Hypothesis Testing For Automobile Prices

Greg DeVore

July 22nd, 2017

Overview

The purpose of this assignment is to explore some of the basic concepts and methods associated with hypothesis testing in statistics. This will be done by looking at automobile data, where the response variable is the price of the automobile, and the explanatory variables are different attributes of each automobile. Specifically, we'll do the following:

1. Test the normality of the distribution of the price versus the logarithm of the price of the automobiles.
2. Apply tests of significance (Welch's t-test) to the price of an automobile when stratified by binary categories such as fuel type, engine aspiration, and rear versus front wheel drive.
3. Apply tests of significance (ANOVA, Tukey HSD) to the price of a automobile when stratified by a multivariate category such as body type.

After exploring these basic concepts and methods, it will be found that:

1. The distribution of the price of the automobiles in the data set has positive skew, and taking the log of the price helps to restore some symmetry to the distribution.
2. A statistically significant relationship exists for the price of an automobile when stratified by both aspiration and drive type, but not fuel type.
3. A statistically significant relationship exists for the price of a sedan versus a hatchback, but not between other body types.

Data Preparation

First, let's look at an overview of the automobile data:

```r
str(auto)
```

```
## 'data.frame':    205 obs. of  26 variables:
##  $ symboling        : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ normalized.losses: chr  "?" "?" "?" "164" ...
##  $ make             : chr  "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
##  $ fuel.type        : chr  "gas" "gas" "gas" "gas" ...
##  $ aspiration       : chr  "std" "std" "std" "std" ...
##  $ num.of.doors     : chr  "two" "two" "two" "four" ...
##  $ body.style       : chr  "convertible" "convertible" "hatchback" "sedan" ...
##  $ drive.wheels     : chr  "rwd" "rwd" "rwd" "fwd" ...
##  $ engine.location  : chr  "front" "front" "front" "front" ...
##  $ wheel.base       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ length           : num  169 169 171 177 177 ...
##  $ width            : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ height           : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curb.weight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
##  $ engine.type      : chr  "dohc" "dohc" "ohcv" "ohc" ...
##  $ num.of.cylinders : chr  "four" "four" "six" "four" ...
##  $ engine.size      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ fuel.system      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
##  $ bore             : chr  "3.47" "3.47" "2.68" "3.19" ...
##  $ stroke           : chr  "2.68" "2.68" "3.47" "3.40" ...
##  $ compression.ratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower       : chr  "111" "111" "154" "102" ...
##  $ peak.rpm         : chr  "5000" "5000" "5000" "5500" ...
##  $ city.mpg         : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highway.mpg      : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price            : chr  "13495" "16500" "16500" "13950" ...
```

Note that price is currently a string. We certainly want this to be a numeric variable, so let's convert it. Additionally, we'll want several of these 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)

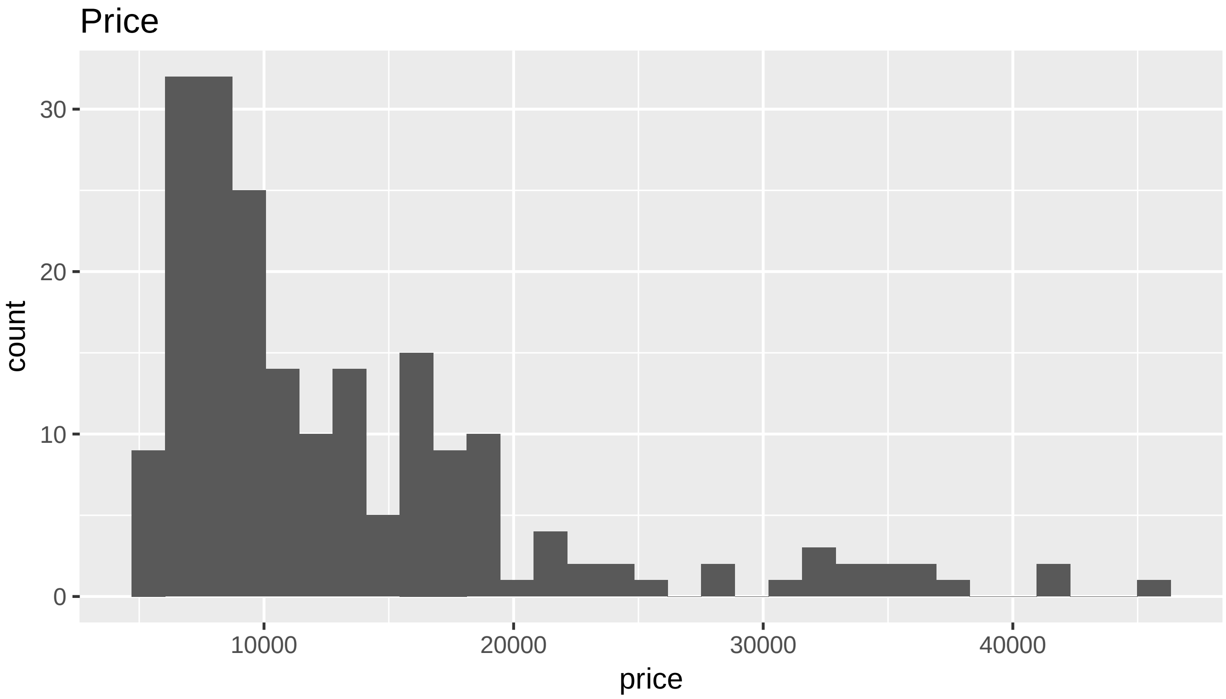
```
# Coerce some factors to numeric and retain only complete observations
auto$price <- as.numeric(auto$price)
auto$fuel.type <- as.factor(auto$fuel.type)
auto$aspiration <- as.factor(auto$aspiration)
auto$body.style <- as.factor(auto$body.style)
auto <- auto[complete.cases(auto),]
```

Testing The Normality of Automobile Price

The significance tests used in this write-up rely on the assumption that the population being sampled is normally distributed. Because of this, it is essential to check whether the variable for automobile price follows a normal distribution. If it does not, techniques exist that can help restore some symmetry to the distribution.

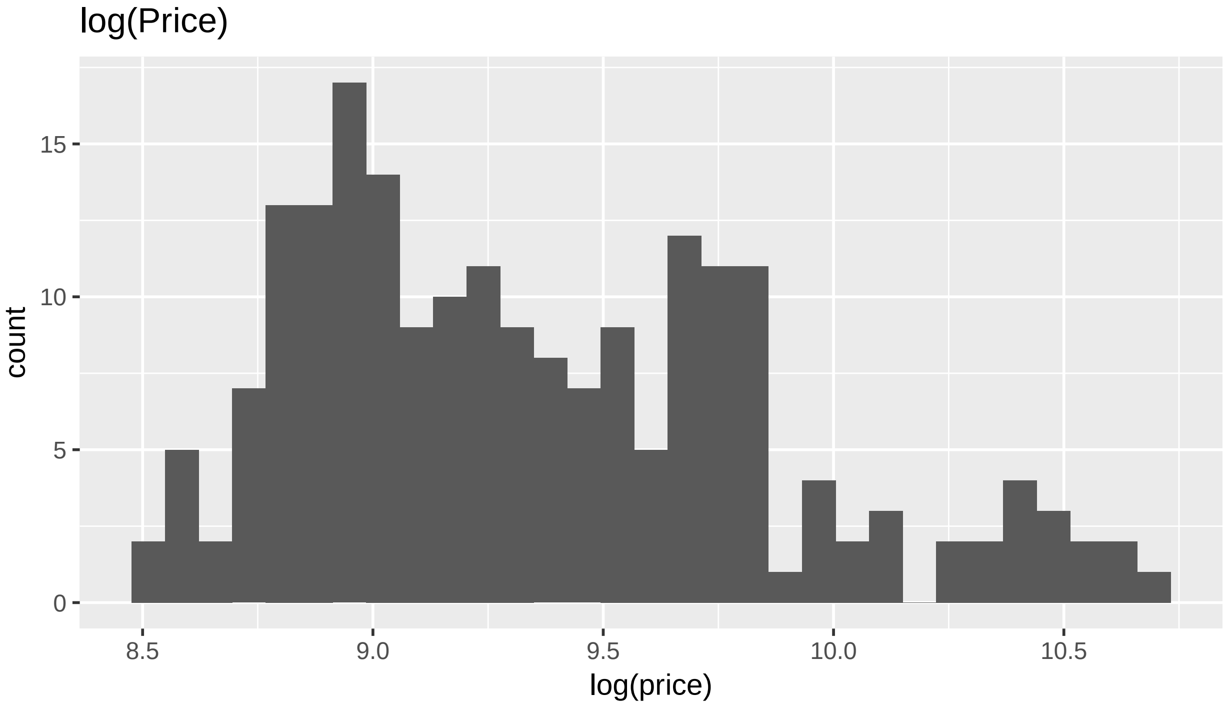
Let's start by looking at the distribution of automobile price:

```
# Histogram of price
bw <- (max(auto$price) - min(auto$price))/30
ggplot(auto, aes(price)) + geom_histogram(binwidth = bw) + ggtitle('Price') +
  theme(aspect.ratio = 1/2)
```



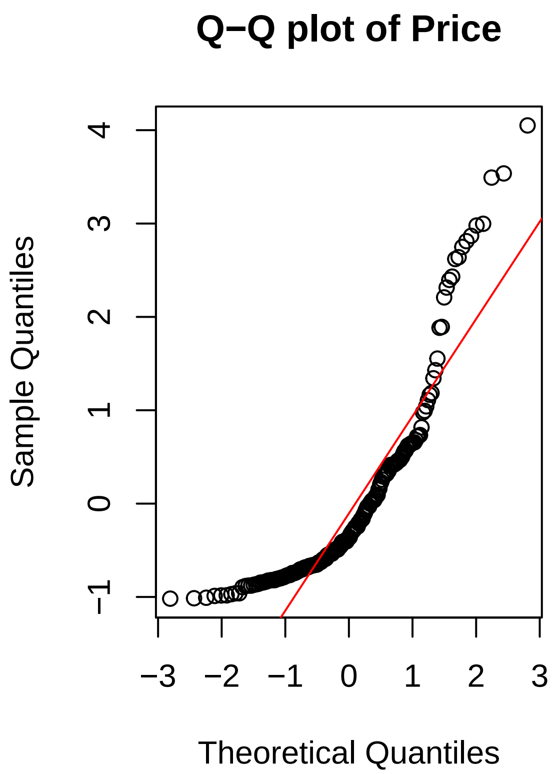
This distribution appears to be quite heavily right-skewed. Let's try taking the logarithm of the price and comparing the results.

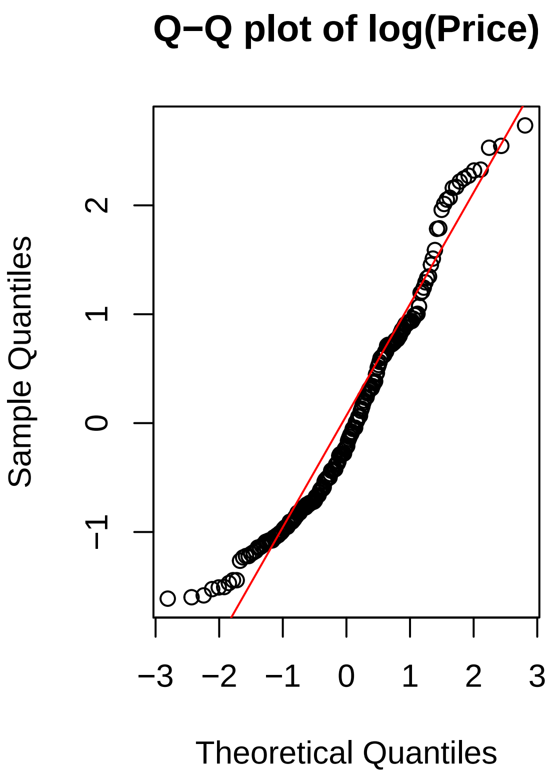
```
# Histogram of log(price)
bw <- (max(log(auto$price)) - min(log(auto$price)))/30
ggplot(auto, aes(log(price))) + geom_histogram(binwidth = bw) + ggtitle('log(Price)') +
  theme(aspect.ratio = 1/2)
```



The distribution of *log*(*price*) is slightly more centered, but actual tests should be performed to determine if it is different enough from the base *price* variable before moving forward. First, let's create a QQ plot to compare the two distributions. This plot will compare the quantiles of the two distributions to a theoretical normal distribution (a red line will be added to each plot that follows this distribution). To make this comparison meaningful, the two distributions of automobile price must be normalized so that they have zero mean and unit standard deviation. The goal is to see which distribution better aligns with a theoretical normal distribution (red line).

```
# Q-Q Plot
par(mfrow = c(1, 2))
# Normalize price
price_norm <- (auto$price - mean(auto$price))/sd(auto$price)
qqnorm(price_norm, main = 'Q-Q plot of Price')
# Add theoretical line (red)
qqline(rnorm(100), col = 2)
# Normalize log(price)
logprice_norm <- (log(auto$price) - mean(log(auto$price)))/sd(log(auto$price))
qqnorm(logprice_norm, main = 'Q-Q plot of log(Price)')
# Add theoretical line (red)
qqline(rnorm(100), col = 2)
```





The QQ plot of *log*(*price*) (right plot) aligns a bit better with the theoretical normal distribution, suggesting that its distribution is more centered than that of *price*.

Finally, let's use a formal test to assess the normality of both distributions. We'll use the Shapiro-Wilk test, which is specifically for normal distributions.

```
shapiro.test(auto$price)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  auto$price
## W = 0.79852, p-value = 2.216e-15
```

```
shapiro.test(log(auto$price))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(auto$price)
## W = 0.949, p-value = 1.432e-06
```

For this test, the larger the p-value, the higher the likelihood that the distribution being tested is normal. Here, both values are small, so both distributions deviate from normality, but the value for *log*(*price*) is significantly larger than the value for *price*. This means that the *log*(*price*) distribution is more normal than the *price* distribution.

As stated previously, the significance tests used here rely heavily on an assumption of normality. The results will be more accurate if the *log*(*price*) is used instead of *price*. Moving forward, all calculations will involve the log of the automobile price. Let's add it as a new column to the data.

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

Significance Tests For Two Samples

In this section, we'll use Welch's t-test to test whether or not there is a significant difference in automobile price when stratified by fuel type, aspiration, and drive wheels. We're using Welch's t-test here, rather than Student's t-test, because we cannot guarantee that the samples being tested are the same size or that they have equal variance. 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 level 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.

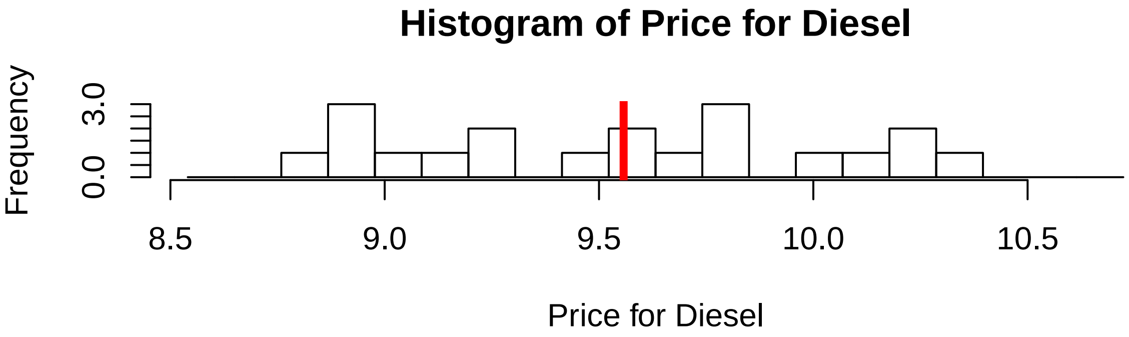
Fuel Type

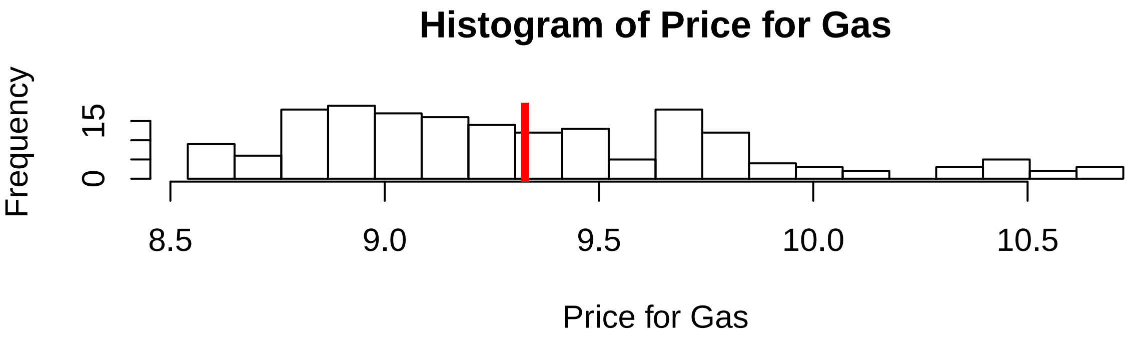
First, let's look at price stratified by fuel type. To do that, we'll create variables by splitting the price of automobiles by the two fuel types, gas and diesel.

```
diesel_price <- auto[auto$fuel.type == 'diesel','logPrice']
gas_price <- auto[auto$fuel.type == 'gas','logPrice']
```

Next, let's look at the histogram of each population

```
plot.t(diesel_price, gas_price, c('Price for Diesel','Price for Gas'))
```





The populations look different enough, but let's apply a formal test.

```r
# Welch's t-test
t.test(diesel_price, gas_price, alternative = 'two.sided')
```

```
## 
##  Welch Two Sample t-test
## 
## data:  diesel_price and gas_price
## t = 1.9971, df = 23.627, p-value = 0.05746
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007901153  0.468325711
## sample estimates:
## mean of x mean of y 
##  9.557420  9.327208
```

The p-value is just over 0.05, and the confidence interval just barely includes zero. However, the cutoff was 0.05, so we cannot reject the null hypothesis and cannot say that a statistically significant relationship exists between the price of an automobile and the fuel type.

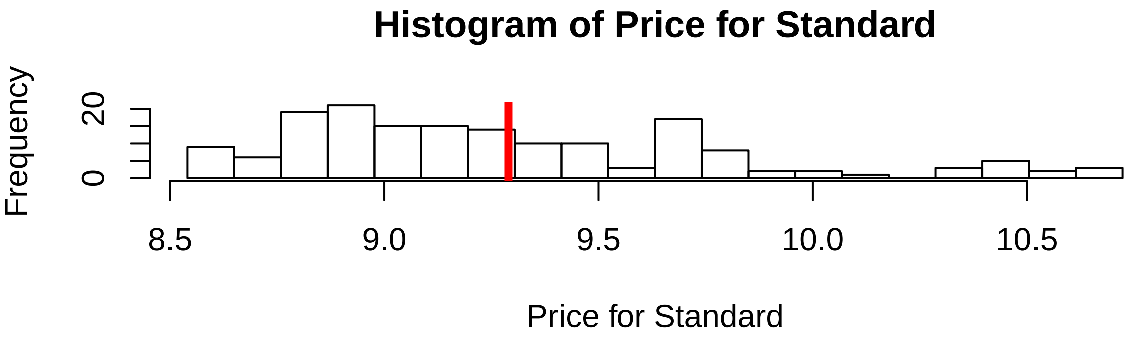
Aspiration

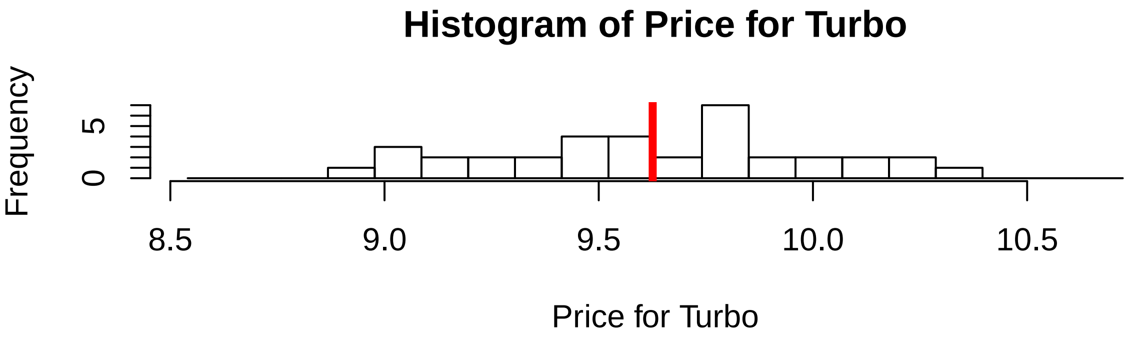
Next, let's look at price stratified by aspiration. To do that, we'll create variables by splitting the price of automobiles by the two types of aspiration, standard and turbo.

```
std_price <- auto[auto$aspiration == 'std','logPrice']
turbo_price <- auto[auto$aspiration == 'turbo','logPrice']
```

Next, let's look at the histogram of each population

```
plot.t(std_price, turbo_price, c('Price for Standard','Price for Turbo'))
```





Once again, the populations look different enough, but let's apply a formal test.

```
# Welch's t-test
t.test(std_price, turbo_price, alternative = 'two.sided')
```

```
##
##  Welch Two Sample t-test
##
## data:  std_price and turbo_price
## t = -4.4777, df = 64.681, p-value = 3.137e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4858704 -0.1861209
## sample estimates:
## mean of x mean of y
##  9.289936  9.625932
```

Now, the p-value is significantly less than 0.05, and the confidence interval is much farther from zero. In this case, we can reject the null hypothesis and say that a statistically significant relationship does exist between the price of an automobile and the type of aspiration.

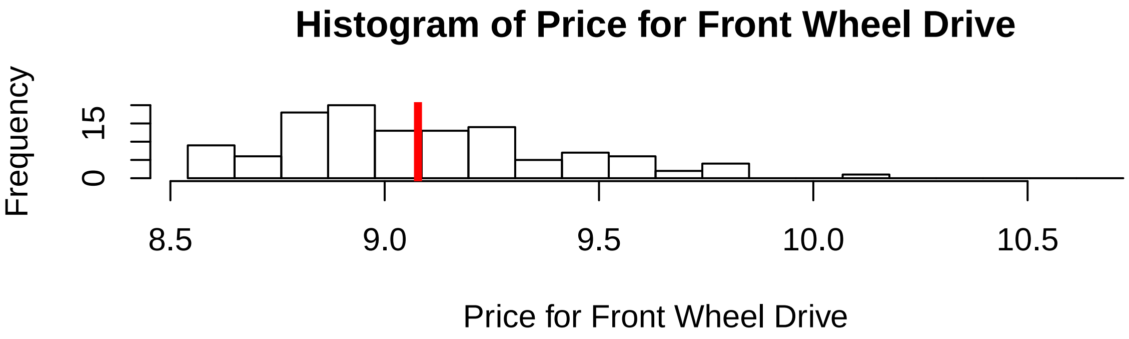
Drive Wheels

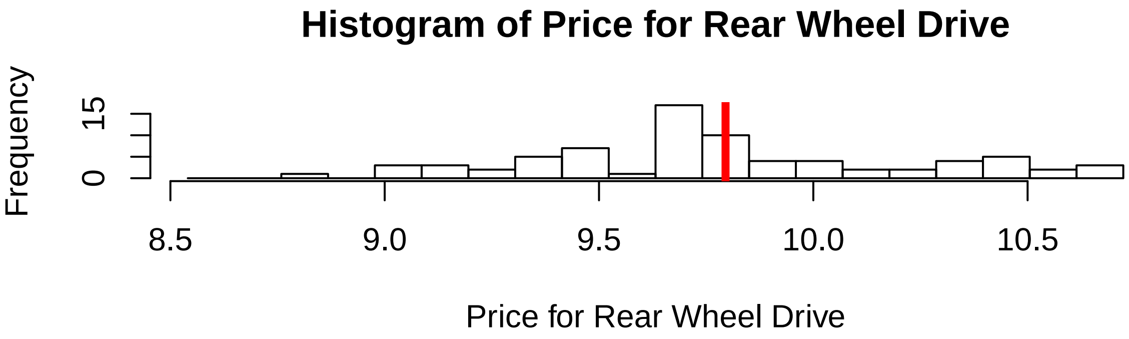
Finally, let's look at price stratified by drive wheels. To do that, we'll create variables by splitting the price of automobiles by the front and rear wheel drive.

```
fwd_price <- auto[auto$drive.wheels == 'fwd','logPrice']
rwd_price <- auto[auto$drive.wheels == 'rwd','logPrice']
```

Next, let's look at the histogram of each population

```
plot.t(fwd_price, rwd_price, c('Price for Front Wheel Drive','Price for Rear Wheel Drive'))
```





This time, the populations look quite different, but let's still apply a formal test.

```
# Welch's t-test
t.test(fwd_price, rwd_price, alternative = 'two.sided')
```

```
##
##  Welch Two Sample t-test
##
## data:  fwd_price and rwd_price
## t = -12.273, df = 123.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8332867 -0.6018405
## sample estimates:
## mean of x mean of y
##  9.077632  9.795195
```

Now, the p-value is nearly zero, and the confidence interval is quite far from zero. In this case, we can once again reject the null hypothesis and say that a statistically significant relationship does exist between the price of an automobile and the drive wheels.

Multi-Sample Significance Tests

Now, we will look at significance tests for more than two samples. Specifically, we'll use ANOVA (Analysis of Variance) and Tukey's HSD to determine if there is a significant difference in automobile price when stratified by body type. 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 level will be used.

To start, let's look at the different body types:

```
summary(auto$body.style)
```

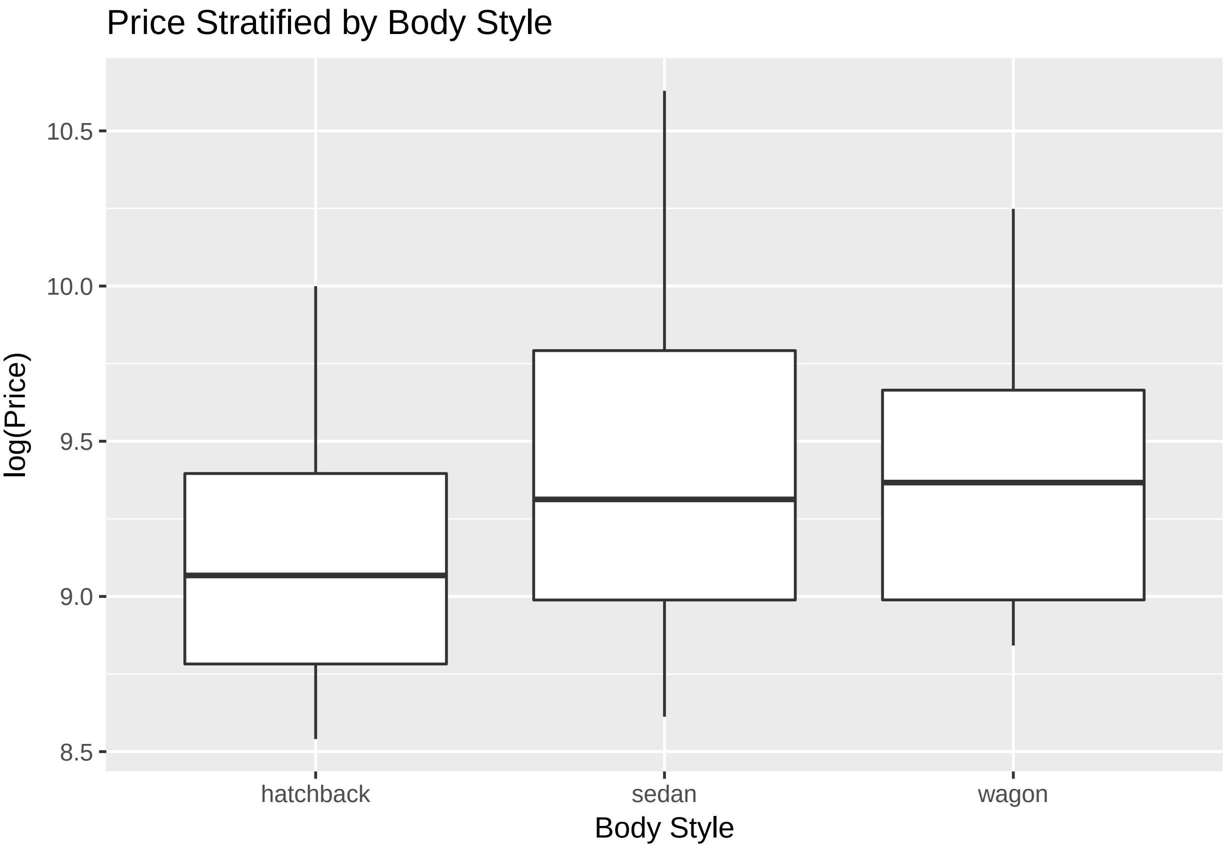
```
## convertible     hardtop   hatchback       sedan       wagon
##           6           8          68          94          25
```

Convertible and hardtop automobiles do not have sufficient data, so they will be excluded. A subset will be created that contains the other body types.

```
# Create subset by dropping convertible and hardtop (insufficient data)
auto_bodystyle <- subset(auto, auto$body.style %in% c('hatchback','sedan','wagon'))
```

Before performing formal tests, let's look at the price distributions for each body type

```
ggplot(auto_bodystyle, aes(body.style,logPrice)) + geom_boxplot() + xlab('Body Style') +
  ylab('log(Price)') + ggtitle('Price Stratified by Body Style')
```



There certainly appear to be differences in price based on body type, so let's apply some formal tests.

ANOVA

First, we'll use ANOVA to see if any of the three price distributions are different from each other in a statistically significant way. This test won't say which groups are different, only if a difference exists.

```
# ANOVA
df_aov <- aov(logPrice ~ body.style, data = auto_bodystyle)
summary(df_aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## body.style    2   3.85  1.9264   9.274 0.000145 ***
## Residuals   184  38.22  0.2077                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is much less than 0.05 (0.000145), so we can say that a significant difference exists between at least two of the groups. To figure out which groups in particular, we can use Tukey's HSD.

Tukey's HSD

This test will tell us which groups are statistically different by performing pairwise comparisons between each of the groups. The input is actually the output from the ANOVA function (an ANOVA model).

```
# Tukey HSD
tukey_anova <- TukeyHSD(df_aov)
tukey_anova
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = logPrice ~ body.style, data = auto_bodystyle)
## 
## $body.style
##                         diff         lwr       upr     p adj
## sedan-hatchback   0.31088820  0.13945380 0.4823226 0.0000870
## wagon-hatchback   0.22291209 -0.02895758 0.4747818 0.0944466
## wagon-sedan      -0.08797611 -0.33030122 0.1543490 0.6675193
```

From this test, it is clear that the sedan and hatchback groups are the most different, as they have the smallest p-value, and the only one that is below the cutoff of 0.05. The wagon and hatchback groups are somewhat different, but not significant enough, and the sedan and wagon groups are the least different. The boxplot shown on the previous page agrees with this assessment, and helps to corroborate Tukey's HSD.

Summary and Conclusion

The purpose of this assignment was to explore some of the basic concepts and methods associated with hypothesis testing in statistics. After exploring the automobile data set, it was found that:

1. The distribution of the price of the automobiles in the data set has positive skew, and taking the log of the price helps to restore some symmetry to the distribution.
2. Using Welch's t-test, it was determined that a statistically significant relationship exists between the price of an automobile and both aspiration and drive type, but not fuel type.
3. Using ANOVA and Tukey's HSD, it was found that a statistically significant relationship exists between the price of a sedan versus a hatchback, but not between other body types.