls' website and used the ratings and number of ratings on Google Maps for each of the stores as our sample.

Methods

As there isn't a known posterior distribution for our average store rating data, we use a Monte Carlo approximation to analyze our data. Our prior is an uninformative uniform distribution with parameters 1 to 5 because we wanted to ensure a prior that would least effect our posterior distribution. This gives equal weight to all possible values. 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 instead the stores overall rating, we can confidently say that any 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.

For both malls, our models will be the same.

Our model for the relationship between rating and number of ratings is:

$$Y_i \sim \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\text{Where } \epsilon_i \sim N(\mu_1, \sigma_1^2)$$

Our model for the store averages would be:

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

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

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

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

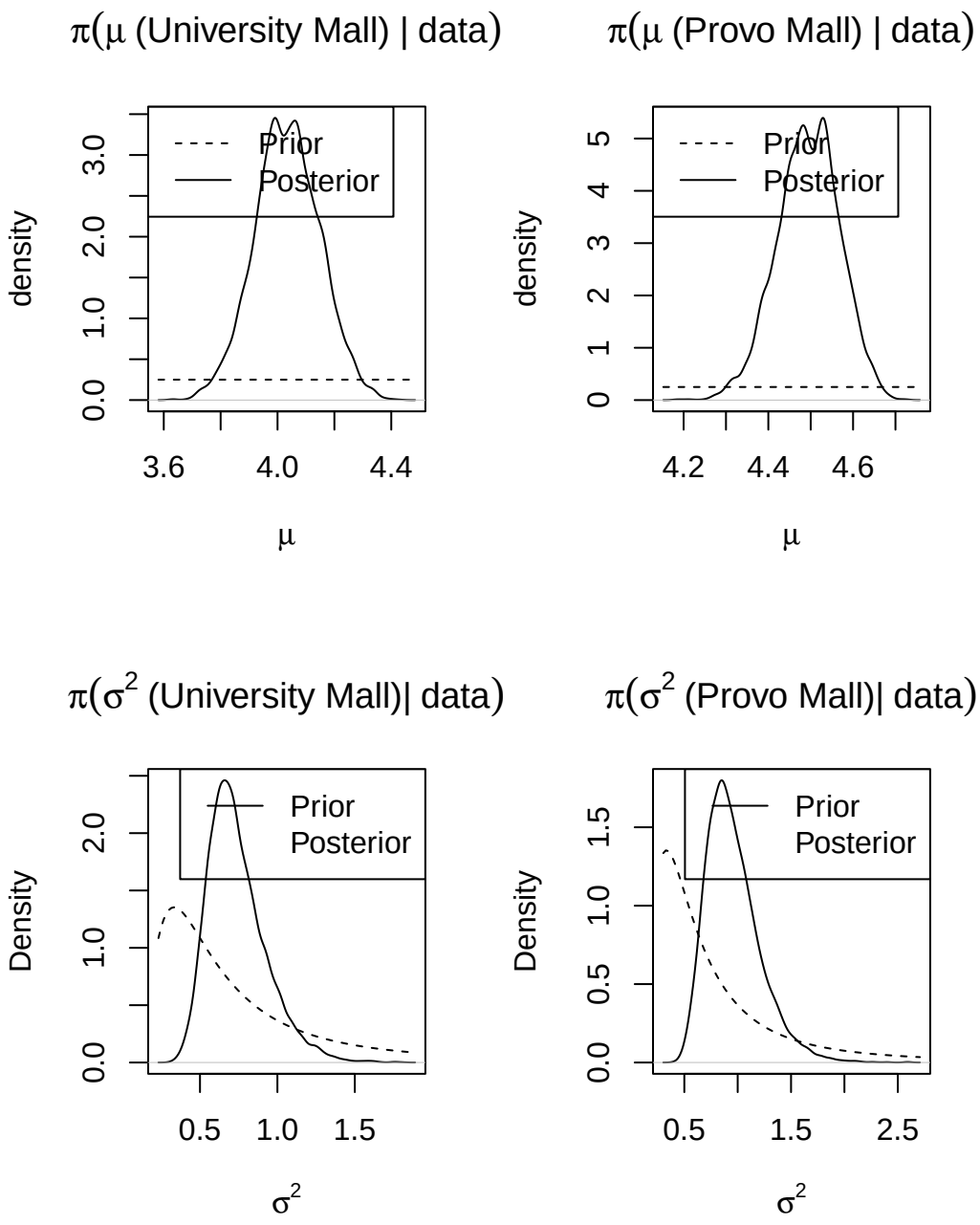
Data Prep

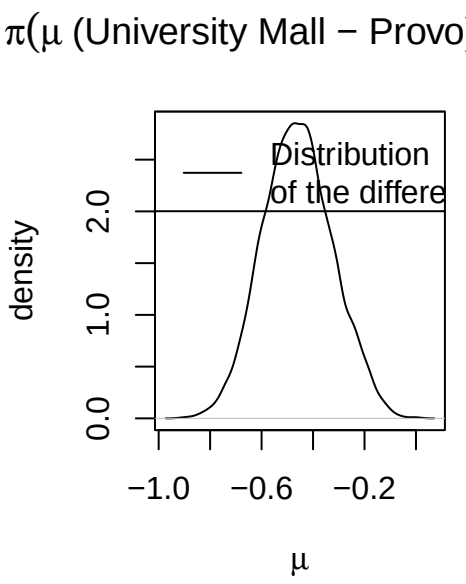
Given the output of the confidence intervals, we can not confidently say that the log(n) has any significant correlative effect on the ratings. Our 95% confidence interval of our beta value for log(N) is between -.56 and .70.

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 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.

We are not able to derive the posterior/full conditional distributions for the parameters we are estimating since there is no known distribution for a normal prior/beta likelihood or a double inverse gamma distribution. However we are able to plot them as a 3D graph, and not that there doesn't appear to be any significant correlation between the two for either populations.


π(μ (University Mall) | data)
π(μ (Provo Mall) | data)
Prior
Posterior
density
μ
π(σ² (University Mall)| data)
π(σ² (Provo Mall)| data)
Density
σ²



We are able to conclude that, since our 95% CIs of our average ratings don't overlap, they are statistically significant. The true average rating for University Mall is between 3.81 and 4.25, and the true average rating for Provo Towne Centre Mall is between 4.34 and 4.63.

However, we aren't able to conclude the same for our variance, since our 95% CIs of our average ratings overlap. The true average variation of ratings for University Mall is between .455 and 1.20, and the true average variation of ratings for Provo Towne Centre Mall is between 0.581 and 1.60.

Conclusions

It's decided. If you have no particular store destination in mind, then the better better is to spend your Friday night at the Provo Towne Centre Mall. You maybe be concerned that their stores may not be as consistently well rated, however, there is no significant difference of ratings given our prior assumptions and data collected.

Before we collected our data, we assumed a very uninformative prior given that we didn't know how stores would be rated in general (hospitals tend to rate super low, as compared to Chick-fil-a…). Using the uniform distribution allowed our posterior density to be nearly completely dependent on our data. Because of the aforementioned problems with the Beta distribution being non-inclusive on the extremes, our data is admittedly very slightly doctored. However, our research of different distributions didn't come up with a better alternative, so it's the best that we got.

Additionally, while the log(N) didn't have any initial significance in our linear regression, given some more data and perhaps some prior knowledge of how stores encourage people to rate, we may be able to discover if there is a more complicated underlying relationship between them all.

Appendix (all code)