

# Bayesian Analysis of New York City's Uber Rides Data

David Arthur, Krishnan Raman

12/10/2020

# Introduction

Uber is a popular ridesharing platform that services millions of customers each day. Uber drivers are interested in making the most money. This involves answering the following questions:

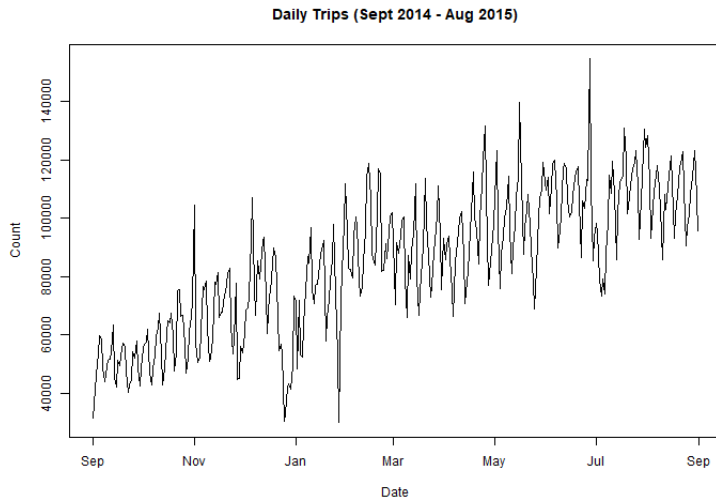
- ▶ What times of the year and week are drivers most needed?
- ▶ What times of day are the busiest and does this vary by location?
- ▶ For an optimal route, how much can a driver expect to make?
- ▶ Is the optimal route really that much better than a random route?

## Two Datasets

	orig	dest	pickup_datetime	distance	duration	fare
1	7C	6A	2014-09-01 09:00:00	4.25	15.183	15.30
2	7B	15	2014-09-01 18:00:00	10.17	34.083	32.28
3	11	2A	2014-09-01 17:00:00	4.02	17.100	15.57
4	3B	4A	2014-09-01 13:00:00	1.46	6.533	8.00
5	2A	10	2014-09-01 14:00:00	8.31	26.283	26.29
6	5B	4C	2014-09-01 12:00:00	1.04	8.583	8.00

	Date.Time	Lat	Lon	Base
1	9/1/2014 0:01:00	40.2201	-74.0021	B02512
2	9/1/2014 0:01:00	40.7500	-74.0027	B02512
3	9/1/2014 0:03:00	40.7559	-73.9864	B02512
4	9/1/2014 0:06:00	40.7450	-73.9889	B02512
5	9/1/2014 0:11:00	40.8145	-73.9444	B02512
6	9/1/2014 0:12:00	40.6735	-73.9918	B02512

# First Question: The Data



## First Question: The Model

The model for the number of trips on day  $t$  is:

$$y(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + f_1(t) + f_2(t) + \epsilon(t) \text{ where } \epsilon(t) \sim \mathcal{N}(0, \sigma^2)$$

This can be more compactly written in matrix/vector notation as:

$$Y = X\beta + f_1 + f_2 + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Here,  $f_1$  and  $f_2$  are Gaussian Processes that represent a smooth, yearly trend and a shorter, periodic trend respectively. The  $X\beta$  is a structural component that accounts for an increasing trend over time.

## First Question: Gaussian Process

We assume that:

$$f_1 \sim \mathcal{N}(0, \tau_1^2 K_1), \quad f_2 \sim \mathcal{N}(0, \tau_2^2 K_2)$$

$$K_1(t, t') = \exp\left(-\frac{|t - t'|^2}{(2)(7^2)}\right), \quad K_2(t, t') = \exp\left(-\frac{-2 \sin^2(\pi|t - t'|/7)}{7^2}\right)$$

$$\pi(\beta) \propto 1$$

$$\sigma^2, \tau_1^2 \sim \text{InverseGamma}(3, 5)$$

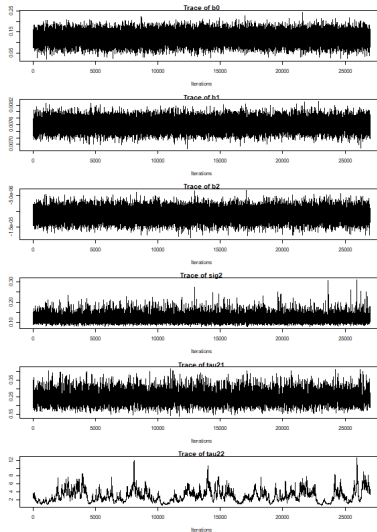
$$\tau_2^2 \sim \text{Gamma}(1.5, 0.2)$$

A Gibbs Sampler with an MH-step for  $\tau_2^2$  is implemented to sample from the posterior.

## First Question: Model Checking

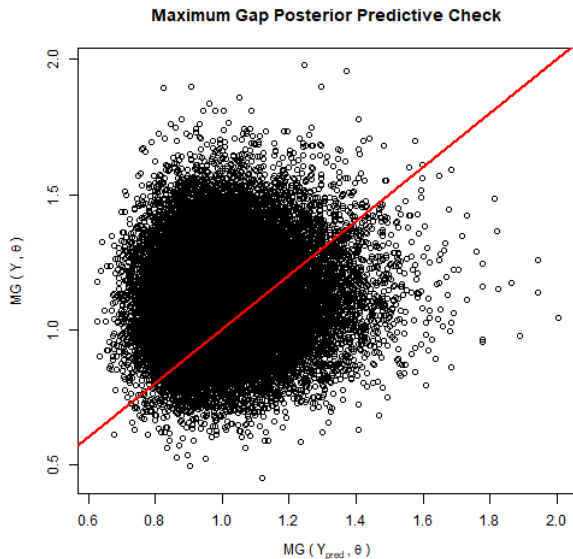
	GR Estimate	GR Upper CI	ESS
$\beta_0$	1.0001	1.0006	26641.7164
$\beta_1$	1.0001	1.0005	23901.4395
$\beta_2$	1.0002	1.0007	36927.7226
$\sigma^2$	0.9999	0.9999	25271.5523
$\tau_1^2$	1.0004	1.0018	11983.7929
$\tau_2^2$	1.0172	1.0278	98.0828

# First Question: Model Checking

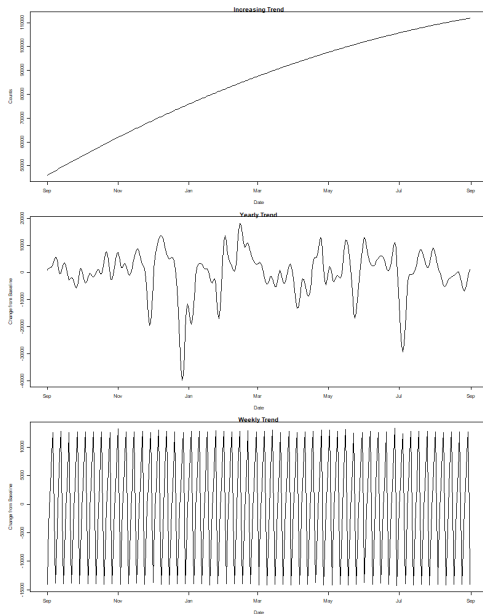




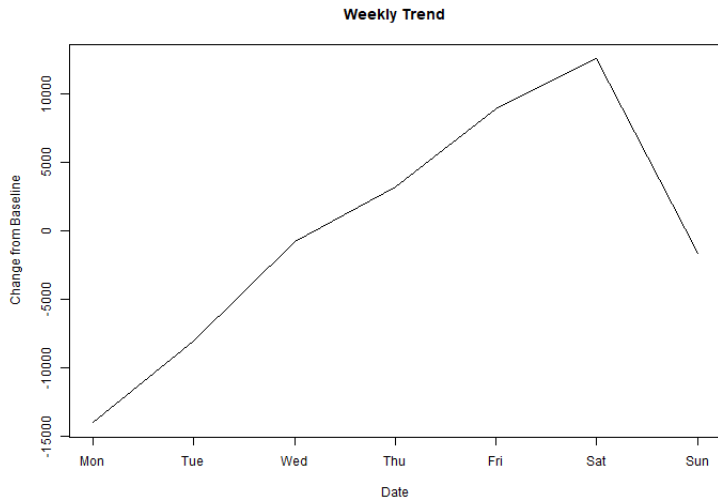
# First Question: Posterior Predictive Check



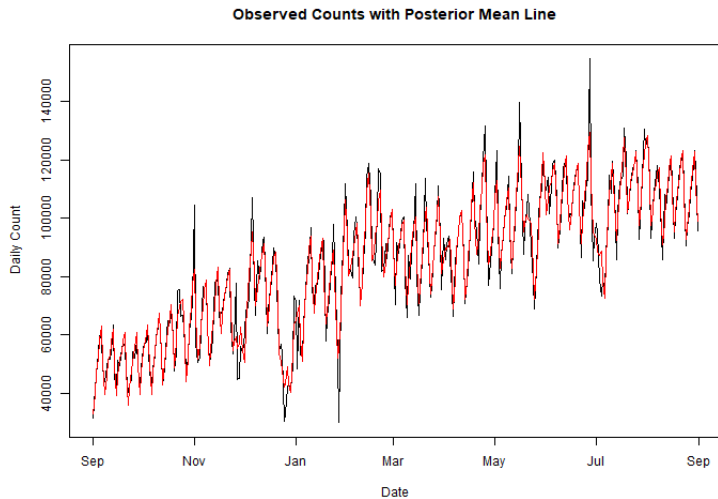
# First Question: Posterior Estimates of Trends



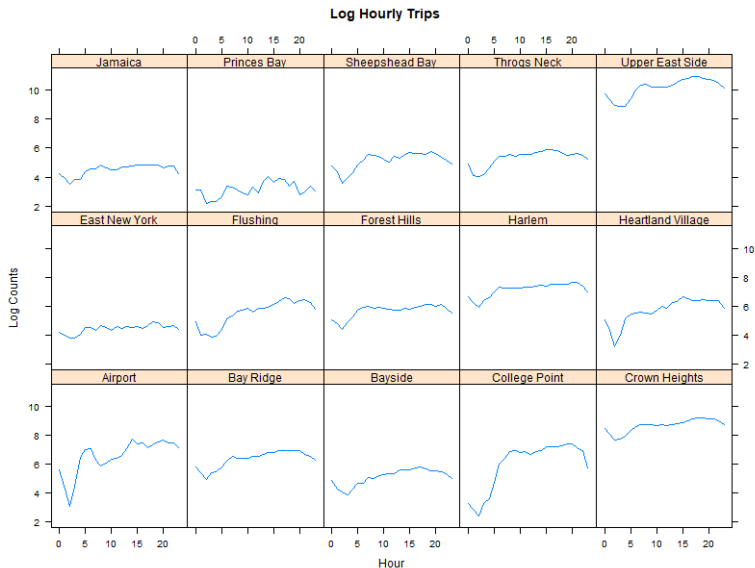
## First Question: Weekly Trend



# First Question: Predictions



## Second Question: The Data



## Second Question: The Model

The model for number of trips at location  $s$  at time  $t$  is:

$$y(s, t) = f(s, t) + \epsilon(s, t) \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Or more compactly:

$$Y = f + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Here,  $f$  just represents a spatio-temporal Gaussian Process that allows us to understand trends in trips across space and time simultaneously instead of considering them separately.

## Question 2: Spatio-Temporal Gaussian Process

We assume that:

$$f \sim \mathcal{N}(0, K)$$

$$K(s, s', t, t') = \tau_s^2 \exp\left(-\frac{\|s - s'\|_2^2}{(2)(0.005^2)}\right) \times \tau_t^2 \exp\left(-\frac{|t - t'|^2}{(2)(12^2)}\right)$$

$$\sigma^2, \tau_s^2, \tau_t^2 \sim \text{InverseGamma}(3, 5)$$

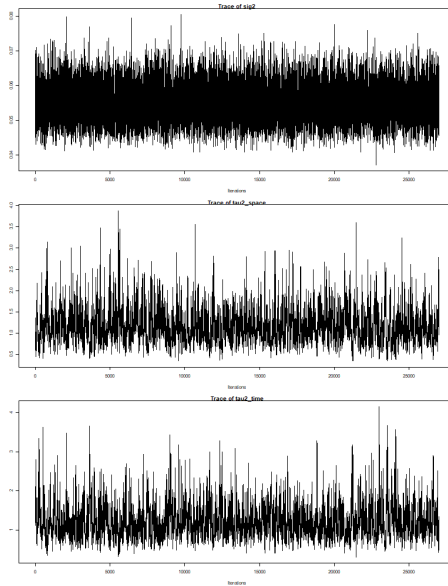
A Gibbs Sampler is used to sample from the posterior distribution.

## Second Question: Model Checking

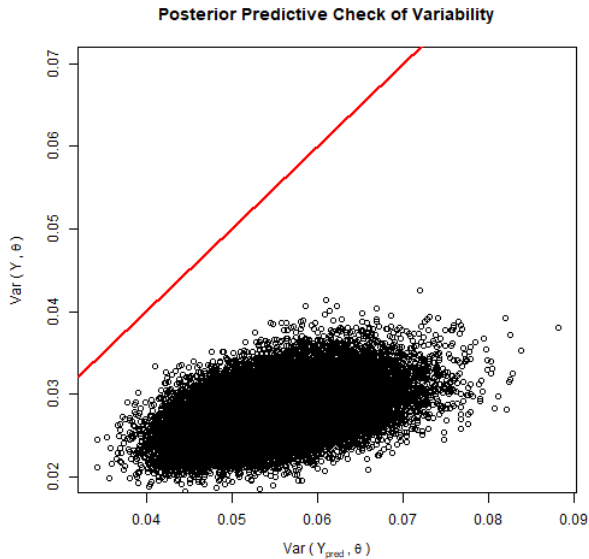
	Point est.	Upper C.I.	ESS
$\sigma^2$	1.0000	1.0001	16561.97
$\tau_s^2$	1.0200	1.0571	2727.01
$\tau_t^2$	1.0193	1.0517	1115.583



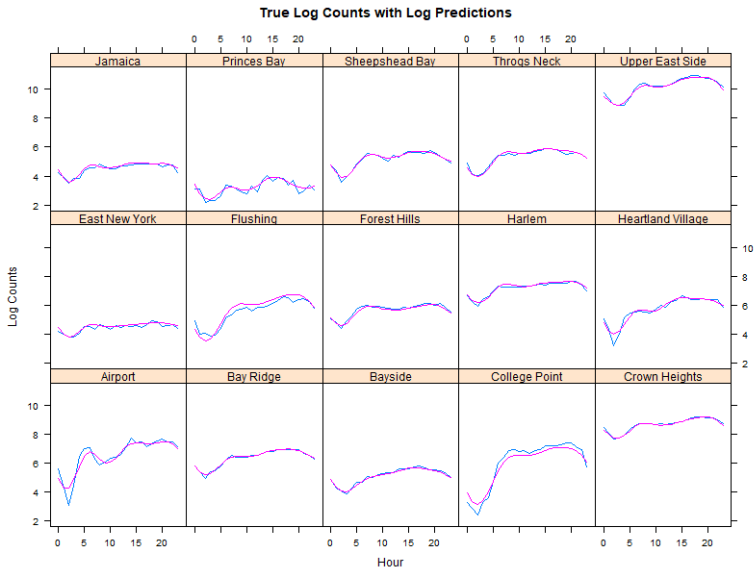
## Second Question: Model Checking



## Second Question: Posterior Predictive Check

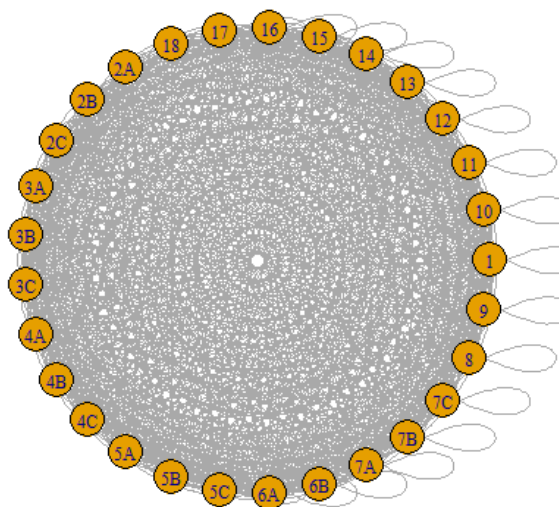


## Second Question: Predictions



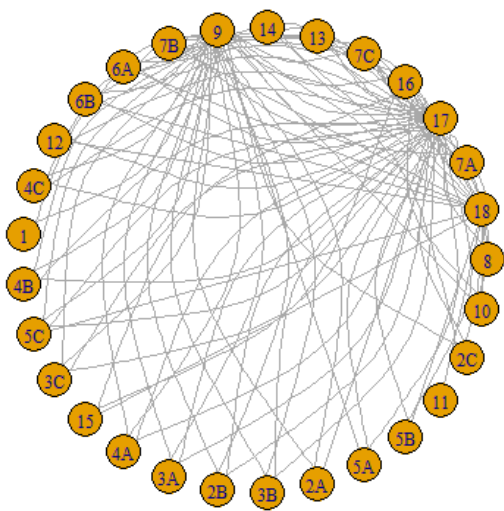
### Third Question: Uber Graphs & Subgraphs

**Uber as a graph, with 29 vertices, 812 edges.**



### Third Question: Graph of Most Lucrative Edges

**Uber top-100 subgraph, Vertices: 29 Edges: 100**



### Third Question: What's an Optimal Circuit?

Some criterion for the optimal circuit:

- ▶ Circuit: Collection of edges
- ▶ Want a closed circuit aka Cycle or loop
- ▶ Want a “simple path” : each node visited only once

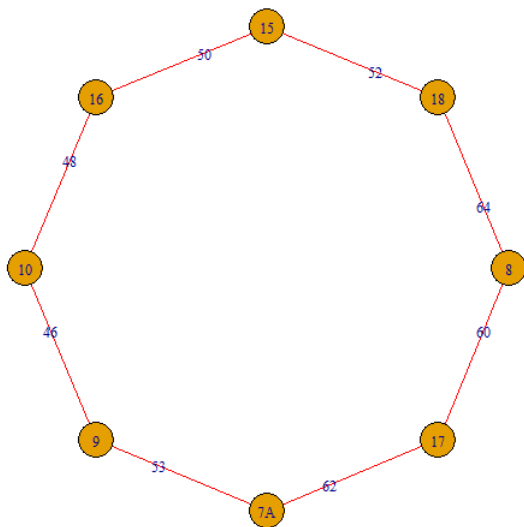
We want to find:

- ▶ Circuit that makes the most money for the Uber Driver

We are assuming:

- ▶ Driver works 9AM-5PM: 8 HOURS
- ▶ Make 1 Uber trip per hour
- ▶ Want an 8-cycle aka Simple Path with 8 Nodes aka 8-gon

### Third Question: The Optimal Circuit



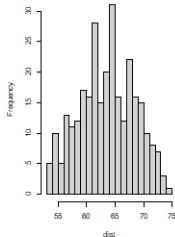
### Third Question: Summary Statistics for Optimal Circuit

Src	Dest	Mean	Median	Sigma	Var
8	18	63.54	63.69	4.90	23.97
18	15	53.06	52.44	6.71	45.06
15	16	50.18	49.72	8.65	74.84
16	10	47.70	47.86	7.69	59.07
10	9	47.20	46.39	7.40	54.79
9	7A	54.00	53.42	4.66	21.70
7A	17	61.56	61.96	6.22	38.72
17	8	60.19	59.89	6.04	36.47

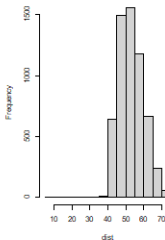


# Third Question: Fare Distributions for Each Edge

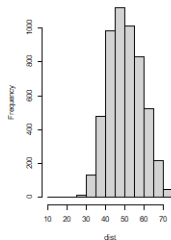
Edge 8 → 18



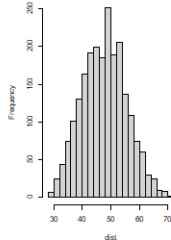
Edge 18 → 15



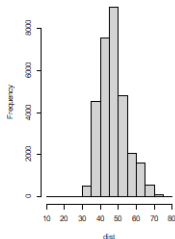
Edge 15 → 16



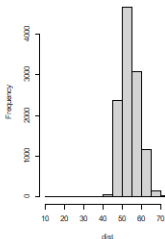
Edge 16 → 10



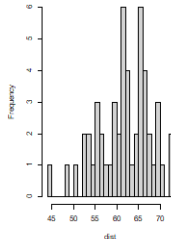
Edge 10 → 9



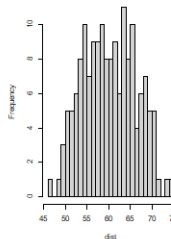
Edge 9 → 7A



Edge 7A → 17



Edge 17 → 8



## Third Question: Two Models

Edge Fare  $X_i$  = Observed fares for edge  $X_i$  over a year.

Circuit Fare  $Y = \sum X_i, i=1:8$ ,  $Y$  is unobserved.

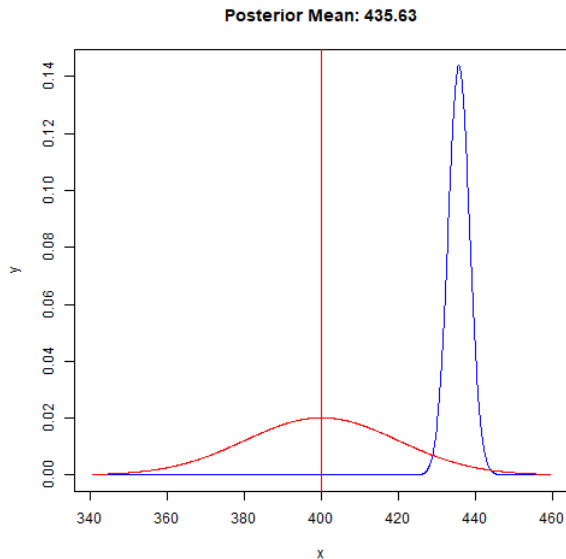
Model 1 : Gaussian Conjugate Prior

- ▶  $X_i$  approx Gaussian, sample mean \$50, sample sd \$7
- ▶ Prior on  $\mu$  is  $\mu \sim N(400, 392)$
- ▶ We attribute a Gaussian prior  $Y \sim N(50 * 8, 8 * (7^2))$
- ▶ Posterior Dist of  $Y$  is product of Prior & Likelihood
- ▶ Gaussian Conjugate prior yields Gaussian posterior

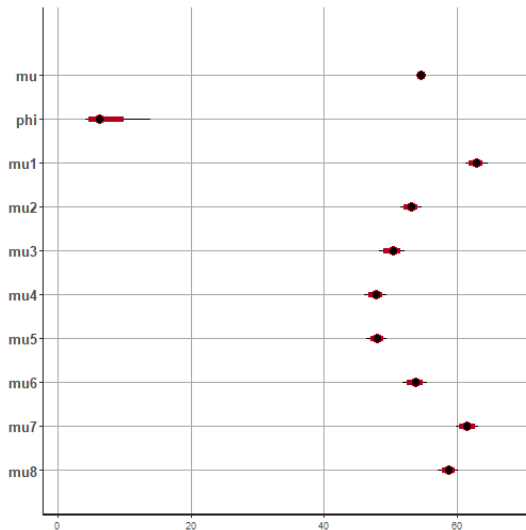
Model 2 : Hierarchical Model

- ▶  $X_i \sim N(\mu_i, \sigma)$
- ▶  $\mu_i \sim N(\mu, \phi)$
- ▶  $Y \sim N(8\mu, \sqrt{8}\sigma)$
- ▶ Estimate posterior mean of  $Y$  via MCMC

### Third Question: Model 1 Posterior of Mean Fare

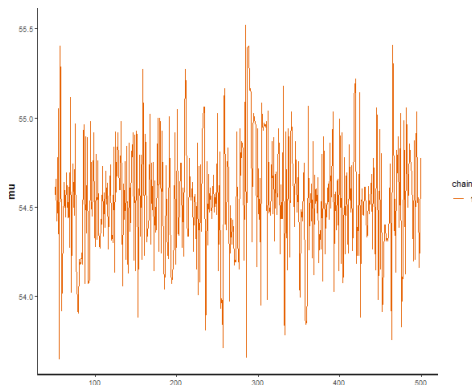


## Third Question: Model 2 Posterior Distributions

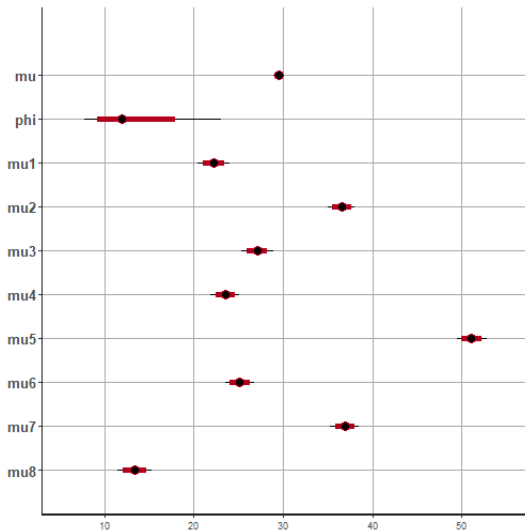


## Third Question: Model 2 Posterior of Mean Fare

Mean Fare from Optimal Daily Circuit: \$437.63



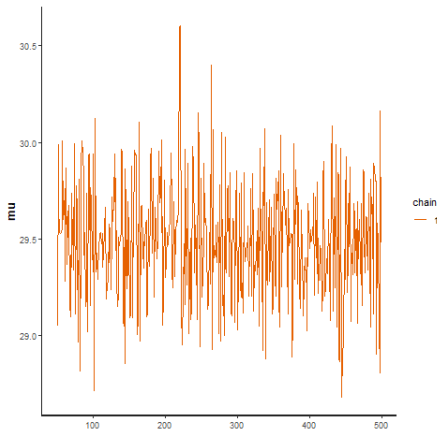
## Fourth Question: Random Circuit Posterior Distributions



## Fourth Question: Random Circuit Mean Fare Posterior

Mean Fare from Random Daily Circuit: \$235.82

|Optimal-Random| Median Loss: \$201.81

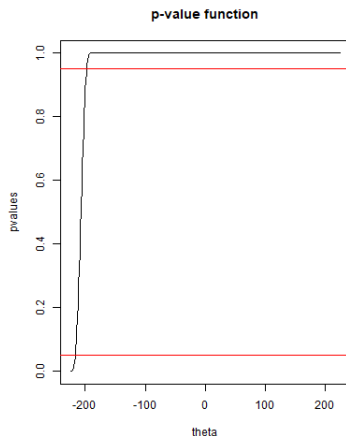
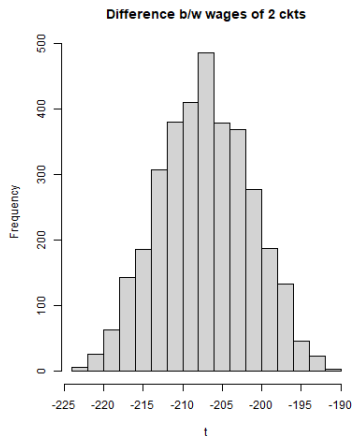


## Fourth Questions: Confidence Distributions

- ▶ Professor Don Rubin: “Fisher Randomization Test is a Stochastic Proof-By-Contradiction”
- ▶ FRT = distribution-free test to compare two (multimodal) Uber circuits
- ▶ Fisher Sharp Null compares individual potential outcomes  $Y_i(1)$  vs  $Y_i(0)$  for every observation.
- ▶ An assignment is a Boolean vector over two weeks.
- ▶ An assignment simply means on the given day, the Uber driver drove Circuit c2.
- ▶ A non-assignment means the wages would come from Ckt c1 aka Null Hypothesis.
- ▶ Compute the biweekly average and compare with the null hypothesis biweekly average
- ▶ This comparison is a simple difference test statistic.



## Fourth Question: Confidence Distributions



# Conclusions

What we confirmed:

- ▶ Holidays are not times of high demand
- ▶ Weekends are when most people need an Uber
- ▶ Central New York is the busiest
- ▶ Most places see a spike in demand in the morning and evening
- ▶ You'll make a lot more money driving the optimal circuit

Room for Improvement:

- ▶ Fix estimation problems with length-scale parameters
- ▶ Address model inadequacies found with posterior predictive checks