Bayesian Analysis of Auto Prices

Greg DeVore

August 6th, 2017

Overview

The purpose of this assignment is to explore the use of Bayesian analysis for hypothesis testing in statistics. The results will be compared to both classical and bootstrap methods. Bayesian analysis is unique with respect to classical (frequentist) methods because it creates a statistical model that can be easily updated and refined as new data is obtained. We start with a prior belief about the distribution of some parameter before any data is collected. Then, we take some measurements and compute the likelihood of observing those measurements using an assumed distribution. The combination of our prior belief and the evidence we collected is called the posterior distribution. The relationship can be summarized as

$$Posterior \propto Likelihood \times Prior$$

or, more formally

$$P(parameters|data) \propto P(data|parameters) \times P(parameters)$$

In other words, the probability of the parameter of interest given the data observed is proportional to the probability of the data observed given the parameter of interest multiplied by our prior belief about the parameter of interest. In our case, this formula will be applied to automobile data, where the parameter of interest is the price of the automobile, and the data are different attributes of each automobile. In particular, Bayesian analysis will be used to investigate the following:

1. Is there a significant difference between the price of an automobile when stratified both fuel type (gas vs. diesel) or aspiration (standard vs. turbo)?
2. Is there a significant different between the price of an automobile when stratified by body type (sedans, wagons, and hatchbacks)?

After Bayesian methods are applied, the following will be found:

1. A statistically significant relationship exists for the price of an automobile when stratified by aspiration (standard vs. turbo) but not by fuel type (gas vs. diesel).
2. A statistically significant relationship exists between the price of sedans and hatchbacks, but not between wagons and hatchbacks, or sedans and wagons.

In both bivariate and multivariate tests, Bayesian and classical methods produce identical results, while bootstrap methods often produce different results for the automobile data.

Data Preparation

As in the previous write-ups, we'll need to convert the automobile price to a numeric variable. Also, we'll want several of the explanatory variables to be factors, as this will be helpful later on. Finally, we'll remove any incomplete observations (rows containing one or more NA's).

```r
# Coerce some factors to numeric and retain only complete observations
auto$price <- as.numeric(auto$price)
auto$fuel.type <- as.factor(auto$fuel.type)
auto$aspiration <- as.factor(auto$aspiration)
auto$body.style <- as.factor(auto$body.style)
auto <- auto[complete.cases(auto),]
```

Additionally, in the previous write-ups, we determined that the log of the automobile price more closely resembled a normal distribution when compared to the price alone. We'll add it as a new column to the data.

```
# Add logPrice as additional column
auto$logPrice <- log(auto$price)
```

Also, we'll need to set some global parameters for our Bayesian analysis.

```
N = 1000 # Use N = 1000 for parameter range
nSampsQI = 100000 # Samples for credible interval calculation
qs = c(0.025, 0.975) # Vector for 95% qi
```

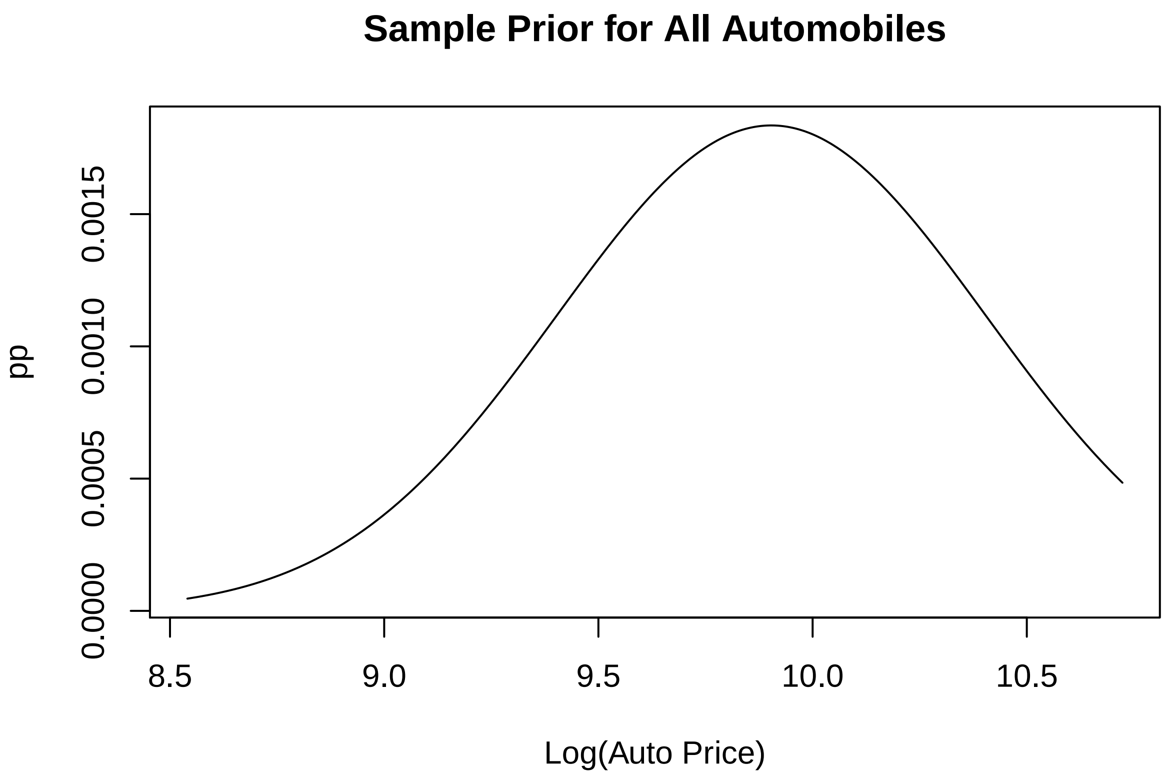
Finally, we'll need some summary statistics of the automobile data to use in our prior distribution. For our prior, we will guess that the average price of a car is $20,000. This initial guess is not based in any way on the actual automobile data. Rather, it's what we believe the average price of an automobile to be. Also, the last variable calculated is the range for our parameter of interest (automobile price) from the minimum price to the maximum price in the data set using N intervals.

```
# Calculate summary stats for all cars
auto_mean = log(20000)
auto_range = seq(min(auto$logPrice), max(auto$logPrice), length = N)
```

Prior Distribution

For all tests, our prior distribution will be a normal distribution with mean equal to the log of our guess for the average automobile price ($20,000), and standard deviation equal to that of the log of the automobile price of the particular sample being tested. As an example, the prior distribution using the standard deviation for all automobile prices is shown below. Because of their similarity (identical mean, slightly different standard deviation), the prior will not be shown for every test sample.

```
pp = dnorm(auto_range, mean = auto_mean, sd = sd(auto$logPrice))
pp = pp / sum(pp)
plot.dist(auto_range, pp, 'Sample Prior for All Automobiles', 'Log(Auto Price)')
```



Significance Tests For Two Samples

In this section, we'll use Bayesian methods to test whether or not there is a significant difference in automobile price when stratified by fuel type and aspiration, and compare the results to those obtained using both classical (Welch's t-test) and bootstrap methods. In all cases, the null hypothesis will be that there is no relationship between the price of an automobile when stratified by the given explanatory variable (the difference in the means of the two samples is zero). The alternative hypothesis will be that a relationship exists (the difference in the means of the two samples is not zero). Since the direction of the difference is not specified (greater than or less than), a two-tailed test will be used. Also, for all tests a confidence interval (or credible interval, in the case of Bayesian analysis) of 95% will be used. This means $\alpha = 0.05$, and the null hypothesis will be rejected in favor of the alternative hypothesis if the p-value is less than α. In the Bayes case, the null hypothesis will be rejected if the credible intervals for each sample overlap. In all plots, a solid red line indicates the mean of the distribution, and dotted red lines indicate the 95% confidence/credible interval for the mean.

Fuel Type

First, let's look at price stratified by fuel type. To do that, we'll create variables by splitting the price of automobiles by the two fuel types, gas and diesel. Also, we'll create the prior distributions for the price using the overall mean price and standard deviation from each sample.

```
# Create vectors of diesel and gas prices
diesel_price <- sort(filter(auto, fuel.type == 'diesel') %>% pull('logPrice'))
gas_price <- sort(filter(auto, fuel.type == 'gas') %>% pull('logPrice'))
# Create prior distributions
pp_diesel = dnorm(auto_range, mean = auto_mean, sd = sd(diesel_price))
pp_diesel = pp_diesel / sum(pp_diesel)
```

```
pp_gas = dnorm(auto_range, mean = auto_mean, sd = sd(gas_price))
pp_gas = pp_gas / sum(pp_gas)
```

The likelihood function will be a normal distribution with mean and standard deviation computed from each sample.

```
# Calculate likelihoods from sampled data
like.diesel = comp.like(auto_range, diesel_price)
```

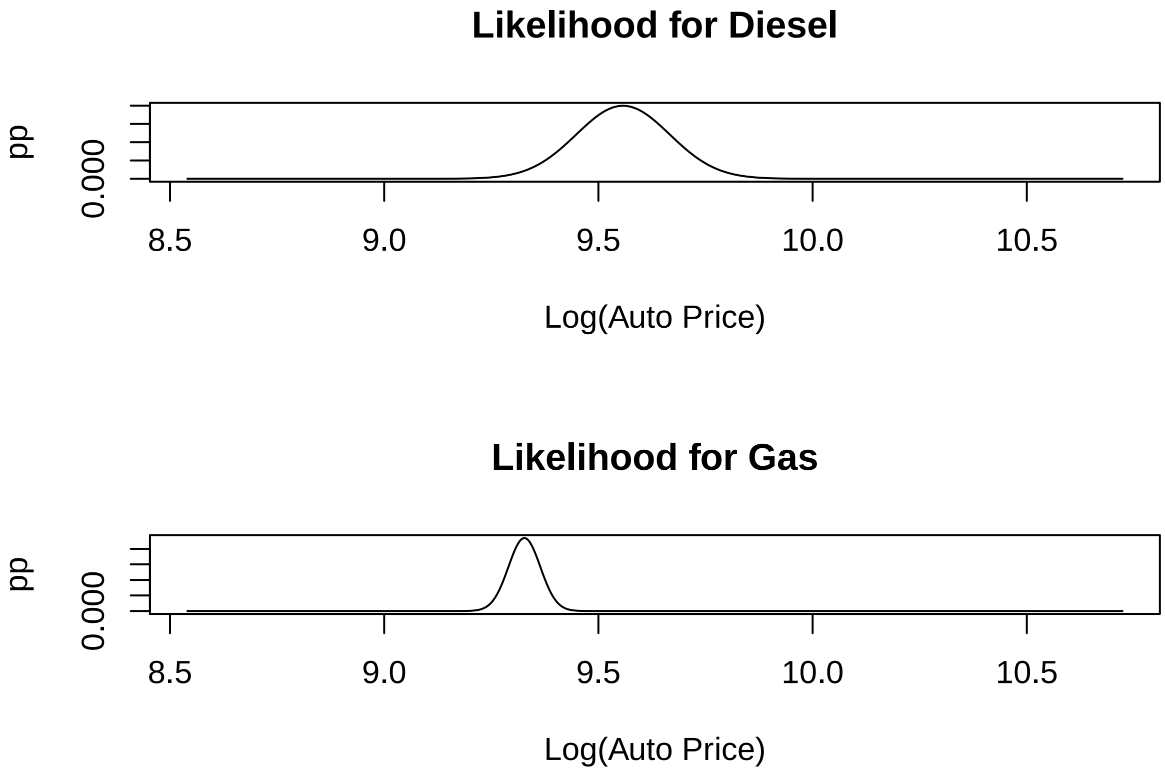
```
##  Mean = 9.55742 Standard deviation = 0.4880124
```

```
like.gas = comp.like(auto_range, gas_price)
```

```
##  Mean = 9.327208 Standard deviation = 0.4998436
```

Note the sample means and standard deviations. Next, let's plot the likelihood functions.

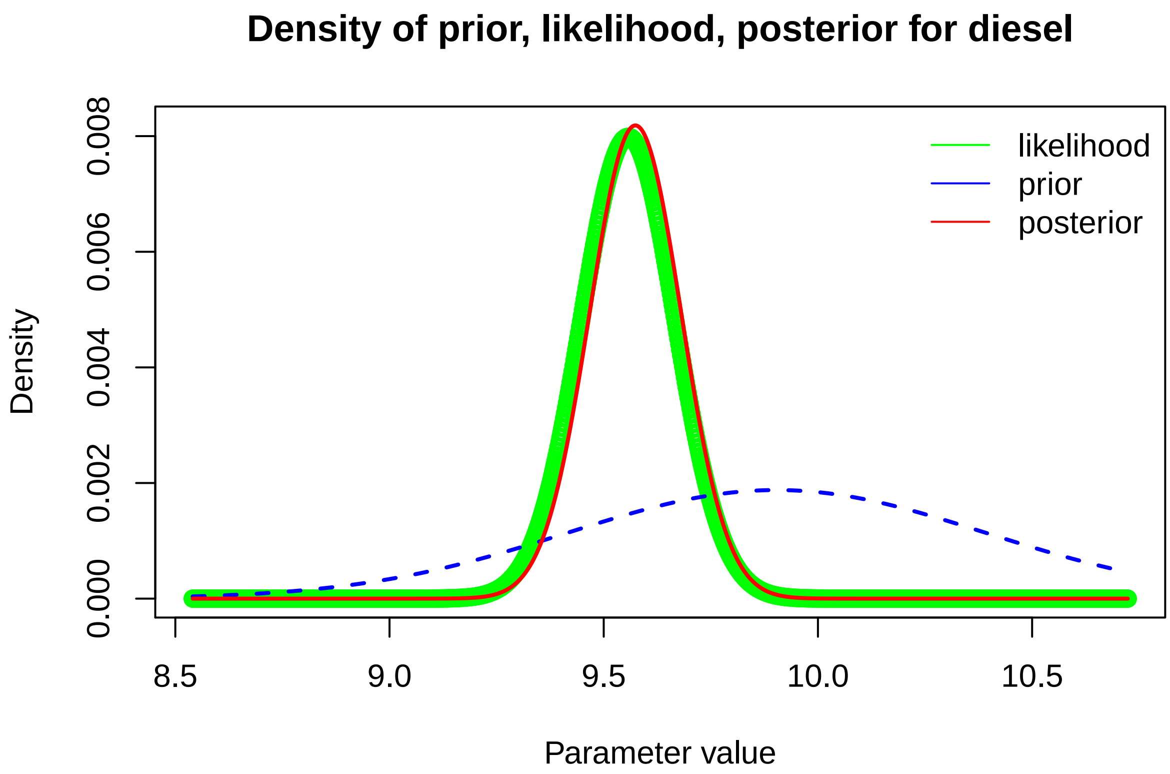
```
par(mfrow = c(2, 1))
plot.dist(auto_range, like.diesel, 'Likelihood for Diesel', 'Log(Auto Price)')
plot.dist(auto_range, like.gas, 'Likelihood for Gas', 'Log(Auto Price)')
```



```
par(mfrow = c(1, 1))
```

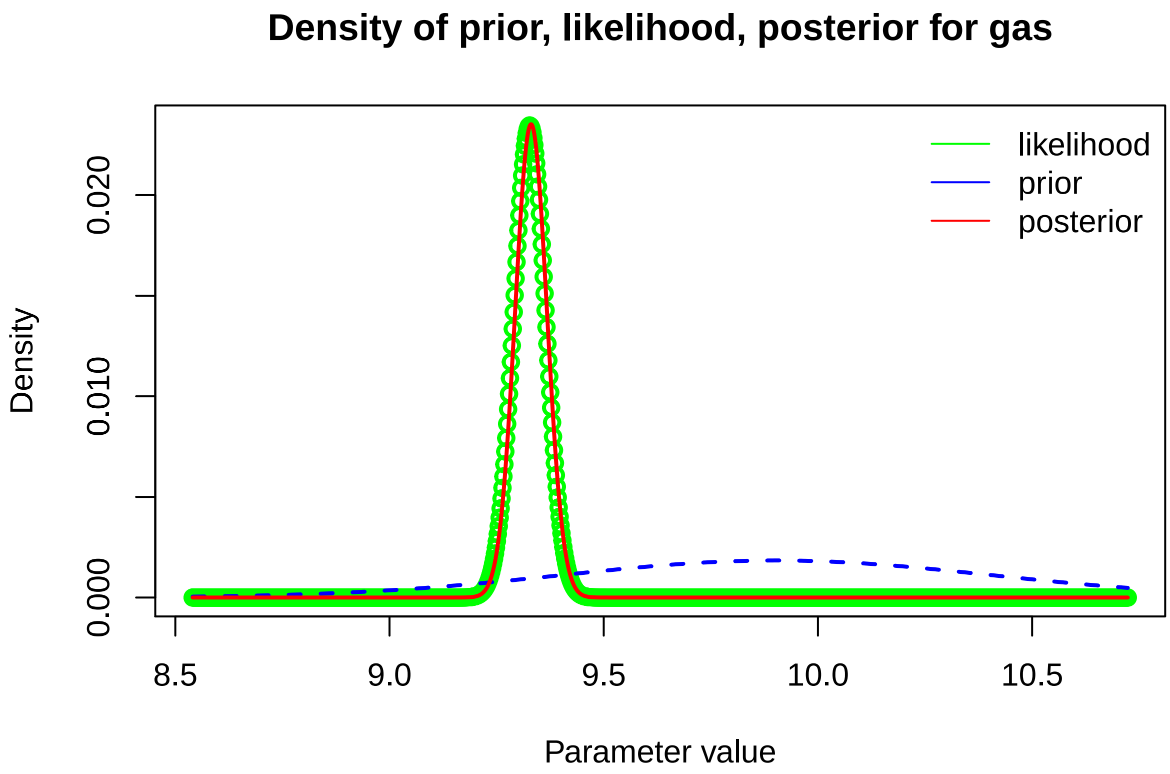
The size of the diesel sample is quite a bit smaller than the gas sample (20 automobiles versus 181), so there is much more spread in the distribution for the diesel likelihood function. Now, let's calculate and plot the posterior distribution from the prior and likelihood.

```
# Calculate posterior distribution
post.diesel = posterior(pp_diesel, like.diesel)
post.gas = posterior(pp_gas, like.gas)
plot.post(pp_diesel, like.diesel, post.diesel, auto_range, 'diesel')
```



```
##  Maximum of prior density = 9.904
##  Maximum likelihood = 9.557
##  MAP = 9.574
```

```
plot.post(pp_gas, like.gas, post.gas, auto_range, 'gas')
```



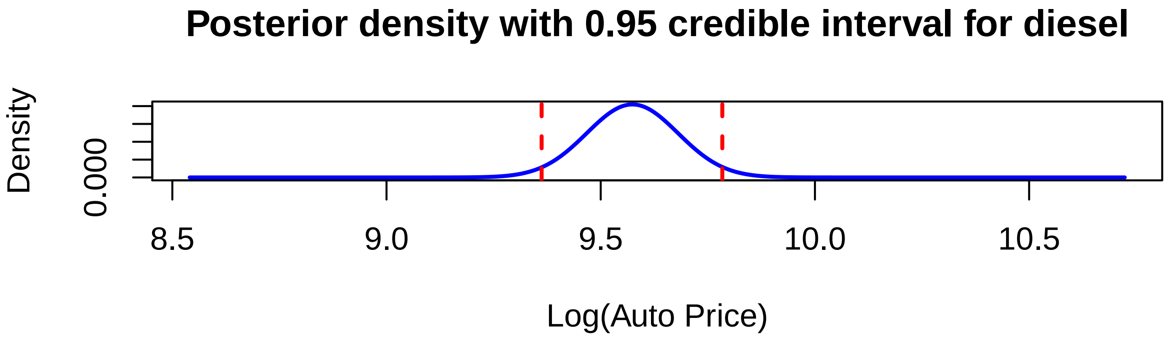
```
##  Maximum of prior density = 9.904
##  Maximum likelihood = 9.327
##  MAP = 9.331
```

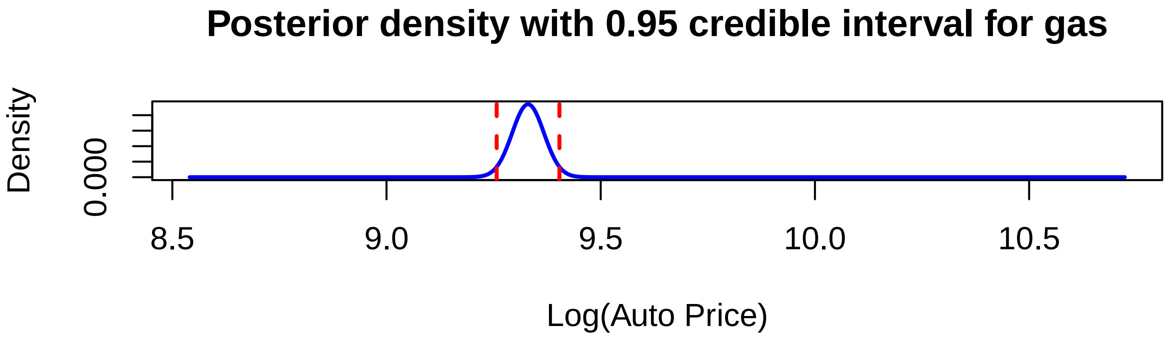
Note that because the size of the diesel sample is much smaller than the gas sample, the prior distribution has a larger effect on the shape of the posterior distribution. This represents the greater uncertainty in the diesel sample and the prior helping to account for that. There appears to be a difference in means for each sample, so let's compute and plot the credible intervals to see if it's a significant difference.

```
# Calculate 95% credible interval
par(mfrow = c(2, 1))
plot.ci(auto_range, post.diesel, nSampsQI, qs, 'diesel')
```

```
## The 0.95 credible interval for diesel is 9.36 to 9.78
```

```
plot.ci(auto_range, post.gas, nSampsQI, qs, 'gas')
```





```
## The 0.95 credible interval for gas is 9.26 to 9.4
```

```
par(mfrow = c(1, 1))
```

Note that there is some overlap in the credible intervals for gas and diesel. This means we can not reject the null hypothesis and can not say that a significant relationship does exist for the price of an automobile when stratified by fuel type.

The classical method (Welch's t-test) rejected the alternative hypothesis with a p-value of 0.05746, while the bootstrap method accepted it (the 95% confidence interval for the difference in sample means did not include zero). In this case, the Bayesian and classical results agree, while the bootstrap results are at odds.

Aspiration

Next, let's look at price stratified by aspiration. To do that, we'll create variables by splitting the price of automobiles by the two types, standard and turbo. Also, we'll create the prior distributions for the price using the overall mean price and standard deviation from each sample.

```
# Create vectors of standard and turbo prices
std_price <- sort(filter(auto, aspiration == 'std') %>% pull('logPrice'))
turbo_price <- sort(filter(auto, aspiration == 'turbo') %>% pull('logPrice'))
# Create prior distributions
pp_std = dnorm(auto_range, mean = auto_mean, sd = sd(std_price))
pp_std = pp_std / sum(pp_std)
pp_turbo = dnorm(auto_range, mean = auto_mean, sd = sd(turbo_price))
pp_turbo = pp_turbo / sum(pp_turbo)
```

As before, the likelihood function will be a normal distribution with mean and standard deviation computed from each sample.

```
# Calculate likelihoods from sampled data
like.std = comp.like(auto_range, std_price)
```

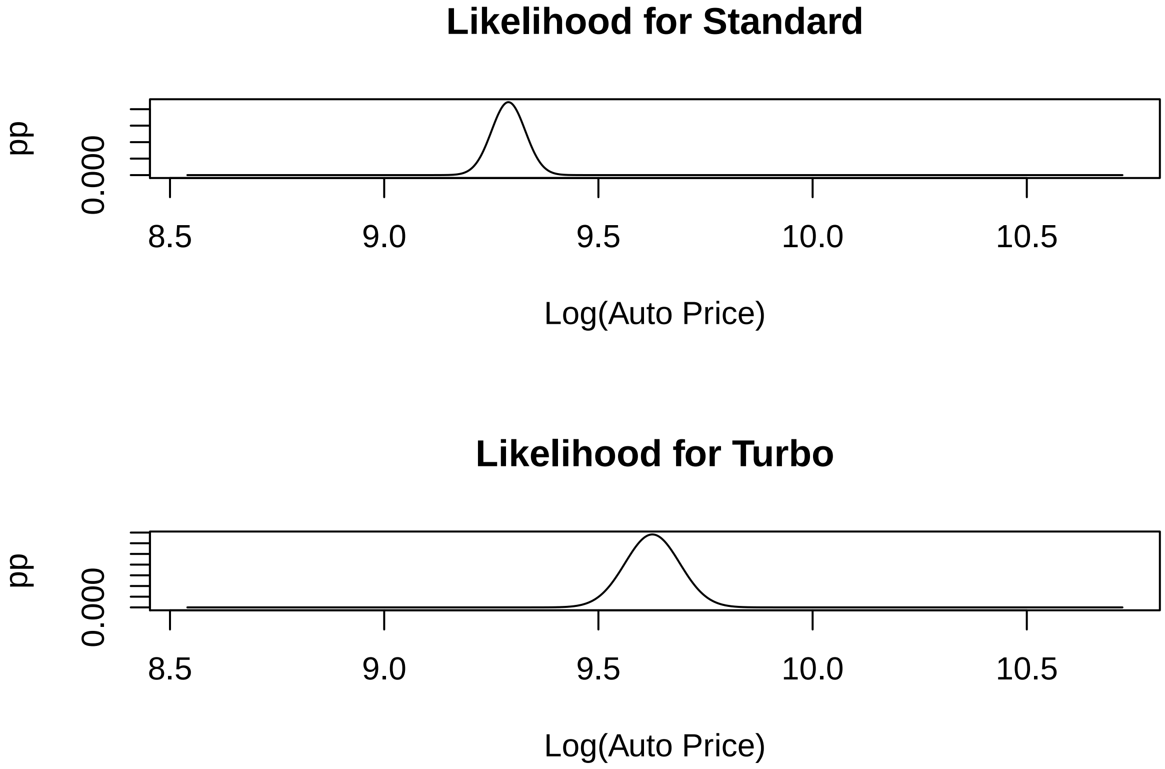
```
##  Mean = 9.289936 Standard deviation = 0.5059097
```

```
like.turbo = comp.like(auto_range, turbo_price)
```

```
##  Mean = 9.625932 Standard deviation = 0.3832251
```

Note the sample means and standard deviations. Next, let's plot the likelihood functions.

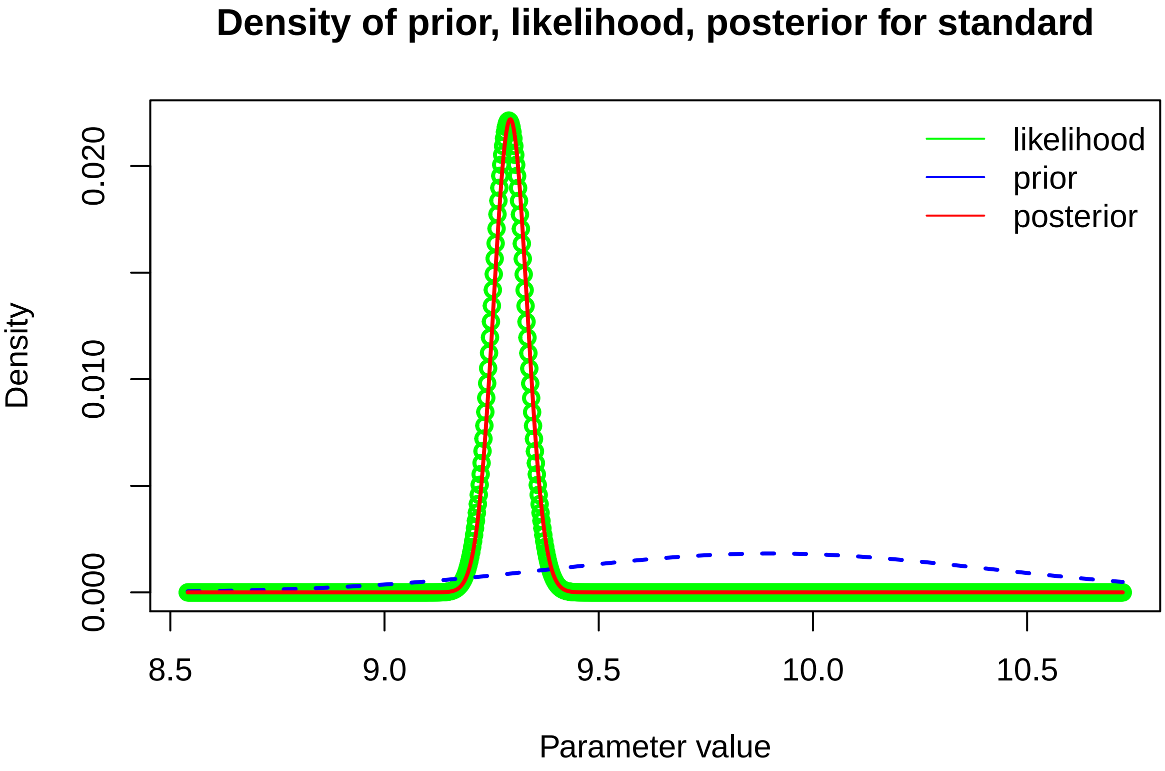
```
par(mfrow = c(2, 1))
plot.dist(auto_range, like.std, 'Likelihood for Standard', 'Log(Auto Price)')
plot.dist(auto_range, like.turbo, 'Likelihood for Turbo', 'Log(Auto Price)')
```



```
par(mfrow = c(1, 1))
```

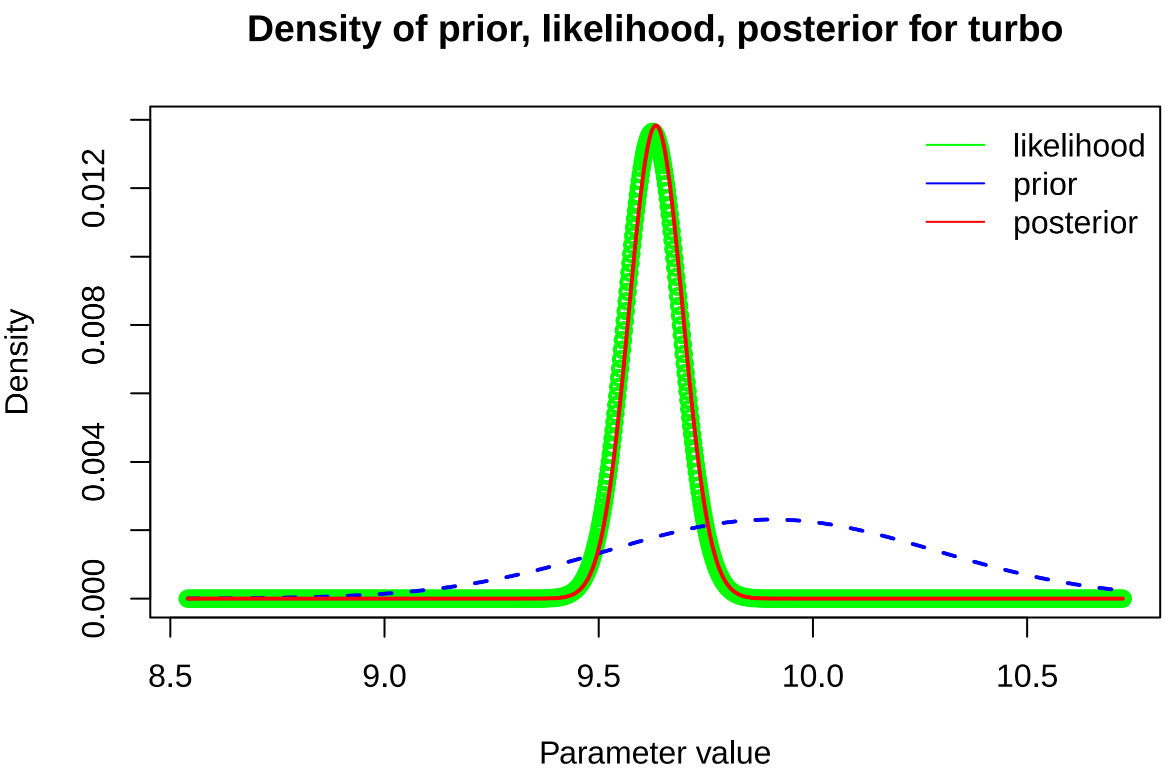
The size of the turbo sample is quite a bit smaller than the standard sample (36 automobiles versus 165), so there is much more spread in the distribution for the turbo likelihood function. Now, let's calculate and plot the posterior distribution from the prior and likelihood.

```
# Calculate posterior distribution
post.std = posterior(pp_std, like.std)
post.turbo = posterior(pp_turbo, like.turbo)
plot.post(pp_std, like.std, post.std, auto_range ,'standard')
```



```
##  Maximum of prior density = 9.904
##  Maximum likelihood = 9.29
##  MAP = 9.294
```

```
plot.post(pp_turbo, like.turbo, post.turbo, auto_range, 'turbo')
```



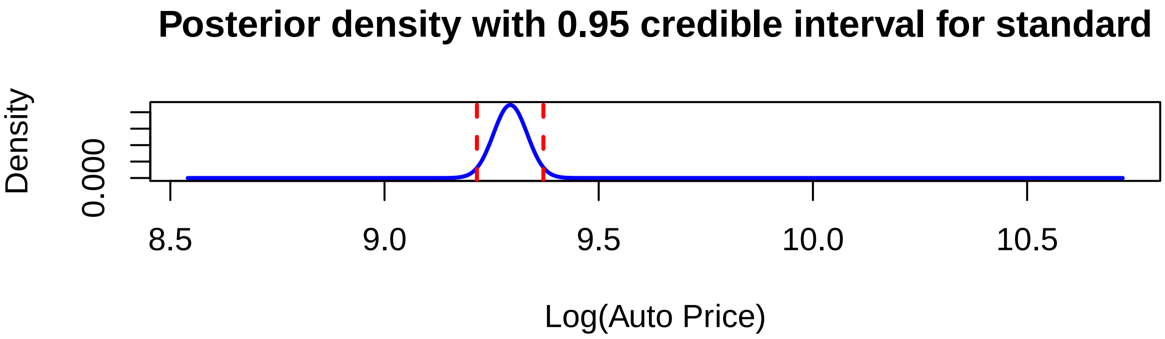
```
##  Maximum of prior density = 9.904
##  Maximum likelihood = 9.626
##  MAP = 9.633
```

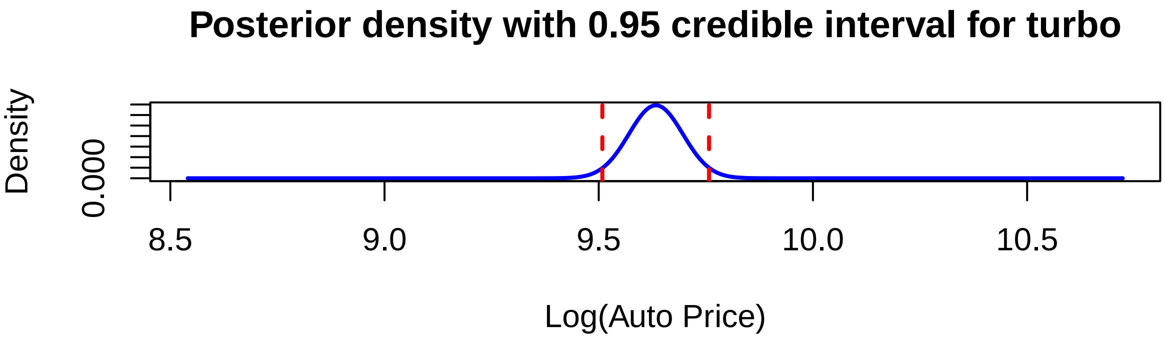
As before, because the size of the turbo sample is much smaller than the standard sample, the prior distribution has a larger effect on the shape of the posterior distribution. This represents the greater uncertainty in the turbo sample and the prior helping to account for that. There appears to be a difference in means for each sample, so let's compute and plot the credible intervals to see if it's a significant difference.

```
# Calculate 95% credible interval
par(mfrow = c(2, 1))
plot.ci(auto_range, post.std, nSampsQI, qs, 'standard')
```

```
## The 0.95 credible interval for standard is 9.22 to 9.37
```

```
plot.ci(auto_range, post.turbo, nSampsQI, qs, 'turbo')
```





```
## The 0.95 credible interval for turbo is 9.51 to 9.76
```

```
par(mfrow = c(1, 1))
```

Now, there is no overlap in the credible intervals for standard and turbo. This means we can reject the null hypothesis and say that a significant relationship does exist for the price of an automobile when stratified by aspiration.

The classical method (Welch's t-test) also accepted the alternative hypothesis with a p-value of 3.137e-05, as did the bootstrap method (the 95% confidence interval for the difference in sample means did not include zero). In this case, all three methods agree.

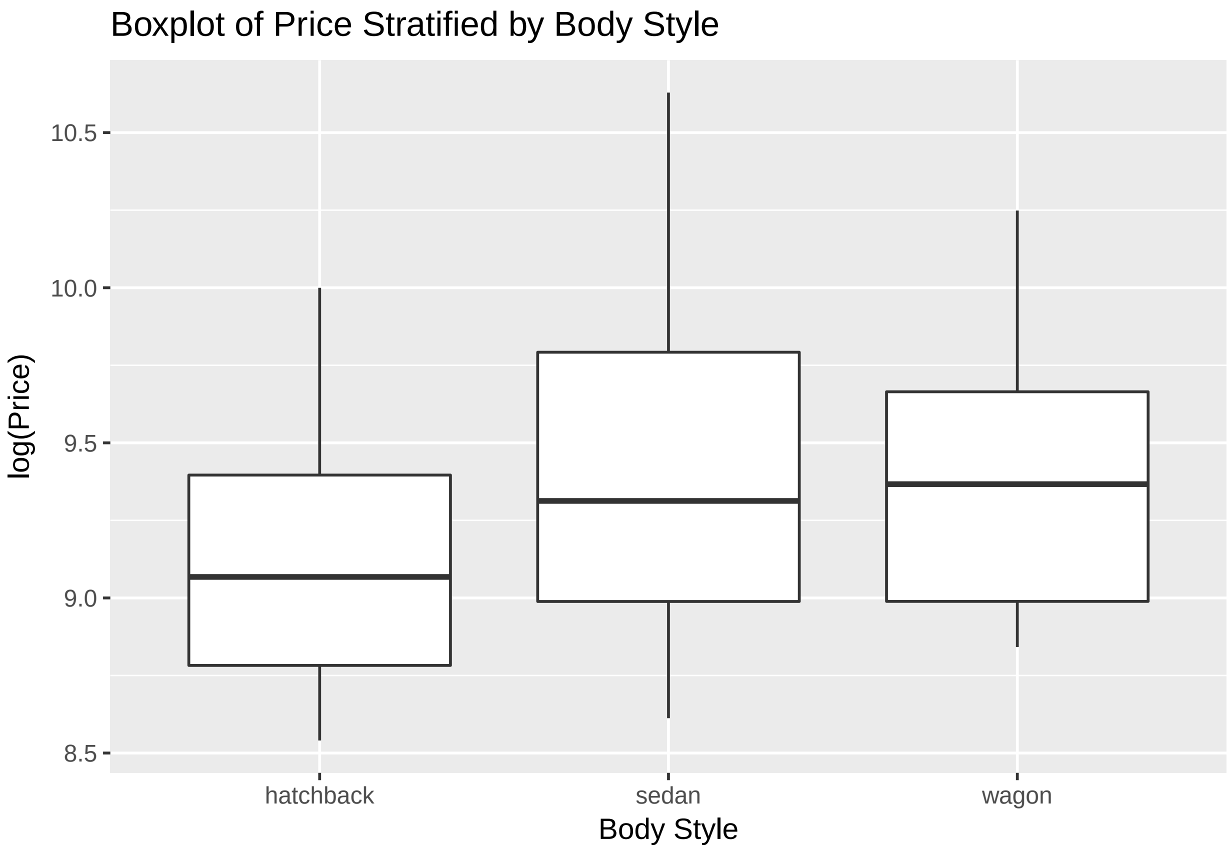
Multi-Sample Significance Tests

In this section, we'll use Bayesian methods to see whether or not there is a significant difference in automobile price when stratified into more than two groups (using body style). Once again, we'll compare the results to those obtained using classical (Tukey's HSD) and bootstrap methods. As with the two-sample tests, the null hypothesis will be that all of the groups come from the same population (i.e., there is no difference in means), and the alternative hypothesis will be that at least one of the groups has a statistically different mean. Also, as before a 95% confidence/credible interval will be used.

In the previous write-ups, we determined that we could only perform comparisons between the body styles of sedan, hatchback, and wagon, as only they had sufficient data for comparison.

To review, let's look at the boxplots of price for each of three body types.

```
# Create subset by dropping convertible and hardtop (insufficient data)
auto_bodystyle <- subset(auto, auto$body.style %in% c('hatchback','sedan','wagon'))
# Visual
ggplot(auto_bodystyle, aes(body.style,logPrice)) + geom_boxplot() + xlab('Body Style') +
  ylab('log(Price)') + ggtitle('Boxplot of Price Stratified by Body Style')
```



Note that hatchbacks appear to be the most different when compared to sedans and wagons. Before we can conduct the pairwise comparisons, we need to create variables by splitting the price by body style. Then, we need to compute the prior distributions and likelihood functions for each of the body types using the sample data.

```
# Get price for each body style
hatch_price <- sort(filter(auto, body.style == 'hatchback') %>% pull('logPrice'))
sedan_price <- sort(filter(auto, body.style == 'sedan') %>% pull('logPrice'))
wagon_price <- sort(filter(auto, body.style == 'wagon') %>% pull('logPrice'))
# Create prior distributions
pp_hatch = dnorm(auto_range, mean = auto_mean, sd = sd(hatch_price))
pp_hatch = pp_hatch / sum(pp_hatch)
pp_sedan = dnorm(auto_range, mean = auto_mean, sd = sd(sedan_price))
pp_sedan = pp_sedan / sum(pp_sedan)
pp_wagon = dnorm(auto_range, mean = auto_mean, sd = sd(wagon_price))
pp_wagon = pp_wagon / sum(pp_wagon)
# Calculate likelihoods from sampled data
like.hatch = comp.like(auto_range, hatch_price)
```

```
##  Mean = 9.127714 Standard deviation = 0.3922469
```

```
like.sedan = comp.like(auto_range, sedan_price)
```

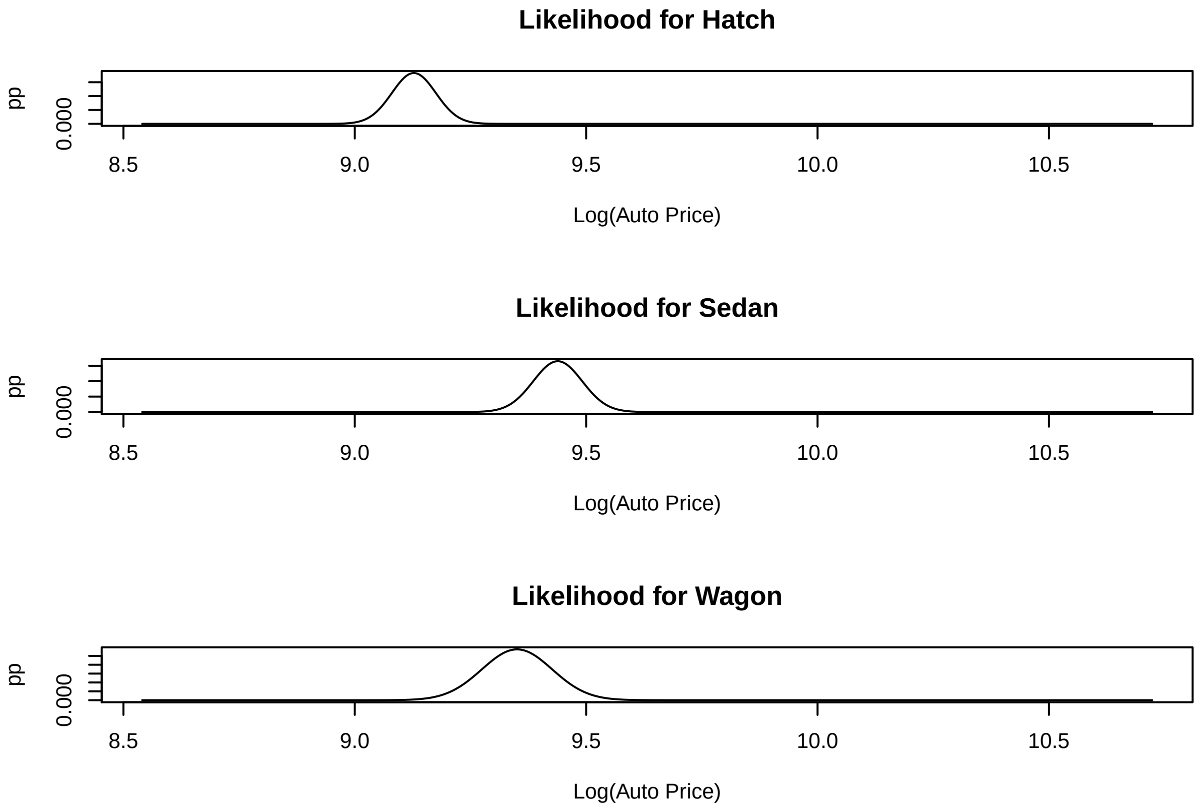
```
##  Mean = 9.438602 Standard deviation = 0.5126516
```

```
like.wagon = comp.like(auto_range, wagon_price)
```

```
##  Mean = 9.350626 Standard deviation = 0.3801369
```

Note the sample means and standard deviations. Let's plot the likelihood functions for each body style.

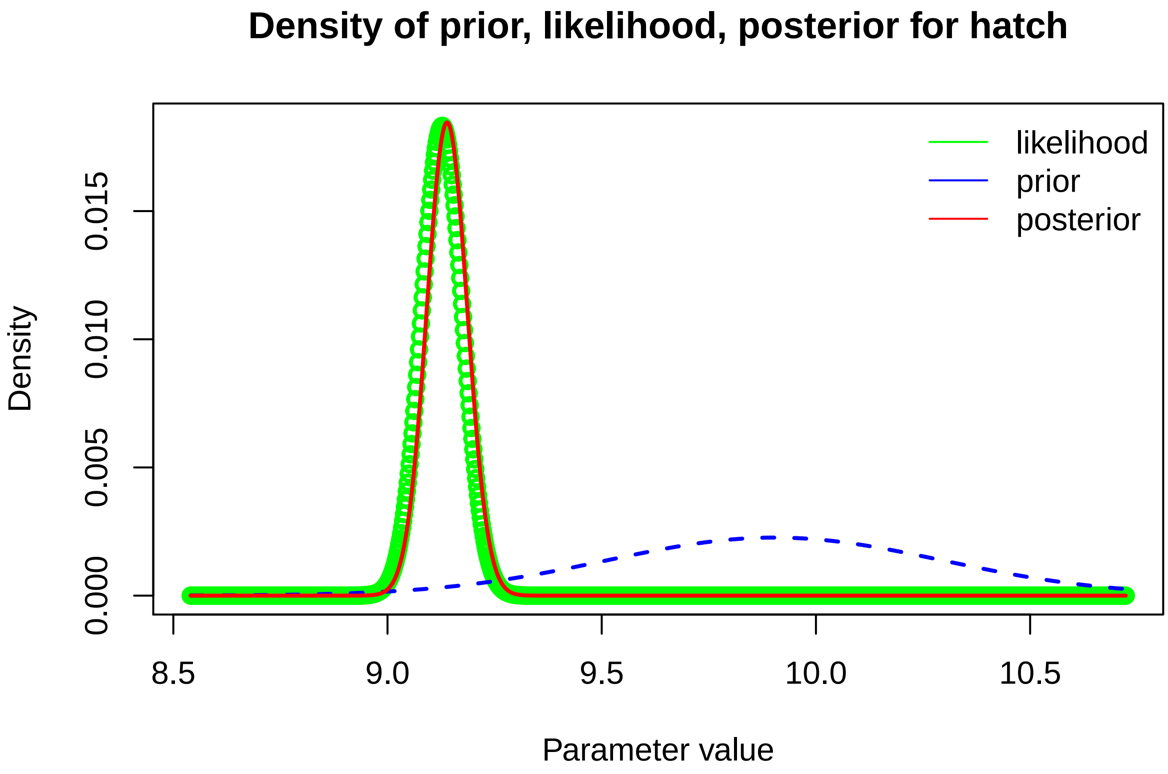
```r
par(mfrow = c(3, 1))
plot.dist(auto_range, like.hatch, 'Likelihood for Hatch', 'Log(Auto Price)')
plot.dist(auto_range, like.sedan, 'Likelihood for Sedan', 'Log(Auto Price)')
plot.dist(auto_range, like.wagon, 'Likelihood for Wagon', 'Log(Auto Price)')
```



```r
par(mfrow = c(1, 1))
```

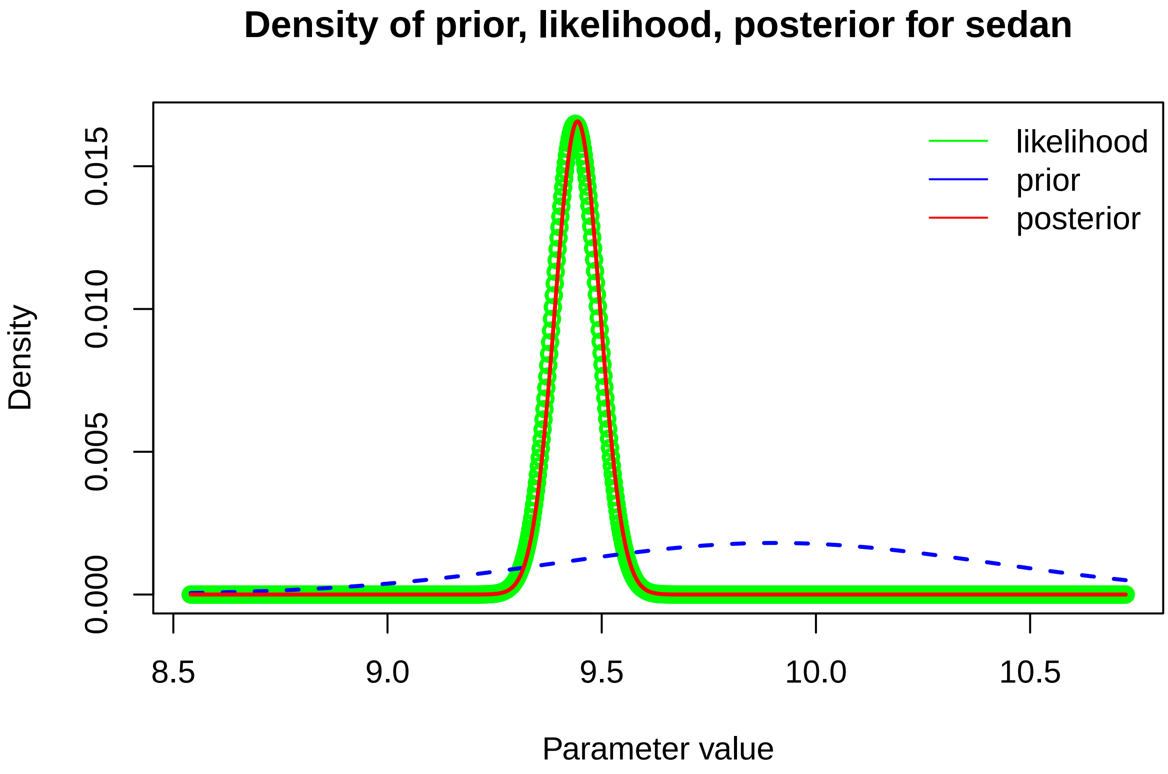
As with the boxplot, the Bayesian estimate for the mean of the hatchback stands out from the other body types. Also, since the size of the wagon sample is quite a bit smaller than the others (25 automobiles versus 94 for sedans and 68 for hatchbacks), there is much more spread in the distribution for the wagon likelihood function. Let's plot the posterior distributions for each body type.

```r
# Calculate posterior distribution
post.hatch = posterior(pp_hatch, like.hatch)
post.sedan = posterior(pp_sedan, like.sedan)
post.wagon = posterior(pp_wagon, like.wagon)
plot.post(pp_hatch, like.hatch, post.hatch, auto_range ,'hatch')
```



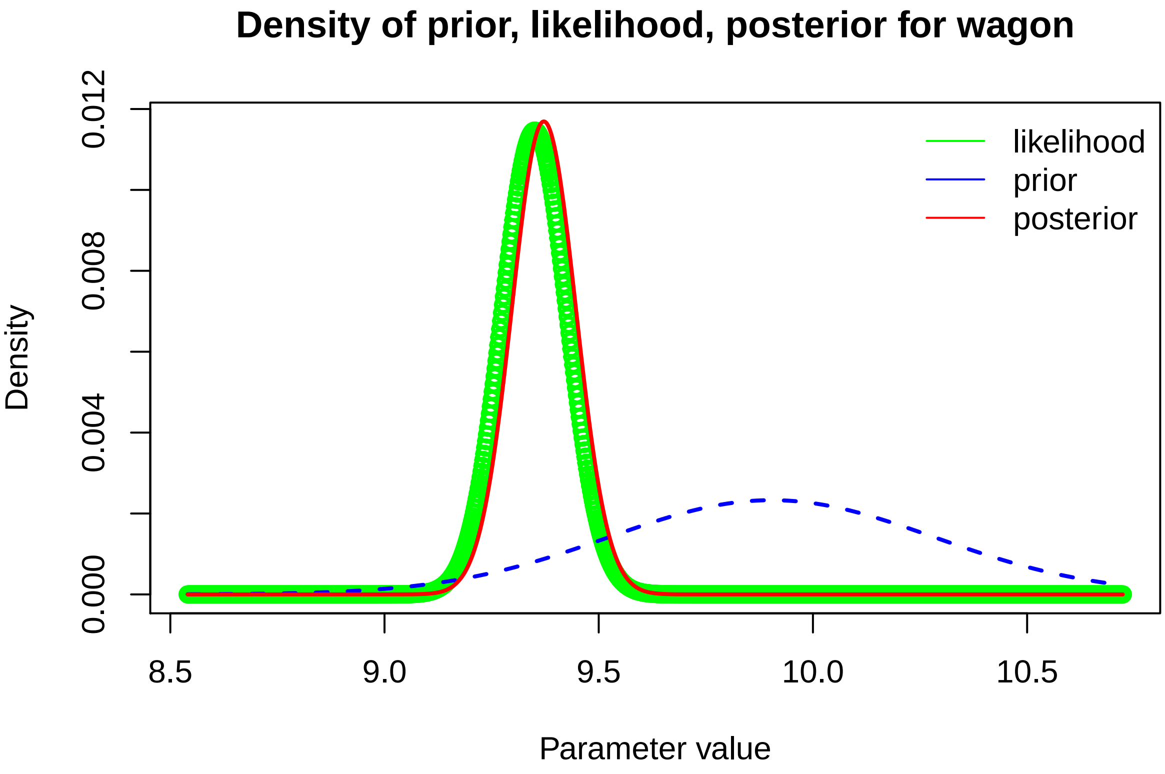
```
##  Maximum of prior density = 9.904
##  Maximum likelihood = 9.128
##  MAP = 9.139
```

```
plot.post(pp_sedan, like.sedan, post.sedan, auto_range, 'sedan')
```



```
##  Maximum of prior density = 9.904
##  Maximum likelihood = 9.439
##  MAP = 9.443
```

```
plot.post(pp_wagon, like.wagon, post.wagon, auto_range, 'wagon')
```



```
##  Maximum of prior density = 9.904
##  Maximum likelihood = 9.351
##  MAP = 9.373
```

Here, none of the sample sizes are particularly large, so the prior distribution has a noticeable effect on the shape of the posterior distributions for all three samples. This represents the greater uncertainty in the samples and the prior helping to account for that. Let's move on to the pairwise comparisons.

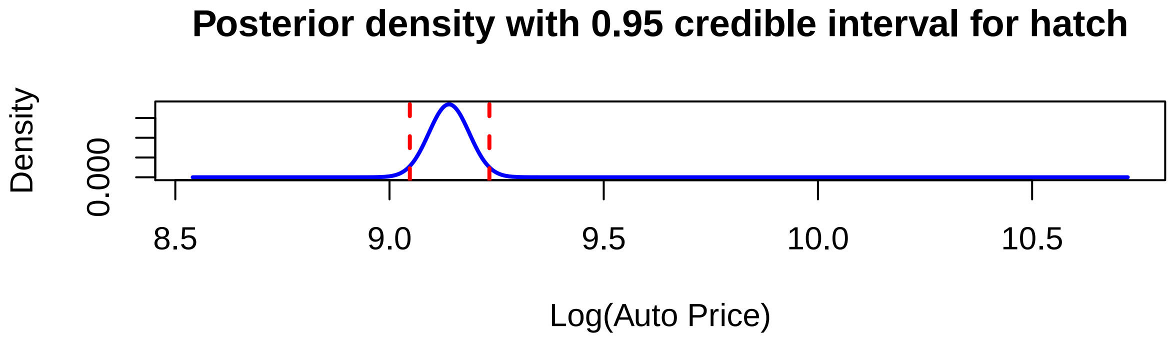
Sedan versus Hatchback

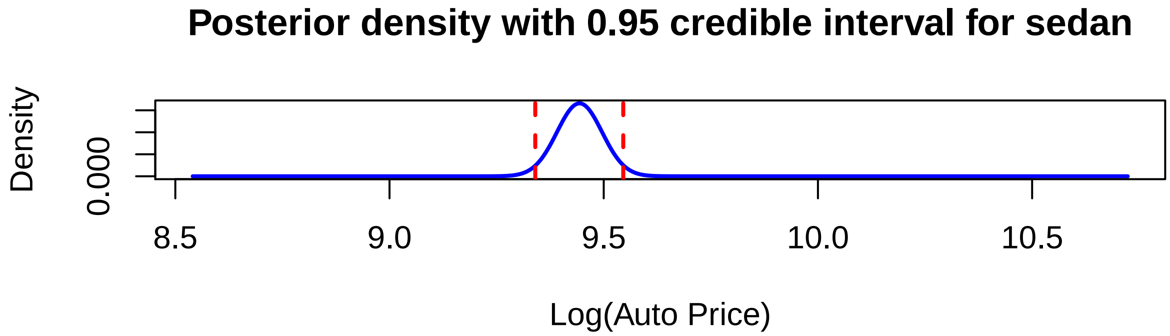
First, let's compare sedans to hatchbacks and plot their 95% credible intervals.

```
# Calculate 95% credible interval
par(mfrow = c(2, 1))
plot.ci(auto_range, post.hatch, nSampsQI, qs, 'hatch')
```

```
## The 0.95 credible interval for hatch is 9.05 to 9.23
```

```
plot.ci(auto_range, post.sedan, nSampsQI, qs, 'sedan')
```





```
## The 0.95 credible interval for sedan is 9.34 to 9.55
```

```
par(mfrow = c(1, 1))
```

The 95% credible intervals do not overlap, so we can reject the null hypothesis and say that a statistically significant relationship exists between the price of a sedan versus a hatchback.

The classical method (Tukey's HSD) also rejected the null hypothesis with a p-value of 0.0000870. The 95% confidence interval computed by the bootstrap method did not include zero, so we could also reject the null hypothesis. In this case, all three methods agree.

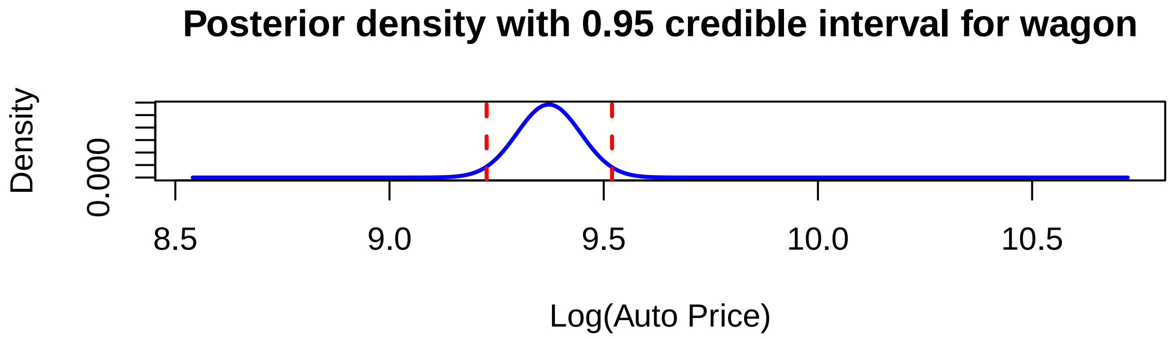
Wagon versus Hatchback

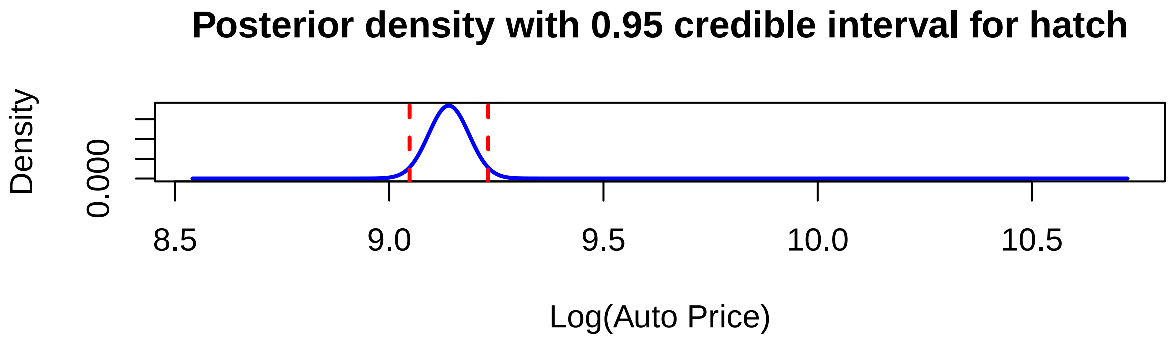
Next, let's compare wagons to hatchbacks and plot their 95% credible intervals.

```
# Calculate 95% credible interval
par(mfrow = c(2, 1))
plot.ci(auto_range, post.wagon, nSampsQI, qs, 'wagon')
```

```
## The 0.95 credible interval for wagon is 9.23 to 9.52
```

```
plot.ci(auto_range, post.hatch, nSampsQI, qs, 'hatch')
```





```
## The 0.95 credible interval for hatch is 9.05 to 9.23
```

```
par(mfrow = c(1, 1))
```

Here, the 95% credible intervals share an endpoint, so we can not reject the null hypothesis and can not say that a statistically significant relationship exists between the price of a wagon versus a hatchback.

The classical method (Tukey's HSD) could not reject the null hypothesis due to a p-value of 0.094. As before, the 95% confidence interval computed by the bootstrap method did not include zero, so we could reject the null hypothesis. In this case, the Bayesian and classical results agree, while the bootstrap results are at odds.

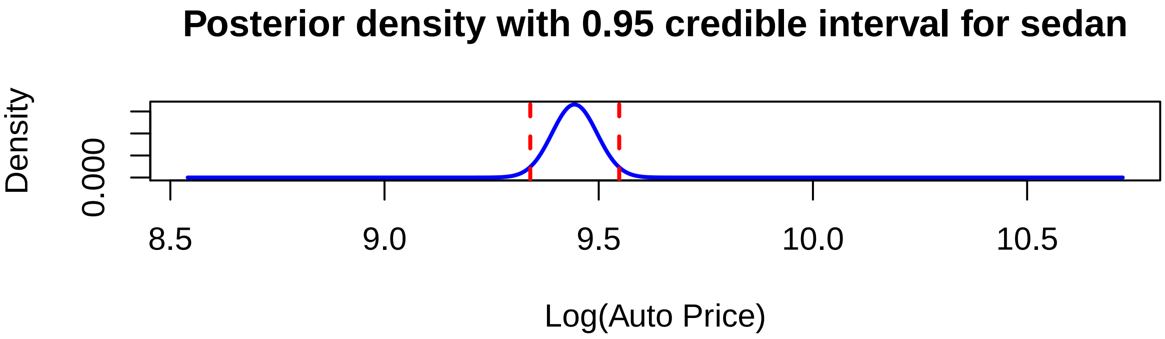
Wagon versus Sedan

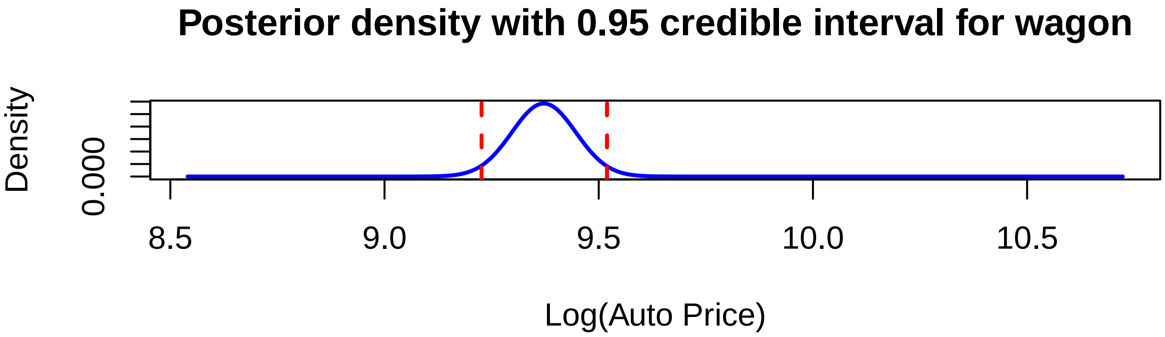
Finally, let's compare sedans to wagons and plot their 95% credible intervals.

```
# Calculate 95% credible interval
par(mfrow = c(2, 1))
plot.ci(auto_range, post.sedan, nSampsQI, qs, 'sedan')
```

```
## The 0.95 credible interval for sedan is 9.34 to 9.55
```

```
plot.ci(auto_range, post.wagon, nSampsQI, qs, 'wagon')
```





```
## The 0.95 credible interval for wagon is 9.23 to 9.52
```

```
par(mfrow = c(1, 1))
```

In this case, the 95% credible intervals do overlap, so we can not reject the null hypothesis and can not say that a statistically significant relationship exists between the price of a wagon versus a sedan.

The classical method (Tukey's HSD) could not reject the null hypothesis due to a p-value of 0.668. Also, the 95% confidence interval computed by the bootstrap method did include zero, so we could also not reject the null hypothesis. In this case, all three methods agree.

Summary and Conclusion

The purpose of this assignment was to compare Bayesian methods to classical and bootstrap methods for hypothesis testing in statistics. Specifically, we compared Bayesian methods to Welch's t-test and bootstrap methods for computing the difference in means between two populations (automobile price stratified by fuel type and aspiration). Also, we compared Bayesian methods to Tukey's HSD and bootstrap methods for computing the difference in means between three populations (automobile price stratified by body style). After comparing these various methods, it was found that:

1. Using Bayesian methods and Welch's t-test, a statistically significant relationship exists for the price of an automobile when stratified by aspiration (standard vs. turbo), but not fuel type (gas vs. diesel). Using bootstrap methods, a statistically significant relationship exists for both.
2. Using Bayesian methods and Tukey's HSD, a statistically significant relationship exists between the price of sedans and hatchbacks only. Using bootstrap methods, a statistically significant relationship exists between the price of sedans and hatchbacks, and wagons and hatchbacks, but not wagons and sedans.

In both bivariate and multivariate tests, Bayesian and classical methods produce different results when compared to bootstrap methods for the automobile data. For the classical results, a possible reason for the difference is that in the initial write-up, it was found that the distribution of the log of automobile price is right skewed. Welch's t-test and Tukey's HSD assume the populations being tested follow a normal distribution, and perhaps the violation of that assumption is the reason different outcomes were obtained. Non-parametric methods such as bootstrapping do not assume any particular distribution. For the Bayesian results, because some of our samples sizes were small (for diesel, turbo, and wagons, for example), our prior distribution had a noticeable effect on the posterior distribution. This represented the uncertainty inherent in a small sample size (larger standard deviation). Because of this, the quality and accuracy of our prior became more important. If our guess was not very accurate, we could end up with posterior distributions for automobile price that were not as accurate as the estimates from the bootstrap methods, and perhaps that is why different outcomes were obtained.